Chapter 13

Principal Component (EOF) Analysis

13.1. BASICS OF PRINCIPAL COMPONENT ANALYSIS

Principal component analysis, often denoted as PCA, is possibly the most widely used multivariate statistical technique in the atmospheric sciences. The technique was introduced into the atmospheric science literature by Obukhov (1947), and became popular for analysis of atmospheric data following the papers by Lorenz (1956), who called the technique *empirical orthogonal function* (EOF) analysis, and Davis (1976). Both the names PCA and EOF analysis are commonly used, and both 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. This chapter 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) and Navarra and Simoncini (2010), which are oriented specifically toward geophysical data; and in Jolliffe (2002), which describes PCA more generally. Hannachi et al. (2007) provide a comprehensive review. In addition, most textbooks on multivariate statistical analysis contain chapters on PCA.

13.1.1. Definition of PCA

PCA reduces a data set containing a large number of variables to a data set containing fewer (hopefully many fewer) new variables. These new variables are linear combinations of the original ones, and these linear combinations are chosen to represent the maximum possible fraction of the variability contained in the original data while being uncorrelated with each other. That is, given multiple observations of a $(K \times 1)$ data vector $\boldsymbol{x}$, PCA finds $(M \times 1)$ vectors $\boldsymbol{u}$ whose elements are linear combinations of the elements of the $\boldsymbol{x}'$s, and which contain most of the information in the original collection of $\boldsymbol{x}'$s. PCA is most effective when this data compression can be achieved with $M \ll K$. This situation occurs when there are substantial correlations among the variables within $\boldsymbol{x}$, in which case $\boldsymbol{x}$ contains redundant information. The elements of these new vectors $\boldsymbol{u}$ are called the *principal components* (PCs).

Data for atmospheric and other geophysical fields generally 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, PCA can be a very useful tool for exploring large multivariate data sets, including those consisting of geophysical fields. Here PCA has the potential for yielding insights into both the spatial and temporal variations exhibited by the field or fields being analyzed, and new interpretations of the original data $\boldsymbol{x}$ can be suggested by the nature of the linear combinations that are most effective in compressing those data.

Usually it is convenient to calculate the PCs as linear combinations of the anomalies $\boldsymbol{x}' = \boldsymbol{x} - \overline{\boldsymbol{x}}$. The first PC, u_1, is that linear combination of $\boldsymbol{x}'$ having the largest variance. The subsequent principal components u_m, $m = 2, 3, \ldots$, are the linear combinations having the largest possible variances, subject to the

Statistical Methods in the Atmospheric Sciences. https://doi.org/10.1016/B978-0-12-815823-4.00013-4

condition that they are uncorrelated with the principal components having lower indices. The result is that all the PCs are mutually uncorrelated.

The new variables or PCs—that is, the elements u_m of $\boldsymbol{u}$ that will account successively for the maximum amount of the joint variability of $\boldsymbol{x}'$ (and therefore also of $\boldsymbol{x}$)—are uniquely defined (except for sign) by the eigenvectors of the covariance matrix of $\boldsymbol{x}$, $[S]$. In particular, the mth principal component, u_m is obtained as the projection of the data vector $\boldsymbol{x}'$ onto the mth eigenvector, $\boldsymbol{e}_m$,

$$u_m = \boldsymbol{e}_m^T \boldsymbol{x}' = \sum_{k=1}^{K} e_{k,m} x_k', \quad m = 1, \ldots, M. \tag{13.1}$$

Notice that each of the M eigenvectors contains one element pertaining to each of the K variables, x_k. Similarly, each realization of the mth principal component in Equation 13.1 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 that are the elements of a particular data vector $\boldsymbol{x}$. Although the weights (the $e_{k,m}$'s) do not sum to 1, their squares do because of the scaling convention $||\boldsymbol{e}_m|| = 1$. (Note that a fixed scaling convention for the weights $\boldsymbol{e}_m$ of the linear combinations in Equation 13.1 allows the maximum variance constraint defining the PCs to be meaningful.) If the data sample consists of n observations (and therefore of n data vectors $\boldsymbol{x}$, or n rows in the data matrix $[X]$), there will be n values for each of the principal components, or new variables, u_m. Each of these constitutes a single-number index of the resemblance between the eigenvector $\boldsymbol{e}_m$ and the corresponding individual data vector $\boldsymbol{x}$.

Geometrically, the first eigenvector, $\boldsymbol{e}_1$, points in the direction (in the K-dimensional space of $\boldsymbol{x}'$) 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 $\boldsymbol{e}_2$, associated with the second-largest eigenvalue λ_2, is constrained to be perpendicular to $\boldsymbol{e}_1$ (Equation 11.50), but subject to this constraint it will align in the direction in which the $\boldsymbol{x}'$ vectors exhibit their next strongest variations. Subsequent eigenvectors $\boldsymbol{e}_m$, $m = 3, 4, \ldots, M$, are similarly numbered according to decreasing magnitudes of their associated eigenvalues, and in turn will be perpendicular to all the previous eigenvectors. Subject to this orthogonality 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. In particular, the orthogonal matrix $[E]$ whose columns are the eigenvectors (Equation 11.51) defines the rigid rotation

$$\boldsymbol{u} = [E]^T \boldsymbol{x}', \tag{13.2}$$

which is the simultaneous matrix-notation representation of $M = K$ linear combinations of the form of Equation 13.1 (i.e., here the matrix $[E]$ is square, with K eigenvector columns). This new coordinate system is oriented such that each consecutively numbered 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 covariance matrix $[S_x]$ 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 empirical orthogonal functions (EOFs). The implied distinction is with theoretical orthogonal functions, such as Fourier harmonics or Tschebyschev polynomials, which also can 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 property implies that the covariances between pairs of the u_m's are all zero, so that the corresponding covariance matrix is diagonal. In fact, the covariance matrix for the principal components is obtained by the diagonalization of $[S_x]$ (Equation 11.56) and is thus simply the diagonal matrix $[\Lambda]$ of the eigenvalues of $[S]$:

$$[S_u] = \mathrm{Var}\left([E]^T \boldsymbol{x}\right) = [E]^T [S_x] [E] = [E]^{-1} [S_x] [E] = [\Lambda]. \tag{13.3}$$

That is, the variance of the mth principal component u_m is the mth eigenvalue λ_m. Equation 11.54 then implies that each PC represents a share of the total variation in $\boldsymbol{x}$ that is proportional to its eigenvalue,

$$R_m^2 = \frac{\lambda_m}{\sum_{k=1}^{K} \lambda_k} \times 100\% = \frac{\lambda_m}{\sum_{k=1}^{K} s_{k,k}} \times 100\%. \tag{13.4}$$

Here R^2 is used in the same sense that is familiar from linear regression (Section 7.2.4). The total variation exhibited by the original data is completely represented in (or accounted for by) the full set of K u_m's, in the sense that the sum of the variances of the centered data $\boldsymbol{x}'$ (and therefore also of the uncentered variables $\boldsymbol{x}$), $\Sigma_k\, s_{k,k}$, is equal to the sum of the variances $\Sigma_m \lambda_m$. of the principal component variables $\boldsymbol{u}$.

Equation 13.2 expresses the transformation of a $(K \times 1)$ data vector $\boldsymbol{x}'$ to a vector $\boldsymbol{u}$ of PCs. If $[E]$ contains all K eigenvectors of $[S_x]$ (assuming $[S_x]$ is nonsingular) as its columns, the resulting vector $\boldsymbol{u}$ will also have dimension $(K \times 1)$. Equation 13.2 sometimes is called the *analysis formula* for $\boldsymbol{x}'$, expressing that the data can be analyzed, or summarized in terms of the principal components. Reversing the transformation in Equation 13.2, the data $\boldsymbol{x}'$ can be reconstructed from the principal components according to

$$\underset{(K\times 1)}{\boldsymbol{x}'} = \underset{(K\times K)}{[E]}\ \underset{(K\times 1)}{\boldsymbol{u}}, \tag{13.5}$$

which is obtained from Equation 13.2 by multiplying on the left by $[E]$ and using the orthogonality property of this matrix (Equation 11.44). The reconstruction of $\boldsymbol{x}'$ expressed by Equation 13.5 is sometimes called the *synthesis formula*. If the full set of $M = K$ PCs is used in the synthesis, the reconstruction is complete and exact, since $\Sigma_m R_m^2 = 1$ (cf. Equation 13.4). If $M < K$ PCs (usually those corresponding to the M largest eigenvalues) are used, the reconstruction is approximate,

$$\underset{(K\times 1)}{\boldsymbol{x}'} \approx \underset{(K\times M)}{[E]}\ \underset{(M\times 1)}{\boldsymbol{u}}, \tag{13.6a}$$

or

$$x_k' \approx \sum_{m=1}^{M} e_{k,m} u_m, \quad k = 1, \ldots, K, \tag{13.6b}$$

but the approximation improves as the number M of PCs used (or, more precisely, as the sum of the corresponding eigenvalues, because of Equation 13.4) increases. Because $[E]$ in Equation 13.6a has only M columns, and operates on a truncated PC vector $\boldsymbol{u}$ of dimension $(M \times 1)$, Equation 13.6 is called the

truncated synthesis formula. The original (in the case of Equation 13.5) or approximated (for Equation 13.6) uncentered data $\boldsymbol{x}$ can easily be obtained by adding back the vector of sample means, that is, by reversing Equation 11.33.

Because each principal component u_m is a linear combination of the original variables x_k (Equation 13.1), and vice versa (Equation 13.5), pairs of principal components and original variables will be correlated unless the eigenvector element $e_{k,m}$ relating them is zero. It can sometimes be informative to calculate these correlations, which are given by

$$r_{u,x} = \mathrm{Corr}(u_m, x_k) = e_{k,m}\sqrt{\frac{\lambda_m}{s_{k,k}}}. \tag{13.7}$$

Example 13.1. 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 13.1a shows a scatterplot of centered (at zero) January 1987 Ithaca minimum temperatures (x_1') and Canandaigua minimum temperatures (x_2') from Table A.1. This is the same scatterplot that appears in the middle of the bottom row of Figure 3.31. It is apparent that the Ithaca temperatures are more variable than the Canandaigua temperatures, with the two standard deviations being $\sqrt{s_{1,1}} = 13.62°\mathrm{F}$ and $\sqrt{s_{2,2}} = 8.81°\mathrm{F}$, respectively. The two variables are clearly strongly correlated and have a Pearson correlation of +0.924 (see Table 3.5). The covariance matrix $[S]$ for these two variables is given as $[A]$ in Equation 11.59. The two eigenvectors of this matrix are ${e_1}^T = [0.848, 0.530]$ and ${e_2}^T = [-0.530, 0.848]$, so that the eigenvector matrix $[E]$ is that shown in Equation 11.60. The corresponding eigenvalues are $\lambda_1 = 254.76$ and $\lambda_2 = 8.29$. These are the same data used to fit the bivariate normal probability ellipses shown in Figures 12.1 and 12.7.

