Table 8.7
Average Monthly Temperature Data for New Delhi, India

Month	J	F	M	A	M	J	J	A	S	O	N	D
Average temperature, °F	57	62	73	82	92	94	88	86	84	79	68	59

harmonics. Also include the original data points in this plot, and visually compare the goodness of fit.

(d) Represent these harmonics using a harmonic dial, with January 1 in the 12-o'clock position, and 1 cm representing 10°F.

8.7. Use the two-harmonic equation for the annual cycle from Exercise 8.6 to estimate the mean daily temperatures for
(a) April 10
(b) October 27

8.8. The amplitudes of the third, fourth, fifth, and sixth harmonics, respectively, of the data in Table 8.7 are 1.4907, 0.5773, 0.6311, and 0.0001°F.
(a) Plot a periodogram for this data. Explain what it shows.
(b) What proportion of the variation in the monthly average temperature data is described by the first two harmonics?

8.9. How many "tic-marks" for frequency are missing from the horizontal axis of Fig. 8.18?

8.10. Suppose that the minor peak in Fig. 8.18 at $f = 13/256 = 0.0508\ \text{mo}^{-1}$ resulted in part from aliasing.
(a) Compute a frequency that could have produced this spurious signal in the spectrum.
(b) How often would the underlying sea-level pressure data need to be recorded and processed in order to resolve this frequency explicitly?

8.11. Derive and plot the theoretical spectra for the two autoregressive processes in Exercise 8.2, assuming unit white-noise variance, and $n = 100$.

Chapter 9 Methods for Multivariate Data

9.1 Background

Much of the material in the first eight 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, one finds in many practical situations that data sets are composed of vector observations. In this situation each data record consists of simultaneous observations of multiple quantities. Such data sets are known as *multivariate*. Examples of multivariate atmospheric data include simultaneous observations of different variables at one location, or an atmospheric field as represented by a set of gridpoint values at a particular time.

Univariate methods can be, and are, applied to individual scalar elements of multivariate data observations. The distinguishing attribute of the methods of multivariate analysis is that both the joint behavior of the multiple simultaneous observations, as well as the variations of the individual data elements, are considered.

This chapter presents introductions to some of the multivariate methods that are used most commonly with atmospheric data. These include approaches to data reduction and structural simplification, summarization of multiple dependencies, and grouping and classification. Absent from this chapter is a presentation of multivariate hypothesis testing. Many multivariate testing methods require the often restrictive assumption of multivariate normality, and are covered extensively in textbooks devoted to multivariate statistics (e.g., Johnson and Wichern, 1982, or Morrison, 1967).

Multivariate methods are more difficult to understand and implement than univariate methods. Notationally, they require use of matrix algebra to make the presentation tractable. The elements of matrix algebra that are necessary to the subsequent material are presented in Section 9.2. The complexities of multivariate data and the methods that have been devised to deal with them dictate that all except the most trivial multivariate analyses will be implemented by 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

"canned" packages for this purpose, and the material in this chapter should help ful in understanding what the programs are doing, and why.

9.2 Matrix Algebra Notation

The mathematics of dealing simultaneously with multiple variables and their mutual correlations is greatly simplified by use of a notation called *matrix algebra*, or *linear algebra*. A brief review of this subject, sufficient for the multivariate techniques described in the remainder of this chapter, is presented in this section. More complete introductions are readily available elsewhere (e.g., Lipschutz, 1968).

9.2.1 Vectors

The *vector* is a fundamental component of the notation of matrix algebra. It is essentially nothing more than an ordered list of scalar variables, or ordinary numbers. These numbers, also called *elements* of the vector, pertain to different aspects of a phenomenon or situation being measured or described. The number of elements, also called the vector's *dimensionality*, will depend on the situation at hand. A familiar meteorological example is the two-dimensional horizontal wind vector, whose two elements are the eastward windspeed u, and the northward windspeed v.

Here vectors will be denoted using boldface lowercase letters, and their individual elements will be distinguished by subscripts running from 1 to K. A vector with only $K = 1$ element is just an ordinary number, or scalar. Most commonly, vectors are *column vectors*, which means that their elements are arranged vertically. For example, the column vector $\mathbf{x}$ would consist of the elements x_1, x_2, $x_3, \ldots, x_K$, arranged as

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_K \end{bmatrix}. \tag{9.1}$$

These same elements can be arranged horizontally, as a *row vector*, through an operation called *transposing* the vector. The transpose operation is denoted by the superscript "T," so that one can write the vector $\mathbf{x}$ in Eq. (9.1) as the row vector

$$\mathbf{x}^T = [x_1, x_2, x_3, \ldots, x_K], \tag{9.2}$$

which is pronounced "x-transpose." Using the transpose of a column vector is useful for notational consistency with certain matrix operations. It is also useful

ographic purposes, as it allows one to write out a vector on a horizontal line

ddition of two or more vectors with of the same dimension is straightforward. ctor addition is accomplished by adding the corresponding elements of the two ctors; for example

$$\mathbf{x} + \mathbf{y} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_K \end{bmatrix} + \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_K \end{bmatrix} = \begin{bmatrix} x_1 + y_1 \\ x_2 + y_2 \\ x_3 + y_3 \\ \vdots \\ x_K + y_K \end{bmatrix}. \tag{9.3}$$

Subtraction is accomplished analogously. Addition and subtraction of vectors with different dimensions is 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 $\mathbf{x}$ in Eq. (9.1) by a constant c yields

$$c\mathbf{x} = \begin{bmatrix} cx_1 \\ cx_2 \\ cx_3 \\ \vdots \\ cx_K \end{bmatrix}. \tag{9.4}$$

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:

$$\mathbf{x}^T\mathbf{y} = [x_1, x_2, x_3, \ldots, x_K]\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_K \end{bmatrix} = x_1y_1 + x_2y_2 + x_3y_3 + \cdots + x_Ky_K$$

$$= \sum_{k=1}^{K} x_k y_k. \tag{9.5}$$

This vector multiplication has been written as the product of a row vector and a column vector for consistency with the operation of matrix multiplication, presented in the following section. As will be seen, the dot product is in fact a special case of matrix multiplication. Equation (9.5) also shows that vector multiplication can also be expressed in *component form* using summation notation. Expanding vector and matrix operations in component form is useful if the calculation is to be programmed for a computer.

It is possible and often convenient to think of vectors geometrically. In this way

of thinking, a vector locates a point in the K-dimensional space whose Cartesian coordinates are each of the K elements of the vector. This is easiest to see in $K = 2$ dimensions, where the two-dimensional space is a plane. In this case x_1 is the horizontal coordinate, and x_2 is the vertical coordinate of the point $\mathbf{x}$. Similarly, one also sometimes thinks of a vector as an arrow "pointing" from the origin (located at $x_1 = x_2 = x_3 = \cdots = x_K = 0$) to the point $\mathbf{x}$. This way of thinking can be extended to higher dimensions, although for K greater than three dimensions the geometry can be difficult or impossible to visualize explicitly.

The length of a vector in its K-dimensional space is a scalar quantity that can be computed using the dot product, as

$$\|\mathbf{x}\| = \sqrt{\mathbf{x}^T\mathbf{x}} = \left[\sum_{k=1}^{K} x_k^2\right]^{1/2}. \tag{9.6}$$

Equation (9.6) is sometimes known as the *Euclidean norm* of the vector $\mathbf{x}$. In $K = 2$ dimensions it is easy to see that this length is simply an application of the Pythagorean theorem, or $\|\mathbf{x}\| = [x_1^2 + x_2^2]^{1/2}$. A common application of this equation is in the computation of the total horizontal windspeed from the horizontal velocity vector $\mathbf{v}^T = [u, v]$, according to $v_H = (u^2 + v^2)^{1/2}$. However, Eq. (9.6) generalizes to arbitrarily high K as well.

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

$$\theta = \cos^{-1}\left[\frac{\mathbf{x}^T\mathbf{y}}{\|\mathbf{x}\|\,\|\mathbf{y}\|}\right]. \tag{9.7}$$

This relationship implies that two vectors are perpendicular if their product is zero, since $\cos[0] = 90°$. Mutually perpendicular vectors are also called *orthogonal*.

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

$$L_{\mathbf{x},\mathbf{y}} = \frac{\mathbf{x}^T\mathbf{y}}{\|\mathbf{y}\|}. \tag{9.8}$$

The geometric interpretations of these three computations of length, angle, and projection are illustrated in Fig. 9.1, for the vectors $\mathbf{x}^T = [1, 1]$ and $\mathbf{y}^T = [2, 0.8]$. The length of $\mathbf{x}$ is simply $\|\mathbf{x}\| = (1^2 + 1^2)^{1/2} = \sqrt{2}$, and the length of $\mathbf{y}$ is $\|\mathbf{y}\| = (2^2 + 0.8^2)^{1/2} = 2.154$. Since the dot product of the two vectors is $\mathbf{x}^T\mathbf{y} = 1 \cdot 2 + 1 \cdot 0.8 = 2.8$, the angle between them is $\theta = \cos^{-1}[2.8/(\sqrt{2} \cdot 2.154)] = 49°$, and the projection of $\mathbf{x}$ onto $\mathbf{y}$ is $2.8/2.154 = 1.302$.

9.2.2 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 the number of rows and columns. A matrix dimension is written "(I × J)", and pronounced "I by J." Matrices are denoted here by boldface uppercase letters surrounded by square brackets.

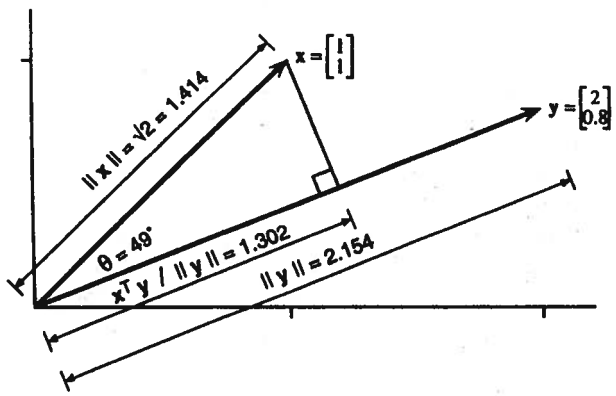

Fig. 9.1 Illustration of the concepts of vector length [Eq. (9.6)], the angle between two vectors [Eq. (9.7)], and the projection of one vector onto another [Eq. (9.8)]; for the two vectors $\mathbf{x}^T = [1, 1]$ and $\mathbf{y}^T = [2, 0.8]$.

The elements of a matrix are the individual variables or numerical values occupying the rows and columns. The matrix elements are identified notationally by two subscripts; the first of these identifies the row number and the second identifies the column number. For example, the (4 × 4) matrix [A] is a collection of the 16 individual elements $a_{i,j}$, arranged as

$$[\mathbf{A}] = \begin{bmatrix} a_{1,1} & a_{1,2} & a_{1,3} & a_{1,4} \\ a_{2,1} & a_{2,2} & a_{2,3} & a_{2,4} \\ a_{3,1} & a_{3,2} & a_{3,3} & a_{3,4} \\ a_{4,1} & a_{4,2} & a_{4,3} & a_{4,4} \end{bmatrix}. \tag{9.9}$$

A matrix with the same number of rows and columns, such as [A] in Eq. (9.9), is called a *square matrix*. The elements of a square matrix for which $i = j$ are arranged on the diagonal line from the upper left to the lower right corners, and are called *diagonal elements*. Correlation matrices [R] (Fig. 3.17) are square matrices having all 1's on the diagonal. A matrix for which $a_{i,j} = a_{j,i}$ for all values of i and j is called *symmetric*. Correlation matrices are also symmetric because the correlation between variable i and variable j is identical to the correlation between variable j and variable i. Another important square, symmetrical matrix is the *identity matrix* [I], consisting of 1's on the diagonal and zeros everywhere else:

$$[\mathbf{I}] = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}. \tag{9.10}$$

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.

Square matrices are special cases for which the number of rows equals the number of columns, i.e., $I = J$. Not all matrices are square. For example, vectors are special cases of matrices where one of the dimensions is 1. A K-dimensional column vector is a (K $\times$ 1) matrix, and a K-dimensional row vector is a (1 $\times$ K) matrix.

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 Eq. (9.2). For example, the relationship between the (3 $\times$ 4) matrix [B] and its transpose, the (4 $\times$ 3) matrix $[\mathbf{B}]^{\mathrm{T}}$ is illustrated by comparing

$$[\mathbf{B}] = \begin{bmatrix} b_{1,1} & b_{1,2} & b_{1,3} & b_{1,4} \\ b_{2,1} & b_{2,2} & b_{2,3} & b_{2,4} \\ b_{3,1} & b_{3,2} & b_{3,3} & b_{3,4} \end{bmatrix} \tag{9.11a}$$

and

$$[\mathbf{B}]^{\mathrm{T}} = \begin{bmatrix} b_{1,1} & b_{2,1} & b_{3,1} \\ b_{1,2} & b_{2,2} & b_{3,2} \\ b_{1,3} & b_{2,3} & b_{3,3} \\ b_{1,4} & b_{2,4} & b_{3,4} \end{bmatrix}. \tag{9.11b}$$

If the matrix [A] is symmetric, then $[\mathbf{A}]^{\mathrm{T}} = [\mathbf{A}]$.

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:

$$c[\mathbf{D}] = c\begin{bmatrix} d_{1,1} & d_{1,2} \\ d_{2,1} & d_{2,2} \end{bmatrix} = \begin{bmatrix} cd_{1,1} & cd_{1,2} \\ cd_{2,1} & cd_{2,2} \end{bmatrix}. \tag{9.12}$$

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

$$\begin{aligned} [\mathbf{D}] + [\mathbf{E}] &= \begin{bmatrix} d_{1,1} & d_{1,2} \\ d_{2,1} & d_{2,2} \end{bmatrix} + \begin{bmatrix} e_{1,1} & e_{1,2} \\ e_{2,1} & e_{2,2} \end{bmatrix} \\ &= \begin{bmatrix} d_{1,1} + e_{1,1} & d_{1,2} + e_{1,2} \\ d_{2,1} + e_{2,1} & d_{2,2} + e_{2,2} \end{bmatrix}. \end{aligned} \tag{9.13}$$

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]), but multiplication of two matrices in reverse order is not even defined unless both are square and of the same dimension. 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. In other words, multiplying a ($I \times J$) matrix [A] and a ($J \times K$) matrix [B] 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

$$\underset{(2 \times 3)}{\begin{bmatrix} a_{1,1} & a_{1,2} & a_{1,3} \\ a_{2,1} & a_{2,2} & a_{2,3} \end{bmatrix}} \underset{(3 \times 2)}{\begin{bmatrix} b_{1,1} & b_{1,2} \\ b_{2,1} & b_{2,2} \\ b_{3,1} & b_{3,2} \end{bmatrix}} = \underset{(2 \times 2)}{\begin{bmatrix} c_{1,1} & c_{1,2} \\ c_{2,1} & c_{2,2} \end{bmatrix}}, \tag{9.14a}$$

where

$$\begin{aligned} [\mathbf{C}] &= \begin{bmatrix} c_{1,1} & c_{1,2} \\ c_{2,1} & c_{2,2} \end{bmatrix} \\ &= \begin{bmatrix} a_{1,1}b_{1,1} + a_{1,2}b_{2,1} + a_{1,3}b_{3,1} & a_{1,1}b_{1,2} + a_{1,2}b_{2,2} + a_{1,3}b_{1,3} \\ a_{2,1}b_{1,1} + a_{2,2}b_{2,1} + a_{2,3}b_{3,1} & a_{2,1}b_{1,2} + a_{2,2}b_{2,2} + a_{2,3}b_{3,2} \end{bmatrix}. \end{aligned} \tag{9.14b}$$

The individual components of [C] as written out in Eq. (9.14b) 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 Eq. (9.5), 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* making up the ith row of [A] and the column *vector* constituting the kth column of [B]. Equivalently, matrix multiplication can be written in terms of the individual matrix elements using summation notation,

$$c_{i,k} = \sum_{j=1}^{J} a_{i,j}b_{j,k}; \qquad i = 1, \ldots, I; \quad k = 1, \ldots, K. \tag{9.15}$$

The analog 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 the 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 square matrices are *invertable*; for example, a matrix [A] is invertable if another matrix [B] exists such that

$$[A][B] = [B][A] = [I]. \tag{9.16}$$

It is then said that [B] is the inverse of [A], or $[B] = [A]^{-1}$; and that [A] is the inverse of [B], or $[A] = [B]^{-1}$. Inverses of (2×2) matrices are easy to compute by hand, using

$$[A]^{-1} = \frac{1}{(a_{1,1}a_{2,2}) - (a_{2,1}a_{1,2})} \begin{bmatrix} a_{2,2} & -a_{1,2} \\ -a_{2,1} & a_{1,1} \end{bmatrix}. \tag{9.17}$$

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.

Table 9.1 lists some properties of arithmetic operations with matrices that have not been specifically mentioned in the foregoing.

Example 9.1. Computation of the Correlation Matrix

The correlation matrix [R] was introduced in Chapter 3 as a device for compactly representing the mutual correlations among K variables. A generalized correlation matrix is shown in Fig. 3.17, and the correlation matrix for the January 1987 data in Appendix A.1 (with the unit diagonal elements and the symmetry implicit) is shown in Table 3.3. The computation of these correlations given in Eq. (3.18) can also be expressed in notation of matrix algebra.

The computation begins with the $(n \times K)$ matrix [X] of data values whose correlations are to be computed. Each row of this matrix is a vector, consisting of

Table 9.1
Some Elementary Properties of Arithmetic Operations with Matrices

Multiplication by the identity matrix	$[A][I] = [I][A] = [A]$
Distributive multiplication by a scalar	$c([A][B]) = (c[A])[B]$
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	$([A][B])^{-1} = [B]^{-1}[A]^{-1}$
Transpose of a matrix product	$([A][B])^T = [B]^T[A]^T$
Combining matrix transpose and inverse	$([A]^{-1})^T = ([A]^T)^{-1}$

one observation each of K variables. The number of these rows is the same as the sample size, n, since each of these row vectors is a sample point (in K dimensions). Thus, [X] is essentially 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 January 4.

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

$$[X'] = [X] - \frac{1}{n}[1][X]. \tag{9.18}$$

The second term in Eq. (9.18) is a $(n \times K)$ matrix containing the sample means. Its n rows are all the same, and each 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 an important matrix called the *variance–covariance matrix*, or sometimes simply the *variance matrix*,

$$[S] = \frac{1}{n-1}[X']^T[X']. \tag{9.19}$$

The variance–covariance matrix [S] is a symmetric $(K \times K)$ matrix whose diagonal elements are the sample variances of the K variables, and whose other elements are the covariances among the K variables. For example, $s_{1,1}$ is the sample variance of the first variable, and $s_{1,2}$ is the sample covariance between the first and the second variables. Because variances are so often expressed in terms of variance–covariance matrices, it is fairly common to see the notation $s_{i,i}$ for the sample variance, which is equivalent to the more common and familiar s_i^2. The operation in Eq. (9.19) corresponds to the summation in the numerator of Eq. (3.18).

The variance–covariance matrix is also known as the *dispersion matrix*, because it describes how the observations 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 axes, 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 [Σ], which appears in the probability density function for the multivariate normal distribution [Eq. (4.50)].

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]. That is, $d_{k,k} = \sqrt{s_{k,k}}$, $k = 1, \ldots, K$. The correlation matrix can then be computed from the variance–covariance matrix using

$$[\mathbf{R}] = [\mathbf{D}]^{-1}[\mathbf{S}][\mathbf{D}]^{-1}. \tag{9.20}$$

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 Eq. (9.20) corresponds to division by the standard deviations in Eq. (3.18).

Note that the correlation matrix [R] is equivalently the variance–covariance matrix of the standardized variables (or *standardized anomalies*) z_k [Eq. (3.16)]. 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 Eq. (9.19) into Eq. (9.20) to yield

$$\begin{aligned}[\mathbf{R}] &= \frac{1}{n-1}[\mathbf{D}]^{-1}[\mathbf{X}']^{\mathrm{T}}[\mathbf{X}'][\mathbf{D}]^{-1} \\ &= \frac{1}{n-1}[\mathbf{Z}]^{\mathrm{T}}[\mathbf{Z}],\end{aligned} \tag{9.21}$$

where [Z] is the $(n \times K)$ matrix whose rows are the vectors of standardized variables $\mathbf{z}$, analogously to the matrix [X'] of the anomalies. The first line of Eq. (9.21) converts the matrix [X'] to the matrix [Z] by dividing each element by its standard deviation, $d_{k,k}$. Comparison of Eqs. (9.21) and (9.19) shows that [R] is, indeed, the variance–covariance matrix for the standardized variables $\mathbf{z}$. ¤

Example 9.2. Multiple Regression Expressed in Matrix Notation

The discussion of multiple linear regression in Chapter 6 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 [Eq. (6.22)] becomes

$$\mathbf{y} = [\mathbf{X}]\mathbf{b}, \tag{9.22a}$$

or

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{1,1} & x_{1,2} & \cdots & x_{1,K} \\ 1 & x_{2,1} & x_{2,2} & \cdots & x_{2,K} \\ 1 & x_{3,1} & x_{3,2} & \cdots & x_{3,K} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n,1} & x_{n,2} & \cdots & x_{n,K} \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ \vdots \\ b_K \end{bmatrix}. \tag{9.22b}$$

Here $\mathbf{y}$ 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 predictor variables, and $\mathbf{b}^{\mathrm{T}} = [b_0, b_1, b_2, \ldots, b_K]$ is a $(K + 1 \times 1)$ vector of the regression parameters. The data matrix in the regression context is similar to that used in Example 9.1, except that it has $K + 1$ rather than K columns. This extra column is the leftmost column of [X], and consists entirely of 1's. Thus, Eq. (9.22) is a vector equation, with dimension $(n \times 1)$ on each side. It is actually n repetitions of Eq. (6.22), once each for the n data records.