The orientations of the two eigenvectors are shown in Figure 13.1a, although their lengths have been exaggerated for clarity. It is evident that the first eigenvector is aligned in the direction in which the data jointly exhibit maximum variation. That is, the point cloud is inclined at the same angle as is $\boldsymbol{e}_1$, which is 32° from the horizontal (i.e., from the vector [1, 0], according to Equation 11.15). 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 $-{e_2}^T = [0.530, -0.848]$). This last eigenvector locates the direction in which data jointly exhibit their smallest variations.

Figure 13.1b illustrates another property of the first eigenvector, which is that it is the direction that minimizes the sum of squared distances (dashed lines) perpendicular to $\boldsymbol{e}_1$, connecting it to the points. The leading eigenvector is thus different from the least-squares regression lines relating the two variables, which would minimize either the sum of squared vertical distances (if the Canandaigua temperatures were being predicted) or the squared horizontal distances (if the Ithaca temperatures were being predicted). In three dimensions, the second eigenvector would be the direction perpendicular to $\boldsymbol{e}_1$, defining the $\boldsymbol{e}_1$–$\boldsymbol{e}_2$ plane minimizing the perpendicular squared distances from the plane to the points, and so on, in progressively higher dimensions.

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° clockwise while looking at Figure 13.1a. Within this rotated coordinate system, each point is defined by a principal component vector

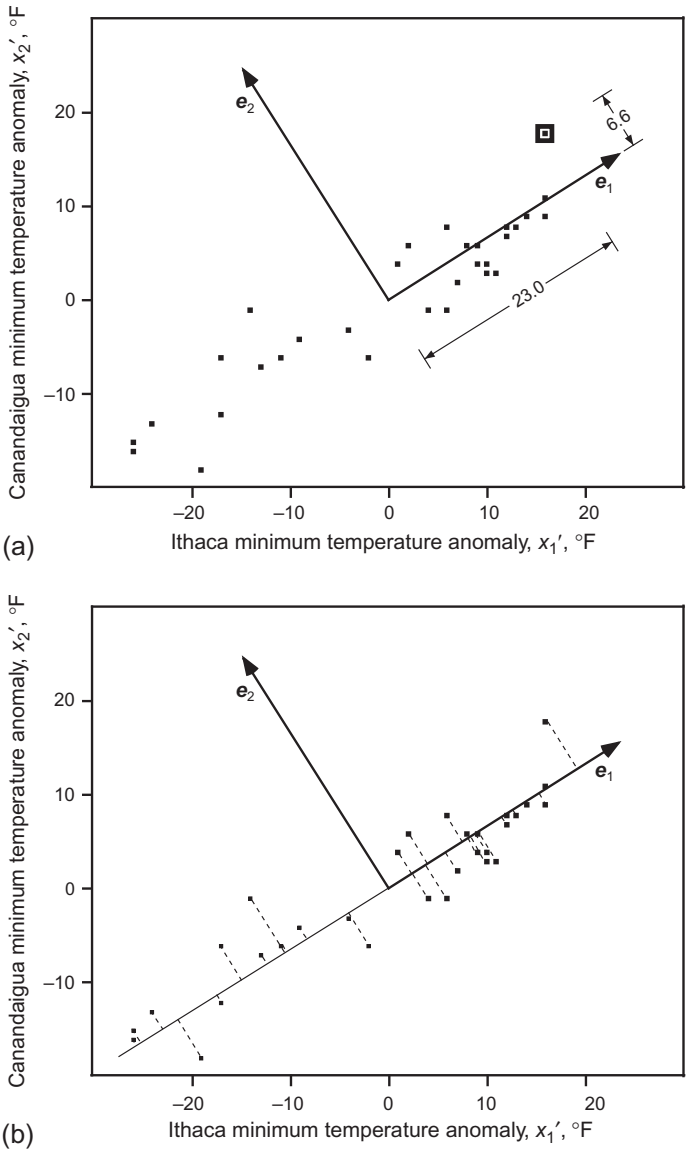

FIGURE 13.1 (a) Scatterplot of January 1987 Ithaca and Canandaigua minimum temperatures (converted to anomalies, or centered), illustrating the geometry of PCA in two dimensions. The eigenvectors $\boldsymbol{e}_1$ and $\boldsymbol{e}_2$ of the covariance matrix $[S]$ for these two variables, as computed in Example 11.3, have been plotted with lengths exaggerated for clarity. The data stretch out in the direction of $\boldsymbol{e}_1$ 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 $\boldsymbol{x}'^{\mathrm{T}}$ [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 vector $\boldsymbol{u}^{\mathrm{T}} = [23.0, 6.6]$ locates the same point as $\boldsymbol{x}'^{\mathrm{T}} = [16.0, 17.8]$. (b) The first eigenvector $\boldsymbol{e}_1$ is also the direction that minimizes the sum of squared lengths of the dashed lines between the points and $\boldsymbol{e}_1$, that are perpendicular to $\boldsymbol{e}_1$.

$\boldsymbol{u}^{\mathrm{T}} = [u_1, u_2]$ of new transformed variables, whose elements consist of the projections of the original data onto the eigenvectors, according to the dot product in Equation 13.1. Figure 13.1a illustrates this projection for the 15 January data point $\boldsymbol{x}'^{\mathrm{T}} = [16.0, 17.8]$, 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 $\boldsymbol{e}_1$). This dispersion is evidently greater than the dispersion of the data along either of the original axes, and indeed it is larger than the dispersion of the data in any other direction in this plane. This maximum sample variance of u_1 is equal to the eigenvalue $\lambda_1 = 254.76°\mathrm{F}^2$. The points in the data set tend to exhibit quite different values of u_1, whereas they have more similar values for u_2. That is, they are much less variable in the $\boldsymbol{e}_2$ direction, and the sample variance of u_2 is only $\lambda_2 = 8.29°\mathrm{F}^2$.

Since $\lambda_1 + \lambda_2 = s_{1,1} + s_{2,2} = 263.05°\mathrm{F}^2$, the new variables jointly retain all the variation exhibited by the original variables. 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_2 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 $[\Lambda]$, as shown in Equation 11.62.

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 alone. 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, or about 80 km apart), with the second principal component describing local variations departing from the overall regional value.

Since so much of the joint variance of the two temperature series is captured by the first principal component, resynthesizing the series using only the first principal component will yield a good approximation to the original data. Using the synthesis Equation 13.6 with only the first ($M = 1$) principal component yields

$$\boldsymbol{x}'(t) = \begin{bmatrix} x_1'(t) \\ x_2'(t) \end{bmatrix} \approx \boldsymbol{e}_1\, u_1(t) = \begin{bmatrix} .848 \\ .530 \end{bmatrix} u_1(t). \tag{13.8}$$

The temperature data $\boldsymbol{x}$ are time series, and therefore so are the principal components $\boldsymbol{u}$. The time dependence for both has been indicated explicitly in Equation 13.8. On the other hand, the eigenvectors are fixed by the covariance structure of the entire series and do not change through time. Figure 13.2 compares the original series (black) and the reconstructions using the first principal component $u_1(t)$ only (gray) for the (a) Ithaca and (b) Canandaigua anomalies. The discrepancies are small because $R^2{}_1 = 96.8\%$. The residual differences would be captured by u_2. The two gray series are exactly proportional to each other, since each is a scalar multiple of the same first principal component time series. Since $\mathrm{Var}(u_1) = \lambda_1 = 254.76$, the variances of the reconstructed series are $(0.848)^2\, 254.76 = 183.2$ and $(0.530)^2\, 254.76 = 71.6°\mathrm{F}^2$, respectively, which are close to but smaller than the corresponding diagonal elements of the original covariance matrix (Equation 11.59). The larger variance for the Ithaca temperatures is also visually evident in Figure 13.2. Using Equation 13.7, the correlations between the first principal component series $u_1(t)$ and the original temperature variables are $0.848(254.76/185.47)^{1/2} = 0.994$ for Ithaca, and $0.530\ (254.76/77.58)^{1/2} = 0.960$ for Canandaigua. ◇

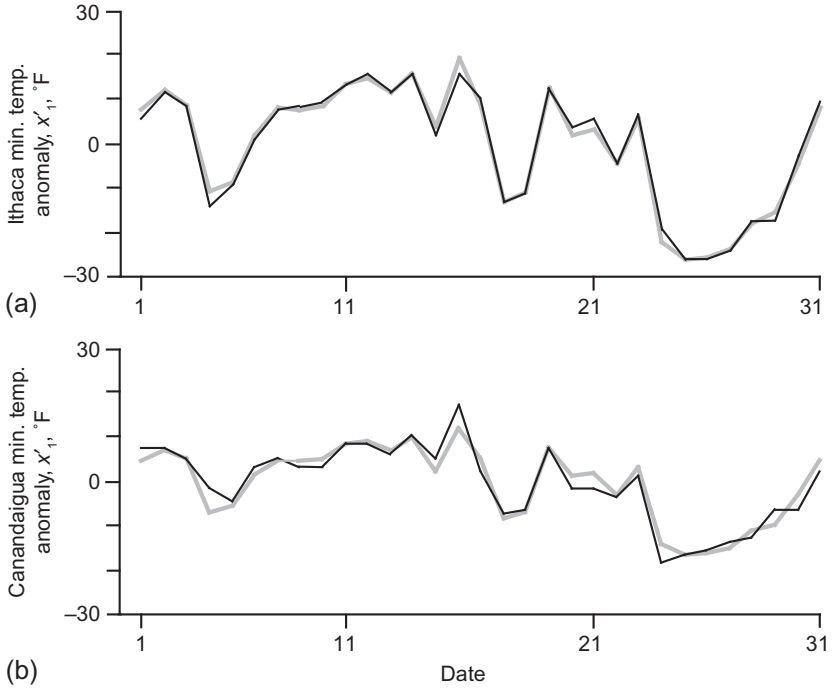

FIGURE 13.2 Time series of January 1987 (a) Ithaca and (b) Canandaigua minimum temperature anomalies (black), and their reconstruction using the first principal component only (gray), through the synthesis Equation 13.8.