The "normal equations" [presented in Eq. (6.6) for the simple case of $K = 1$] are obtained by multiplying each side of Eq. (9.22) by $[\mathbf{X}]^{\mathrm{T}}$:

$$[\mathbf{X}]^{\mathrm{T}}\mathbf{y} = [\mathbf{X}]^{\mathrm{T}}[\mathbf{X}]\mathbf{b}, \tag{9.23a}$$

or

$$\begin{bmatrix} \Sigma\, y \\ \Sigma\, x_1 y \\ \Sigma\, x_2 y \\ \vdots \\ \Sigma\, x_K y \end{bmatrix} = \begin{bmatrix} n & \Sigma\, x_1 & \Sigma\, x_2 & \cdots & \Sigma\, x_K \\ \Sigma\, x_1 & \Sigma\, x_1^2 & \Sigma\, x_1 x_2 & \cdots & \Sigma\, x_1 x_K \\ \Sigma\, x_2 & \Sigma\, x_2 x_1 & \Sigma\, x_2^2 & \cdots & \Sigma\, x_2 x_K \\ \vdots & \vdots & \vdots & & \vdots \\ \Sigma\, x_K & \Sigma\, x_K x_1 & \Sigma\, x_K x_2 & \cdots & \Sigma\, x_K^2 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ \vdots \\ b_K \end{bmatrix}, \tag{9.23b}$$

where all the summations are over the n data points. Here the $[\mathbf{X}]^{\mathrm{T}}[\mathbf{X}]$ matrix has dimension $(K + 1 \times K + 1)$. Each side of Eq. (9.23) has dimension $(K + 1 \times 1)$, and this equation actually represents $K + 1$ simultaneous equations containing the $K + 1$ unknown regression coefficients. Matrix algebra is very commonly used to solve sets of simultaneous linear equations such as these. One way to obtain the solution is to multiply both sides of Eq. (9.23) by the inverse of the $[\mathbf{X}]^{\mathrm{T}}[\mathbf{X}]$ matrix. This operation is analogous to dividing both sides by this quantity, and yields

$$\begin{aligned}([\mathbf{X}]^{\mathrm{T}}[\mathbf{X}])^{-1}[\mathbf{X}]^{\mathrm{T}}\mathbf{y} &= ([\mathbf{X}]^{\mathrm{T}}[\mathbf{X}])^{-1}[\mathbf{X}]^{\mathrm{T}}[\mathbf{X}]\mathbf{b} \\ &= [\mathbf{I}]\mathbf{b} \\ &= \mathbf{b},\end{aligned} \tag{9.24}$$

which is the vector of regression parameters. ¤

9.2.3 Eigenvalues and Eigenvectors of a Square Matrix

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

$$[\mathbf{B}]\mathbf{e} = \lambda\mathbf{e}, \tag{9.25a}$$

or equivalently

$$([\mathbf{B}] - \lambda[\mathbf{I}])\mathbf{e} = \mathbf{0}, \tag{9.25b}$$

where $\mathbf{0}$ is a vector consisting entirely of zeros. For every eigenvalue and eigen-

vector pair that can be found to satisfy Eq. (9.25), any scalar multiple of the eigenvector, $c\mathbf{e}$, will also satisfy the equation together with that eigenvalue. Consequently, it is usual to require that the eigenvectors have unit length:

$$\|\mathbf{e}\| = 1. \tag{9.26}$$

This restriction removes the ambiguity only up to a change in sign, since if a vector $\mathbf{e}$ satisfies Eq. (9.25), its negative, $-\mathbf{e}$ will, also.

If [B] is nonsingular there will be $M = K$ nonzero eigenvalue/eigenvector pairs λ_m and $\mathbf{e}_m$, where K is the number of rows and columns in [B]. Each eigenvector will be dimensioned $(M \times 1)$. 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*.

For many statistical applications, eigenvalues and eigenvectors are calculated for real (not containing complex or imaginary numbers), symmetric matrices. Eigenvalues and eigenvectors of such matrices have a number of important and remarkable properties, some of which also hold for the eigenvalues and eigenvectors of square matrices more generally. The first of these properties is that the eigenvectors of symmetric matrices are orthogonal. That is, their dot products with each other are zero. Furthermore, since each eigenvector is defined to have unit length, the dot product of any eigenvector with itself is one:

$$\mathbf{e}_i^T\mathbf{e}_j = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases}. \tag{9.27}$$

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 Eq. (8.71), expressing the orthogonality of the sine and cosine functions. Often the $(M \times M)$ matrix [E] is formed, the M columns of which are the eigenvectors $\mathbf{e}_k$. Because of the orthogonality and unit length of the eigenvectors,

$$[\mathbf{E}]^T[\mathbf{E}] = [\mathbf{I}], \tag{9.28a}$$

which implies

$$[\mathbf{E}]^{-1} = [\mathbf{E}]^T. \tag{9.28b}$$

Matrices exhibiting these properties are said to be *orthogonal*.

The matrix of eigenvectors [E] has the property that it *diagonalizes* the original matrix [B] from which the eigenvectors and eigenvalues were calculated. This diagonalization is expressed mathematically as

$$[\mathbf{E}]^{-1}[\mathbf{B}][\mathbf{E}] = [\Lambda] = \begin{bmatrix} \lambda_1 & 0 & 0 & \cdots & 0 \\ 0 & \lambda_2 & 0 & \cdots & 0 \\ 0 & 0 & \lambda_3 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & \lambda_M \end{bmatrix}. \tag{9.29}$$

Multiplication of [B] on the left by $[\mathbf{E}]^{-1}$ and on the right by [E] produces the diagonal matrix [Λ], the elements of which are the eigenvalues λ_m corresponding to the eigenvectors $\mathbf{e}_m$ in each of the K columns of [E]. The eigenvalues themselves have the property that their sum is equal to the sum of the diagonal elements of the original matrix [B]. That is

$$\sum_{m=1}^{M} \lambda_m = \sum_{k=1}^{K} b_{k,k}. \tag{9.30}$$

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 $K = M = 2, 3$, or 4. This calculation first requires 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, different algorithms are used. Computer routines for calculating numerical approximations to eigenvalues and eigenvectors are widely available, particularly for the real and symmetrical matrices that are most often of interest in statistical work.

Example 9.3. Eigenvalues and Eigenvectors of a (2 × 2) Symmetric Matrix

The symmetric matrix

$$[\mathbf{B}] = \begin{bmatrix} 185.47 & 110.84 \\ 110.84 & 77.58 \end{bmatrix} \tag{9.31}$$

has as its eigenvalues $\lambda_1 = 254.76$ and $\lambda_2 = 8.29$, with corresponding eigenvectors $\mathbf{e}_1^T = [0.848, 0.530]$ and $\mathbf{e}_2^T = [-0.530, 0.848]$. 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

$$[\mathbf{E}] = \begin{bmatrix} 0.848 & -0.530 \\ 0.530 & 0.848 \end{bmatrix}, \tag{9.32}$$

and this matrix of eigenvectors diagonalizes the original matrix [B] according to

$$[\mathbf{E}]^{-1}[\mathbf{B}][\mathbf{E}] = \begin{bmatrix} 0.848 & 0.530 \\ -0.530 & 0.848 \end{bmatrix}\begin{bmatrix} 185.47 & 110.84 \\ 110.84 & 77.58 \end{bmatrix}\begin{bmatrix} 0.848 & -0.530 \\ 0.530 & 0.848 \end{bmatrix}$$

$$= \begin{bmatrix} 254.76 & 0 \\ 0 & 8.29 \end{bmatrix} = [\Lambda]. \tag{9.33}$$

Because of the orthogonality of the eigenvectors, the inverse of [E] is equal to its transpose as indicated in Eq. (9.33). Finally, the sum of the eigenvectors, 254.76 +

8.29 = 263.05, equals the sum of the diagonal elements of the original [B] matrix, 185.47 + 77.58 = 263.05. ¤

Example 9.4. Confidence Ellipses for the Bivariate Normal Distribution

Recall from Chapter 4 that curves of constant height on the probability density function (PDF) for the bivariate normal distribution [Eq. (4.45)] are elliptical, and centered on the vector mean. The equations for these ellipses are of interest because the volumes under the density function enclosed by them correspond to particular probabilities. The ellipses are sometimes known as *confidence ellipses*.

The directions of the major and minor axes of these ellipses are given by the eigenvectors of the variance–covariance matrix of the two variables. The degrees to which these ellipses are stretched in these two directions are proportional to the square roots of the respective eigenvalues. These concepts are illustrated for the minimum temperature data in Table A.1, the variance–covariance matrix for which is given in Eq. (9.31). The direction of the major axis is given by the first eigenvector, $\mathbf{e}_1^T = [0.848, 0.530]$, whose angle with the x_1 (Ithaca temperature) axis is 32° [from Eq. (9.7)]. The ellipses are stretched much more in the $\mathbf{e}_1$ than in the $\mathbf{e}_2$ direction because the first eigenvalue is so much larger than the second.

The constants multiplying the square roots of the eigenvalues depend on the amount of probability each ellipse encloses. These constants can be obtained from cumulative probabilities of the χ^2 distribution. This distribution is a special case of the gamma distribution, for which cumulative probabilities can be obtained using Table B.2. For the bivariate normal distribution, the distance from the center to the curve of the ellipse, along the direction of the axes, is given by the product $[\lambda_m \chi_2^2(F)]^{1/2}$, where $\chi_2^2(F)$ is the value of the chi-square variable with $\nu = 2$ degrees of freedom corresponding to a cumulative probability F. Equivalently, this is the value from the gamma distribution with $\alpha = 1$ and $\beta = 2$ at the same cumulative probability. In Fig. 9.2, the 50% and 90% confidence ellipses are derived from $F = 0.5$ and $F = 0.9$. Doubling the values in Table B.2 (because $\beta = 2$), these distances are $[1.39\lambda_m]^{1/2}$ and $[4.60\lambda_m]^{1/2}$, respectively.

These ideas generalize directly for the multivariate normal distribution [Eq. (4.50)]. Here, there are $K > 2$ variables in the data vector $\mathbf{x}$, and the axes of the confidence (hyper) ellipses are oriented in the directions of the eigenvectors of the $(K \times K)$ variance-covariance matrix. The equation defining these ellipses is

$$(\mathbf{x} - \boldsymbol{\mu})^T [\mathbf{S}]^{-1} (\mathbf{x} - \boldsymbol{\mu}) = \chi_K^2(F) = F_\Gamma^{-1}\left(\frac{K}{2}, 2\right), \qquad (9.34)$$

which holds for the bivariate ($K = 2$) normal distribution as well. Here $F_\Gamma(\alpha, \beta)$ denotes the cumulative gamma distribution function. In general, the constant by which the square roots of the eigenvalues must be multiplied relates to the

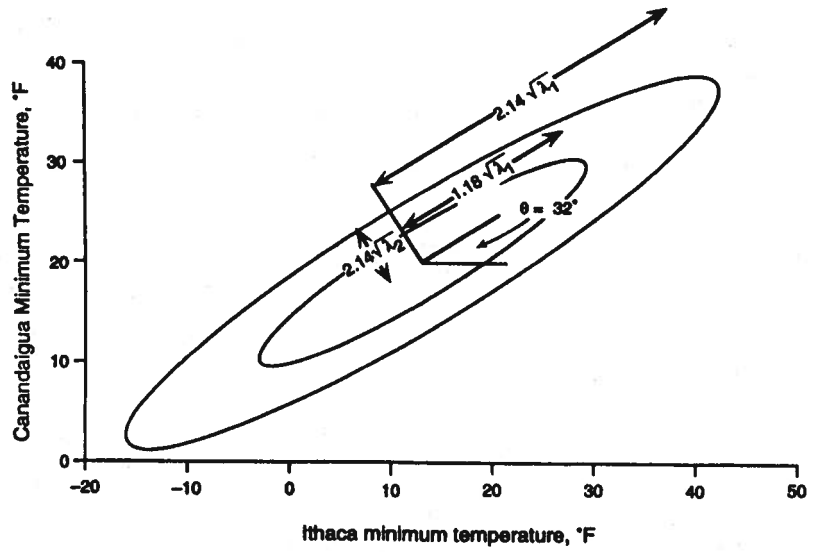

Fig. 9.2 (50%) Inner and (90%) outer confidence ellipses for the bivariate normal distribution representing the minimum temperature data in Table A.1. The major and minor axes of the ellipses are oriented in the directions of the two eigenvectors of the variance–covariance matrix in Eq. (9.31), and are stretched in these directions by an amount proportional to the square roots of the respective eigenvalues. The ellipses are centered at the point defined by the two mean values.

χ^2 distribution with $\nu = K$ degrees of freedom, or the gamma distribution with $\alpha = K/2$. ¤

9.3 Principal-Component (EOF) Analysis

Possibly the most widely used multivariate statistical technique in the atmospheric sciences is *principal-component analysis* (PCA). The technique became popular for analysis of atmospheric data following the paper by Lorenz (1956), who called the technique *empirical orthogonal function* (EOF) analysis. Both names are commonly used, and refer to the same set of procedures. Sometimes the method is incorrectly referred to as *factor analysis*, which is a related but distinct multivariate statistical method.

The purpose of PCA is to reduce a data set containing a large number of variables to a data set containing fewer (hopefully many fewer) new variables, but that nevertheless represent a large fraction of the variability contained in the original data. That is, given multiple observations of a $(K \times 1)$ data vector $\mathbf{x}$, one hopes to find $(M^* \times 1)$ vectors $\mathbf{u}$ whose elements are functions of the elements of the $\mathbf{x}$ values, that contain most of the information in the original collection of $\mathbf{x}$ values,

and whose dimensionality $M^* << K$. This goal can be achieved if there are substantial correlations among the variables contained in $\mathbf{x}$, in which case $\mathbf{x}$ contains redundant information. The elements of these new vectors $\mathbf{u}$ are called the *principal components*. In addition to constituting a compact representation of the original data $\mathbf{x}$, the new variables comprising $\mathbf{u}$ exhibit the very desirable attribute of being mutually uncorrelated.

Usually the data for atmospheric, and other geophysical, fields exhibit many large correlations among the variables x_k, and a PCA results in a much more compact representation of their variations. Beyond mere data compression, however, a PCA can be a very useful tool for exploring large multivariate data sets, especially those consisting of geophysical fields. Here PCA has the potential for yielding substantial insights into both the spatial and temporal variations exhibited by the field or fields being analyzed.

This section is intended to provide a basic introduction to what has become a very large subject. Book-length treatments of PCA are given in Preisendorfer (1988), which is oriented specifically toward geophysical data; and in Jolliffe (1986), which describes PCA more generally. In addition, most textbooks on multivariate statistical analysis contain chapters on PCA.

9.3.1 Basics of PCA

At root, a PCA is based on analysis of the variance–covariance matrix [S] [Eq. (9.19)]. This matrix contains the sample variances of the K elements of the data vector $\mathbf{x}$ on its diagonal, and the covariances among these variables in the off-diagonal positions. Thus, [S] contains a wealth of information about the nature of the multiple variations exhibited by $\mathbf{x}$. The variance–covariance matrix is related to the correlation matrix [R] through Eq. (9.20), which shows that [R] is a nondimensionalized version of [S]. That is, [R] is produced by dividing each element of [S] by the standard deviations $\sqrt{s_{i,i}}$ and $\sqrt{s_{j,j}}$ of the variables occupying the ith row and jth column.

PCAs are conducted on *centered data*, or *anomalies* in the terminology of the atmospheric sciences. An anomaly x'_k is obtained simply by subtracting the sample mean of the x_k values. Similarly a vector of anomalies $\mathbf{x}'$ is computed simply as

$$\mathbf{x}' = \mathbf{x} - \overline{\mathbf{x}} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_K \end{bmatrix} - \begin{bmatrix} \overline{x}_1 \\ \overline{x}_2 \\ \overline{x}_3 \\ \vdots \\ \overline{x}_K \end{bmatrix}. \tag{9.35}$$

Anomalies differ from the original data only with respect to their mean values, which are all zero. Anomalies retain the same physical dimensions as the original variables, and exhibit the same variances and mutual correlations. Indeed, Eq. (9.19) shows that the variance–covariance matrices of $\mathbf{x}$ and $\mathbf{x}'$ are identical.

The new variables, u_m, that will account successively for the maximum amount of the joint variability of $\mathbf{x}'$ (and therefore also of $\mathbf{x}$) are found using the eigenvectors of [S]. In particular, the mth principal component, u_m is obtained as the projection of the data vector $\mathbf{x}'$ onto the mth eigenvector, $\mathbf{e}_m$,

$$u_m = \mathbf{e}_m^T \mathbf{x}' = \sum_{k=1}^{K} e_{km} x'_k, \qquad m = 1, \ldots, M. \tag{9.36}$$

Notice that each of the M eigenvectors contains one element pertaining to each of the K variables, x_k. Similarly, the mth principal component in Eq. (9.36) is computed from a particular set of observations of the K variables x_k. That is, each of the M principal components is a sort of weighted average of the x'_k values. Although the weights (the e_{km} values) do not sum to 1, their squares do because of the scaling convention $\|\mathbf{e}_m\| = 1$. If the data sample consists of n observations (and therefore of n data records, or n rows in the data matrix [X]), there will be n values for each of the principal components, or new variables, u_m.

Geometrically, the first eigenvector, $\mathbf{e}_1$, points to the direction (in the K-dimensional space of $\mathbf{x}'$) in which the data vectors jointly exhibit the most variability. This first eigenvector is the one associated with the largest eigenvalue, λ_1. The second eigenvector $\mathbf{e}_2$, associated with the second-largest eigenvalue λ_2, is constrained to be perpendicular to $\mathbf{e}_1$ [Eq. (9.27)], but subject to this constraint it will align in the direction that the $\mathbf{x}'$ vectors exhibit their next strongest variations. Subsequent eigenvectors $\mathbf{e}_m$, $m = 3, 4, \ldots, M$, are similarly numbered in decreasing magnitudes of their associated eigenvalues, and will in turn be perpendicular to all the previous eigenvectors. Subject to this constraint these eigenvectors will continue to locate directions in which the original data jointly exhibit maximum variability.

Put another way, the eigenvectors define a new coordinate system in which to view the data. This coordinate system is oriented such that each new axis is aligned along the direction of the maximum joint variability of the data, consistent with that axis being orthogonal to the preceding ones. These axes will turn out to be different for different data sets, because they are extracted from the sample variance-covariance matrix [S] particular to a given data set. That is, they are orthogonal functions, but are defined empirically according to the particular data set at hand. This observation is the basis for the eigenvectors being known in this context as EOFs. The implied distinction is with theoretical orthogonal functions, such as Fourier harmonics or Tschebyschev polynomials, which can also be used to define alternative coordinate systems in which to view a data set.

It is a remarkable property of the principal components that they are uncorrelated. That is, the correlation matrix for the new variables u_m is simply [I]. This implies that the covariances between pairs of the u_m values are also zero, so that

the corresponding variance–covariance matrix is diagonal. In fact, the variance–covariance matrix for the principal components is obtained by the diagonalization of [S] [Eq. (9.29)], and is thus simply the diagonal matrix [Λ] of the eigenvalues of [S]. That is, the variance of the mth principal component u_m is the mth eigenvalue λ_m. Equation (9.30) then implies that the total variation exhibited by the x'_k values is completely represented in (or "accounted for" by) the u_m values, in the sense that the sum of the variances of the centered data $\mathbf{x}'$ (and therefore also of the uncentered variables $\mathbf{x}$) is equal to the sum of the variances of the new variables $\mathbf{u}$.

The expression for the $(M \times 1)$ vector of principal components $\mathbf{u}$, corresponding to Eq. (9.36), is

$$\mathbf{u} = [\mathbf{E}]^T \mathbf{x}', \tag{9.37}$$

where, as before, [E] is the $(M \times K)$ square matrix whose columns are the eigenvectors of [S]. The matrix [E] is said to transform the data vector $\mathbf{x}'$ to the vector of new variables $\mathbf{u}$. This equation is sometimes called the *analysis formula* for $\mathbf{x}'$, expressing that the data can be analyzed in terms of the principal components. Conversely, the original data can be recovered from the principal components through the reverse transformation

$$\mathbf{x}' = [\mathbf{E}]\mathbf{u}, \tag{9.38a}$$

or

$$x'_k = \sum_{m=1}^{M} e_{km} u_m, \qquad k = 1, \ldots, K, \tag{9.38b}$$

which is obtained from Eq. (9.37) by multiplying on the left by [E] and using the orthogonality property of this matrix [Eq. (9.28)]. Equation (9.38) is sometimes called the *synthesis formula*, as it expresses the fact that the original centered data can be reconstructed from the principal components. The original uncentered data $\mathbf{x}$ can easily be obtained by adding back the vector of sample means, that is, by reversing Eq. (9.35).

Example 9.5. PCA in Two Dimensions

The basics of PCA are most easily appreciated in a simple example where the geometry can be visualized. If $K = 2$, the space of the data is two-dimensional, and can be graphed on a page. Figure 9.3 shows a scatterplot of centered January 1987 Ithaca minimum temperatures (x'_1) and Canandaigua minimum temperatures (x'_2) from Table A.1. This is the same scatterplot as appears in the middle of the bottom row of Fig. 3.18. Both variables are centered at zero. It is also apparent that the Ithaca temperatures are more variable than the Canandaigua temperatures, with the two standard deviations being $\sqrt{s_{1,1}} = 13.62°\text{F}$ and $\sqrt{s_{2,2}} = 8.81°\text{F}$, respectively. The two variables are clearly strongly correlated, and have a Pearson

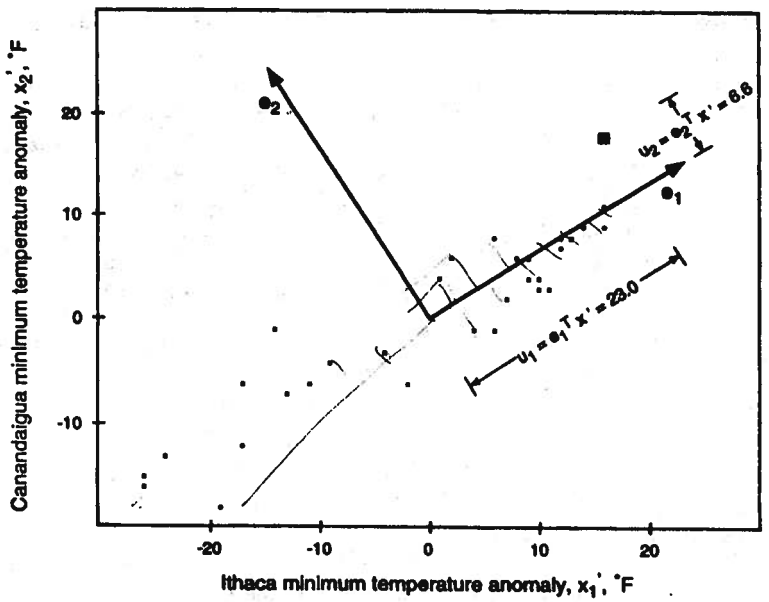

Fig. 9.3 Scatterplot of January 1987 Ithaca and Canandaigua minimum temperatures (converted to anomalies, or "centered"), illustrating the geometry of PCA in two dimensions. The eigenvectors $\mathbf{e}_1$ and $\mathbf{e}_2$ of the variance–covariance matrix [S] for these two variables, as computed in Example 9.3, have been plotted with exaggerated lengths for clarity. The data stretch out in the direction of $\mathbf{e}_1$ to the extent that 96.8% of the joint variance of these two variables occurs along this axis. The coordinates u_1 and u_2, corresponding to the data point $\mathbf{x}'^T = [16.0, 17.8]$, recorded on January 15 and indicated by the large square symbol, are shown by lengths in the directions of the new coordinate system defined by the eigenvectors. That is, the coordinations $\mathbf{u}^T = [23.0, 6.6]$ locate the same data point as $\mathbf{x}'^T = [16.0, 17.8]$.

correlation of +0.924 (Table 3.3). The covariance matrix [S] for these two variables is given as [B] in Eq. (9.31). The two eigenvectors of this matrix are $\mathbf{e}_1^T = [0.848, 0.530]$ and $\mathbf{e}_2^T = [-0.530, 0.848]$, so that the eigenvector matrix [E] is as shown in Eq. (9.32). The corresponding eigenvalues are $\lambda_1 = 254.76$ and $\lambda_2 = 8.29$. Notice also that this is the same data used to fit the bivariate normal distribution for which confidence ellipses are shown in Fig. 9.2.

The orientations of the two eigenvectors are shown in Fig. 9.3, although their lengths have been exaggerated for clarity. It is evident that the first eigenvector is aligned parallel to the direction in which the data jointly exhibit maximum variation. That is, the point cloud is inclined at the same angle as is $\mathbf{e}_1$. This angle is 32° from the horizontal (i.e., from the vector [1, 0]), according to Eq. (9.7). Since the data in this simple example exist in only $K = 2$ dimensions, the constraint that the second eigenvector must be perpendicular to the first determines its direction

up to sign (i.e., it could as easily be $-\mathbf{e}_2^T = [0.530, -0.848]$). This last eigenvector locates the direction in which data jointly exhibit their smallest variations.