13.1.2. PCA Based on the Covariance Matrix vs. the Correlation Matrix

A PCA can be computed as easily using 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 $\boldsymbol{z}$ (Equation 11.32). The vector of standardized variables $\boldsymbol{z}$ is related to the vectors of original variables $\boldsymbol{x}$ and their centered counterparts $\boldsymbol{x}'$ according to the scaling transformation (Equation 11.34). Therefore PCA on the correlation matrix amounts to analysis of the joint variance structure of the standardized variables z_k, as computed using either Equation 11.34 or (in scalar form) Equation 3.27.

The difference between a PCA performed using the variance–covariance and correlation matrices will be one of emphasis. Since PCA seeks to find variables successively maximizing the 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's 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 13.1, 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 the standardized variables z_k equally, since all have equal (unit) variance.

If the PCA is computed using the correlation matrix, the analysis formula, Equations 13.1 and 13.2, will pertain to the standardized variables, z_k and $\boldsymbol{z}$, respectively. Similarly the synthesis formulae, Equations 13.5 and 13.6 will pertain to $\boldsymbol{z}$ and z_k rather than to $\boldsymbol{x}'$ and x_k'. In this case the original data

$\boldsymbol{x}$ can be recovered from the result of the synthesis formula by reversing the standardization given by Equations 11.33 and 11.34, that is,

$$\boldsymbol{x} = [D]\boldsymbol{z} + \bar{\boldsymbol{x}}. \tag{13.9}$$

Although $\boldsymbol{z}$ and $\boldsymbol{x}'$ can easily be obtained from each other using Equation 11.34, 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 eigenvectors and principal components of one knowing only the eigenvectors and principal components of the other. This fact implies that these two alternatives for PCA do not yield equivalent information and that an intelligent choice of one over the other must be made 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 alternative usually will be PCA using the covariance matrix, although the choice 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, and that correlation-based PCA was more effective at locating the primary storm tracks.

However, if the analysis is of unlike variables—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 °F will be $(1.8)^2 = 3.24$ times as large as the variance of the same temperatures expressed in °C. If the PCA has been done using the correlation matrix, the analysis formula, Equation 13.2, pertains to the vector $\boldsymbol{z}$ rather than $\boldsymbol{x}'$, and the synthesis in Equation 13.5 will yield the standardized variables z_k (or approximations to them if Equation 13.6 is used for the reconstruction). The summations in the denominators of Equation 13.4 will equal the number of standardized variables, since each has unit variance.

Example 13.2 Correlation-Versus Covariance-Based PCA for Arbitrarily Scaled Variables

The importance of basing a PCA on the correlation matrix when the variables being analyzed are not measured on comparable scales is presented in Table 13.1. 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 (6x6) arrays in the upper-right portions of parts (a) and (b) of this table constitute the matrices $[E]$ whose columns are the eigenvectors.

The sample variances of each of the variables are shown, as are the six eigenvectors $\boldsymbol{e}_m$ arranged in decreasing order of their eigenvalues λ_m. The cumulative percentage of variance represented is calculated according to Equation 13.4. The much smaller variances of the precipitation variables in (a) are an artifact 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.

Because of the different 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

TABLE 13.1 Comparison of PCA Computed Using (a) the Covariance Matrix, and (b) the Correlation Matrix, of the Data in Table A.1

Variable	Sample Variance	$\boldsymbol{e}_1$	$\boldsymbol{e}_2$	$\boldsymbol{e}_3$	$\boldsymbol{e}_4$	$\boldsymbol{e}_5$	$\boldsymbol{e}_6$
(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
	Cum. % 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
	Cum. % variance	58.9	92.0	97.7	98.9	99.5	100.0

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 °C, 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 calculated from 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. An application of the truncated synthesis formula (Equation 13.6) with the leading $M = 4$ eigenvector therefore would result in reconstructed precipitation data very near their average values. That is, essentially none of the variation in precipitation would be represented.

Since the correlation matrix is the covariance matrix for comparably scaled variables z_k, each has equal variance. Unlike the analysis on the covariance matrix, this 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 $\boldsymbol{e}_1$ for the four temperature variables. 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 precipitation variations would not be lost in a truncated data representation including at least the first $M = 2$ eigenvectors, but rather would be very nearly completely reconstructed. ◇

13.1.3. 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 13.2 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) into the literature as another name for the eigenvectors of a PCA. The terms *modes of variation* and *pattern vectors* also are used primarily by analysts of geophysical data, especially in relation to analysis of fields, to be described in Section 13.2. 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 $\boldsymbol{e}_m$ through the individual element $e_{k,m}$. 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 $\boldsymbol{x}_i'$. In the statistical literature these are

TABLE 13.2 A Partial Guide to Synonymous Terminology Associated With PCA

Eigenvectors, $\boldsymbol{e}_m$	Eigenvector elements, $e_{k,m}$	Principal Components, $\boldsymbol{u}_m$	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			

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, which 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.

13.1.4. 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 chapter assumes that the eigenvectors are scaled to unit length, that is, $||\boldsymbol{e}_m|| \equiv 1$. Recall that vectors of any length will satisfy Equation 11.48 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. 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 by the same factor. That is, multiplying the eigenvector $\boldsymbol{e}_m$ 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 (Equations 13.1 and 13.2) to remain valid. The expected values of the principal component scores for centered data $\boldsymbol{x}'$ 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.

Table 13.3 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 $||\boldsymbol{e}_m|| \equiv 1$ adopted in this presentation. Under this scaling, the expected value (mean) of each of the principal components is zero (because it is the data anomalies x' that have been projected onto the eigenvectors), 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 (Equation 11.56) produced by adopting the rigidly rotated geometric coordinate system defined by the eigenvectors. When scaled in this way, the

TABLE 13.3 Three common eigenvector scalings used in PCA; their consequences for the properties of the principal components, u_m; and their relationship to the original variables, x_k; and the standardized original variables, z_k

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

correlation between a principal component u_m and a variable x_k is given by Equation 13.7. The 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 sometimes are rescaled by multiplying each element by the square root of the corresponding eigenvalue. This rescaling produces vectors of differing lengths, $||\boldsymbol{e}_m|| \equiv (\lambda_m)^{1/2}$, but which point in exactly the same directions as the original eigenvectors having unit lengths. Consistency in the analysis formula implies that the principal components are also changed by the factor $(\lambda_m)^{1/2}$, with the result that the variance of each u_m increases to ${\lambda_m}^2$. 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.

The last scaling presented in Table 13.3, resulting in $||\boldsymbol{e}_m|| \equiv (\lambda_m)^{-1/2}$, 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 is more awkward, but this scaling has the advantage that all the principal components have equal, unit variance. This property can be useful in the detection of outliers.

13.1.5. Connections to the Multivariate Normal Distribution

The distribution of the data $\boldsymbol{x}$, whose sample covariance matrix $[S]$ is used to calculate a PCA, need not be multivariate normal in order for the PCA to be valid. Regardless of the joint distribution of $\boldsymbol{x}$, the resulting principal components u_m will uniquely be those uncorrelated linear combinations that successively maximize the represented fractions of the variances on the diagonal of $[S]$. However, if in addition $\boldsymbol{x} \sim \mathrm{N}_K(\boldsymbol{\mu}_x, [\Sigma_x])$, then as linear combinations of the multinormal $\boldsymbol{x}'$s, the joint distribution of the principal components will also have a multivariate normal distribution,

$$\boldsymbol{u} \sim N_M\Big([E]^T \boldsymbol{\mu}_x, [\Lambda]\Big). \tag{13.10}$$

Equation 13.10 is valid both when the matrix $[E]$ contains the full number $M = K$ of eigenvectors as its columns, or some fewer number $1 \le M < K$. If the principal components are calculated from the centered data $\boldsymbol{x}'$, then $\boldsymbol{\mu}_u = \boldsymbol{\mu}_{x'} = \boldsymbol{0}$.

If the joint distribution of $\boldsymbol{x}$ is multivariate normal, then the transformation of Equation 13.2 is a rigid rotation to the principal axes of the probability ellipses of the distribution of $\boldsymbol{x}$, yielding the uncorrelated and mutually independent u_m. With this background it is not difficult to understand Equations 12.5 and 12.34, which say that the distribution of Mahalanobis distances to the mean of a multivariate normal distribution follows the χ^2_K distribution. One way to view the χ^2_K is as the distribution of K squared independent standard Gaussian variables z_k^2 (see Section 4.4.5). Calculation of the Mahalanobis distance (or, equivalently, the Mahalanobis transformation, Equation 12.21) produces uncorrelated values with zero mean and unit variance, and a (squared) distance involving them is then simply the sum of the squared values.

It was noted in Section 12.3 that an effective way to search for multivariate outliers when assessing multivariate normality is to examine the distribution of linear combinations formed using eigenvectors associated with the smallest eigenvalues of $[S]$ (Equation 12.18). These linear combinations are, of course, the last principal components. Figure 13.3 illustrates why this idea works, in the easily visualized

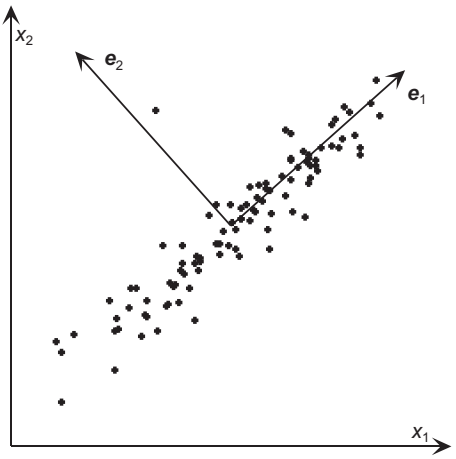

FIGURE 13.3 Identification of a multivariate outlier by examining the distribution of the last principal component. The projection of the single outlier onto the first eigenvector yields a quite ordinary value for its first principal component u_1, but its projection onto the second eigenvector yields a prominent outlier in the distribution of the u_2 values.

$K = 2$ situation. The point scatter shows a strongly correlated pair of Gaussian variables, with one multivariate outlier. The outlier is not especially unusual within either of the two univariate distributions, but it stands out in two dimensions because it is inconsistent with the strong positive correlation of the remaining points. The distribution of the first principal component u_1, obtained geometrically by projecting the points onto the first eigenvector $\boldsymbol{e}_1$, is at least approximately Gaussian, and the projection of the outlier is a very ordinary member of this distribution. On the other hand, the distribution of the second principal component u_2, obtained by projecting the points onto the second eigenvector $\boldsymbol{e}_2$, is concentrated near the origin except for the single large outlier. Other than the outlier, this distribution is also approximately Gaussian. This approach is effective in identifying the multivariate outlier because its existence has distorted the PCA only slightly, so that the leading eigenvector continues to be oriented in the direction of the main data scatter. Because a small number of outliers contribute only slightly to the full variability, it is the last (low-variance) principal components that represent them.