The two eigenvectors determine an alternative coordinate system in which to view the data. This fact may become more clear if you rotate this book 32° to the right. Within this rotated coordinate system, each point is defined by a principal component vector $\mathbf{u}^T = [u_1, u_2]$ of new transformed variables, whose elements consist of the projections of the data onto the eigenvectors, according to the dot product in Eq. (9.36). Figure 9.3 illustrates this projection for the 15 January data point $\mathbf{x}'^T = [16.0, 17.8]$, which is indicated by the large square symbol. For this datum, $u_1 = (0.848)(16.0) + (0.530)(17.8) = 23.0$, and $u_2 = (-0.530)(16.0) + (0.848)(17.8) = 6.6$.

The sample variance of the new variable u_1 is an expression of the degree to which it spreads out along its axis (i.e., along the direction of $\mathbf{e}_1$). This dispersion is evidently greater than the dispersion of the data along either of the original axes, and in fact the sample variance of u_1 is equal to the eigenvalue $\lambda_1 = 254.76°\text{F}^2$. While the points in the data set tend to exhibit quite different values of u_1, they have more similar values for u_2. That is, they are much less variable in the $\mathbf{e}_2$ direction, and the sample variance of u_2 is only $\lambda_2 = 8.29°\text{F}^2$.

Since $\lambda_1 + \lambda_2 = s_{1,1} + s_{2,2} = 263.05°\text{F}^2$, the new variables retain all of the variation exhibited by the original data. However, the fact that the point cloud seems to exhibit no slope in the new coordinate frame defined by the eigenvectors indicates that u_1 and u_1 are uncorrelated. Their lack of correlation can be verified by transforming the 31 pairs of minimum temperatures in Table A.1 to principal components and computing the Pearson correlation, which is zero. The variance–covariance matrix for the principal components is therefore $[\Lambda]$, shown in Eq. (9.33).

The two original temperature variables are so strongly correlated that a very large fraction of their joint variance, $\lambda_1/(\lambda_1 + \lambda_2) = 0.968$, is represented by the first principal component. It would be said that the first principal component describes 96.8% of the total variance. The first principal component might be interpreted as reflecting the regional minimum temperature for the area including these two locations (they are about 50 miles apart), with the second principal component describing random variations departing from the overall regional value. ¤

9.3.2 Truncation of the Principal Components

Mathematically, there are as many eigenvectors of [S] as there are elements of the data vector $\mathbf{x}'$. That is, $M = K$. However, it is typical of atmospheric data that substantial covariances (or correlations) exist among the original K variables, and as a result there are few or no off-diagonal elements of [S] (or [R]) that are near zero. This situation implies that there is redundant information in $\mathbf{x}$, and that the first few eigenvectors of its dispersion matrix will locate directions in which the joint variability of the data is greater than the variability of any single element, of $\mathbf{x}'$. Similarly, the last few eigenvectors will point to directions in the K-dimensional space of $\mathbf{x}'$ where the data jointly exhibit very little variation. This feature was illustrated in Example 9.5 for daily temperature values measured at nearby locations.

To the extent that there is redundancy in the original data $\mathbf{x}'$, it is possible to capture most of their variations by considering only the most important directions of their joint variations. That is, most of the information content of the data may be represented using some smaller number $M^* < M$ of the principal components u_m. In effect, the original data set containing the K variables x_k is approximated by the smaller set of new variables u_m. If $M^* << M$, retaining only the first M^* of the principal components results in a much smaller data set. This "data compression" capability of PCA is often a primary motivation for its use.

The truncated representation of the original data can be expressed mathematically by the analysis formula, Eq. (9.37), but in this case the dimension of the truncated $\mathbf{u}$ is $(M^* \times 1)$, and [E] is the (nonsquare, $K \times M^*$) matrix whose columns consist only of the first M^* eigenvectors of [S]. The synthesis formula, Eq. (9.38), is then only approximately true because the original data cannot be exactly resynthesized without using all M eigenvectors:

$$x'_k \approx \sum_{m=1}^{M^*} e_{km} u_m, \qquad k = 1, \ldots, K. \tag{9.39}$$

Regardless of the choice of how many principal components are retained, the proportion of the total joint variation in the data represented by the M^* retained principal components can be computed from the eigenvalues as

$$\text{Proportion of total variance described} = \frac{\sum_{m=1}^{M^*} \lambda_m}{\sum_{m=1}^{M} \lambda_m} = \frac{\sum_{m=1}^{M^*} \lambda_m}{\sum_{k=1}^{K} s_{k,k}}, \tag{9.40}$$

where $s_{k,k}$ is the kth diagonal element of [S], or the sample variance of x'_k.

An important application of the truncation of the full set of M principal components is in the setting of multiple regression (Chapter 6). It often occurs in multiple regression that the predictor (x) variables exhibit very large mutual correlations, which is referred to as *multicollinearity*. In this situation, it turns out that the sampling distributions of the estimated regression coefficients can become very broad, with the practical consequence that a forecasting equation may perform very badly when implemented on future data independent of the training sample. The problem can be rectified by first subjecting the predictor variables to a PCA, and then using the first M^* principal components $\mathbf{u}$ as predictors in place of the original variables $\mathbf{x}$. This approach is called *principal-components regression*. Because the principal components are uncorrelated, they do not exhibit the multicollinearity problem, and their regression coefficients will be estimated with

much better precision. As is the case with predictors that are harmonic functions [Eq. (8.69)], principal-component predictors can be added or deleted from a candidate regression without changing the regression coefficients of other principal-component predictors. This property can aid in the understanding of regression results if the principal components themselves admit of physical interpretations. If M^* is chosen to be large enough, most of the information contained jointly in the full set of the original correlated variables will also be represented in the principal components, so that, in general, little information will be sacrificed. Principal components regression is discussed more fully in Jolliffe (1986) and Draper and Smith (1981).

9.3.3 How Many Principal Components Should Be Retained?

The question, of course, arises as to how few principal components can be retained without discarding important information carried in the original data. There is not a single clear criterion that can be used to choose this number M^* of principal components that are best retained in a given circumstance. While the choice of the truncation level can be aided by one or more of the many available principal-component selection rules, it is ultimately a subjective choice that will depend in part on the data at hand and the purposes of the PCA.

Many of the established principal component selection rules are rooted in the statistical literature, but a substantial amount of work has been done in this area with a specific orientation toward atmospheric and oceanographic data sets (especially Preisendorfer *et al.*, 1981). In this latter context the problem is often viewed in the electronic signal processing paradigm of separating "signal" from "noise." That is, one imagines that there are some number of distinct signals, or modes of variation, embedded in the data, and that this number (M^*) is less than the number of measurements (K) composing each data vector $\mathbf{x}$. In this view, the goal is then to find and discard the $K - M^*$ "noise" components, which are regarded as random-number contaminants that have been added to the interesting data (the "signal"). It is hoped that selection of the first M^* principal components will "filter" the noise and leave behind the meaningful portion of the original data.

Most principal-component selection rules can be viewed as *dominant variance* rules, in that it is the M^* principal components with the largest eigenvalues (which therefore represent the largest fraction of variance in the original data) that are selected. The differences among these rules are with respect to how and where the line is drawn between the retained and discarded principal components.

Figure 9.4 illustrates two graphically based principal-component selection rules for a PCA involving $K = 6$ observations in each data vector $\mathbf{x}$. Figure 9.4a shows the plot of the eigenvalues vs. the corresponding principal component number, called the *scree graph* (or, sometimes, the "scree test"). Figure 9.4b shows the same data with the eigenvalues plotted logarithmically, which is called the *log-eigenvalue* (LEV) *diagram*. Use of either the scree graph or the LEV diagram

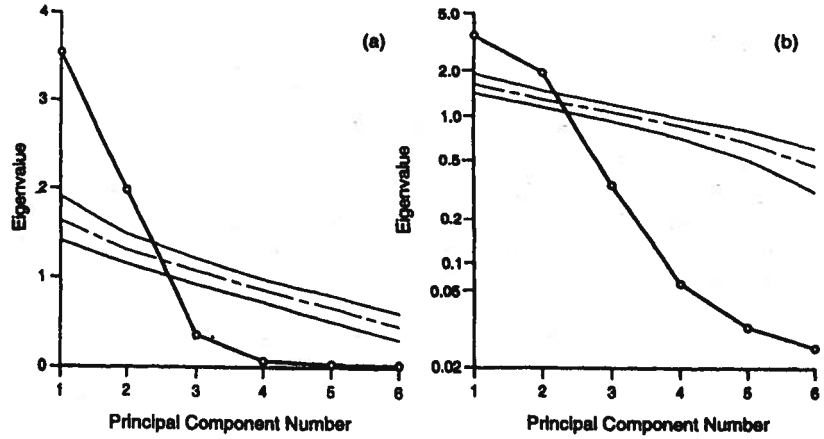

Fig. 9.4 Graphical displays of eigenvalue magnitudes as a function of the principal-component number (heavier lines connecting circled points), for a K = 6-dimensional analysis: (a) linear scaling, or "scree" graph; (b) logarithmic scaling, or LEV diagram. Both the scree and LEV criteria would lead to retention of the first three principal components in this analysis. Lighter lines in both panels show results of the resampling tests necessary to apply rule N of Preisendorfer *et al.* (1981). Dashed line is median of eigenvalues for 1000 (6 × 6) dispersion matrices of independent Gaussian variables, constructed using the same sample size as the data being analyzed. Solid lines indicate the 5th and 95th percentiles of these simulated eigenvalue distributions. Rule *N* would indicate retention of only the first two principal components, as these are the ones significantly larger than what would be expected from data with no correlation structure.

requires a subjective judgment about the existence and location of a "break" in the plotted curve.

When using the scree graph qualitatively, the goal is to locate a point separating a steeply sloping portion to the left, and a more shallowly sloping portion to the right. The principal-component number at which the separation occurs is then taken as M^*. In Figure 9.4a this point would probably be chosen as $M^* = 3$. On the other hand, the LEV diagram is motivated by the idea that, if the last $K - M^*$ principal components represent uncorrelated noise, then the magnitudes of their eigenvalues should decay exponentially with increasing principal-component number. This behavior should be identifiable in the logarithmic LEV diagram as a straight-line portion on its right-hand side. The M^* retained principal components would then be the ones whose log-eigenvalues lie above the leftward extrapolation of this line. In Fig. 9.4b, $M^* = 3$ would probably be chosen by most viewers of this LEV diagram, although this choice is not unambiguous. Similarly, not all scree graphs exhibit an obvious "elbow."

Another group of dominant-variance selection rules involves comparing each eigenvalue (and therefore the variance described by its principal component) to the amount of the joint variance reflected in the "average" eigenvalue. Principal

components whose eigenvalues are above this threshold are retained. This set of selection rules can be summarized by the criterion

$$\text{Retain } \lambda_m \quad \text{if} \quad \lambda_m > \frac{T}{K} \sum_{k=1}^{K} s_{k,k}, \tag{9.41}$$

where $s_{k,k}$ is the sample variance of the kth element of $\mathbf{x}$. The simplest of these, called *Kaiser's rule*, retains the principal components accounting for more than the average amount of the total variance. Thus, Kaiser's rule uses Eq. (9.41) with the threshold parameter $T = 1$. Jolliffe (1972) has argued that Kaiser's rule is too strict, and suggested the alternative of $T = 0.7$. A third alternative is to use the *broken stick model*, so called because it is based on the expected length of the mth longest piece of a randomly broken unit line segment. According to this criterion, the threshold parameter in Eq. (9.41) is taken to be

$$T = \frac{1}{M} \sum_{j=m}^{M} \frac{1}{j}. \tag{9.42}$$

This rule yields a different threshold for each principal component: $T = T(m)$. All three criteria described above lead to choosing $M^* = 2$ for the data in Fig. 9.4.

Of the many principal-component selection rules devised and investigated by Preisendorfer *et al.* (1981), the most commonly used is their *rule N*. Rule *N* identifies the largest M^* principal components to be retained on the basis of resampling (or Monte Carlo) tests, involving the distribution of eigenvalues of randomly generated dispersion matrices. The procedure involves repeatedly generating sets of vectors of independent Gaussian random numbers with the same dimension (K) and sample size (n) as the data $\mathbf{x}$ being analyzed, and then computing the eigenvalues of their dispersion matrices. These randomly generated eigenvalues are then scaled in a way that makes them comparable to the eigenvalues λ_m to be tested, for example, by requiring that the sum of each set of randomly generated eigenvalues will equal the sum of the eigenvalues computed from the original data. Each λ_m from the real data is then compared to the empirical distribution of its synthetic counterparts, and is retained if it is larger than 95% of these.

The light lines in the panels of Fig. 9.4 illustrate the use of rule *N* to select a principal-component truncation level. The dashed lines reflect the medians of 1000 sets of eigenvalues computed from 1000 (6×6) dispersion matrices of independent Gaussian variables, constructed using the same sample size as the data being analyzed. The solid lines show 95th and 5th percentiles of those distributions for each of the six eigenvalues. The first two eigenvalues λ_1 and λ_2 are larger than more than 95% of their synthetic counterparts, and for these the null hypothesis that the corresponding principal components represent only "noise" would therefore be rejected at the 5% level. Accordingly, rule *N* would choose $M^* = 2$ for this data. As an aside, note that the nearly straight-line relationships exhibited by the synthetic eigenvalue quantiles in Fig. 9.4b support the premise of

the LEV criterion, that the eigenvalues of dispersion matrices of uncorrelated noise should decrease exponentially.

A table of 95% critical values for rule *N*, for selected sample sizes n and dimensions K, is presented in Overland and Preisendorfer (1982). Corresponding large-sample tables are given in Preisendorfer *et al.* (1981) and Preisendorfer (1988). Preisendorfer (1988) notes that if there is substantial autocorrelation (i.e., time correlation) present in the individual variables x_k, that it may be more appropriate to construct the resampling distributions for rule *N* (or to use the tables just mentioned) using the smallest effective sample size [Eq. (5.12)] among the x_k, rather than using n independent vectors of Gaussian variables to construct each synthetic dispersion matrix. Another potential problem with rule *N*, and other similar procedures, is that the data $\mathbf{x}$ may not be approximately Gaussian. For example, one or more of the x_k values could be precipitation variables. To the extent that the original data are not Gaussian, the resampling procedure will not simulate accurately the physical process that generated them, and the results of the tests may be misleading. A possible remedy for the problem of non-Gaussian data might be to use a bootstrap version of rule *N*, although this approach seems not to have been tried in the literature to date.

9.3.4 PCA Based on the Covariance Matrix versus the Correlation Matrix

PCA can be conducted as easily on the correlation matrix [**R**] as it can on the covariance matrix [**S**]. The correlation matrix is the variance–covariance matrix of the vector of standardized variables $\mathbf{z}$ [Eq. (9.21)]. The vector of standardized variables $\mathbf{z}$ is related to the vectors of original variables $\mathbf{x}$ and centered variables $\mathbf{x}'$ according to

$$\mathbf{z} = [\mathbf{D}]^{-1}\mathbf{x}' = [\mathbf{D}]^{-1}(\mathbf{x} - \bar{\mathbf{x}})$$

$$= \begin{bmatrix} \frac{1}{\sqrt{s_{1,1}}} & 0 & 0 & \cdots & 0 \\ 0 & \frac{1}{\sqrt{s_{2,2}}} & 0 & \cdots & 0 \\ 0 & 0 & \frac{1}{\sqrt{s_{3,3}}} & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & \frac{1}{\sqrt{s_{K,K}}} \end{bmatrix} \begin{bmatrix} x_1 - \bar{x}_1 \\ x_2 - \bar{x}_2 \\ x_3 - \bar{x}_3 \\ \vdots \\ x_K - \bar{x}_K \end{bmatrix}, \tag{9.43}$$

where the diagonal matrix [**D**] was defined in conjunction with Eq. (9.20), and contains diagonal elements that are the sample standard deviations of x_k (and of x'_k). Therefore, PCA on the correlation matrix amounts to analysis of the joint variance structure of the standardized variables z_k, as computed using either Eq. (9.43) or (in scalar form) Eq. (3.16).

The difference between PCAs performed using the variance–covariance and correlation matrices will be one of emphasis. Since the PCA seeks to find variables successively maximizing proportion of the total variance, $\Sigma_k s_{k,k}$, represented, analyzing the covariance matrix [S] results in principal components that emphasize the x_k values having the largest variances. Other things equal, the tendency will be for the first few eigenvectors to align near the directions of the variables having the biggest variances. In Example 9.5, the first eigenvector points more toward the Ithaca minimum temperature axis because the variance of the Ithaca minimum temperatures is larger than the variance of the Canandaigua minimum temperatures. Conversely, PCA applied to the correlation matrix [R] weights all of the standardized variables z_k equally, since all have equal (unit) variance.

If the PCA is conducted using the correlation matrix, the analysis formula, Eqs. (9.36) and (9.37), will pertain to the standardized variables, z_k and $\mathbf{z}$, respectively. Similarly, the synthesis formula, Eq. (9.38), will pertain to $\mathbf{z}$ and z_k rather than to $\mathbf{x}'$ and x'_k. In this case the original data $\mathbf{x}$ can be recovered from the result of the synthesis formula by reversing the standardization given by Eq. (9.43):

$$\mathbf{x} = [\mathbf{D}]\mathbf{z} + \bar{\mathbf{x}}. \tag{9.44}$$

Although $\mathbf{z}$ and $\mathbf{x}'$ can be easily obtained from each other using Eq. (9.43), the eigenvalue/eigenvector pairs of [R] and [S] do not bear simple relationships to one another. In general, it is not possible to compute the principal components of one knowing only the principal components of the other. This fact implies that these two alternatives for PCA do not yield equivalent information, and that it is important to make an intelligent choice of one over the other for a given application. If an important goal of the analysis is to identify or isolate the strongest variations in a data set, the better choice will usually be PCA using the covariance matrix. This will be particularly so if all the x_k values are measured in the same physical units. In this situation, the choice between PCA of the covariance or the correlation matrix will depend on the judgment of the analyst and the purpose of the study. For example, in analyzing gridded numbers of extratropical cyclones, Overland and Preisendorfer (1982) found that PCA on their covariance matrix better identified regions having the highest variability in cyclone numbers, whereas correlation-based PCA was more effective at locating the primary storm tracks.

However, if the analysis is of unlike variables, that is, variables not measured in the same units, it will almost always be preferable to compute the PCA using the correlation matrix. Measurement in unlike physical units yields arbitrary relative scalings of the variables, which results in arbitrary relative magnitudes of the variances of these variables. To take a simple example, the variance of a set of temperatures measured in degrees Fahrenheit will be $(1.8)^2 = 3.24$ times as large as the variance of the same temperatures expressed in degrees Celsius. If the PCA has been done using the correlation matrix, the analysis formula, Eq. (9.37), pertains to the vector $\mathbf{z}$ rather than $\mathbf{x}'$; and the truncated synthesis in Eq. (9.39) will

yield approximately the standardized variables z_k. The summations in the denominators of Eq. (9.40) will equal the number of standardized variables since each has unit variance.

Example 9.6. Correlation-versus-Covariance-Based PCA for Arbitrarily Scaled Quantities

The importance of basing a PCA on the correlation matrix when the variables being analyzed are not measured on comparable scales is illustrated in Table 9.2. This table summarizes PCAs of the January 1987 data in Table A.1 in (a) unstandardized (covariance matrix) and (b) standardized (correlation matrix) forms. Sample variances of the variables are shown, as are the six eigenvectors, the six eigenvalues, and the cumulative percentages of variance accounted for by the principal components. The (6 × 6) arrays in the upper-right portions ("Canan-

Table 9.2

Comparison of PCA Computed Using (a) the Covariance Matrix and (b) the Correlation Matrix of the Data in Table A.1[a]

Variable	Sample variance	$\mathbf{e}_1$	$\mathbf{e}_2$	$\mathbf{e}_3$	$\mathbf{e}_4$	$\mathbf{e}_5$	$\mathbf{e}_6$
		a. Covariance Results					
Ithaca ppt	0.059 in.^{2}	.003	.017	.002	−.028	.818	−.575
Ithaca T_{max}	892.2°F^2	.359	−.628	.182	−.665	−.014	−.003
Ithaca T_{min}	185.5°F^2	.717	.527	.456	.015	−.014	.000
Canandaigua ppt	0.028 in^2	.002	.010	.005	−.023	.574	.818
Canandaigua T_{max}	61.8°F^2	.381	−.557	.020	.737	.037	.000
Canandaigua T_{min}	77.6°F^2	.459	.131	−.871	−.115	−.004	.003
Eigenvalues, λ_k		337.7	36.9	7.49	2.38	0.065	0.001
Cumulative % variance		87.8	97.4	99.3	99.9	100.0	100.0
		b. Correlation Results					
Ithaca ppt	1.000	.142	.677	.063	−.149	−.219	.668
Ithaca T_{max}	1.000	.475	−.203	.557	.093	.587	.265
Ithaca T_{min}	1.000	.495	.041	−.526	.688	−.020	.050
Canandaigua ppt	1.000	.144	.670	.245	.096	.164	−.658
Canandaigua T_{max}	1.000	.486	−.220	.374	−.060	−.737	−.171
Canandaigua T_{min}	1.000	.502	−.021	−.458	−.695	−.192	−.135
Eigenvalues, λ_k		3.532	1.985	0.344	0.074	0.038	0.027
Cumulative % variance		58.9	92.0	97.7	98.9	99.5	100.0

[a] The sample variances of each variable are shown, as are the six eigenvectors $\mathbf{e}_m$ arranged in decreasing order of their eigenvalues λ_m. The cumulative percentage of variance represented is calculated according to Eq. (9.40) multiplied by 100%. The much smaller variances of the precipitation (ppt) variables are artifacts of the measurement units, but result in precipitation being unimportant in the first four principal components computed from the covariance matrix, which collectively account for 99.9% of the total variance of the data set. Computing the principal components from the correlation matrix ensures that variations of the temperature and precipitation variables are weighted equally.

daigua" and "Ithaca" rows, columns 3–8) of this table constitute the matrix of eigenvectors, [E], in each case.

Because of the magnitudes of the variations of the data in relation to their measurement units, the variances of the unstandardized precipitation data are tiny in comparison to the variances of the temperature variables. This is purely an artifact of the measurement unit for precipitation (inches) being relatively large in comparison to the range of variation of the data (about 1 in.), and the measurement unit for temperature (°F) being relatively small in comparison to the range of variation of the data (about 40°F). If the measurement units had been millimeters and degrees Celsius, respectively, the differences in variances would have been much smaller. If the precipitation had been measured in micrometers, the variances of the precipitation variables would dominate the variances of the temperature variables.

Because the variances of the temperature variables are so much larger than the variances of the precipitation variables, the PCA of the covariance matrix is dominated by the temperatures. The eigenvector elements corresponding to the two precipitation variables are negligibly small in the first four eigenvectors, so these variables make negligible contributions to the first four principal components. However, these first four principal components collectively describe 99.9% of the joint variance. No reasonable principal-component truncation rule would retain the last two principal components, which carry almost all of the precipitation information. An application of the truncated synthesis formula, Eq. (9.39), would therefore result in reconstructed precipitation values very near their average values. That is, essentially none of the variation in precipitation would be represented.