13.2. APPLICATION OF PCA TO GEOPHYSICAL FIELDS

13.2.1. PCA for a Single Field

The overwhelming majority of applications of PCA to atmospheric data have involved analyses of fields (i.e., spatial arrays of variables) such as geopotential heights, temperatures, precipitation, and so on. In these cases the full data set consists of multiple observations of a field or set of fields. Frequently these multiple observations take the form of time series, for example, a sequence of daily hemispheric 500 mb height maps. 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) physical space, the data from these locations at a given observation time are arranged in the K-dimensional vector $\boldsymbol{x}$. That is, regardless of their geographical arrangement, each location is assigned a number (as in Figure 9.27) from 1 to K, which refers to the appropriate element

in the data vector $\boldsymbol{x} = [x_1, x_2, x_3, \ldots, x_K]^{\mathrm{T}}$. 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 successive times.

To emphasize that the original data consists of K time series, the analysis equation (13.1 or 13.2) is sometimes written with an explicit time index:

$$\boldsymbol{u}(t) = [E]^T \boldsymbol{x}'_t, \tag{13.11a}$$

or, in scalar form,

$$u_m(t) = \sum_{k=1}^{K} e_{k,m} x'_k(t), m = 1, \ldots, M. \tag{13.11b}$$

Here the time index t runs from 1 to n. The synthesis equations (13.5 or 13.6) can be written using the same notation, as was done in Equation 13.8. Equation 13.11 emphasizes that if the data $\boldsymbol{x}$ consist of a set of time series, then the principal components $\boldsymbol{u}$ 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 individual principal component time series are sometimes analyzed using the tools presented in Chapter 10. However, each of the time series of principal components will be uncorrelated with the time series of all the other principal components.

When the K elements of $\boldsymbol{x}$ 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 (Equation 13.11b). Each eigenvector element $e_{k,m}$ 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 summarized using smooth contours in the same way as an ordinary meteorological field. 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 corresponding principal components. These geographic displays of eigenvectors sometimes also are interpreted as representing uncorrelated modes of variability of the fields from which the PCA was extracted. There are cases where this kind of interpretation can be reasonable (but see Section 13.2.4 for a cautionary counterexample), particularly for the leading eigenvector. However, because of the mutual orthogonality constraints on the eigenvectors, strong interpretations of this sort are often not justified for the subsequent EOFs (North, 1984).

Figure 13.4 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 percentages below and to the right of the panels show the fraction of the total hemispheric variance (Equation 13.4) 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 Figure 3.33, and apparently reflect the same underlying physical processes in the atmosphere. For example, Figure 13.4b evidently reflects the PNA pattern of alternating height anomalies stretching from the Pacific Ocean through northwestern North America to southeastern North America. A positive value for the second principal component of this data set corresponds to negative 500 mb height anomalies (troughs) in the northeastern Pacific and in the southeastern United States, and to positive height anomalies (ridges) in the western part of the continent,

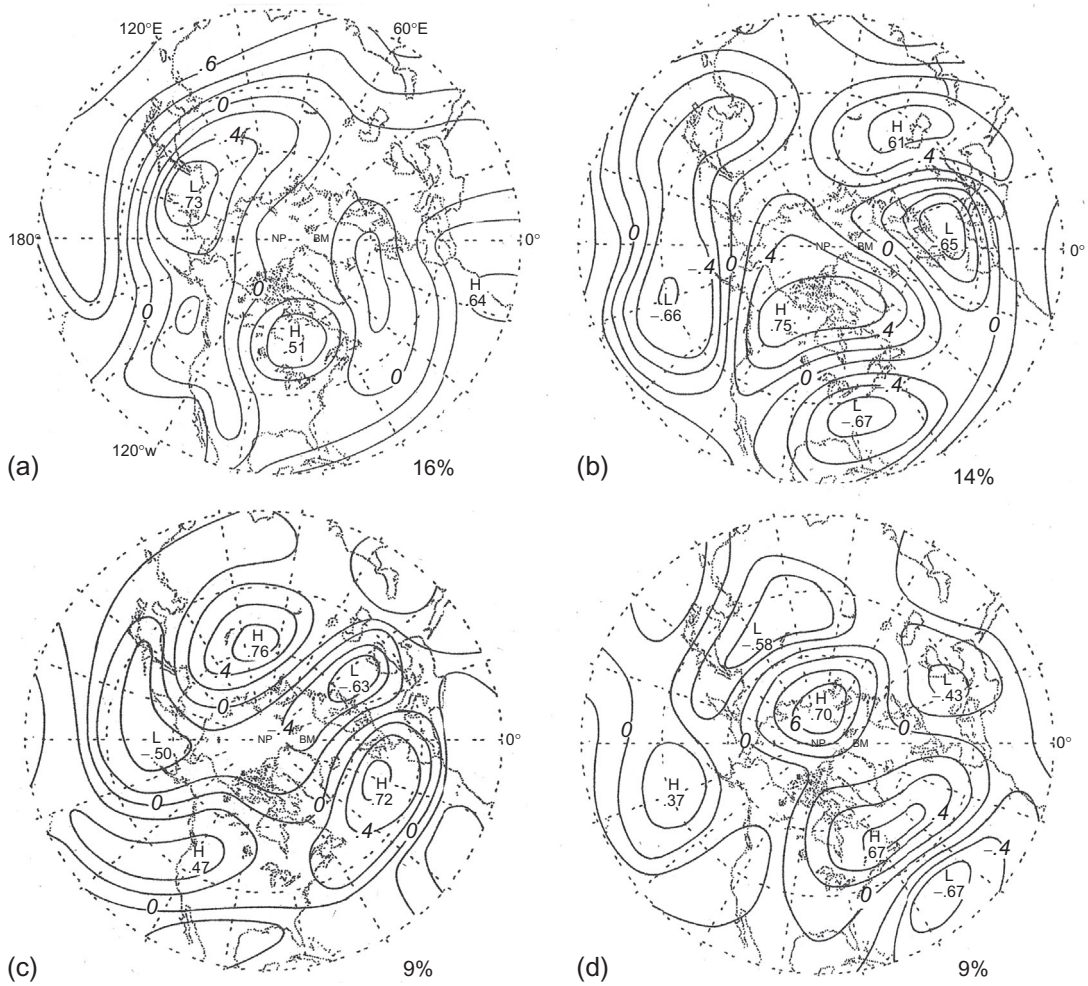

FIGURE 13.4 Spatial displays of the first four eigenvectors of gridded winter monthly mean 500 mb heights for the northern hemisphere, 1962–1977. This PCA was computed using the correlation matrix of the height data, and scaled so that $||\boldsymbol{e}_m|| = \lambda_m^{1/2}$. Percentage values below and to the right of each map are proportion of total variance x100% (Equation 13.4). The patterns resemble the teleconnectivity patterns for the same data (Figure 3.33). *From Wallace and Gutzler (1981). © American Meteorological Society. Used with permission.*

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.

Principal component analyses are most frequently structured as just described, by computing the eigenvalues and eigenvectors from the $(K \times K)$ covariance or correlation matrix of the $(n \times K)$ data matrix $[X]$. However, this usual approach, known as *S-mode* PCA, is not the only possibility. An alternative, known as *T-mode* PCA is based on the eigenvalues and eigenvectors of the $(n \times n)$ covariance or correlation matrix of the data matrix $[X]^{\mathrm{T}}$. Thus in a T-mode PCA the eigenvector elements correspond to the individual data samples (which often form a time series), and the principal components $\boldsymbol{u}$ relate to the K variables (which may be spatial points), so that the two approaches portray different aspects of a

data set in complementary ways. Compagnucci and Richman (2008) compare these two approaches for representing atmospheric circulation fields. The eigenvalues and eigenvectors produced by the two approaches will be different, because the S-mode anomalies are computed by subtracting the K column means of $[X]$, whereas the T-mode anomalies are computed by subtracting the n row means, which will be the column means of $[X]^{\mathrm{T}}$. Accordingly the number of nonzero eigenvalues in an S-mode analysis will be the smaller of $n-1$ and K, and the number of nonzero eigenvalues of a T-mode analysis will be the smaller of n and $K-1$.

13.2.2. Simultaneous PCA for Multiple Fields

It is also possible to apply PCA to vector-valued fields, which are fields with data for more than one variable at each location or gridpoint. This kind of analysis is equivalent to simultaneous PCA of two or more fields. If there are L such variables at each of the K gridpoints, then the dimensionality of the data vector $\boldsymbol{x}$ is given by the product KL. The first K elements of $\boldsymbol{x}$ are observations of the first variable, the second K elements are observations of the second variable, and the last K elements of $\boldsymbol{x}$ will be observations of the Lth variable. Since the L different variables generally will be measured in unlike units, it will almost always be appropriate 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 $(KL \times KL)$. Application of PCA to this kind of correlation matrix will produce principal components successively maximizing the joint variance of the L standardized variables in a way that considers the correlations both between and among these variables at the K locations. This joint PCA procedure is sometimes called *combined PCA*, (CPCA), or *extended EOF* (EEOF) analysis.

Figure 13.5 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 the correlations between the first of the L variables at these locations and all of the KL variables. Rows $K+1$ to $2K$ similarly contain the correlations between the second of the L variables and all the KL 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

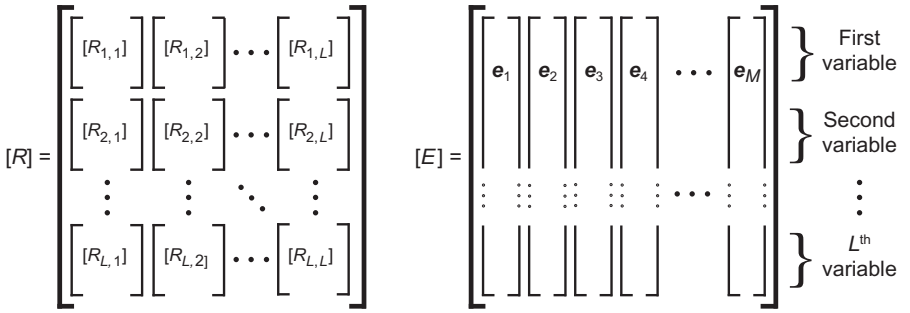

FIGURE 13.5 Illustration of the structures of the correlation matrix and of the matrix of eigenvectors for PCA of vector field data. The basic data consist of multiple observations of L variables at each of K locations, so the dimensions of both $[R]$ and $[E]$ are $(KL \times KL)$. The correlation matrix 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 $[R]$ are the 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 the diagonals. Each eigenvector column of $[E]$ similarly consists of L segments, each of which contains K elements pertaining to the individual locations.

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 $[R_{i,j}] = [R_{j,i}]^{\mathrm{T}}$. 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 13.6, from the classic paper by Kutzbach (1967), illustrates this process for the case of $L = 2$ data values 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 joint variance of the $KL = 23 \times 2 = 46$ standardized variables.

In addition to effectively condensing very much information, the patterns shown in Figure 13.6 are consistent with the underlying atmospheric physical processes. In particular, the temperature anomalies are consistent with patterns of thermal advection implied by the pressure anomalies. If the first principal component u_1 is positive for a particular January, 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

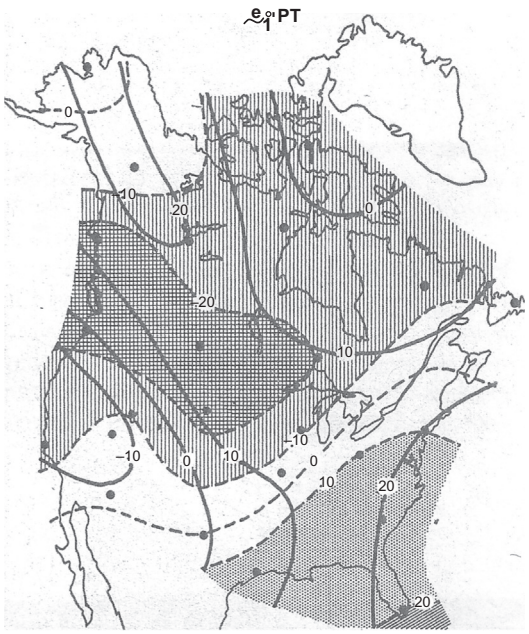

FIGURE 13.6 Spatial display of the elements of the first eigenvector of the (46×46) correlation matrix of average January sea-level pressures and temperatures at 23 locations in North America (dots). The first principal component of this correlation matrix accounts for 28.6% of the joint (standardized) 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). © American Meteorological Society. Used with permission.*

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 the dashed contours and shading in Figure 13.6, when their signs are also reversed.

Figure 13.6 is a simple example involving familiar variables. Its interpretation is easy and obvious if we are 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 about possibly unknown underlying physical mechanisms may be hidden in the complex relationships among several fields.

Example 13.3 Characterization of the Madden–Julian Oscillation

The Madden–Julian oscillation (MJO, Madden and Julian, 1972) is a travelling pattern of enhanced and suppressed tropical convection that propagates eastward from the western Indian to the eastern Pacific ocean basins on a one- to two-month timescale. It is a prominent element of subseasonal tropical atmospheric variability that exhibits characteristic signatures in satellite-observed outgoing longwave radiation (OLR, which is a proxy for cold, high convective cloud tops), coupled with upper- and lower-tropospheric zonal (east-west) wind convergence and divergence.

Monitoring of the MJO is typically done using a diagram derived from extended EOF analysis of OLR, 850 mb zonal winds and 200 mb zonal winds, averaged between 15°S and 15°N, as a function of longitude (Wheeler and Hendon, 2004). Figure 13.7 shows the two leading EOFs for the three combined variables. The traces for each of the three variables are shown as scalar functions of longitude rather than as maps because north–south variations have been collapsed by the meridional averaging.

It is notable that these two eigenvectors in Figure 13.7 have eigenvalues of comparable magnitude, which are well separated from the magnitudes of the third and subsequent eigenvalues (see Section 13.4), and that the traces for each of the three variables are in approximate quadrature (the loadings for EOF2 lag those in EOF1 by approximately a quarter cycle). These characteristics, which are spatial counterparts to the properties that can emerge from PCAs computed for time series (Section 13.7.1), allow the propagating nature of the MJO to be portrayed in a two-dimensional phase space defined by the corresponding principal components (Figure 13.8). In the context of the MJO diagram, these are conventionally denoted as RMM1 and RMM2, respectively. The diagram is divided into octants, with Phase 1 corresponding to convection over the western Indian ocean, and Phase 8 corresponding to convection over the eastern Pacific. An MJO cycle is portrayed as daily points in the diagram trace out a counterclockwise orbit in this phase space, an example of which for January through March of 2009 is shown in Figure 13.8. Figure 13.9 shows composites of December through February OLR and 850 mb zonal wind anomalies, averaged over the eight MJO phases jointly defined by the two principal components in the Wheeler–Hendon diagram (Figure 13.8), showing that the extended EOF analysis has been very effective at portraying this propagating phenomenon. ◇

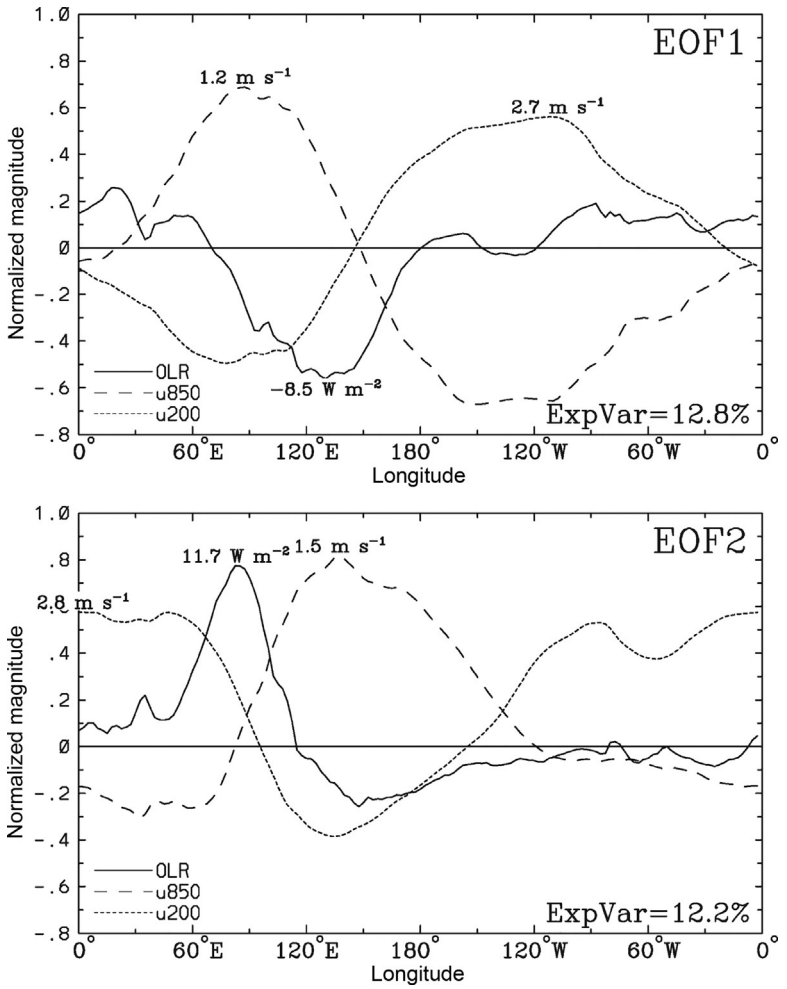

FIGURE 13.7 The leading two EOFs from an extended EOF analysis of OLR, 850 mb zonal wind, and 200 mb zonal wind, averaged within 15° of the equator, as functions of longitude. *From Wheeler and Hendon (2004). © American Meteorological Society. Used with permission.*

13.2.3. Scaling Considerations and Equalization of Variance

A complication arises in PCA of fields in which the geographical distribution of data locations is not uniform (Baldwin et al., 2009; Karl et al., 1982; North et al., 1982). The problem is that the PCA has no information about the spatial distributions of the locations, or even that the elements of the data vector $\boldsymbol{x}$ may pertain to different locations, but nevertheless finds linear combinations that maximize the joint variance. Regions that are overrepresented in $\boldsymbol{x}$, in the sense that data locations are concentrated in that region, will tend to dominate the analysis, whereas data-sparse regions will be underweighted. In contrast, the goal of PCA on geophysical fields is usually to approximate the *intrinsic EOFs* (Baldwin et al., 2009; North et al., 1982; Stephenson, 1997), which are properties of the actual underlying continuous field(s), and are independent of any spatial sampling pattern.

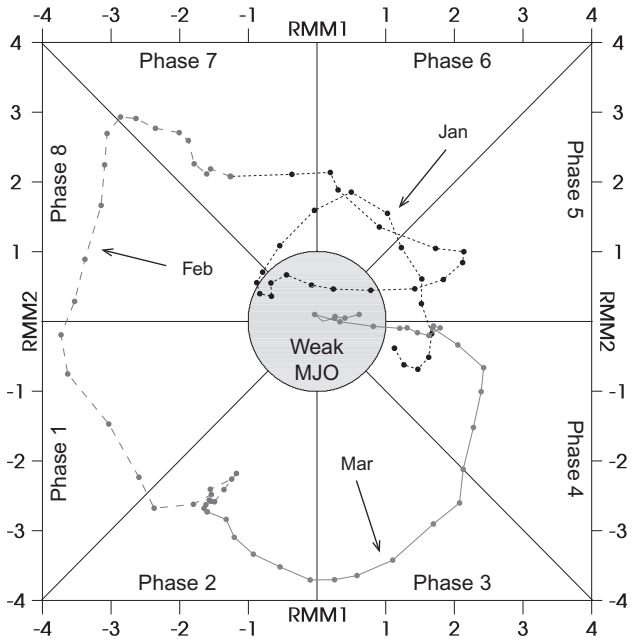

FIGURE 13.8 An example Wheeler–Hendon (2004) diagram, defined by the principal components RMM1 and RMM2, corresponding to the leading eigenvectors shown in Figure 13.7. The counterclockwise trace indicates progression of the MJO during January (dotted), February (dashed), and March (solid), 2009. *Modified from Peatman et al. (2015).*

Data available on a regular latitude–longitude grid is a common cause of this problem. In this case the number of gridpoints per unit area increases with increasing latitude because the meridians converge at the poles, so that a PCA for this kind of gridded data will emphasize high-latitude features and deemphasize low-latitude features. One approach to geographically equalizing the variances is to multiply the data by $\sqrt{\cos\phi}$, where ϕ is the latitude (North et al., 1982). The same effect can be achieved by multiplying each element of the covariance or correlation matrix being analyzed by $\sqrt{\cos\phi_k}\,\sqrt{\cos\phi_\ell}$, where k and ℓ are the indices for the two locations (or location/variable combinations) corresponding to that element of the matrix. The square roots are necessary even though the areas that are proportional to the cosines of the latitudes, because it is the variances and covariances of the analyzed quantities that need to be equalized for the PCA. Baldwin et al. (2009) formulate this process more generally by defining a weighting matrix that can concisely represent the effects of different spatial sampling arrays. Of course these rescalings must be reversed when recovering the original data from the principal components, as in Equations 13.5 and 13.6. An alternative procedure is to interpolate irregularly or nonuniformly distributed data onto an equal-area grid (Araneo and Compagnucci, 2004; Karl et al., 1982). This latter approach is also applicable when the data pertain to an irregularly spaced network, such as climatological observing stations.