Since the correlation matrix is the variance–covariance matrix for comparably scaled variables z_k, each has equal variance. Unlike the analysis on the covariance matrix, the PCA does not ignore the precipitation variables when the correlation matrix is analyzed. Here the first (and most important) principal component represents primarily the closely intercorrelated temperature variables, as can be seen from the relatively larger elements of $\mathbf{e}_1$ for the temperatures. However, the second principal component, which accounts for 33.1% of the total variance in the scaled data set, represents primarily the precipitation variations. The eigenvalues for this PCA are shown in Fig. 9.4. The first two principal components were retained according to all of the selection rules discussed in the previous section, and would almost surely be retained in any reasonable truncation rule. The result would be that the precipitation variations would not be lost in the truncated data representation, but rather would be very nearly completely reconstructed by an application of Eq. (9.39). □

9.3.5 Application of PCA to Fields

The overwhelming majority of applications of PCA to atmospheric data have involved analysis of fields (i.e. spatial arrays of variables) such as geopotential

heights, temperatures, and precipitation. In these cases the full data set consists of multiple observations of a field or of a set of fields. Frequently these multiple observations take the form of time series, for example, a sequence of daily hemispheric 500-mb heights. Another way to look at this kind of data is as a collection of K mutually correlated time series that have been sampled at each of K gridpoints or station locations. The goal of PCA as applied to this type of data is usually to explore, or to express succinctly, the joint space/time variations of the many variables in the data set.

Even though the locations at which the field is sampled are spread over a two-dimensional (or possibly three-dimensional) space, the data from these locations at a given observation time are arranged in the one-dimensional vector $\mathbf{x}$. That is, regardless of their geographic arrangement, each location is assigned a number (as in Fig. 7.11) from 1 to K, which refers to the appropriate element in the data vector $\mathbf{x}^T = [x_1, x_2, x_3, \ldots, x_K]$. In this most common application of PCA to fields, the data matrices [X] and [X′] are thus dimensioned ($n \times K$), or (time × space), since data at K locations in space have been sampled at n different times.

To emphasize that the original data consists of K time series, the analysis equation [(9.36) or (9.37)] is sometimes written with an explicit time index:

$$\mathbf{u}(t) = [\mathbf{E}]^T \mathbf{x}'(t), \tag{9.45a}$$

or, in scalar form,

$$u_m(t) = \sum_{k=1}^{K} e_{km} x'_k(t), \qquad m = 1, \ldots, M. \tag{9.45b}$$

Here the time index t runs from 1 to n. The synthesis equation [(9.38) or (9.39)] can be written using the same notation. Equation (9.45) emphasizes that, if the data $\mathbf{x}$ consist of a set of time series, then the principal components $\mathbf{u}$ are also time series. The time series of one of the principal components, $u_m(t)$, may very well exhibit serial correlation (correlation with itself through time), and the principal-component time series are sometimes analyzed using tools presented in Chapter 8. However, each principal-component time series will be uncorrelated with the time series of all the other principal components.

When the K elements of $\mathbf{x}$ are measurements at different locations in space, the eigenvectors can be displayed graphically in a quite informative way. Notice that each eigenvector contains exactly K elements, and that these elements have a one-to-one correspondence with each of the K locations in the dot product from which the corresponding principal component is calculated [Eq. (9.45b)]. Each eigenvector element e_{km} can be plotted on a map at the same location as its corresponding data value x'_k, and this "field" of eigenvector elements can itself be displayed with smooth contours in the same way as ordinary meteorological fields. Such maps depict clearly which locations are contributing most strongly to the respective principal components. Looked at another way, such maps indicate the geographic distribution of simultaneous data anomalies represented by the corre-

sponding principal component. These geographic displays of eigenvectors are also sometimes interpreted as representing uncorrelated modes of variability of the field from which the PCA was extracted.

Figure 9.5, from Wallace and Gutzler (1981), shows the first four eigenvectors of a PCA of the correlation matrix for winter monthly-mean 500-mb heights at gridpoints in the Northern Hemisphere. The numbers below and to the right of the panels show the percentage of the total hemispheric variance [Eq. (9.40) × 100%] represented by each of the corresponding principal components. Together, the first four principal components account for nearly half of the (normalized) hemispheric winter height variance. These patterns resemble the *teleconnectivity* patterns for the same data shown in Fig. 3.20, and apparently reflect the same underlying physical processes in the atmosphere. For example, Fig. 9.5b evidently reflects the "PNA" pattern of alternating height anomalies stretching from the Pacific Ocean to northwestern North America to southeastern North America. A positive value of the second principal component of this data set corresponds to negative 500-mb height anomalies (troughs) in the northwestern Pacific and in the southeastern

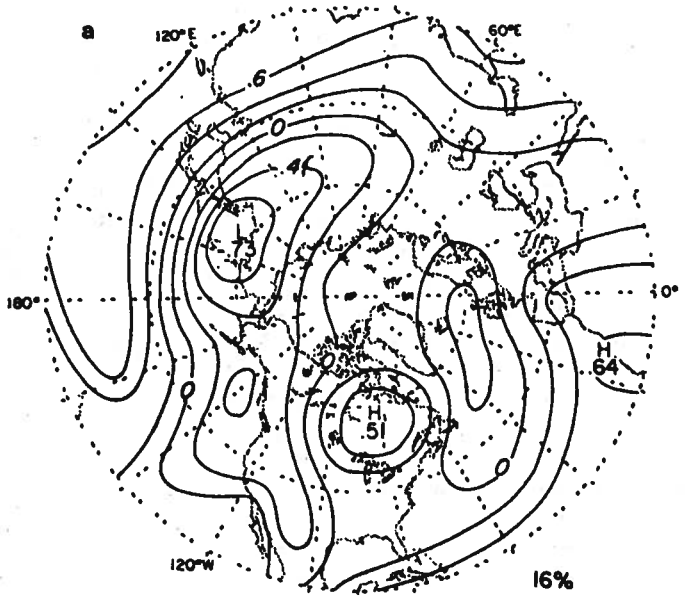

Fig. 9.5 Spatial displays of the first four eigenvectors of gridded winter monthly-mean 500-mb heights for the Northern Hemisphere, 1962–1977. This PCA was computed using the correlation matrix of the height data. Percentage values below and to the right of each map are proportion of total variance [Eq. (9.40)] × 100%. The patterns resemble the "teleconnectivity" patterns for the same data (Fig. 3.20). [From Wallace and Gutzler (1981), reproduced with permission of the American Meteorological Society.]

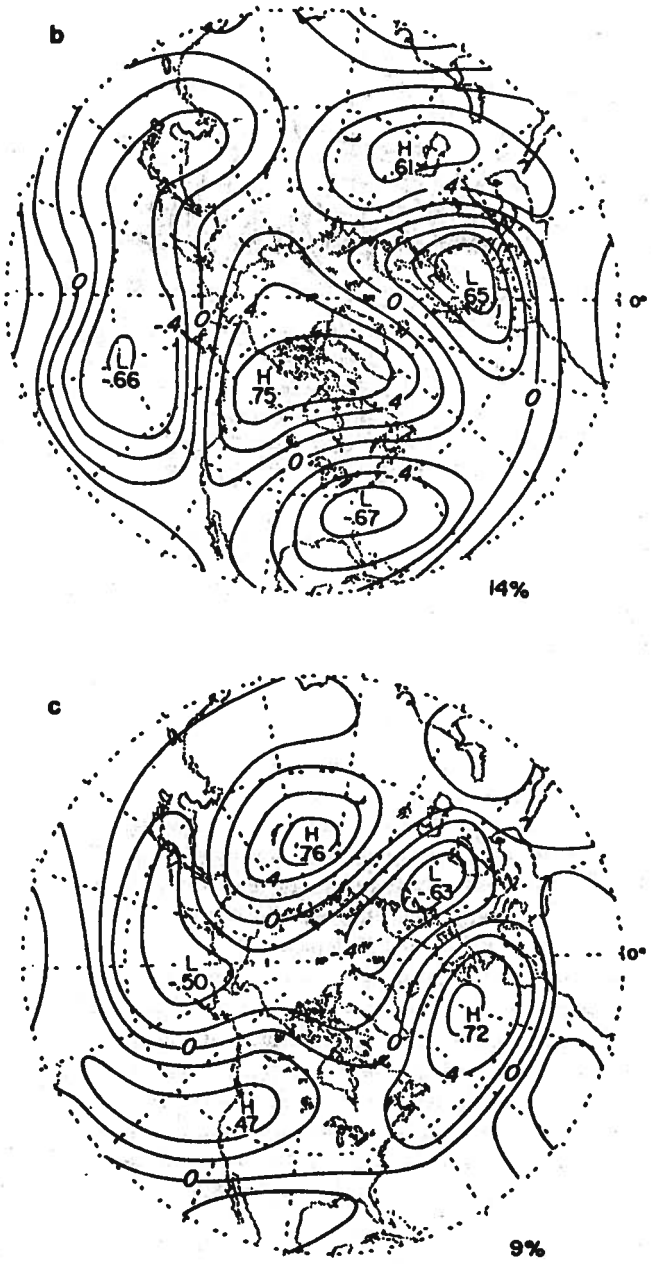

Fig. 9.5 *Continued.*

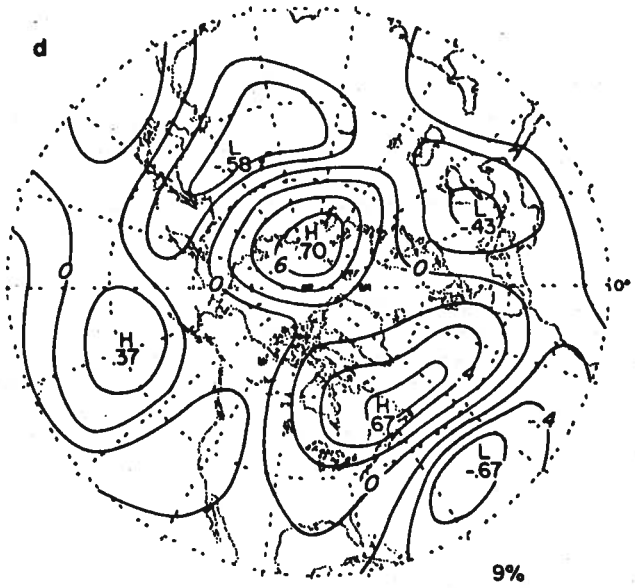

Fig. 9.5 *Continued.*

United States, and to positive height anomalies (ridges) the western part of the continent, and over the central tropical Pacific. A negative value of the second principal component yields the reverse pattern of anomalies, and a more zonal 500-mb flow over North America.

It is also possible to apply PCA to vector-valued fields, which are fields with observations of more than one variable at each location or gridpoint. This is equivalent to simultaneous analysis of two or more fields. If there are L such variables at each of the K gridpoints, then the dimensionality of the data vector $\mathbf{x}$ is given by the product KL. The first K elements of $\mathbf{x}$ are observations of the first variable, the second K elements are observations of the second variable, and the last K elements of $\mathbf{x}$ will be observations of the Lth variable. Since the L different variables will generally be measured in unlike units, it will almost always be desirable to base the PCA of such data on the correlation matrix. The dimension of [R], and of the matrix of eigenvectors [E], will then be $(M \times M)$, where $M = KL$. Application of PCA to this kind of correlation matrix will produce principal components successively maximizing the joint variance of the L variables in a way that considers the correlations both between and among these variables at the K locations.

Figure 9.6 illustrates the structure of the correlation matrix (left) and the matrix of eigenvectors (right) for PCA of vector field data. The first K rows of [R] contain

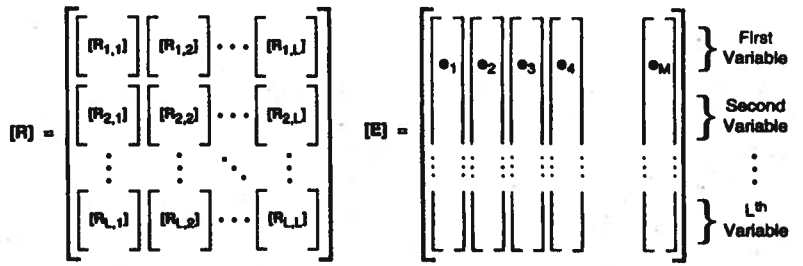

Fig. 9.6 Structures of the correlation matrix and the matrix of eigenvectors for PCA of vector field data. The basic data consist of multiple observations of L variables at each of K locations, so the dimensions of both [R] and [E] are $(M \times M)$, where $M = KL$. The correlation matrix consists of $(K \times K)$ submatrices containing the correlations between sets of the L variables jointly at the K locations. The submatrices located on the diagonal of [R] thus contain ordinary correlation matrices for each of the L variables. The off-diagonal submatrices contain correlation coefficients, but are not symmetrical and will not contain 1's on their diagonals. Each eigenvector column of [E] similarly consists of L segments, each of which contains K elements pertaining to the individual locations.

the correlations between the first of the L variables at these locations and all of the M variables. Rows $K + 1$ to $2K$ similarly contain the correlations between the second of the L variables and all the M variables, and so on. Another way to look at the correlation matrix is as a collection of L^2 submatrices, each dimensioned $(K \times K)$, which contain the correlations between sets of the L variables jointly at the K locations. The submatrices located on the diagonal of [R] thus contain ordinary correlation matrices for each of the L variables. The off-diagonal submatrices contain correlation coefficients, but are not symmetric and will not contain 1's on their diagonals. However, the overall symmetry of [R] implies that $[\mathbf{R}_{i,j}] = [\mathbf{R}_{j,i}]^T$. Similarly, each column of [E] consists of L segments, and each of these segments contains the K elements pertaining to each of the individual locations.

The eigenvector elements resulting from a PCA of a vector field can be displayed graphically in a manner that is similar to the maps drawn for ordinary, scalar fields. Here, each of the L groups of K eigenvector elements is either overlaid on the same base map, or plotted on separate maps. Figure 9.7, from Kutzbach (1967), illustrates this process for the case of $L = 2$ observations at each location. The two variables are average January surface pressure and average January temperature, measured at $K = 23$ locations in North America. The heavy lines are an analysis of the (first 23) elements of the first eigenvector that pertain to the pressure data, and the dashed lines with shading show an analysis of the temperature (second 23) elements of the same eigenvector. The corresponding principal component accounts for 28.6% of the total variance of the $23 \times 2 = 46$ variables.

In addition to effectively condensing an enormous amount of information, the patterns shown in Fig. 9.7 are physically consistent with atmospheric processes. In particular, the temperature anomalies are consistent with patterns of thermal

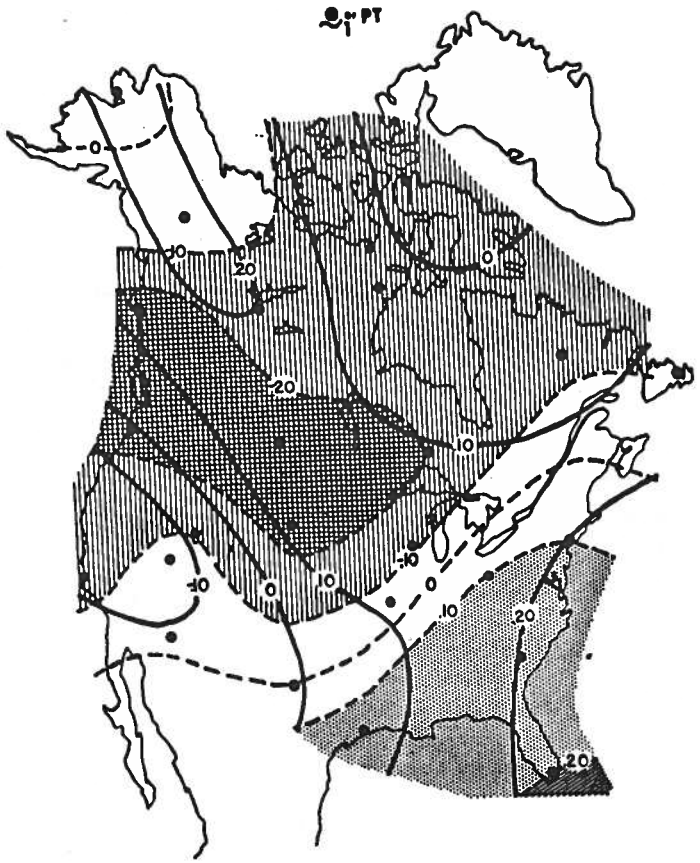

Fig. 9.7 Spatial display of the elements of the first eigenvector of the (46 × 46) correlation matrix of average January sea-level pressures and temperatures at 23 locations in North America. The first principal component of this correlation matrix accounts for 28.6% of the joint (scaled) variance of the pressures and temperatures. Heavy lines are a hand analysis of the sea-level pressure elements of the first eigenvector, and dashed lines with shading are a hand analysis of the temperature elements of the same eigenvector. The joint variations of pressure and temperature depicted are physically consistent with temperature advection in response to the pressure anomalies. [From Kutzbach (1967), reproduced with permission of the American Meteorological Society.]

advection implied by the pressure anomalies. If the first principal component u_1 is positive for a particular month, the solid contours imply positive pressure anomalies in the north and east, with lower than average pressures in the southwest. On the west coast, this pressure pattern would result in weaker than average westerly surface winds and stronger than average northerly surface winds. The

resulting advection of cold air from the north would produce colder temperatures, and this cold advection is reflected by the negative temperature anomalies in this region. Similarly, the pattern of pressure anomalies in the southeast would enhance southerly flow of warm air from the Gulf of Mexico, resulting in positive temperature anomalies as shown. Conversely, if u_1 is negative, reversing the signs of the pressure eigenvector elements implies enhanced westerlies in the west, and northerly wind anomalies in the southeast, which are consistent with positive and negative temperature anomalies, respectively. These temperature anomalies are indicated by Fig. 9.7, when the signs on the temperature contours are also reversed.

Figure 9.7 is a simple example involving familiar variables. Its interpretation is easy and obvious if one is conversant with the climatological relationships of pressure and temperature patterns over North America in winter. However, the physical consistency exhibited in this example (where the "right" answer is known ahead of time) is indicative of the power of this kind of PCA to uncover meaningful joint relationships among atmospheric (and other) fields in an exploratory setting, where clues to possibly unknown underlying physical mechanisms may be hidden in the complex relationships among several fields.

9.3.6 Rotation of the Eigenvectors

When the PCA eigenvector elements are plotted geographically, there is a strong tendency to try to ascribe physical interpretations to the corresponding principal components. The results shown in Fig. 9.7 indicate that it can be both appropriate and informative to do so. However, the orthogonality constraint on the eigenvectors [Eq. (9.27)] can lead to problems with these interpretations, especially for the second and subsequent principal components. While the orientation of the first eigenvector in its M-dimensional space is determined solely by the direction of the maximum variation in the data, subsequent vectors must be orthogonal to previously determined eigenvectors, regardless of the nature of the physical processes that gave rise to the data. To the extent that the underlying physical processes are not independent, interpretation of the corresponding principal components as being independent "modes of variability" will not be justified. Although the first principal component may represent an important mode of variability or physical process, it may well also include aspects of other correlated modes or processes. Thus, the orthogonality constraint on the eigenvectors can result in the influences of several distinct physical processes being jumbled together in a single principal component.

When physical interpretation, rather than data compression, is a primary goal of PCA, it is often desirable to rotate a subset of the initial eigenvectors to a second set of new coordinate vectors. Usually it is some number M^* of the leading eigenvectors (i.e. eigenvectors with largest corresponding eigenvalues) of the original PCA that are rotated, with M^* determined using a truncation criterion such as

Eq. (9.41). The review article by Richman (1986) explains in some detail why rotation may be useful, and also gives a technical exposition of the subject.

As a consequence of rotation of the eigenvectors, a second set of new variables is produced, called *rotated principal components*. The rotated principal components are obtained from the original data in a manner similar to the PCA analysis equation, as the dot product of data vectors and the rotated eigenvectors. Depending on the method used to rotate the eigenvectors, the resulting rotated principal components may or may not be mutually uncorrelated.

A number of procedures for rotating the original eigenvectors exist, but all seek to produce what is known as *simple structure* in the resulting analysis. Roughly speaking, simple structure is achieved if a large number of the elements of the resulting rotated vectors are near zero, and few of the remaining elements correspond to (have the same index k as) elements that are also not near zero in the other rotated vectors. The result is that the rotated vectors represent mainly the few original variables corresponding the elements not near zero, and that the representation of the original variables is split between as few of the rotated principal components as possible.

Figure 9.8, from Horel (1981), shows spatial displays of the first two rotated eigenvectors of monthly-averaged hemispheric winter 500-mb heights. Using the truncation criterion of Eq. (9.41) with $T = 1$, the first 19 eigenvectors of the correlation matrix of these data were rotated. The two patterns in Fig. 9.8 are similar to the first two unrotated eigenvectors derived from the same data (Figs. 9.5a and 9.5b), although the signs have been (arbitrarily) reversed. However, the rotated vectors conform more to the idea of simple structure in that more of the hemispheric fields are fairly flat (near zero) in Fig. 9.8, and each panel emphasizes more uniquely a particular feature of the variability of the 500-mb heights corresponding to the teleconnection patterns in Fig. 3.20. The rotated vector in Fig. 9.8a focuses primarily on height differences in the northwestern and western tropical Pacific, called the *western Pacific* teleconnection pattern. It thus represents variations in the 500-mb jet at these longitudes, with positive values of the corresponding rotated principal component indicating weaker-than-average westerlies, and negative values indicating the reverse. Similarly, the PNA pattern stands out exceptionally clearly in Fig. 9.8b, where the rotation has separated it from the eastern hemisphere pattern evident in Fig. 9.5b.

9.3.7 The Varied Terminology of PCA

The subject of PCA is sometimes regarded as a difficult and confusing one, but much of this confusion derives from a proliferation of the associated terminology, especially in writings by analysts of atmospheric data. Table 9.3 organizes the more common of these in a way that may be helpful in deciphering the PCA literature.

Lorenz (1956) introduced the term *empirical orthogonal function* (EOF; men-

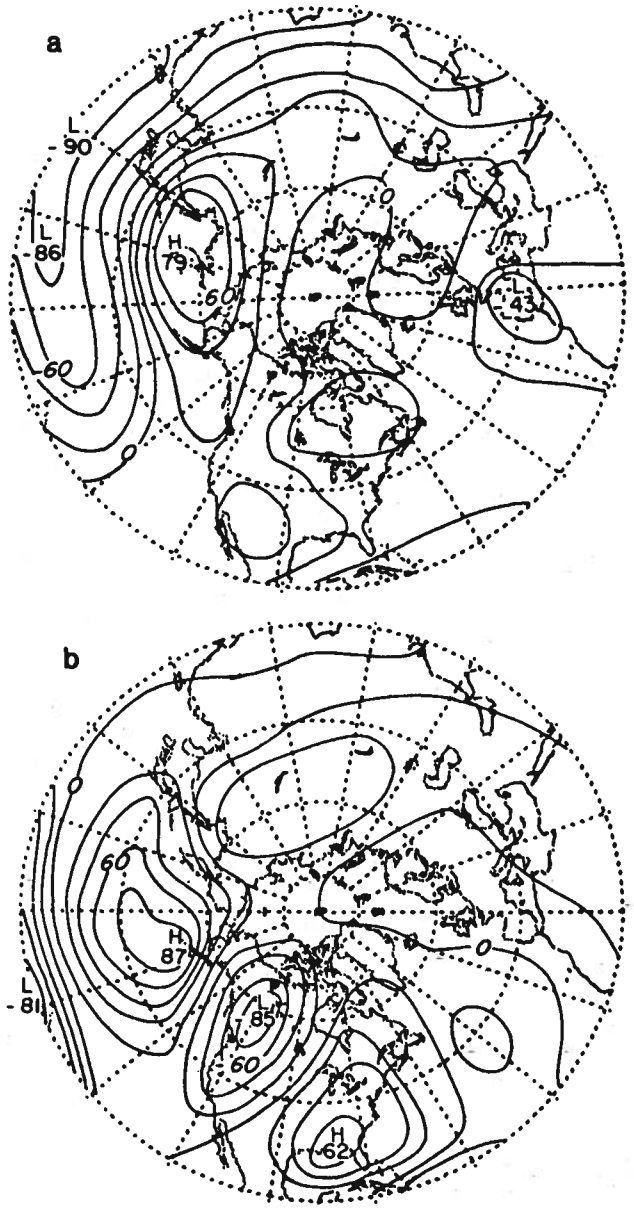

Fig. 9.8 Spatial displays of the first two rotated eigenvectors of monthly-averaged hemispheric winter 500-mb heights. The data is the same as that underlying Fig. 9.6, but the rotation has better isolated the patterns of variability, allowing a clearer interpretation in terms of the teleconnection patterns in Fig. 3.20. [From Horel (1981), reproduced with permission of the American Meteorological Society.]