Use of extended EOF analysis is not limited to settings involving multiple variables at a common set of locations. A slightly more complicated scaling problem arises when multiple fields with different spatial

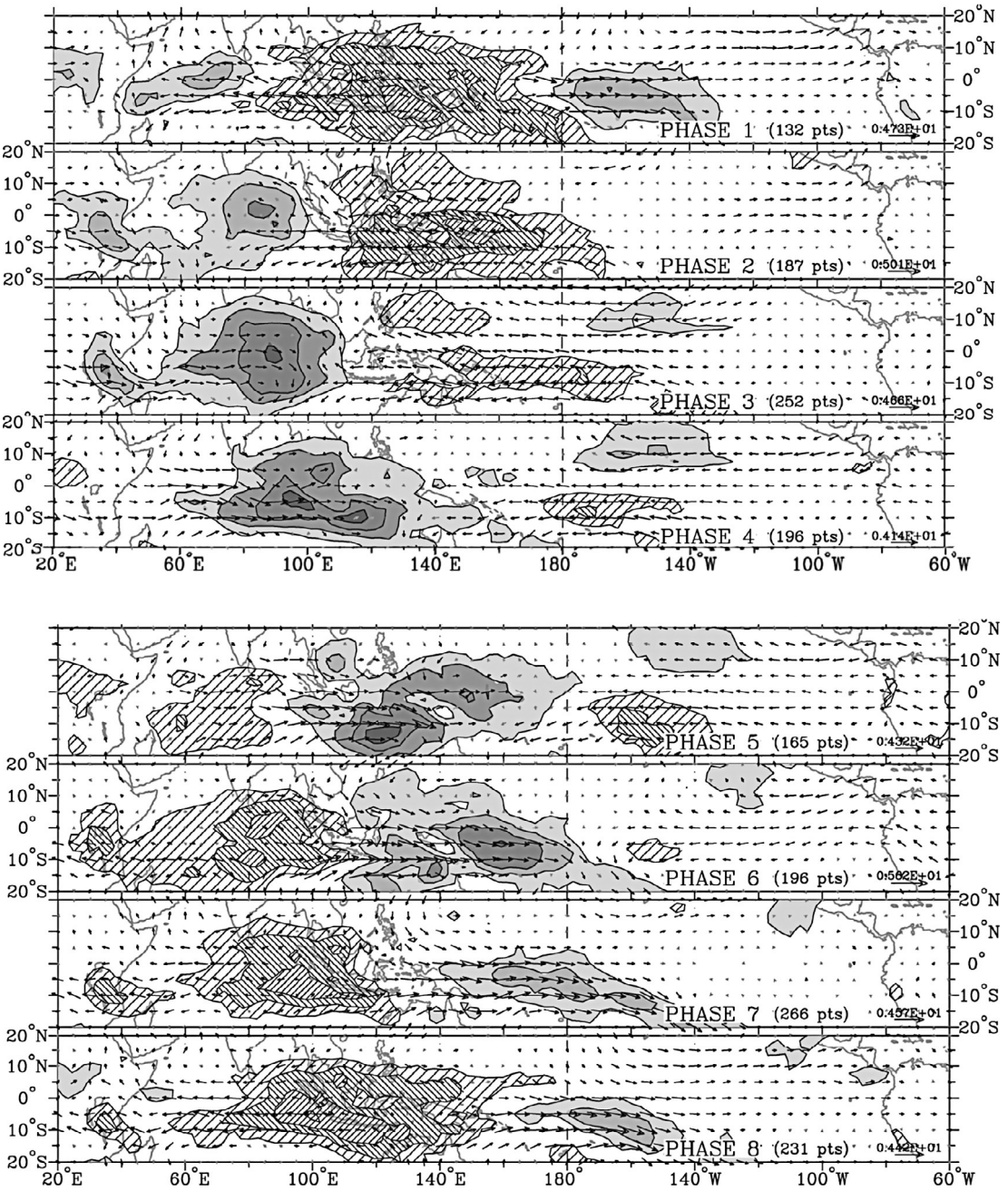

FIGURE 13.9 Composites of December–January–February MJO conditions derived from the eight sectors defined in Figure 13.8, showing regions of enhanced (shaded) and suppressed (hatched) convection, and corresponding 850 mb zonal wind anomalies, as the phenomenon propagates from west to east. *From Wheeler and Hendon (2004). © American Meteorological Society. Used with permission.*

resolutions or spatial domains are simultaneously analyzed with PCA. Here an additional rescaling is necessary to equalize the sums of the variances in each field. Otherwise fields with more gridpoints will dominate the PCA, even if all the fields pertain to the same geographic area.

13.2.4. Domain Size Effects: Buell Patterns

In addition to providing an efficient data compression, results of a PCA are sometimes interpreted in terms of underlying physical processes. For example, the spatial eigenvector patterns in Figure 13.4 have been interpreted as teleconnected modes of atmospheric variability, and the eigenvector displayed in Figure 13.6 reflects the connection between pressure and temperature fields that is expressed as thermal advection. The possibility that informative or at least suggestive interpretations may result can be a strong motivation for computing a PCA.

One problem that can occur when making such interpretations of a PCA for field data arises when the spatial scale (or one or more of the important spatial scales) of the data variations is comparable to or larger than the spatial domain being analyzed. In such cases the space/time variations in the data are still efficiently represented by the PCA, and PCA is still a valid approach to data compression. But the resulting spatial eigenvector patterns take on characteristic shapes that are nearly independent of the underlying variations in the data. These characteristic shapes are called *Buell patterns*, after the author of the paper that first pointed out their existence (Buell, 1979).

Consider, as an artificial but simple example, a 5×5 array of $K = 25$ points representing a square spatial domain. Define the correlations among data values observed at these points to be functions only of their spatial separation d, according to $r(d) = \exp(-d/2)$. The separations of adjacent points in the horizontal and vertical directions are $d = 1$, and so would exhibit correlation $r(1) = 0.61$; points adjacent diagonally would exhibit correlation $r(\sqrt{2}/2) = 0.49$, and so on. This correlation function is shown in Figure 13.10a. It is unchanging across the domain, and produces no spatially distinct features, or preferred patterns of variability. Its spatial scale is comparable to the domain size, which is 4×4 distance units vertically and horizontally, corresponding to $r(4) = 0.14$.

Even though there are no preferred regions of variability within the 5×5 domain, the eigenvectors of the resulting (25×25) correlation matrix $[R]$ appear to indicate that there are. The first of these eigenvectors, which accounts for 34.3% of the variance, is shown in Figure 13.10b. It appears to indicate generally in-phase variations throughout the domain, but with larger amplitude (greater magnitudes of variability) near the center. This first characteristic Buell pattern is an artifact of the mathematics behind the eigenvector calculation if all the correlations are positive and does not merit interpretation beyond its suggestion that the scale of variation of the data is comparable to or larger than the size of the spatial domain.

The dipole patterns in Figures 13.10c and 13.10d are also characteristic Buell patterns and result from the constraint of mutual orthogonality among the eigenvectors. They do not reflect "dipole oscillations" or "seesaws" in the underlying data, whose correlation structure (by virtue of the way this artificial example has been constructed) is homogeneous and isotropic. Here the patterns are oriented diagonally, because opposite corners of this square domain are further apart than opposite sides, but the characteristic dipole pairs in the second and third eigenvectors might instead have been oriented vertically and horizontally in a differently shaped domain. Notice that the second and third eigenvectors account for equal proportions of the variance, and so are actually oriented arbitrarily within the two-dimensional space that they span (see Section 13.4). Additional Buell patterns are sometimes seen in subsequent eigenvectors, the next of which typically suggest tripole patterns of the form – + – or + – +.

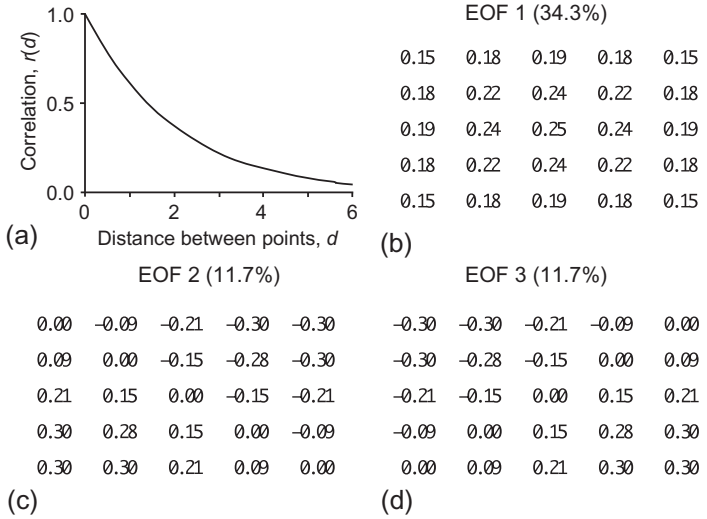

FIGURE 13.10 Artificial example of Buell patterns. Data on a 5×5 square grid with unit vertical and horizontal spatial separation exhibit correlations according to the function of their spatial separations shown in (a). Panels (b)–(d) show the first three eigenvectors of the resulting correlation matrix, displayed in the same 5×5 spatial arrangement. The resulting single central hump (b), and pair of orthogonal dipole patterns (c) and (d), are characteristic artifacts of the domain size being comparable to or smaller than the spatial scale of the underlying data variations.

13.3. TRUNCATION OF THE PRINCIPAL COMPONENTS

13.3.1. Why Truncate the Principal Components?

Mathematically, there are as many eigenvectors of $[S]$ or $[R]$ as there are elements of the data vector $\boldsymbol{x}$, provided $K \leq n-1$. 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 $\boldsymbol{x}$, 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 $\boldsymbol{x}$. Similarly, the last few eigenvectors will point to directions in the K-dimensional space of $\boldsymbol{x}$ where the data jointly exhibit very little variation. This property was illustrated in Example 13.1 for daily temperature values measured at two nearby locations.

To the extent that there is redundancy in the original data $\boldsymbol{x}$, it is possible to capture most of their variance 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 $M < K$ 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 $M << K$, 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. If $M \approx K$ principal components are required to capture a usefully large proportion of the variance in the original data $\boldsymbol{x}$ there is probably little point to computing a PCA.

The truncated representation of the original data can be expressed mathematically by a truncated version of the analysis formula, Equation 13.2, in which the dimension of the truncated $\boldsymbol{u}$ is $(M \times 1)$, and $[E]$ is the (nonsquare, $K \times M$) matrix whose columns consist only of the first M eigenvectors

(i.e., those associated with the largest M eigenvalues) of [S] or [R]. The corresponding synthesis formula, Equation 13.6, is then only approximately true because the original data cannot be exactly resynthesized without using all K eigenvectors.

Where is the appropriate balance between data compression (choosing M to be as small as possible) and avoiding excessive information loss (truncating only a small number, $K - M$, of the principal components)? There is no clear criterion that can be used to choose the number of principal components that are best retained in a given circumstance. The choice of the truncation level can be aided by one or more of the many available principal component selection rules, but it is ultimately a subjective choice that will depend in part on the data at hand and the purpose(s) of the analysis.

13.3.2. Subjective Truncation Criteria

Some approaches to truncating principal components are subjective, or nearly so. Perhaps the most basic criterion is to retain enough of the principal components to represent a "sufficient fraction" of the variances of the original $\boldsymbol{x}$. That is, enough principal components are retained for the total amount of variability represented to be larger than some critical value,

$$\sum_{m=1}^{M} R_m^2 \geq R_{crit}^2, \tag{13.12}$$

where R_m^2 is defined as in Equation 13.4. Of course the difficulty comes in determining how large the fraction R_{crit}^2 must be in order to be considered "sufficient." Ultimately this will be a subjective choice, informed by the analyst's knowledge of the data at hand and the uses to which they will be put. Jolliffe (2002) suggests that $70\% \leq R_{\text{crit}}^2 \leq 90\%$ may often be a reasonable range.

Another essentially subjective approach to principal component truncation is based on the shape of the graph of the eigenvalues λ_m in decreasing order as a function of their index $m = 1, \ldots, K$, known as the *eigenvalue spectrum*. Since each eigenvalue measures the variance represented in its corresponding principal component, this graph is analogous to the power spectrum (see Section 10.5.2), further extending the parallels between EOF and Fourier analyses.

Plotting the eigenvalue spectrum with a linear vertical scale produces what is known as the *scree graph*. 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 the truncation cutoff, M. There is no guarantee that the eigenvalue spectrum for a given PCA will exhibit a single slope separation, or that it (or they) will be sufficiently abrupt to unambiguously locate a cutoff M. Sometimes this approach to principal component truncation is called the scree "test," although this name implies more objectivity and theoretical justification than is warranted: the scree-slope criterion does not involve quantitative statistical inference. Figure 13.11a shows the scree graph (circles) for the PCA summarized in Table 13.1b. This is a relatively well-behaved example, in which the last three eigenvalues are quite small, leading to a fairly distinct bend at $K = 3$, and so to a truncation after the first $M = 3$ principal components would be suggested.

An alternative but similar approach is based on the log-eigenvalue spectrum, or *log-eigenvalue (LEV) diagram*. Choosing a principal component truncation based on 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 LEV diagram as an approximately straight-line portion on its right-hand side. The M

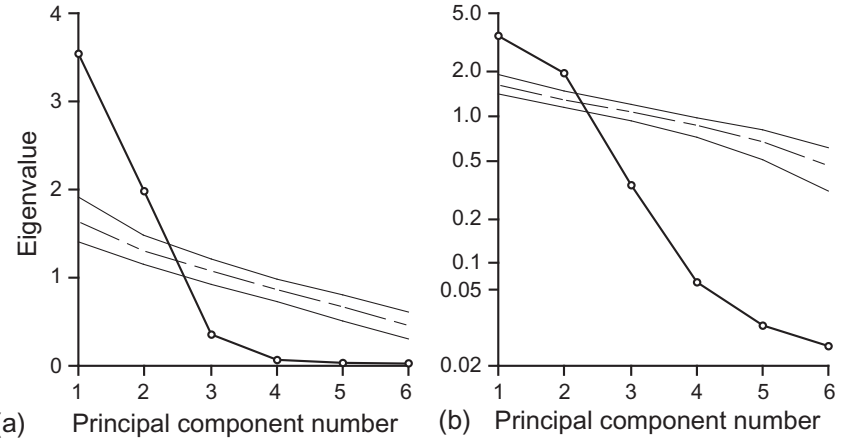

FIGURE 13.11 Graphical displays of eigenvalue spectra, that is, eigenvalue magnitudes as a function of the principal component number (heavier lines connecting circled points), for a $K = 6$-dimensional analysis (see Table 13.1b): (a) Linear scaling, or scree graph, (b) logarithmic scaling, or LEV diagram. 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, on the grounds that only these are significantly larger than what would be expected from data with no correlation structure.

retained principal components would then be the ones whose log-eigenvalues lie above the leftward extrapolation of this line. As before, depending on the data set there may no, or more than one, quasi-linear portions, and their limits may not be clearly defined. Figure 13.11b shows the LEV diagram for the PCA summarized in Table 13.1b. Here $M = 3$ would probably be chosen by most viewers of this LEV diagram, although the choice is not unambiguous.

13.3.3. Rules Based on the Size of the Last Retained Eigenvalue

Another class of principal-component selection rules involves focusing on how small an "important" eigenvalue can be. This set of selection rules can be summarized by the criterion

$$\text{Retain } \lambda_m \text{ if } \lambda_m > \frac{T}{K} \sum_{k=1}^{K} s_{k,k}, \tag{13.13}$$

where $s_{k,k}$ is the sample variance of the kth element of $\boldsymbol{x}$, and T is a threshold parameter.

A simple application of this idea, known as *Kaiser's rule*, involves comparing each eigenvalue (and therefore the variance described by its principal component) to the amount of the joint variance reflected by the average eigenvalue. Principal components whose eigenvalues are above this threshold are retained. That is, Kaiser's rule uses Equation 13.13 with the threshold parameter $T = 1$. Jolliffe (1972, 2002) has argued that Kaiser's rule is too strict (i.e., typically seems to discard too many principal components). He suggests that the alternative $T = 0.7$ often will provide a roughly correct threshold, which allows for the effects of sampling variations.

A third alternative in this class of truncation rules 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 Equation 13.13 is taken to be

$$T(m) = \sum_{j=m}^{K} \frac{1}{j}. \tag{13.14}$$

This rule yields a different threshold for each candidate truncation level, that is, $T = T(m)$, so that the truncation is made at the smallest m for which Equation 13.13 is not satisfied, according to the threshold in Equation 13.14.

All of the three criteria described in this subsection would lead to choosing $M = 2$ for the eigenvalue spectrum in Figure 13.11.

13.3.4. Rules Based on Hypothesis Testing Ideas

Faced with a subjective choice among sometimes vague and possibly conflicting truncation criteria, it is natural to hope for a more objective approach based on the sampling properties of PCA statistics. Section 13.4 describes some large-sample results for the sampling distributions of eigenvalue and eigenvector estimates that have been calculated from multivariate normal samples. Based on these results, Mardia et al. (1979) and Jolliffe (2002) describe tests for the null hypothesis that the last K–M eigenvalues are all equal, and so correspond to noise that should be discarded in the principal component truncation. One problem with this approach occurs when the data being analyzed do not have a multivariate normal distribution and/or are not independent, in which case inferences based on those assumptions may produce serious errors. But a more difficult problem with this approach is that it usually involves examining sequences of tests that are not independent: Are the last two eigenvalues plausibly equal, and if so, are the last three equal, and if so, are the last four equal…? The true test level for a random number of correlated tests will bear an unknown relationship to the nominal level at which each test in the sequence is conducted. This procedure can be used to choose a truncation level, but it will be as much a rule of thumb as the other possibilities already presented in this section, and not a quantitative choice based on a known small probability for falsely rejecting a null hypothesis.

Resampling counterparts to testing-based truncation rules have been used frequently with atmospheric data. The most common of these is known as *Rule N* (Overland and Preisendorfer, 1982; Preisendorfer et al., 1981). Rule N identifies the largest M principal components to be retained on the basis of a sequence of resampling 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 $\boldsymbol{x}$ 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 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 Figure 13.11 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 accordingly, Rule N would choose $M = 2$ for this data.

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 temporal 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 (e.g., Bretherton et al., 1999; Preisendorfer et al., 1981)

$$n' \approx n\frac{1-r_1^2}{1+r_1^2} \tag{13.15}$$

appropriate to eigenvalues and other second moment quantities among the x_k, rather than using n independent vectors of Gaussian variables to construct each synthetic dispersion matrix. Equation 13.15 is analogous to Equation 5.12 pertaining to inferences about means, but the squaring of the lag-1 autocorrelation in Equation 13.15 renders the result much less sensitive to autocorrelation effects.

Another potential problem with Rule N, and other similar procedures, is that the data $\boldsymbol{x}$ may not be approximately Gaussian. For example, one or more of the x_k's could be precipitation variables. To the extent that the original data are not Gaussian, the random number generation procedure will not simulate accurately the underlying physical process, and the results of the test 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.

The primary weakness of the Rule N procedure derives from the fact that only its test for the leading eigenvalue is correct. The reason is that, having rejected the proposition that λ_1 is not different from the others, the Monte Carlo sampling distributions for the remaining eigenvalues are no longer meaningful because they are conditional on all K eigenvalues reflecting noise. That is, these synthetic sampling distributions will imply too much variance if it has been inferred that λ_1 has more than a random share, since the sum of the eigenvalues is constrained to equal the total variance. Accordingly, Priesendorfer (1988) notes that Rule N tends to retain too few principal components.