Table 9.3
A Partial Guide to Synonymous Terminology Associated with PCA

Eigenvectors, $\mathbf{e}_m$	Eigenvector elements, $e_{k,m}$	Principal components, u_m	Principal-component elements, $u_{i,m}$
EOFs	Loadings	Empirical orthogonal variables	Scores
Modes of variation	Coefficients		Amplitudes
Pattern vectors	Pattern coefficients		Expansion coefficients
Principal axes	Empirical orthogonal weights		Coefficients
Principal vectors			
Proper functions			
Principal directions			

tioned earlier) into the literature as another name for the eigenvectors of a PCA. The terms *modes of variation* and *pattern vectors* are also used primarily by analysts of geophysical data, especially in relation to analysis of fields as described previously. The remaining terms for the eigenvectors derive from the geometric interpretation of the eigenvectors as basis vectors, or axes, in the K-dimensional space of the data. These terms are used in the literature of a broader range of disciplines.

The most common name for individual elements of the eigenvectors in the statistical literature is *loading*, connoting the weight of the kth variable x_k that is borne by the mth eigenvector $\mathbf{e}_m$ through the individual element e_{km}. The term *coefficient* is also a usual one in the statistical literature. The term *pattern coefficient* is used mainly in relation to PCA of field data, where the spatial patterns exhibited by the eigenvector elements can be illuminating. *Empirical orthogonal weights* is a term that is sometimes used to be consistent with the naming of the eigenvectors as EOFs.

The new variables u_m defined with respect to the eigenvectors are almost universally called *principal components*. However, they are sometimes known as *empirical orthogonal variables* when the eigenvectors are called EOFs. There is more variation in the terminology for the individual values of the principal components $u_{i,m}$ corresponding to particular data vectors $\mathbf{x}'_i$ [or, in the notation of Eq. (9.45), of $u_m(t)$ corresponding to a particular $\mathbf{x}'(t)$]. In the statistical literature these are most commonly called *scores*, which has a historical basis in the early and widespread use of PCA in psychometrics. In atmospheric applications, the principal component elements are often called *amplitudes* by analogy to the amplitudes of a Fourier series, that multiply the (theoretical orthogonal) sine and cosine functions. Similarly, the term *expansion coefficient* is also used for this meaning. Sometimes expansion coefficient is shortened simply to *coefficient*, although this can be the source of some confusion since it is more standard for the term *coefficient* to denote an eigenvector element.

9.3.8 Scaling Conventions in PCA

Another contribution to confusion in the literature of PCA is the existence of alternative scaling conventions for the eigenvectors. The presentation in this section has assumed that the eigenvectors are scaled to unit length: $\|\mathbf{e}_m\| = 1$. Recall that vectors of any length will satisfy Eq. (9.25) if they point in the appropriate direction, and as a consequence it is common for the output of eigenvector computations to be expressed with this scaling.

However, it is sometimes useful to express and manipulate PCA results using alternative scalings of the eigenvectors. When this is done, each element of an eigenvector is multiplied by the same constant, so their relative magnitudes and relationships remain unchanged. Thus, for example, the overall patterns in maps such as those shown in Figs. 9.5, 9.7, and 9.8 are unchanged under different scalings, and it is only the magnitudes of the contour labels that change. Therefore, the qualitative results of an exploratory analysis based on PCA do not depend on the scaling selected, but if different, related analyses are to be compared it is important to be aware of the scaling convention used in each.

Rescaling the lengths of the eigenvectors changes the magnitudes of the principal components correspondingly. That is, multiplying the eigenvector $\mathbf{e}_m$ by a constant requires that the principal component scores u_m be multiplied by the same constant in order for the analysis formulas that define the principal components [Eqs. (9.36) and (9.37)] to remain valid. The expected values of the principal component scores for centered data $\mathbf{x}'$ are zero, and multiplying the principal components by a constant will produce rescaled principal components whose means are also zero. However, their variances will change by a factor of the square of the scaling constant, and the magnitudes of their correlations with the original variables will be affected as well.

Table 9.4 summarizes the effects of three common scalings of the eigenvectors on the properties of the principal components. The first row indicates their properties under the scaling convention $\|\mathbf{e}_m\| = 1$ adopted in this presentation. The eigenvectors plotted in Fig. 9.8 conform to this scaling. Under this scaling, the expected value (mean) of each of the principal components is zero, and the variance of each is equal to the respective eigenvalue, λ_m. This result is simply an expression of the diagonalization of the variance–covariance matrix [Eq. (9.29)] produced by adopting the geometric coordinate system defined by the eigenvectors. When scaled in this way, the correlation between a principal component u_m and a variable x_k is proportional to the eigenvector element that connects them, $e_{k,m}$, with the constant of proportionality being the ratio of the square root of the corresponding eigenvalue $(\lambda_m)^{1/2}$ to the standard deviation of the variable s_k. The

Table 9.4

Three Common Eigenvector Scalings Used in PCA and Their Consequences for the Properties of the Principal Components u_m, and Relationship to Original Variables x_k and Standardized Original Variables z_k

Eigenvector scaling	$E[u_m]$	$\mathrm{Var}[u_m]$	$\mathrm{Corr}[u_m, x_k]$	$\mathrm{Corr}[u_m, z_k]$
$\|\mathbf{e}_m\| = 1$	0	λ_m	$e_{k,m}(\lambda_m)^{1/2}/s_k$	$e_{k,m}(\lambda_m)^{1/2}$
$\|\mathbf{e}_m\| = (\lambda_m)^{1/2}$	0	λ_m^2	$e_{k,m}/s_k$	$e_{k,m}$
$\|\mathbf{e}_m\| = (\lambda_m)^{-1/2}$	0	1	$e_{k,m}\lambda_m/s_k$	$e_{k,m}\lambda_m$

correlation between u_m and the standardized variable z_k is given by the product of the eigenvector element and the square root of the eigenvalue, since the standard deviation of a standardized variable is one.

The eigenvectors are commonly rescaled by multiplying each element by the square root of the corresponding eigenvalue. This rescaling produces vectors of differing lengths, $\|\mathbf{e}_m\| \equiv (\lambda_m)^{1/2}$, but that point in exactly the same directions as the original eigenvectors with unit lengths. The eigenvectors plotted in Figs. 9.5 and 9.7 conform to this convention. Consistency in the analysis formula requires that the principal components also be multiplied by $(\lambda_m)^{1/2}$, with the result that the variance of each u_m increases to λ_m^2. A major advantage of this rescaling, however, is that the eigenvector elements are more directly interpretable in terms of the relationship between the principal components and the original data. Under this rescaling, each eigenvector element $e_{k,m}$ is numerically equal to the correlation $r_{u,z}$ between the mth principal component u_m and the kth standardized variable z_k. Also, this scaling is usually applied if a subset of the eigenvectors are to be rotated.

The last scaling shown in Table 9.4, resulting in $\|\mathbf{e}_m\| \equiv (\lambda_m)^{-1/2}$, is less commonly used. This scaling is achieved by dividing each element of the original unit-length eigenvectors by the square root of the corresponding eigenvalue. The resulting expression for the correlations between the principal components and the original data are more awkward, but this scaling has the advantage that all principal components have equal, unit variance. This property can be useful in the detection of outliers.

9.4 Canonical Correlation Analysis

Canonical correlation analysis (CCA) is a statistical technique that identifies a sequence of pairs of patterns in two multivariate data sets, and constructs sets of transformed variables by projecting the original data onto these patterns. The approach thus bears some similarity to PCA, which searches for patterns within a single multivariate data set that represent maximum amounts of the variation in the data. In CCA, the patterns are chosen such that the new variables defined by projection of the two data sets onto these patterns exhibit maximum correlation, but are uncorrelated with the projections of the data onto any of the other identified patterns. That is, CCA identifies new variables that maximize the interrelationships between two data sets, in contrast to the patterns describing the internal variability within a single data set identified in PCA. It is this sense that CCA has been referred to as a "double-barreled PCA."

Canonical correlation analysis can also be viewed as an extension of multiple regression to the case of a vector-valued predictand variable $\mathbf{y}$. Ordinary multiple regression finds a weighted average, or pattern, of the vector of predictor variables $\mathbf{x}$ such that the correlation between the dot product $\mathbf{b}^T\mathbf{x}$ and the scalar predictand y is maximized. Here the elements of the vector $\mathbf{b}$ are the ordinary least-squares regression coefficients computed using the methods described in Chapter 6, and $\mathbf{b}^T\mathbf{x}$ is a new variable called the "predicted value of y," or $\hat{y}$. Canonical correlation analysis looks for pairs of sets of weights analogous to the regression coefficients, such that the correlations between the new variables defined by the respective dot products with $\mathbf{x}$ and (the vector) $\mathbf{y}$ are maximized.

As is also the case with PCA, CCA has been most widely applied to geophysical data in the form of fields. In this setting the vector $\mathbf{x}$ contains observations of one variable at a collection of gridpoints or locations, and the vector $\mathbf{y}$ contains observations of a different variable at a set of locations that may or may not be the same as those represented in $\mathbf{x}$. Typically the data consist of time series of observations of the two fields. When individual observations of the fields $\mathbf{x}$ and $\mathbf{y}$ are made simultaneously, a CCA can be useful in diagnosing aspects of the coupled variability of the two fields (e.g., Nicholls, 1987). When observations of $\mathbf{x}$ precede observations of $\mathbf{y}$ in time, the CCA may lead to statistical forecasts of the $\mathbf{y}$ field using the $\mathbf{x}$ field as a predictor (e.g., Barnston and Ropelewski, 1992). A more comprehensive comparison between CCA and PCA in the context of atmospheric data analysis is included in Bretherton *et al.* (1992).

9.4.1 Canonical Variates

A CCA transforms pairs of original data vectors $\mathbf{x}$ and $\mathbf{y}$ into sets of new variables, called *canonical variates*, v_m and w_m, defined by the dot products

$$v_m = \mathbf{a}_m^T\mathbf{x} = \sum_{i=1}^{I} a_{m,i}x_i, \qquad m = 1, \ldots, \min(I, J); \tag{9.46a}$$

and

$$w_m = \mathbf{b}_m^T\mathbf{y} = \sum_{j=1}^{J} b_{m,j}y_j, \qquad m = 1, \ldots, \min(I, J). \tag{9.46b}$$

This construction of the canonical variates is similar to that of the principal com-

ponents u_m [Eq. (9.36)], in that each is a linear combination (a sort of weighted average) of elements of the respective data vectors $\mathbf{x}$ and $\mathbf{y}$. Note that the data vectors and their corresponding vectors of weights, $\mathbf{a}_m$ and $\mathbf{b}_m$, need not have the same dimensions. The vectors $\mathbf{x}$ and $\mathbf{a}_m$ each have I elements, and the vectors $\mathbf{y}$ and $\mathbf{b}_m$ each have J elements. The number of pairs, M, of canonical variates that can be extracted from the two data sets is equal to the smaller of the dimensions of $\mathbf{x}$ and $\mathbf{y}$, i.e., $M = \min(I, J)$.

The properties of CCA are such that the canonical vectors $\mathbf{a}_m$ and $\mathbf{b}_m$ are chosen which result in the canonical variates having the properties

$$\text{Corr}[v_1, w_1] \geq \text{Corr}[v_2, w_2] \geq \cdots \geq \text{Corr}[v_M, w_M] \geq 0; \quad (9.47a)$$

$$\text{Corr}[v_k, w_m] = \begin{cases} r_{C_m}, & k = m \\ 0, & k \neq m \end{cases}; \quad (9.47b)$$

and

$$\text{Var}[v_m] = \text{Var}[w_m] = 1, \qquad m = 1, \ldots, M. \quad (9.47c)$$

Equation (9.47a) states that each of the M successive pairs of canonical variates exhibits a weaker correlation than the previous pair. These (Pearson product–moment) correlations between the pairs of canonical variates are called the *canonical correlations*, r_C. The canonical correlations can always be expressed as positive numbers, since either $\mathbf{a}_m$ or $\mathbf{b}_m$ can be multiplied by -1 if necessary. Equation (9.47b) states that each canonical variates is uncorrelated with all of the other canonical variates except its specific counterpart in the mth pair. Finally, Eq. (9.47c) states that each of the canonical variates has unit variance.

9.4.2 Computation of CCA

The information drawn upon by CCA is contained in the joint variance–covariance matrix of the variables $\mathbf{x}$ and $\mathbf{y}$. For purposes of computing this matrix, the two data vectors are concatenated into a single vector $\mathbf{c}^T = [\mathbf{x}^T, \mathbf{y}^T]$. This vector contains $I + J$ elements, the first I of which are the elements of $\mathbf{x}$, and the last J of which are the elements of $\mathbf{y}$. The $(I + J \times I + J)$ variance-covariance matrix of $\mathbf{c}$, $[\mathbf{S}_C]$, is then partitioned into four blocks, in a manner similar to that done for the correlation matrix in Fig. 9.6:

$$[\mathbf{S}_C] = \frac{1}{n-1}[\mathbf{C}']^T[\mathbf{C}'] = \begin{bmatrix} \underset{(I\times I)}{[\mathbf{S}_{xx}]} & \underset{(I\times J)}{[\mathbf{S}_{xy}]} \\ \underset{(J\times I)}{[\mathbf{S}_{yx}]} & \underset{(J\times J)}{[\mathbf{S}_{yy}]} \end{bmatrix}. \quad (9.48)$$

Each of the n rows of the $(n \times I + J)$ matrix $[\mathbf{C}']$ contains one observation of the vector $\mathbf{x}'$ and one observation of the vector $\mathbf{y}'$, where as before the primes indicate centering of the data by subtraction of each of the respective sample means. The

$(I \times I)$ matrix $[\mathbf{S}_{xx}]$ is the variance–covariance matrix of the I variables in $\mathbf{x}$. The $(J \times J)$ matrix $[\mathbf{S}_{yy}]$ is the variance–covariance matrix of the J variables in $\mathbf{y}$. The matrices $[\mathbf{S}_{xy}]$ and $[\mathbf{S}_{yx}]$ contain the covariances between each of the elements of $\mathbf{x}$ and each element of $\mathbf{y}$, and they are related according to $[\mathbf{S}_{xy}] = [\mathbf{S}_{yx}]^T$.

The canonical correlations r_C are given by the square roots of the nonzero eigenvalues of the matrices

$$[\mathbf{M}_x] = [\mathbf{S}_{xx}]^{-1}[\mathbf{S}_{xy}][\mathbf{S}_{yy}]^{-1}[\mathbf{S}_{yx}] \quad (9.49a)$$

and

$$[\mathbf{M}_y] = [\mathbf{S}_{yy}]^{-1}[\mathbf{S}_{yx}][\mathbf{S}_{xx}]^{-1}[\mathbf{S}_{xy}]. \quad (9.49b)$$

The matrix $[\mathbf{M}_x]$ is dimensioned $(I \times I)$, and the matrix $[\mathbf{M}_y]$ is dimensioned $(J \times J)$. The first $M = \min(I, J)$ eigenvalues of these two matrices will be identical, and if $I \neq J$, the remaining eigenvalues of the larger matrix will all be zero. Furthermore, the canonical vectors $\mathbf{a}_m$ and $\mathbf{b}_m$ are the respective eigenvectors of these matrices, satisfying

$$[\mathbf{M}_x]\mathbf{a}_m = r_{C_m}^2\mathbf{a}_m, \qquad m = 1, \ldots, M; \quad (9.50a)$$

and

$$[\mathbf{M}_y]\mathbf{b}_m = r_{C_m}^2\mathbf{b}_m, \qquad m = 1, \ldots, M. \quad (9.50b)$$

Here r_{C_m} is the mth canonical correlation. In general the matrices $[\mathbf{M}_x]$ and $[\mathbf{M}_y]$ are not symmetric, so that finding their eigenvalues and eigenvectors is more demanding computationally than the corresponding calculations required for PCA.

Canonical correlation analysis can be performed on standardized variables as well, in which case the dispersion matrix in Eq. (9.48) will be the $(I + J \times I + J)$ joint correlation matrix of the standardized variables. Basing the CCA on this correlation matrix yields exactly the same canonical correlations, but will produce different values for the canonical vectors $\mathbf{a}_m$ and $\mathbf{b}_m$.

Example 9.7. CCA of the January 1987 Temperature Data

A simple illustration of the mechanics of a small CCA can be provided by again analyzing the January 1987 temperature data for Ithaca and Canandaigua, New York, given in Table A.1. Let the $I = 2$ Ithaca temperature variables be $\mathbf{x} = [T_{max}, T_{min}]^T$, and similarly let the $J = 2$ Canandaigua temperature variables be $\mathbf{y}$. The joint covariance matrix $[\mathbf{S}_C]$ of these quantities is then the (4×4) matrix

$$[\mathbf{S}_C] = \begin{bmatrix} 59.516 & 75.433 & 58.070 & 51.697 \\ 75.433 & 185.467 & 81.633 & 110.800 \\ 58.070 & 81.633 & 61.847 & 56.119 \\ 51.697 & 110.800 & 56.119 & 77.581 \end{bmatrix}. \quad (9.51)$$

This symmetric matrix contains the sample variances of the four variables on the diagonal, and the covariances between the variables in the other positions. This matrix is related to the corresponding elements of the correlation matrix involving the same variables (Table 3.3) through the square roots of the diagonal elements: Dividing each element by the square roots of the diagonal elements in its row and column produces the corresponding correlation matrix. This is the operation shown in Eq. (9.20).

Since $I = J$ in this example, the matrices required to compute $[\mathbf{M}_x]$ and $[\mathbf{M}_y]$ using Eq. (9.49) are obtained by "quartering" $[\mathbf{S}_C]$, to yield

$$[\mathbf{S}_{xx}] = \begin{bmatrix} 59.516 & 75.433 \\ 75.433 & 185.467 \end{bmatrix}, \tag{9.52a}$$

$$[\mathbf{S}_{yy}] = \begin{bmatrix} 61.847 & 56.119 \\ 56.119 & 77.581 \end{bmatrix}, \tag{9.52b}$$

and

$$[\mathbf{S}_{yx}] = \begin{bmatrix} 58.070 & 81.633 \\ 51.697 & 110.800 \end{bmatrix}, \tag{9.52c}$$

with the fourth (upper-right) matrix given by the transpose of Eq. (9.52c): $[\mathbf{S}_{xy}] = [\mathbf{S}_{yx}]^T$. Applying Eq. (9.49) yields

$$[\mathbf{M}_x] = \begin{bmatrix} .8304 & .3771 \\ .0682 & .7004 \end{bmatrix} \tag{9.53a}$$

and

$$[\mathbf{M}_y] = \begin{bmatrix} .8452 & .2589 \\ .0910 & .6856 \end{bmatrix}. \tag{9.53b}$$

The $M = 2$ eigenvectors of $[\mathbf{M}_x]$, $\mathbf{a}_1$ and $\mathbf{a}_2$, and the corresponding eigenvectors of $[\mathbf{M}_y]$, $\mathbf{b}_1$ and $\mathbf{b}_2$, are shown in Table 9.5. The first element of each pertains to the respective maximum temperature variable, and the second elements pertain to the minimum temperature variables. The first eigenvalue of these two matrices is $\lambda_1 = 0.5020$, yielding the first canonical correlation as $\sqrt{\lambda_1} = 0.709$. The second canonical correlation of 0.038 is so small that this canonical pair would not be of practical interest.

The time series of the first pair of canonical variables is given by the dot products of $\mathbf{a}_1$ and $\mathbf{b}_1$ with the pairs of temperature values for Ithaca and Canandaigua, respectively, from Table A.1. The value of v_1 for January 1 would be constructed as (33)(.9613) + (19)(.2753) = 36.95. The time series of v_1 (pertaining to the Ithaca temperatures) would consist of 31 values (one for each day): 36.95, 37.46, 34.90, 27.60, 25.13, . . . , 31.87, 39.02. Similarly, the time series for w_1 (pertaining to the Canandaigua temperatures) is 41.47, 43.35, 37.03, 33.72, 33.64, . . . , 33.91,

Table 9.5

Canonical Vectors $\mathbf{a}_m$ (Corresponding to Ithaca Temperatures) and $\mathbf{b}_m$ (Corresponding to Canandaigua Temperatures) for Partition of Covariance Matrix in Eq. (9.51) with $I = J = 2$[a]

	$\mathbf{a}_1$ (Ithaca)	$\mathbf{b}_1$ (Canandaigua)	$\mathbf{a}_2$ (Ithaca)	$\mathbf{b}_2$ (Canandaigua)
T_{max}	.9613	.9409	−.8457	−.7155
T_{min}	.2753	.3386	−.7155	.6986
λ_m	0.5020		0.00139	
$r_{C_m} = \sqrt{\lambda_m}$	0.709		0.038	

[a] These vectors are the eigenvectors of the matrices in Eq. (9.53). Also shown are the eigenvalues λ_m of $[\mathbf{M}_x]$ and $[\mathbf{M}_y]$, and the canonical correlations which are their square roots.

43.54. The first canonical correlation coefficient, $r_{C_m} = 0.709$, is the correlation between this first pair of canonical variables, v_1 and w_1, and is the largest possible correlation between pairs of linear combinations of these two data sets. ¤

9.4.3 CCA Applied to Fields

Canonical correlation analysis is usually most interesting for atmospheric data when applied to fields. Here the spatially distributed observations (either at gridpoints or observing locations) are encoded into the vectors $\mathbf{x}$ and $\mathbf{y}$ in the same way as for PCA. That is, even though the data are drawn from a two- or three-dimensional field, each location is numbered sequentially and pertains to one element of the corresponding data vector. It is not necessary for the spatial domains encoded into $\mathbf{x}$ and $\mathbf{y}$ to be the same, and, indeed, in the applications of CCA that have appeared in the literature they are usually different.

As is the case with the use of PCA with spatial data, it is often informative to plot maps of the canonical vectors by associating the magnitudes of their elements and the geographic locations to which they pertain. In this context the canonical vectors are sometimes called *canonical patterns*, since the resulting maps show spatial patterns of the ways in which the original variables contribute to the canonical variables. Examining the pairs of maps formed by corresponding vectors $\mathbf{a}_m$ and $\mathbf{b}_m$ can be informative about the nature of the relationship between variations in the data over the two domains encoded in $\mathbf{x}$ and $\mathbf{y}$, respectively.

Figure 9.9, from Wallace *et al.* (1992), shows one such pair of canonical patterns. Figure 9.9a shows the spatial distribution of the elements of a canonical vector $\mathbf{a}$, pertaining to a data vector $\mathbf{x}$ that contains values of average December–January sea–surface temperatures (SSTs) in the north Pacific Ocean. This pattern accounts for 18% of the total variance of the SSTs in the data set analyzed. Figure 9.9b shows the spatial distribution of the elements of the corresponding canonical vector $\mathbf{b}$, which pertains to average hemispheric 500-mb heights $\mathbf{y}$ for

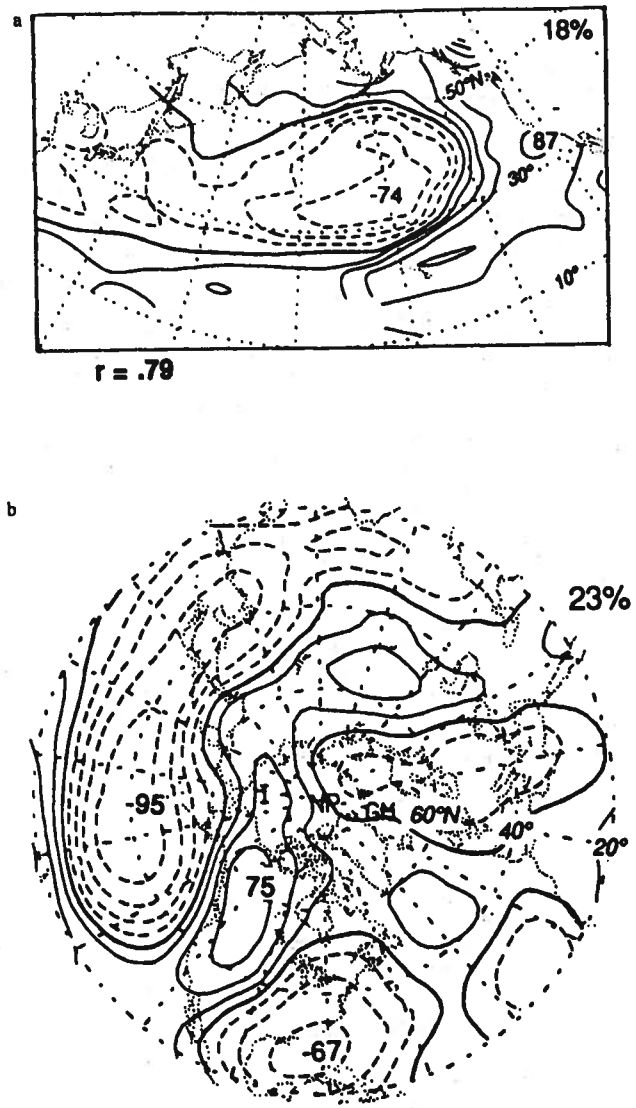

Fig. 9.9 Spatial displays of a pair of canonical vectors pertaining to average winter sea–surface temperatures (SSTs) in the northern Pacific Ocean (a), and hemispheric winter 500-mb heights (b). The pattern of SST anomalies in panel a (and its negative) are associated with the PNA pattern of 500-mb height anomalies shown in panel b. The correlation between SSTs projected onto panel a and 500-mb heights projected onto panel b is the canonical correlation 0.79. [From Wallace *et al.* (1992), reproduced with permission of the American Meteorological Society.]

the same winters included in the SST data in **x**. This pattern accounts for 23% of the total variance of the winter hemispheric height variations. The sea-surface temperature pattern described by the canonical vector on the left is one of either cold water in the central north Pacific and warm water along the west coast of North America, or warm water in the central north Pacific and cold water along the west coast of North America. The pattern of 500-mb height variations in Fig. 9.8b is remarkably similar to the PNA pattern (cf. Figs. 9.8b and 3.20). Taken together, these figures indicate that cold SSTs in the central Pacific simultaneously with warm SSTs in the northeast Pacific tend to coincide with a 500-mb ridge over northwestern North America and a 500-mb trough over southeastern North America. Similarly, warm water in the central north Pacific and cold water in the northwestern Pacific are associated with the more zonal PNA flow. The time series of the canonical variable v is obtained by projecting (i.e., forming the dot products of) the time series of SSTs **x** onto the vector **a** shown in Fig. 9.9a, and the time series of the canonical variable w results from the 500 mb heights **y** being projected onto the vector **b** shown in Fig. 9.9b. The correlation between the two time series v and w is the canonical correlation $r_C = 0.79$.

Figure 9.10, from Barnett and Preisendorfer (1987), illustrates the application of CCA to atmospheric fields where the observations **x** and **y** have not been made simultaneously. Here **y** is average January surface temperature over the United States, and **x** is average sea-level pressures over much of the Northern Hemisphere, during the preceding October, November, and December. The sea-level pressure fields each consisted of 280 gridpoints, so each vector **x** consists of $I = (3)(280) = 840$ elements. Here, the first 280 elements of **x** contain the October pressures, the second 280 elements contain the November pressures, and the third group of 280 elements contain the December pressures. Essentially the same approach to simultaneous consideration of multiple fields has been taken as in the PCA results shown in Fig. 9.7, with data from several fields being concatenated into a single vector.

However, the essential feature of the construction of the CCA shown in Fig. 9.10 is that the **x** fields all precede the **y** field, so that, at least in principle, the CCA can be used to forecast **y**, if **x** is known. The forecast is constructed using the pattern vector shown in Fig. 9.10a to forecast the pattern in Fig. 9.10b. That is, these maps indicate that there is a tendency for the pressure anomaly patterns shown in Fig. 9.10a to be followed by the temperature patterns shown in Fig. 9.10b. As before, these maps actually specify two sets of patterns, depending on which sign is applied. They indicate that low December pressures in the north Pacific and western Atlantic, with high December pressures over western North America (with similar but weaker patterns during November and October) have a tendency to be followed by relatively warm January temperatures in western North America and relatively cool January temperatures in eastern North America. In years when December pressures are relatively high over the north Pacific

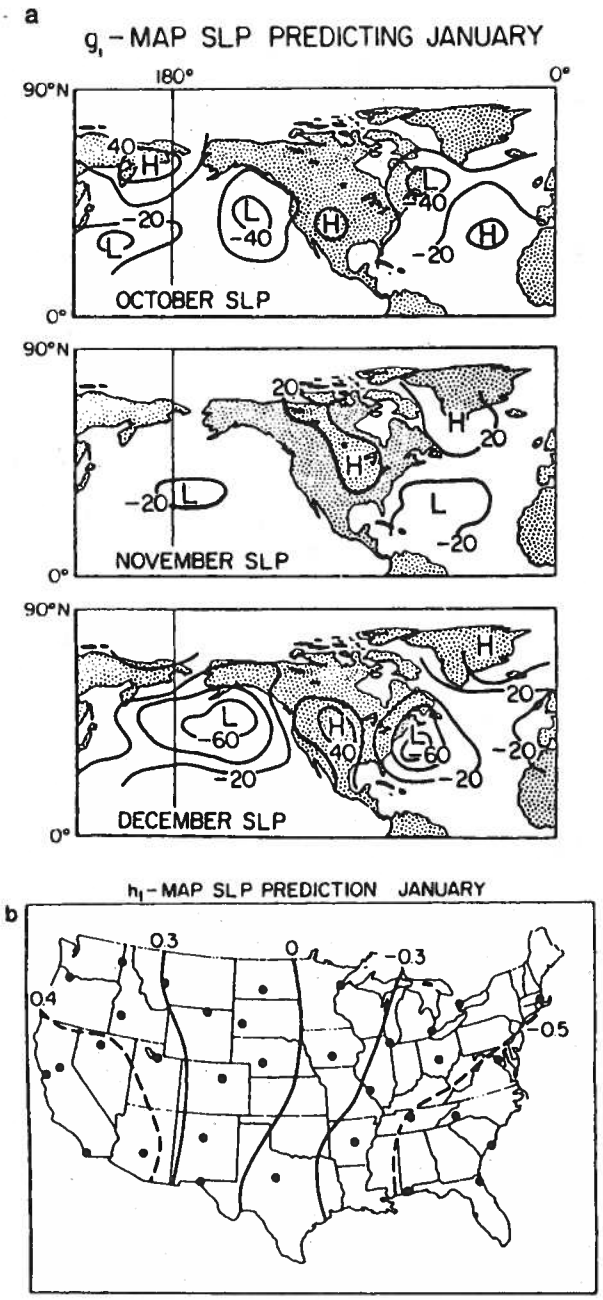

Fig. 9.10 Spatial displays of a pair of canonical vectors pertaining to average sea-level pressures (SLPs) in October, November, and December (a) and average temperatures over the United States during the following January (b). Since the sea-level pressures in **x** are observed prior to the surface temperatures **y**, the link between the two fields made by the CCA may be useful for temperature forecasts. [From Barnett and Preisendorfer (1987), reproduced with permission of the American Meteorological Society.]

and western Atlantic, and are relatively low over western North America, these maps indicate that North American temperatures in the following January tend to be relatively cool in the west and relatively warm in the east.

When using CCA in an exploratory or diagnostic analysis, it may not be of interest to reconstruct contributions to the original data made by a set of canonical patterns. However, when used as part of a forecasting procedure, it is necessary operationally to reconstruct estimated data values predicting the **y** field using some number of the canonical vectors $\mathbf{b}_m$. Usually it is more convenient to work with the fields of anomalies, $\mathbf{x}'$ and $\mathbf{y}'$, but note that the covariance matrices for the anomalies ($[\mathbf{S}_{x'x'}]$ and $[\mathbf{S}_{y'y'}]$) and the covariance matrices for the original data ($[\mathbf{S}_{xx}]$ and $[\mathbf{S}_{yy}]$) are identical. This "synthesis" process is in general more difficult than the corresponding procedure in PCA [Eqs. (9.38) and (9.39)], because the canonical vectors are not orthogonal. However, a synthesis equation can be derived for CCA by manipulating the matrix "analysis" equation

$$\underset{(n \times M^*)}{[\mathbf{W}]} = \underset{(n \times J)}{[\mathbf{Y}']} \underset{(J \times M^*)}{[\mathbf{B}]} \tag{9.54}$$

Here $[\mathbf{Y}']$ is the full matrix of predictand anomaly values; each of the n rows of this matrix contains one data record $\mathbf{y}'^{\mathrm{T}}$. The matrix $[\mathbf{B}]$ has as its columns some number M^* (not necessarily the full number M) of the canonical vectors **b** pertaining to $\mathbf{y}'$, and $[\mathbf{W}]$ is the matrix of the corresponding M^* canonical scores. Equation (9.54) can be solved for $[\mathbf{Y}']$ by multiplying on the right by $[\mathbf{B}]^{\mathrm{T}}$, and then multiplying on the right again by $([\mathbf{B}][\mathbf{B}]^{\mathrm{T}})^{-1}$:

$$[\mathbf{W}][\mathbf{B}]^{\mathrm{T}} = [\mathbf{Y}'][\mathbf{B}][\mathbf{B}]^{\mathrm{T}} \tag{9.55a}$$

$$\begin{aligned} [\mathbf{W}][\mathbf{B}]^{\mathrm{T}}([\mathbf{B}][\mathbf{B}]^{\mathrm{T}})^{-1} &= [\mathbf{Y}'][\mathbf{B}][\mathbf{B}]^{\mathrm{T}}([\mathbf{B}][\mathbf{B}]^{\mathrm{T}})^{-1} \\ &= [\mathbf{Y}'][\mathbf{I}] \\ &= [\mathbf{Y}']. \end{aligned} \tag{9.55b}$$

The relationship between $[\mathbf{V}]$, $[\mathbf{X}']$, and $[\mathbf{A}]$ is analogous. The forecast quantities are recovered from the anomalies simply by adding back the appropriate mean value. If the CCA has been computed using standardized variables, so that Eq. (9.48) is a correlation matrix, the dimensional values of the forecast variables need to be reconstructed by multiplying by the appropriate standard deviation and adding the appropriate mean [i.e., by reversing Eq. (3.16) or (4.20) to yield Eq. (4.22)].

9.4.4 Combining PCA and CCA

In practical applications it is sometimes useful to "prefilter" the two fields of raw data using PCA before subjecting them to a CCA. Thus, rather than directly correlating the fields $\mathbf{x}$ and $\mathbf{y}$, the combined analysis operates on the fields $\mathbf{u}_x$ and $\mathbf{u}_y$, which consist of the first principal components of $\mathbf{x}$ and $\mathbf{y}$. That is, separate PCAs are computed for each of the two fields, and these two analyses may be truncated at different numbers of principal components M_x^* and M_y^*.

This combined approach is not always best, and can be inferior if important information is discarded when truncating the PCA. However, combining CCA with PCA is attractive in circumstances where there are large numbers I and/or J of observations on the original variables, especially if the $\mathbf{x}$ and $\mathbf{y}$ fields are strongly spatially correlated. Using a PCA to reduce the dimensionality of these original fields reduces the number of computations necessary to carry out the CCA while retaining most of the original information. In instances where the sample size, n, is not large compared to I and J, preprocessing the data with a PCA also tends to improve the stability (i.e., sampling properties) of the CCA. This stability consideration is especially important when a forecasting application is contemplated, since good performance on independent data is required.

9.5 Discriminant Analysis

The final two sections in this chapter deal with classification of multivariate data. Given a collection of data vectors $\mathbf{x}$ consisting of K elements, it is desired to separate these into groups on the basis of the individual vector elements x_k. If the groups are specified in advance, the classification problem is one of discriminating among the known groups. Here it is necessary to have available a "training sample" in which it is known that each of the vectors is correctly classified, in order to build rules according to which future observations $\mathbf{x}$ of unknown group origin can be similarly classified. This is the problem of discriminant analysis, presented in this section.

If the criterion according to which the observations $\mathbf{x}$ have been grouped pertains to a time after the observation of $\mathbf{x}$, then the discriminant analysis can be employed as a forecasting procedure. Here each group represents a discrete outcome, and the forecast consists of a categorical statement that one of these outcomes will occur. The forecast is made by "classifying" the current observation $\mathbf{x}$ as belonging to the group that is forecast to occur. Examples of the use of discriminant analysis for forecasting can be found in Lawson and Cerveny (1985), Miller (1962), and Ward and Folland (1991).

For some data sets it is not known to which group each observation belongs, or possibly even how many groups exist. In this case discriminant analysis cannot be applied. However, degrees of similarity and difference among different vector observations $\mathbf{x}$ can be used to construct groups on the basis only of the data at hand.

That is, one looks to see how the data points "clump" or cluster in their K-dimensional space. This is the problem of cluster analysis, presented in Section 9.6.

9.5.1 Linear Discriminant Analysis

The simplest form of discriminant analysis involves distinguishing between two groups on the basis of a K-dimensional vector of observations $\mathbf{x}$. A training sample must exist, consisting of n_1 observations of $\mathbf{x}$ known to have come from group 1, and n_2 observations of $\mathbf{x}$ known to have come from group 2. That is, the basic data are two data matrices, $[\mathbf{X}_1]$, dimensioned $(n_1 \times K)$, and $[\mathbf{X}_2]$, dimensioned $(n_2 \times K)$. The goal is to find a linear function of the K elements of the observation vector, called the *discriminant function*, that will best allow a future K-dimensional vector of observations $\mathbf{y}$ to be classified as belonging to either group 1 or group 2.

Each of the two groups is characterized by a K-dimensional mean vector

$$\overline{\mathbf{x}}_g = \frac{1}{n_g}[\mathbf{X}_g]^{\mathrm{T}}\mathbf{1} = \begin{bmatrix} \dfrac{1}{n_g}\displaystyle\sum_{i=1}^{n_g} x_{i,1} \\ \dfrac{1}{n_g}\displaystyle\sum_{i=1}^{n_g} x_{i,2} \\ \vdots \\ \dfrac{1}{n_g}\displaystyle\sum_{i=1}^{n_g} x_{i,K} \end{bmatrix}, \qquad g = 1, 2; \tag{9.56}$$

where $\mathbf{1}$ is a $(n \times 1)$ vector containing only 1's. That is, the K elements of each of the two mean vectors are computed by averaging each of the K elements of the data vectors $\mathbf{x}$. The averaging in Eq. (9.56) is done separately for the observations belonging to groups 1 and 2.

Nearly always in discriminant analysis, the assumption is made that the populations underlying each of the groups has the same covariance matrix. That is, the mean vectors characterizing each group may be different (discrimination between groups would be very difficult if they were not), but the nature of the dispersion of the data points around these mean vectors is assumed to be the same for each group. The sample covariance matrices, $[\mathbf{S}_1]$ and $[\mathbf{S}_2]$, are computed from the data matrices, $[\mathbf{X}_1]$ and $[\mathbf{X}_2]$, using Eqs. (9.18) and (9.19). Since the covariance structures are assumed to be equal, these two variance–covariance matrices are averaged to yield a pooled estimate of the dispersion of the data around their means,

$$[\mathbf{S}_{\text{pooled}}] = \frac{(n_1 - 1)[\mathbf{S}_1] + (n_2 - 1)[\mathbf{S}_2]}{(n_1 + n_2 - 2)}. \tag{9.57}$$

If $n_1 = n_2$, Eq. (9.57) results in each element of $[\mathbf{S}_{\text{pooled}}]$ being constructed as the simple average of the corresponding elements of $[\mathbf{S}_1]$ and $[\mathbf{S}_2]$.

The fundamental idea behind linear discriminant analysis is to find a direction $\mathbf{d}_1$ in the K-dimensional space of the data, such that the distance between the two mean vectors is maximized when the data are projected onto $\mathbf{d}_1$. Finding this direction reduces the discrimination problem from one of sifting through the relationships among the K elements of the data vectors to looking at a single number. That is, the data vector $\mathbf{x}$ is transformed to a new variable, $\delta_1 = \mathbf{d}_1^T\mathbf{x}$, known as *Fisher's linear discriminant function*. The groups of K-dimensional multivariate data are essentially reduced to groups of univariate data with different means, distributed along the $\mathbf{d}_1$ axis. The discriminant vector locating this direction of maximum separation is given by

$$\mathbf{d}_1 = [\mathbf{S}_{\text{pooled}}]^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2), \tag{9.58a}$$

or, if a unit vector is more convenient,

$$\mathbf{d}_1 = \frac{[\mathbf{S}_{\text{pooled}}]^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)}{\|[\mathbf{S}_{\text{pooled}}]^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)\|}. \tag{9.58b}$$

Either of the scalings in Eq. (9.58), or any other constant multiple of $\mathbf{d}_1$, can be used equally well, since all of these vectors will point in the same direction.

The decision to classify a future observation $\mathbf{y}$ as belonging to either group 1 or group 2 can now be made according to the value of the scalar $\delta_1 = \mathbf{d}_1^T\mathbf{y}$. This dot product is a one-dimensional (i.e., scalar) projection of the vector $\mathbf{y}$ onto the direction of maximum separation, $\mathbf{d}_1$. The discriminant function δ_1 is essentially a new variable, analogous to the new variable u in PCA and the new variables v and w in CCA, produced as a linear combination of the elements of a data vector $\mathbf{y}$. In the simplest application, the observation $\mathbf{y}$ is classified as belonging to group 1 if the projection $\mathbf{d}_1^T\mathbf{y}$ is closer to the projection of the group 1 mean onto the direction $\mathbf{d}_1$, and is classified as belonging to group 2 if $\mathbf{d}_1^T\mathbf{y}$ is closer to the projection of the mean of group 2.

Along the $\mathbf{d}_1$ axis, the midpoint between the means of the two groups is given by the projection of the average of these two mean vectors onto the vector $\mathbf{d}_1$:

$$\hat{\delta}_1 = \mathbf{d}_1^T\frac{\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2}{2}. \tag{9.59}$$

This value defines the dividing line between values of the discriminant function $\mathbf{d}_1^T\mathbf{y}$ for a future observation $\mathbf{y}$ that would result in its being assigned to either group 1 or group 2. That is, a future observation $\mathbf{y}$ whose group membership is unknown can be assigned to a group according to the rule

$$\text{Assign } \mathbf{y} \text{ to group 1 if } \quad \mathbf{d}_1^T\mathbf{y} - \hat{\delta}_1 \geq 0, \tag{9.60a}$$

or

$$\text{Assign } \mathbf{y} \text{ to group 2 if } \quad \mathbf{d}_1^T\mathbf{y} - \hat{\delta}_1 < 0. \tag{9.60b}$$

Example 9.8. A Linear Discriminant Analysis for $K = 2$-Dimensional Data

Table 9.6 shows average July temperature and precipitation for stations in three regions of the United States. The data vectors are composed of $K = 2$ elements each: one temperature element and one precipitation element. Consider the problem of distinguishing between membership in group 1 and group 2. This problem might arise if the stations in Table 9.6 represented the "core" portions of their respective climatic regions, and on the basis of these data one wanted to classify stations not listed in this table as belonging to one or the other of these two groups.

The mean vectors for the $n_1 = 10$ and $n_2 = 9$ data vectors in groups 1 and 2 are

$$\bar{\mathbf{x}}_1 = \begin{bmatrix} 80.6°\text{F} \\ 5.67 \text{ in.} \end{bmatrix} \quad \text{and} \quad \bar{\mathbf{x}}_2 = \begin{bmatrix} 78.7°\text{F} \\ 3.57 \text{ in.} \end{bmatrix}, \tag{9.61a}$$

and the two sample variance–covariance matrices are

$$[\mathbf{S}_1] = \begin{bmatrix} 1.47 & 0.65 \\ 0.65 & 1.45 \end{bmatrix} \quad \text{and} \quad [\mathbf{S}_2] = \begin{bmatrix} 2.08 & 0.06 \\ 0.06 & 0.17 \end{bmatrix}. \tag{9.61b}$$

Since $n_1 \neq n_2$, the pooled estimate for the common variance–covariance matrix is obtained by the weighted average specified by Eq. (9.57). The direction $\mathbf{d}_1$ pointing in the direction of maximum separation of the two sample mean vectors is then computed using Eq. (9.58a) as

$$\begin{aligned} \mathbf{d}_1 &= \begin{bmatrix} 1.76 & 0.37 \\ 0.37 & 0.85 \end{bmatrix}^{-1}\left(\begin{bmatrix} 80.6 \\ 5.67 \end{bmatrix} - \begin{bmatrix} 78.7 \\ 3.57 \end{bmatrix}\right) \\ &= \begin{bmatrix} 0.625 & -0.272 \\ -0.272 & 1.295 \end{bmatrix}\begin{bmatrix} 1.9 \\ 2.10 \end{bmatrix} = \begin{bmatrix} 0.62 \\ 2.20 \end{bmatrix}. \end{aligned} \tag{9.62}$$

Figure 9.11 illustrates the geometry of this problem. Here the data for the warmer and wetter southeastern stations of group 1 are plotted as circles, and the central U.S. stations of group 2 are plotted as X's. The vector means for the two groups are also indicated by the heavy symbols. The projections of these two means onto the direction $\mathbf{d}_1$ are indicated by the lighter dashed lines. The midpoint between these two projections locates the dividing point between the two groups in the one-dimensional discriminant space defined by $\mathbf{d}_1$. The heavy dashed line perpendicular to the discriminant function at this point divides the (temperature, precipitation) plane into two regions. Future points $\mathbf{y}$ of unknown group membership falling above and to the right of this heavy line will be classified as belonging to group 1, and points falling below and to the left will be classified as belonging to group 2.

Since the average of the mean vectors for groups 1 and 2 is $[79.65, 4.62]^T$, the value of the dividing point is $= (0.62)(79.65) + (2.20)(4.62) = 59.55$. Of the 19 points in this training data, only that for Atlanta has been misclassified. For this

Table 9.6

Average July Temperature (Temp., °F) and Precipitation (Ppt., in.) for Locations in Three Regions of the United States[a]

Group 1: southeastern USA			Group 2: central USA			Group 3: northeastern USA		
Station	Temp	Ppt	Station	Temp	Ppt	Station	Temp	Ppt
Athens, GA	79.2	5.18	Concordia, KS	79.0	3.37	Albany, NY	71.4	3.00
Atlanta, GA	78.6	4.73	Des Moines, IA	76.3	3.22	Binghamton, NY	68.9	3.48
Augusta, GA	80.6	4.40	Dodge City, KS	80.0	3.08	Boston, MA	73.5	2.68
Gainesville, FL	80.8	6.99	Kansas City, MO	78.5	4.35	Bridgeport, CT	74.0	3.46
Huntsville, AL	79.3	5.05	Lincoln, NE	77.6	3.20	Burlington, VT	69.6	3.43
Jacksonville, FL	81.3	6.54	Springfield, MO	78.0	3.58	Hartford, CT	73.4	3.09
Macon, GA	81.4	4.46	St. Louis, MO	78.9	3.63	Portland, ME	68.1	2.83
Montgomery, AL	81.7	4.78	Topeka, KS	78.6	4.04	Providence, RI	72.5	3.01
Pensacola, FL	82.3	7.18	Wichita, KS	81.4	3.62	Worcester, MA	69.9	3.58
Savannah, GA	81.2	7.37						
Averages	80.6	5.67		78.7	3.57		71.3	3.17

[a] Averages are for the period 1951–1980. [From Quayle and Presnell (1991).]

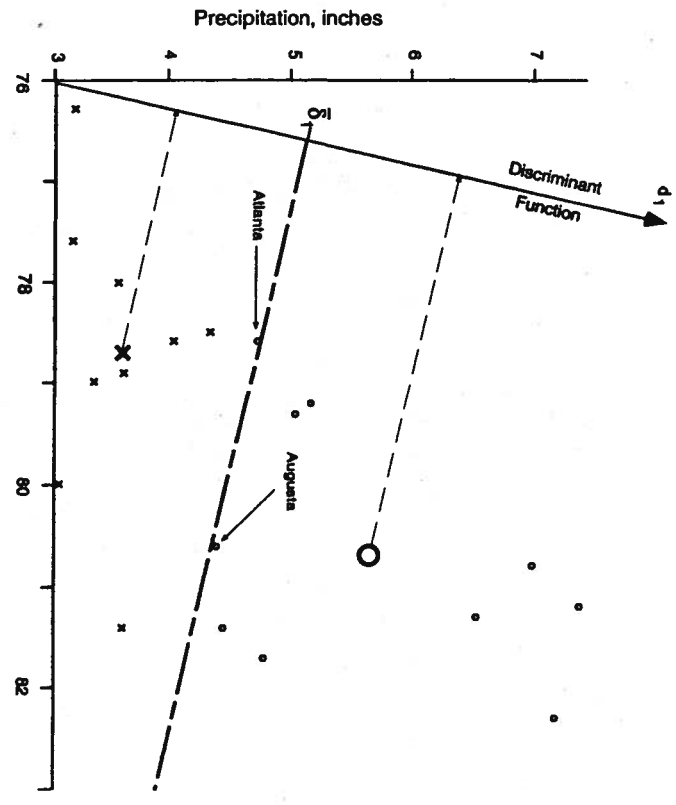

Fig. 9.11 Geometry of linear discriminant analysis applied to the southeastern (circles) and central (Xs) U.S. data in Table 9.6. The (vector) means of the two groups of data are indicated by the heavy symbols, and their projections onto the discriminant function are indicated by the light dashed arrows. The midpoint between these two projections, δ_1, defines the dividing line (heavier dashed line) used to assign future (temperature, precipitation) pairs to the groups. Of these training data, only the data point for Atlanta would be misclassified. Note that the discriminant function has been shifted to the right (i.e., does not pass through the origin) in order to improve the clarity of the plot, but that this does not affect the relative positions of the projections of the data points onto it.

station, $\delta_1 = \mathbf{d}^T\mathbf{x} = (0.62)(78.6) + (2.20)(4.73) = 59.14$. Since this value of δ_1 is slightly less than the midpoint value, Atlanta would be falsely classified as belonging to group 2 [cf. Eq. (9.60)]. By contrast, the point for Augusta lies just to the "group 1 side" of the heavy dashed line. For Augusta, $\delta_1 = \mathbf{d}^T\mathbf{x} = (0.62)(80.6) + (2.20)(4.40) = 59.65$, which is slightly greater than the cutoff value.

Consider now the assignment of two stations not listed in Table 9.6 to either group 1 or group 2. For New Orleans, LA the average July temperature is 82.1°F, and the average July precipitation is 6.73 in. Applying Eq. (9.60), one finds $\mathbf{d}^T\mathbf{y} - \delta_1 = (0.62)(82.1) + (2.20)(6.73) - 59.55 = 65.71 - 59.55 = 6.16 > 0$. Therefore, New Orleans would be classified as belonging to group 1. Similarly,

the average July temperature and precipitation for Columbus, OH are 74.7°F and 3.37 in., respectively. For this station, $\mathbf{d}_1^T\mathbf{y} - \delta_1 = (0.62)(74.7) + (2.20)(3.37) - 59.55 = 53.73 - 59.55 = -5.82 < 0$, which would result in Columbus being classified as belonging to group 2. ¤

Example 9.8 was constructed with $K = 2$ observations in each data vector in order to allow the geometry of the problem to be represented in two dimensions. However, it is not necessary to restrict the use of discriminant analysis to situations with only bivariate observations. In fact, discriminant analysis is potentially most powerful when allowed to operate on higher-dimensional data. For example, it would be possible to extend Example 9.8 to classifying stations according to average temperature and precipitation for all 12 months. If this were done, each data vector $\mathbf{x}$ would consist of $K = 24$ values. However, the discriminant vector $\mathbf{d}_1$ would also consist of $K = 24$ elements, and the dot product $\delta_1 = \mathbf{d}_1^T\mathbf{y}$ would still be a single scalar that could be used to classify the group membership of $\mathbf{y}$.

Usually high-dimensional data vectors of atmospheric data exhibit substantial correlation among the K elements, and thus carry some redundant information. For example, the 12 monthly mean temperatures and 12 monthly mean precipitation values for a location do not provide 24 independent pieces of information. If only for computational economy, it can be a good idea to reduce the dimensionality of this kind of data before subjecting it to a discriminant analysis. This reduction in dimension is most commonly achieved through a principal-component analysis (Section 9.3). Since the groups in discriminant analysis are assumed to have the same covariance structure, it is natural to perform the PCA on the estimate of their common variance–covariance matrix, $[\mathbf{S}_{\text{pooled}}]$. If the data vectors are not of consistent units (e.g., some temperatures and some precipitation amounts), it will make more sense to perform the PCA on the corresponding correlation matrix. The discriminant analysis can then be carried out using M^*-dimensional data vectors containing elements that are the first M^* principal components, rather than the original K-dimensional raw data vectors. The resulting discriminant function will then pertain to the principal components, $\mathbf{u}$, rather than to the original data, $\mathbf{x}$. In addition, if the first two principal components account for a large fraction of the total variance, the data can be visualized in a plot such as that in Fig. 9.11, where the horizontal and vertical axes are the first two principal components.

The point on the discriminant function between the projections of the two sample means is not always the best point at which to make the separation between groups. One might have prior information that the probability of membership in group 1 is higher than that for group 2, perhaps because group 2 members are rather rare overall. If this is so, it would usually be desirable to move the classification boundary toward the group 2 mean, with the result that more future observations $\mathbf{y}$ would be classified as belonging to group 1. Similarly, if misclassifying a group 1 data value as belonging to group 2 were to be a more serious error than

misclassifying a group 2 data value as belonging to group 1, one would again want to move the boundary toward the group 2 mean. If p_1 is the prior probability (the probability according to previous information) that the observation $\mathbf{y}$ belongs to group 1, p_2 is the prior probability that the observation $\mathbf{y}$ belongs to group 2, $L(1 \mid 2)$ is the loss, or penalty, incurred when a group 1 member is incorrectly classified as part of group 2, and $L(2 \mid 1)$ is the loss incurred when a group 2 member is incorrectly classified as part of group 1, the classification rule in Eq. (9.60) can be revised to yield

$$\text{Assign } \mathbf{y} \text{ to group 1 if } \mathbf{d}_1^T\mathbf{y} - \bar{\delta}_1 \geq \ln\left[\frac{L(1 \mid 2)}{L(2 \mid 1)}\frac{p_2}{p_1}\right], \tag{9.63a}$$

or

$$\text{Assign } \mathbf{y} \text{ to group 2 if } \mathbf{d}_1^T\mathbf{y} - \bar{\delta}_1 < \ln\left[\frac{L(1 \mid 2)}{L(2 \mid 1)}\frac{p_2}{p_1}\right]. \tag{9.63b}$$

Note that if the misclassification losses are equal, the adjustment involves only the prior probabilities. If the prior probabilities are equal, the adjustment involves only the losses. If both the losses and the prior probabilities are equal, Eq. (9.63) reduces to Eq. (9.60).

9.5.2 Multiple Discriminant Analysis

The generalization of linear discriminant analysis to the case of more than two groups is called *multiple discriminant analysis*. Here the basic problem is to allocate a K-dimensional data vector $\mathbf{y}$ to one of $G > 2$ groups on the basis of one or more discriminant functions. The discriminant functions are computed on the basis of a training set of G data matrices $[\mathbf{X}_1], [\mathbf{X}_2], [\mathbf{X}_3], \ldots, [\mathbf{X}_G]$ dimensioned, respectively, $(n_g \times K)$. A sample variance-covariance matrix can be computed from each of the G sets of data, $[\mathbf{S}_1], [\mathbf{S}_2], [\mathbf{S}_3], \ldots, [\mathbf{S}_G]$, according to Eq. (9.19).

As in the two-group linear discriminant analysis, it is commonly assumed that the G groups are drawn from populations with different mean vectors, but having the same variance–covariance matrix. This common variance–covariance matrix is estimated by the weighted-average pooled estimate

$$[\mathbf{S}_{\text{pooled}}] = \frac{1}{n - G}\sum_{g=1}^{G}(n_g - 1)[\mathbf{S}_g], \tag{9.64}$$

where there are n_g observations in each group, and the total sample size is

$$n = \sum_{g=1}^{G} n_g. \tag{9.65}$$

The estimated common variance–covariance matrix in Eq. (9.64) is sometimes

called the *within-groups variance matrix*. Equation (9.57) is a special case of Eq. (9.64), with $G = 2$.

Computation of multiple discriminant functions also requires calculation of the *between-groups variance matrix*

$$[\mathbf{S}_B] = \frac{1}{G-1}\sum_{g=1}^{G}(\bar{\mathbf{x}}_g - \bar{\mathbf{x}}_.)(\bar{\mathbf{x}}_g - \bar{\mathbf{x}}_.)^T, \tag{9.66}$$

where

$$\bar{\mathbf{x}}_g = \frac{1}{n_g}[\mathbf{X}_g]^T\mathbf{1} \tag{9.67a}$$

is the sample mean vector of the gth group, and

$$\bar{\mathbf{x}}_. = \frac{1}{n}\sum_{g=1}^{G} n_g\bar{\mathbf{x}}_g \tag{9.67b}$$

is the grand, or overall vector mean of all n observations. The matrix $[\mathbf{S}_B]$ is essentially a variance–covariance matrix describing the dispersion of the G sample means around the overall mean [cf. Eq. (9.19)].

The number of discriminant functions that can be computed is the smaller of $G - 1$ and K. Thus for the two-group case discussed in the previous section, there is only $G - 1 = 1$ discriminant function, regardless of the dimensionality K of the data vectors. In the more general case, the discriminant functions are derived from the eigenvectors of the $(K \times K)$ matrix

$$[\mathbf{D}] = [\mathbf{S}_{pooled}]^{-1}[\mathbf{S}_B] \tag{9.68}$$

corresponding to the $M = \min(G - 1, K)$ nonzero eigenvalues. Usually computer routines for calculating eigenvectors will scale the results to unit length (i.e., $\|\mathbf{e}_m\| = 1$), whereas the discriminant vectors $\mathbf{d}_m$ are conventionally scaled such that

$$\mathbf{d}_m^T[\mathbf{S}_{pooled}]\mathbf{d}_m = 1. \tag{9.69}$$

This scaling is achieved by multiplying the unit-length eigenvectors $\mathbf{e}_m$ by constants c_m,

$$\mathbf{d}_m = c_m\mathbf{e}_m, \tag{9.70a}$$

where

$$c_m = (\mathbf{e}_m^T[\mathbf{S}_{pooled}]\mathbf{e}_m)^{-1/2}. \tag{9.70b}$$

The first discriminant function is then the dot product $\delta_1 = \mathbf{d}_1^T\mathbf{x}$, the second discriminant function is $\delta_2 = \mathbf{d}_2^T\mathbf{x}$, . . . , and the Mth discriminant function is $\delta_M = \mathbf{d}_M^T\mathbf{x}$. The M eigenvectors $\mathbf{d}_m$ define an M-dimensional *discriminant space*, in which the G groups of data exhibit maximum separation. The projections δ_m of the data onto these vectors are sometimes called the *discriminant coordinates* or *canonical variates*. This second appellation is unfortunate and a cause of confusion, since these do not pertain to canonical correlation analysis.

As was also the case when distinguishing between $G = 2$ groups, observations $\mathbf{y}$ can be assigned to groups according to which of the G group means is closest in discriminant space. For the $G = 2$ case the discriminant space is one-dimensional, consisting only of a line. The group assignment rule in Eq. (9.60) is then particularly simple. More generally, it is necessary to evaluate the Euclidean distances between the candidate vector $\mathbf{y}$ and each of the G group means in order to find which is closest. It is actually easier to evaluate these in terms of squared distances, yielding the classification rule

Assign $\mathbf{y}$ to group g if

$$\sum_{m=1}^{M}[\mathbf{d}_m(\mathbf{y} - \bar{\mathbf{x}}_g)]^2 \le \sum_{m=1}^{M}[\mathbf{d}_m(\mathbf{y} - \bar{\mathbf{x}}_h)]^2, \quad \text{for all } h \ne g. \tag{9.71}$$

That is, the sum of the squared distances between $\mathbf{y}$ and each of the group means, along the directions defined by the vectors $\mathbf{d}_m$, are compared in order to find the closest group mean.

Example 9.9. Multiple Discriminant Analysis with G = 3 Groups

Consider discriminating among all three groups of data in Table 9.6. Using Eq. (9.64), the pooled estimate of the common variance-covariance matrix is

$$\begin{aligned}[\mathbf{S}_{pooled}] &= \frac{1}{28-3}\left(9\begin{bmatrix}1.47 & 0.65\\ 0.65 & 1.45\end{bmatrix} + 8\begin{bmatrix}2.08 & 0.06\\ 0.06 & 0.17\end{bmatrix}\right.\\ &\qquad \left. + 8\begin{bmatrix}4.85 & -0.17\\ -0.17 & 0.10\end{bmatrix}\right)\\ &= \begin{bmatrix}2.75 & 0.20\\ 0.20 & 0.61\end{bmatrix},\end{aligned} \tag{9.72a}$$

and using Eq. (9.66) the between-groups variance matrix is

$$\begin{aligned}[\mathbf{S}_B] &= \frac{1}{2}\left(\begin{bmatrix}12.96 & 5.33\\ 5.33 & 2.19\end{bmatrix} + \begin{bmatrix}2.89 & -1.05\\ -1.05 & 0.38\end{bmatrix} + \begin{bmatrix}32.49 & 5.81\\ 5.81 & 1.04\end{bmatrix}\right)\\ &= \begin{bmatrix}24.17 & 5.04\\ 5.04 & 1.81\end{bmatrix}.\end{aligned} \tag{9.72b}$$

The directions of the two discriminant functions are specified by the eigenvectors of the matrix

$$\begin{aligned}[\mathbf{D}] &= [\mathbf{S}_{pooled}]^{-1}[\mathbf{S}_B]\\ &= \begin{bmatrix}0.373 & -0.122\\ -0.122 & 1.685\end{bmatrix}\begin{bmatrix}24.17 & 5.04\\ 5.04 & 1.81\end{bmatrix} = \begin{bmatrix}8.40 & 1.65\\ 5.54 & 2.43\end{bmatrix},\end{aligned} \tag{9.73a}$$

which, when scaled according to Eq. (9.69), are

$$\mathbf{d}_1 = \begin{bmatrix} 0.542 \\ 0.415 \end{bmatrix} \quad \text{and} \quad \mathbf{d}_2 = \begin{bmatrix} -0.282 \\ 1.230 \end{bmatrix}. \tag{9.73b}$$

The eigenvectors $\mathbf{d}_1$ and $\mathbf{d}_2$ define the directions of the first discriminant function $\delta_1 = \mathbf{d}_1^T\mathbf{x}$ and the second discriminant function $\delta_2 = \mathbf{d}_2^T\mathbf{x}$. Figure 9.12 shows the data for all three groups in Table 9.6 plotted in the discriminant space defined by these two functions. Points for groups 1 and 2 are shown by circles and X's, as in Fig. 9.11, and points for group 3 are shown by +'s. The heavy symbols locate the respective vector means for the three groups. Note that the point clouds for groups 1 and 2 appear to be stretched and distorted relative to their arrangement in Fig. 9.11. This is because the matrix [D] is not symmetric, so that the two discriminant vectors in Eq. (9.73b) are not orthogonal.

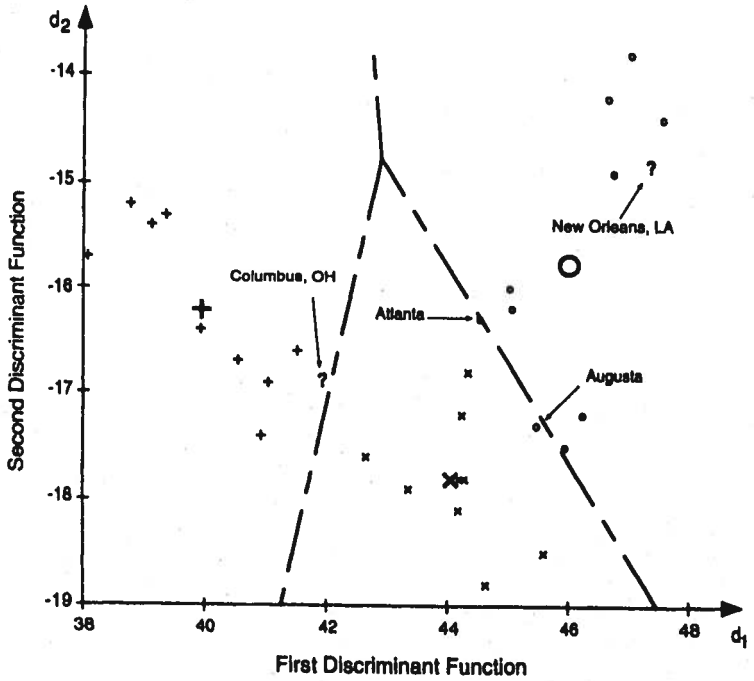

Fig. 9.12 Geometry of multiple discriminant analysis applied to the three groups of data in Table 9.6. Group 1 stations are plotted as circles, group 2 stations are plotted as Xs, and group 3 stations are plotted as +'s. The three vector means are indicated by the corresponding heavy symbols. The two axes are the first and second discriminant functions, and the heavy dashed lines divide sections of this discriminant space allocated to each group. The data for Atlanta and Augusta are misclassified as belonging to group 2. The two stations Columbus and New Orleans, which are not part of the training data in Table 9.6, are shown as question marks, and allocated to groups 3 and 1, respectively.

The heavy dashed lines in Fig. 9.12 divide the portions of the discriminant space that are assigned to each of the three groups by the classification criterion in Eq. (9.71). Here the data for Atlanta and Augusta have both been misclassified as belonging to group 2 rather than group 1. For Atlanta, for example, the squared distance to the group 1 mean is $[.542(78.6 - 80.6) + .415(4.73 - 5.67)]^2 + [-.282(78.6 - 80.6) + 1.230(4.73 - 5.67)]^2 = 2.52$, and the squared distance to the group 2 mean is $[.542(78.6 - 78.7) + .415(4.73 - 3.57)]^2 + [-.282(78.6 - 78.7) + 1.230(4.73 - 3.57)]^2 = 2.31$. A line in this discriminant space could be drawn visually that would include these two stations in the group 1 region. That the discriminant analysis has not specified this line is probably a consequence of the assumption of equal covariance matrices not being well satisfied.

The data points for the two stations Columbus and New Orleans, which are not part of the training data in Table 9.6, are shown by the question marks in Fig. 9.12. The location in the discriminant space of the point for New Orleans is $\delta_1 = (.542)(82.1) + (.415)(6.73) = 47.3$ and $\delta_2 = (-.282)(82.1) + (1.230)(6.73) = -14.9$, which is within the region assigned to group 1. The coordinates in discriminant space for the Columbus data are $\delta_1 = (.542)(74.7) + (.415)(3.37) = 41.9$ and $\delta_2 = (-.282)(74.7) + (1.230)(3.37) = -16.9$, which is within the region assigned to group 3. ¤

9.6 Cluster Analysis

Cluster analysis deals with separating data into groups whose identities are not known in advance. In general, even the "correct" number of groups into which the data should be sorted is not known in advance. Rather, it is the degree of similarity and difference between individual observations $\mathbf{x}$ that are used to define the groups, and to assign group membership. Examples of use of cluster analysis in the climatological literature include grouping daily weather observations into synoptic types (Kalkstein *et al.*, 1987), defining weather regimes from upper-air flow patterns (Mo and Ghil, 1988; Molteni *et al.*, 1990), grouping members of forecast ensembles (Tracton and Kalnay, 1993), grouping regions of the tropical oceans on the basis of ship observations (Wolter, 1987), and defining climatic regions based on surface climate variables (DeGaetano and Shulman, 1990; Fovell and Fovell, 1993; Galliani and Filippini, 1985; Guttman, 1993).

Cluster analysis is primarily an exploratory data analysis tool, rather than an inferential tool. Given a sample of data vectors $\mathbf{x}$ contained in the rows of a $(n \times K)$ data matrix [X], the procedure will define groups and assign group membership at varying levels of aggregation. Unlike discriminant analysis, the procedure does not contain rules for assigning membership to future observations $\mathbf{y}$. However, a cluster analysis can bring out groupings in the data that might otherwise be overlooked, possibly leading to empirically useful stratification of the data, or helping to suggest physical bases for observed structure in the data.

Most commonly implemented cluster analysis procedures are *hierarchical*. That is, they build a hierarchy of sets of groups, each of which is constructed by merging pairs of previously defined groups. The procedure begins by considering that the n observations of $\mathbf{x}$ have no group structure or, equivalently, that the data set consists of n groups containing one observation each. The first step is to find the two groups (i.e., data vectors) that are closest in their K-dimensional space, and to combine them into a new group. There are now $n - 1$ groups, one of which has two members. On each subsequent step, the two groups that are closest are merged to form a larger group. This process continues until, on the final, $(n - 1)$st, step all n observations have been aggregated into a single group.

The n-group clustering at the beginning of this process and the one-group clustering at the end of this process are neither useful nor enlightening. Hopefully, however, a natural clustering of the data into a workable number of informative groups will emerge at some intermediate stage. That is, one hopes that the n data vectors cluster or "clump together" in their K-dimensional space into some number G, $1 < G < n$, groups that reflect similar data generating processes. The ideal result is a division of the data that minimizes differences between members of a given cluster, and maximizes differences between members of different clusters.

9.6.1 Distance Measures and Clustering Methods

Central to the idea of the clustering of data points is the idea of distance. Clusters should consist of points separated by small distances, relative to the distances between clusters. The most intuitive and commonly used distance measure in cluster analysis is the Euclidean distance [Eq. (9.6)] in the K-dimensional space of the data vectors. Thus, the distance between two points $\mathbf{x}_i$ and $\mathbf{x}_j$ is

$$d_{i,j} = \|\mathbf{x}_i - \mathbf{x}_j\| = \left[\sum_{k=1}^{K} (x_{i,k} - x_{j,k})^2\right]^{1/2}. \tag{9.74}$$

Euclidean distance is by no means the only available choice for measuring distance between points or clusters. For example, the squared distance $d_{i,j}^2$ might also be used. The angle between pairs of vectors [Eq. (9.7)] is another possible choice, as are the many other alternatives presented in Romesburg (1984). Tracton and Kalnay (1993) have used the anomaly correlation [Eq. (7.39)] to group members of forecast ensembles, and the ordinary Pearson correlation is sometimes used as a clustering criterion as well. These latter two criteria are inverse distance measures, which should be maximized within groups, and minimized between groups.

While distances between pairs of points can be unambiguously defined, there are a number of alternatives for defining the distance between clusters comprised of groups of points. The criterion used to define cluster-to-cluster (intercluster) distances essentially defines the method of clustering. A few of the most common clustering methods are listed below. Most of these can be used with distance measures other than that in Eq. (9.74) as well.

- *Single-linkage*, or minimum-distance clustering. Here the distance between clusters G_1 and G_2 is the smallest Euclidean distance between one member of G_1 and one member of G_2. Formally,

$$d_{G_1,G_2} = \min_{i \in G_1, j \in G_2} [d_{i,j}]. \tag{9.75}$$

- *Complete-linkage*, or maximum-distance clustering groups data points on the basis of the largest distance between points in the two groups G_1 and G_2,

$$d_{G_1,G_2} = \max_{i \in G_1, j \in G_2} [d_{i,j}]. \tag{9.76}$$

- *Average-linkage* clustering defines cluster-to-cluster distance as the average Euclidean distance between all possible pairs of points in the two groups being compared. If G_1 contains n_1 points and G_2 contains n_2 points, this distance measure is

$$d_{G_1,G_2} = \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} d_{i,j}. \tag{9.77}$$

- *Centroid clustering* compares distances between the centroids, or vector averages, of pairs of clusters. According to this measure, the distance between G_1 and G_2 is

$$d_{G_1,G_2} = \|\bar{\mathbf{x}}_{G_1} - \bar{\mathbf{x}}_{G_2}\|, \tag{9.78}$$

where the vector means are as defined in Eq. (9.67a).

- *Ward's minimum variance method* merges that pair of clusters that will result in the minimum sum of squared distances between the points and the centroids of their respective groups, summed over the resulting groups. That is, for all possible ways of merging two of $G + 1$ groups to make G groups, that merger is made that minimizes

$$W = \sum_{g=1}^{G} \sum_{i=1}^{n_g} \|\mathbf{x}_i - \bar{\mathbf{x}}_g\|^2 = \sum_{g=1}^{G} \sum_{i=1}^{n_g} \sum_{k=1}^{K} (x_{i,k} - \bar{x}_{g,k})^2, \tag{9.79}$$

where the centroid, or group mean, for the newly merged group is recomputed using the data for both of the previously separate groups, before the squared distances are calculated.

Figure 9.13 illustrates single-linkage, complete-linkage, and centroid clustering for two hypothetical groups G_1 and G_2 in a $K = 2$-dimensional space. The open circles denote data points, of which there are $n_1 = 2$ in G_1 and $n_2 = 3$ in G_2. The centroids of the two groups are indicated by the solid circles. The single-linkage distance between G_1 and G_2 is the distance $\|\mathbf{x}_2 - \mathbf{x}_3\|$ between the closest pair of points in the two groups. The complete-linkage distance is that between the most distant pair, $\|\mathbf{x}_1 - \mathbf{x}_5\|$. The centroid distance is the Euclidean distance between

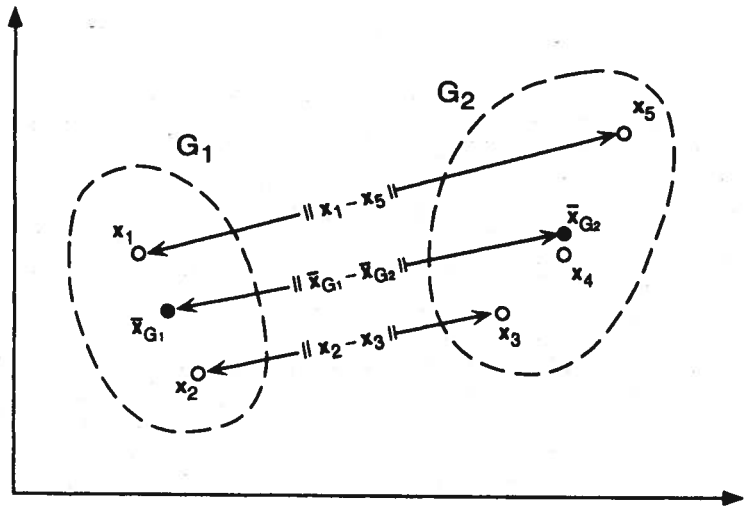

Fig. 9.13 Three measures of the distance in $K = 2$-dimensional space, between a cluster G_1 containing the two elements $\mathbf{x}_1$ and $\mathbf{x}_2$, and a cluster G_2 containing the elements $\mathbf{x}_3$, $\mathbf{x}_4$, and $\mathbf{x}_5$. The data points are indicated by open circles, and centroids of the two groups are indicated by the solid circles. According to the maximum-distance, or complete-linkage, criterion, the distance between the two groups is $\|\mathbf{x}_1 - \mathbf{x}_5\|$, or the greatest distance between all six possible pairs of points in the two groups. The minimum-distance, or single-linkage, criterion computes the distance between the groups as equal to the distance between the nearest pair of points, or $\|\mathbf{x}_2 - \mathbf{x}_3\|$. According to the centroid method, the distance between the two clusters is the distance between the sample means of their respective points.

the two vector means $\|\mathbf{x}_{G_1} - \mathbf{x}_{G_2}\|$. The average-linkage distance can also be visualized in Fig. 9.13, as the average of the six possible distances between individual members of G_1 and G_2: $(\|\mathbf{x}_1 - \mathbf{x}_5\| + \|\mathbf{x}_1 - \mathbf{x}_4\| + \|\mathbf{x}_1 - \mathbf{x}_3\| + \|\mathbf{x}_2 - \mathbf{x}_5\| + \|\mathbf{x}_2 - \mathbf{x}_4\| + \|\mathbf{x}_2 - \mathbf{x}_3\|)/6$.

The results of a cluster analysis can depend strongly on which definition is chosen for the distances between clusters. Single-linkage clustering is susceptible to "chaining," or the production of a few large clusters, which are formed by virtue of nearness of pairs of points at opposite edges of clusters to be merged. At the other extreme, complete-linkage clusters tend to be more numerous, as the criterion for merging clusters is more stringent. Average-distance clustering appears to be intermediate between these two extremes. Kalkstein *et al.* (1987) compared average-linkage, centroid clustering, and Ward's method for clustering surface weather observations. They report that centroid clustering exhibited chaining, Ward's method tended to produce groups of roughly equal size, and that average-linkage clustering seemed to be the most satisfactory.

The average-distance criterion appears to be the most commonly used approach to clustering, but none of the available methods are guaranteed to be best in any

given circumstance. One approach to choosing clusters is to repeat the analysis using more than one clustering criterion, and hope that agreement among different methods reflects real structure in the data rather than artifacts of particular clustering criteria. In any case, the results of a cluster analysis should not be accepted blindly. Rather, they should make sense in terms of other information available to the analyst.

9.6.2 The Dendrogram, or Tree Diagram

The progress and intermediate results of a cluster analysis are conventionally illustrated using the *dendrogram*, or "tree" diagram. Beginning with the "twigs" at the beginning of the analysis, when each of the n observations $\mathbf{x}$ constitutes its own cluster, one pair of "branches" is joined at each step as the closest two clusters are merged. The distances between these clusters before they are merged are also indicated in the diagram by the distance of the points of merger from the initial n-cluster stage of the "twigs."

Figure 9.14 illustrates a simple dendrogram, reflecting the clustering of the five points plotted as open circles in Fig. 9.13. The analysis begins at the bottom of

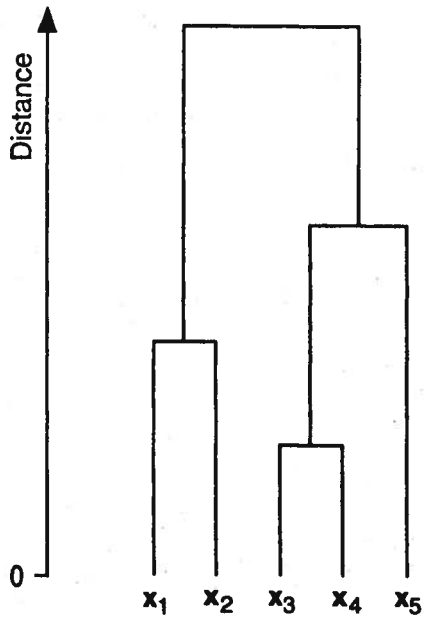

Fig. 9.14 A dendrogram, or tree diagram, for a clustering of the five points plotted as open circles in Fig. 9.13. The results of the four clustering steps are indicated as the original five lines are progressively joined from the bottom to the top of the diagram, with the distances between joined clusters indicated by the heights of the horizontal lines.

Fig. 9.14, when all five points constitute separate clusters. At the first stage, the closest two points, $\mathbf{x}_3$ and $\mathbf{x}_4$ are merged into a new cluster. The distance $d_{3,4} = \|\mathbf{x}_3 - \mathbf{x}_4\|$ is proportional to the distance between the horizontal bar joining these two points and the bottom of the figure. At the next stage, the points $\mathbf{x}_1$ and $\mathbf{x}_2$ are merged into a single cluster because the distance between them is smallest of the six distances between the four clusters that existed at the previous stage. The distance $d_{1,2} = \|\mathbf{x}_1 - \mathbf{x}_2\|$ is necessarily larger than the distance $d_{3,4} = \|\mathbf{x}_3 - \mathbf{x}_4\|$, since $\mathbf{x}_1$ and $\mathbf{x}_2$ were not chosen for merger on the first step, and the horizontal line indicating the distance between them is plotted higher on Fig. 9.14 than the distance between $\mathbf{x}_3$ and $\mathbf{x}_4$. The third step merges $\mathbf{x}_5$ and the pair $(\mathbf{x}_3, \mathbf{x}_4)$, to yield the two-group stage indicated by the dashed lines in Fig. 9.13.

9.6.3 How Many Clusters?

A cluster analysis will produce a different grouping of n observations at each of the $n - 1$ steps. On the first step each observation is in a separate group, and on the last step all the observations are in a single group. An important practical problem in cluster analysis is the choice of which intermediate stage will be chosen as the final solution. Thus, one needs to choose the level of aggregation in the tree diagram at which to stop further merging of clusters. Although the goal guiding this choice is to find that level of clustering that maximizes similarity within clusters and minimizes similarity between clusters, in practice the best number of clusters for a given problem is seldom obvious. Generally the stopping point will require a subjective choice.

One approach to the problem of choosing the best number of clusters is through summary statistics based on ideas drawn from hypothesis testing. Some of these objective stopping criteria that have been developed are discussed in Fovell and Fovell (1993), who also provide references to the broader literature on such methods.

A traditional subjective approach to determination of the stopping level is to inspect a plot of the distances between merged clusters as a function of the stage of the analysis. When similar clusters are being merged early in the process, these distances are small, and they increase relatively little from step to step. Late in the process there may be only a few clusters, separated by large distances. If a point can be discerned where the distances between merged clusters jumps markedly, the process can be stopped just before these distances become large.

Wolter (1987) suggests a Monte Carlo approach, where sets of random numbers simulating the real data are subjected to cluster analysis. The distributions of clustering distances for the random numbers are compared to the actual clustering distances for the data of interest. The idea here is that genuine clusters in the real data should be closer than clusters in the random data, and that the clustering algorithm should be stopped at the point where clustering distances are greater than for the analysis of the random data.

Example 9.10. A Cluster Analysis in $K = 2$ Dimensions

As is the case with discriminant analysis, the mechanics of cluster analysis are easiest to see when the data vectors have only $K = 2$ dimensions. Consider the data in Table 9.6, where these two dimensions are average July temperature and average July precipitation. These data were collected into three groups for use in the discriminant analysis worked out in Example 9.9. However, the point of a cluster analysis is to try to discern group structure within a data set, without prior knowledge or information about the nature of that structure. Therefore, for purposes of the cluster analysis, the data in Table 9.6 should be regarded as consisting of $n = 28$ observations of two-dimensional vectors $\mathbf{x}$, whose natural groupings we would like to discern.

Because the temperature and precipitation values have different physical units, it is probably wise to normalize the data (i.e., convert them to standardized anomalies) before subjecting them to a clustering algorithm. That is, all the temperatures are transformed by subtracting the overall mean of the 28 temperatures (77.0°F), and dividing by the overall standard deviation of the temperature observations (4.42°F). For the precipitation values the corresponding mean and standard deviation are 4.19 and 1.36 in. The reason for this pretreatment of the data is to avoid the same kind of problem that can occur when conducting a PCA (Section 9.3) using unlike data, where a variable with a much higher variance than the others will dominate the analysis even if the high variance is an artifact of the units of measurement. In a cluster analysis, the relative sizes of the units can have a large impact on the Euclidean distance between points. For example, if the precipitation had been reported in millimeters, there would be apparently more distance between points in the direction of the precipitation axis, and a clustering algorithm would focus on precipitation differences to define groups. If the precipitation were reported in meters, there would be essentially no distance between points in the direction of the precipitation axis, and a clustering algorithm would separate points almost entirely on the basis of the temperature data.

Figure 9.15 shows the results of clustering the data in Table 9.6, using the Euclidean distance measure in Eq. (9.74), and the complete-linkage clustering criterion in Eq. (9.76). On the left is a tree diagram for the process, with the individual stations listed at the bottom as the "leaves." There are 27 horizontal lines in this tree diagram, each of which represents a merging of the two clusters it connects. For example, at the first stage of the analysis the two closest points (Springfield and St. Louis) are merged into the same cluster. At the second stage Huntsville and Athens are merged, at the third stage Worcester and Binghamton are merged, and at the fourth stage Macon and Augusta are merged. At the fifth stage Concordia is merged with the cluster consisting of Springfield and St. Louis.

The heights of the horizontal lines indicating group mergers correspond to the distances between the merged clusters. Since the merger at each stage is between the two closest clusters, these distances become greater at later stages.

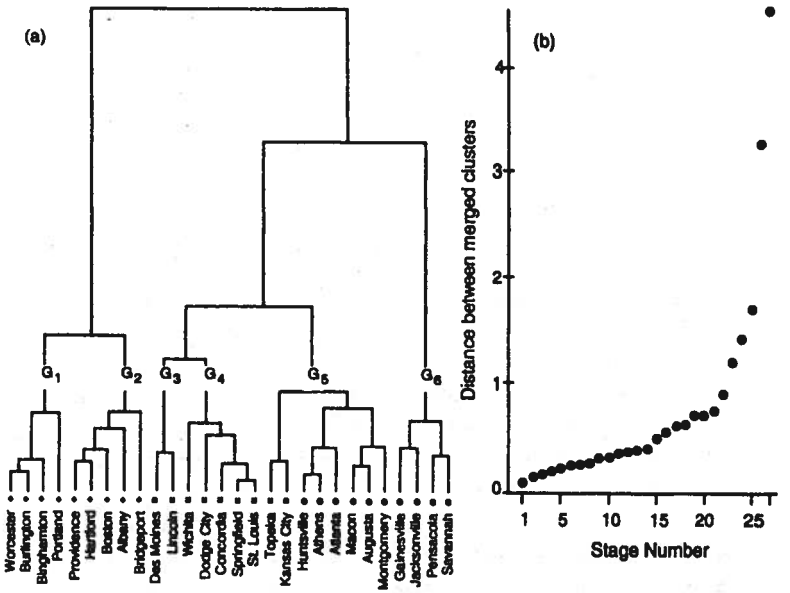

Fig. 9.15 Dendrogram (a) and the corresponding plot of the distance between merged clusters as a function of the stage of the cluster analysis (b) for the data in Table 9.6. Normalized data (standardized anomalies) have been clustered according to the complete-linkage criterion. The distances between merged groups appear to increase markedly at stage 22 or 23, indicating that the analysis should stop after 21 or 22 stages, which for these data would yield seven or six clusters, respectively. The six numbered clusters correspond to the grouping of the data shown in Fig. 9.16. The seven-cluster solution would split Topeka and Kansas City from the Alabama and Georgia stations in G_5. The five-cluster solution would merge G_3 and G_4.

Figure 9.15b shows the distance between merged clusters as a function of the stage in the analysis. Subjectively, these distances climb gradually until perhaps stage 22 or stage 23, where the distances between combined clusters begin to become noticeably larger. A reasonable interpretation of this change in slope is that the "true" clusters have been defined at this point in the analysis, and that the larger distances at later stages indicate mergers of unlike clusters that should be distinct. Note, however, that a single change in slope does not occur in every cluster analysis, so that the choice of where to stop group mergers may not always be so clear-cut. It is possible, for example, for there to be two or more relatively flat regions in the plot of distance versus stage, separated by segments of larger slope. In such a case the choice of where to stop the analysis is more ambiguous.

If Fig. 9.15b is interpreted as exhibiting its first major slope increase between stages 22 and 23, a plausible point at which to stop the analysis would be after stage 22. This stopping point would result in the definition of the six clusters

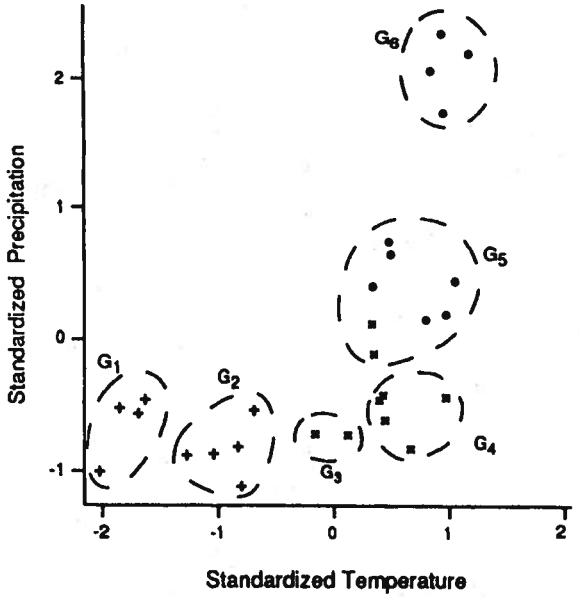

Fig. 9.16 Scatterplot of the data in Table 9.6 expressed as standardized anomalies, with dashed lines showing the six groups defined in the cluster analysis tree diagram in Fig. 9.15. The five-group clustering would merge the central U.S. stations in groups 3 and 4. The seven-group clustering would split the two central U.S. stations in group 5 from six southeastern U.S. stations.

labeled G_1–G_6 on the tree diagram in Fig. 9.15a. This level of clustering assigns the nine northeastern stations (+ symbols) into two groups, assigns seven of the nine central stations (× symbols) into two groups, allocates the central stations Topeka and Kansas City to group 5 with six of the southeastern stations (○ symbols), and assigns the remaining four southeastern stations to a separate cluster.

Figure 9.16 indicates these six groups in the $K = 2$-dimensional space of the data, by separating points in each cluster with dashed lines. If this solution seemed too highly aggregated on the basis of the prior knowledge and information available to the analyst, one could choose the seven-cluster solution produced after stage 21, separating the central U.S. cities Topeka and Kansas City from the six southeastern cities in group 5. If the six-cluster solution seemed too finely split, the five-cluster solution produced after stage 23 would merge the central U.S. stations in groups 3 and 4. ¤

Exercises

9.1. Calculate the matrix product **[B][E]**, using the values in Eqs. (9.31) and (9.32).

9.2. Derive the regression equation produced in Example 6.1, using matrix notation.

9.3. Calculate the angle between the first two eigenvectors in Table 9.2, section a.

9.4. Regard the joint behavior of the maximum temperature data for Ithaca and Canandaigua in Table A.1 as being bivariate normal. The eigenvectors of the variance–covariance matrix for these two variables are $\mathbf{e}_1^T = [.700, .714]$ and $\mathbf{e}_2^T = [-.714, .700]$, where the first element of each vector corresponds to the Ithaca temperature.
   (a) Find the corresponding eigenvalues.
   (b) Plot the 75% and 95% confidence ellipses of this distribution.

9.5. Using information and results from Exercise 9.4:
   (a) Calculate the values of the first principal component for January 1 and January 2.
   (b) Estimate the variance of all 31 values of the first principal component.
   (c) What proportion of the total variability of the maximum temperature data is represented by the first principal component?

9.6. A PCA of the data in Table A.3 yields the three eigenvectors $\mathbf{e}_1^T = [.593, .552, -.587]$, $\mathbf{e}_2^T = [.332, -.831, -.446]$, and $\mathbf{e}_3^T = [.734, -.069, .676]$, where the three elements in each vector pertain to the temperature, precipitation, and pressure data, respectively. The corresponding three eigenvalues are $\lambda_1 = 2.476$, $\lambda_2 = 0.356$, and $\lambda_3 = 0.169$.
   (a) Was this analysis done using the covariance matrix or the correlation matrix? How can you tell?
   (b) How many principal components should be retained according to Kaiser's rule, and according to the broken-stick model?
   (c) Discuss how to find the number of principal components to retain according to rule *N*. If the appropriate computing resources are available, implement your algorithm.

9.7. Using the information in Table 9.5 and the data in Table A.1, calculate the values of the canonical variables v_1 and w_1 for January 6 and January 7.

9.8. Conduct a discriminant analysis using the temperature and pressure data in Table A.3, to classify data points as coming from either El Niño or non–El Niño years.
   (a) What is the discriminant vector, scaled to have unit length?
   (b) Which, if any, of the El Niño years have been misclassified?

Appendix A Example Data Sets

In real applications of climatological data analysis one would hope to use much more data (e.g., all available January daily data, rather than data for just a single year), and would have a computer perform the computations. These small data sets have been used in a number of examples in this book so that the computations can be performed by hand, and a clearer understanding of procedures can be achieved.

Table A.1
Daily Precipitation (inches) and Temperature[a] (°F) Observations at Ithaca and Canadaigua, New York, for January 1987

	Ithaca			Canandaigua		
Date	Precipitation	T_{max}	T_{min}	Precipitation	T_{max}	T_{min}
1	0.00	33	19	0.00	34	28
2	0.07	32	25	0.04	36	28
3	1.11	30	22	0.84	30	26
4	0.00	29	−1	0.00	29	19
5	0.00	25	4	0.00	30	16
6	0.00	30	14	0.00	35	24
7	0.00	37	21	0.02	44	26
8	0.04	37	22	0.05	38	24
9	0.02	29	23	0.01	31	24
10	0.05	30	27	0.09	33	29
11	0.34	36	29	0.18	39	29
12	0.06	32	25	0.04	33	27
13	0.18	33	29	0.04	34	31
14	0.02	34	15	0.00	39	26
15	0.02	53	29	0.06	51	38
16	0.00	45	24	0.03	44	23
17	0.00	25	0	0.04	25	13
18	0.00	28	2	0.00	34	14

continues