A better approach (Wilks, 2016c) is to test the sequence of scaled eigenvalues

$$\lambda_k^* = \frac{\lambda_k}{\dfrac{1}{N_{rank}-k+1}\displaystyle\sum_{i=k}^{N_{rank}}\lambda_k}, \tag{13.16}$$

where $N_{\text{rank}} = \min(n\text{–}1, K)$ is the number of nonzero eigenvalues. Equation 13.16 is the fraction of variance represented by λ_k when the larger eigenvalues have been omitted from the denominator, for which analytic sampling distributions (Tracy and Widom, 1996) are available for sufficiently large n and K. Good approximations to the right tails of these sampling distributions, which are different for each $\lambda^*{}_k$, are provided by Pearson III distributions (Equation 4.55), with parameters

$$\alpha = 46.4 \tag{13.17a}$$

$$\beta_k = \frac{0.186\,\sigma_k}{\max(n_k, K^*)} \tag{13.17b}$$

and

$$\zeta_k = \frac{\mu_k - 9.85\,\sigma_k}{\max(n_k, K^*)}, \tag{13.17c}$$

where

$$\mu_k = \left(\sqrt{n_k - 1/2} + \sqrt{K^* - 1/2}\right)^2 \tag{13.18a}$$

and

$$\sigma_k = \sqrt{\mu_k}\left(\frac{1}{\sqrt{n_k - 1/2}} + \frac{1}{\sqrt{K^* - 1/2}}\right)^{1/3} \tag{13.18b}$$

are the Tracy–Widom location and scale parameters, $K^* = K–k+1$, and $n_k = n–k+1$. The sequence of tests, $k = 1, 2, \ldots, N_{\text{rank}}$ are computed, each pertaining to the null hypothesis H_k:$\{\lambda_k = \lambda_{k+1} = \lambda_{k+2} = \cdots = \lambda_{\text{Nrank}}\}$ that the eigenvalue λ_k and smaller eigenvalues represent only noise and so are equal. The truncation point M is chosen to be one smaller than the first k for which H_k is not rejected (i.e., for which λ^*_k is smaller than the 1–α quantile of the relevant Pearson III distribution defined by Equations 13.17 and 13.18). That is,

$$M = \min\left(k \in \{1, 2, \ldots, N_{rank} - 1\} : p_k > \alpha\right) - 1, \tag{13.19}$$

where p_k is the p value pertaining to the kth test and the null hypothesis H_k. The first (k =1) of these tests will be equivalent to the Rule N test for the first eigenvalue, but the subsequent tests account for the larger-than-random fractions of variance in the lower indexed eigenvalues previously judged to be significant.

Example 13.4 Principal Component Truncation Using Sequential Testing

Table 13.1b presents the PCA computed from the standardized January 1987 data in Table A.1. The scree and LEV diagrams for the eigenvalues λ_k are plotted in Figure 13.11, together with the Monte Carlo distributions for Rule N. However, because the two leading eigenvalues clearly represent more of the variance than would their counterparts derived from purely random data, the Rule-N Monte Carlo distributions for the trailing eigenvalues are too large, which may lead to too few eigenvalues being retained according to this criterion.

Table 13.4 contains the information for this problem derived from Equations 13.16–13.18, leading to a sequence of hypothesis tests allowing estimation of the truncation point M. For the first ($k = 1$) of these tests, $\lambda_k = \lambda^*_k$ because the average over all six eigenvalues (denominator of Equation 13.16) is 1 for this PCA based on a correlation matrix. The quantity $(\lambda^*_1 - \zeta_1)/\beta_1$ should be distributed according to a gamma distribution with $\alpha = 46.4$ (Equation 13.17a) and $\beta = 1$ if the null hypothesis that all the (underlying generating-process) eigenvalues are equal, which is strongly rejected for the $k = 1$ test. Similarly, $(\lambda^*_2 - \zeta_2)/\beta_2$ and $(\lambda^*_3 - \zeta_3)/\beta_3$ are both on the far right tail of this gamma distribution, and so the null hypotheses H_2 and H_3 are strongly rejected as well. The p value p_4, for the $k = 4$ test, is not small, leading to H_4 being the first null hypothesis not rejected, so that the truncation point $M = 3$ is chosen by this procedure according to Equation 13.19. The more conservative Rule N procedure retains only the first two principal components. ◇

13.3.5. Rules Based on Structure in the Retained Principal Components

The truncation rules presented so far all relate to the magnitudes of the eigenvalues. The possibility that physically important principal components need not have the largest variances (i.e., eigenvalues) has

TABLE 13.4 Quantities from Equations 13.16–13.18 Applied to Truncation of the PCA Presented in Table 13.1b

k	n_k	K^*	μ_k	σ_k	β_k	ζ_k	λ_k	λ^*_k	$(\lambda^*_k - \zeta_k)/\beta_k$	p_k Value
1	31	6	61.90	6.663	.04000	−.1203	3.532	3.532	91.31	8×10^{-8}
2	30	5	57.04	6.561	.04068	−.2529	1.985	4.021	105.06	4×10^{-11}
3	29	4	51.97	6.467	.04148	−.4045	0.344	2.849	78.43	3.7×10^{-5}
4	28	3	46.58	6.396	.04249	−.5864	0.074	1.597	51.39	0.227
5	27	2	40.61	6.395	.04406	−.8289	0.038	1.169	45.34	0.543

motivated a class of truncation rules based on expected characteristics of physically important principal component series (Preisendorfer et al., 1981; Preisendorfer, 1988). Since most atmospheric data that are subjected to PCA are time series (e.g., time sequences of spatial fields recorded at K gridpoints), a plausible hypothesis may be that principal components corresponding to physically meaningful processes should exhibit time dependence, because the underlying physical processes are expected to exhibit time dependence. Preisendorfer et al. (1981) and Preisendorfer (1988) proposed several such truncation rules, which test null hypotheses that the individual principal component time series are uncorrelated, using either their power spectra or their autocorrelation functions. The truncated principal components are then those for which this null hypothesis is not rejected. This class of truncation rule seems to have been used very little in practice.

13.4. SAMPLING PROPERTIES OF THE EIGENVALUES AND EIGENVECTORS

13.4.1. Asymptotic Sampling Results for Multivariate Normal Data

Principal component analyses are calculated from finite data samples and are as subject to sampling variations as any other statistical estimation procedure. That is, we rarely if ever know the true covariance matrix $[\Sigma]$ for the population or underlying generating process, but rather estimate it using the sample counterpart $[S]$. Accordingly the eigenvalues and eigenvectors calculated from $[S]$ are also estimates based on the finite sample and are thus subject to sampling variations. Understanding the nature of these variations is quite important to correct interpretation of the results of a PCA.

The equations presented in this section must be regarded as approximate, as they are asymptotic (large-n) results, and are based also on the assumption that the underlying $\boldsymbol{x}$'s have a multivariate normal distribution. It is also assumed that no pair of the population eigenvalues are equal, implying (in the sense to be explained in Section 13.4.2) that all the population eigenvectors are well defined. The validity of these results is therefore approximate in most circumstances, but they are nevertheless quite useful for understanding the nature of sampling effects on the uncertainty about estimated eigenvalues and eigenvectors.

Eigenvalue Results

The basic result for the sampling properties of estimated eigenvalues is that, in the limit of very large sample size, their sampling distribution is unbiased, and multivariate normal,

$$\sqrt{n}(\hat{\boldsymbol{\lambda}}-\boldsymbol{\lambda}) \sim N_K\left(\mathbf{0}, 2[\Lambda]^2\right), \tag{13.20a}$$

or

$$\hat{\boldsymbol{\lambda}} \sim N_K\left(\boldsymbol{\lambda}, \frac{2}{n}[\Lambda]^2\right). \tag{13.20b}$$

Here $\hat{\boldsymbol{\lambda}}$ is the $(K \times 1)$ vector of estimated eigenvalues; $\boldsymbol{\lambda}$ is its true value; and the $(K \times \mathrm{K})$ matrix $[\Lambda]^2$ is the square of the diagonal, population eigenvalue matrix, having elements ${\lambda_k}^2$. Because $[\Lambda]^2$ is diagonal the sampling distributions for each of the K estimated eigenvalues are (approximately) independent univariate Gaussian distributions,

$$\sqrt{n}\left(\hat{\lambda}_k - \lambda_k\right) \sim N\left(0, 2\lambda_k^2\right), \tag{13.21a}$$

or

$$\hat{\lambda}_k \sim N\left(\lambda_k, \frac{2}{n}\lambda_k^2\right). \tag{13.21b}$$

Equations 13.20 and 13.21 are large-sample approximations, and there is a bias in the sample eigenvalues for finite sample size. In particular, the largest eigenvalues will be overestimated (will tend to be larger than their population counterparts) and the smallest eigenvalues will tend to be underestimated, and these effects increase with decreasing sample size (Quadrelli et al., 2005; Von Storch and Hannoschock, 1985). These biases can be understood as a consequence of the sorting of the sample eigenvalues, so that the largest sample eigenvalue will be labeled as λ_1 regardless of the rank of its generating-process counterpart, and similarly the smallest sample eigenvalue will be labeled as λ_K. The Monte Carlo Rule N distributions in Figure 13.11 illustrate the results of this phenomenon for a small sample ($n = 31$, $K = 6$) situation where the true underlying eigenvalue spectrum is completely flat, with all generating-process $\lambda_k = 1$.

Using Equation 13.21a to construct a standard Gaussian variate provides an expression for the distribution of the relative error of the eigenvalue estimate,

$$z = \frac{\sqrt{n}\left(\hat{\lambda}_k - \lambda_k\right) - 0}{\sqrt{2}\lambda_k} = \sqrt{\frac{n}{2}}\left(\frac{\hat{\lambda}_k - \lambda_k}{\lambda_k}\right) \sim N(0,1). \tag{13.22}$$

Equation 13.22 implies

$$\Pr\left\{\left|\sqrt{\frac{n}{2}}\left(\frac{\hat{\lambda}_k - \lambda_k}{\lambda_k}\right)\right| \le z(1-\alpha/2)\right\} = 1-\alpha, \tag{13.23}$$

which leads to the (1–α)•100% confidence interval for the kth eigenvalue,

$$\frac{\hat{\lambda}_k}{1+z(1-\alpha/2)\sqrt{2/n}} \le \lambda_k \le \frac{\hat{\lambda}_k}{1-z(1-\alpha/2)\sqrt{2/n}}. \tag{13.24}$$

Eigenvector Results

The elements of each sample eigenvector are approximately unbiased, and their sampling distributions are approximately multivariate normal. But the variances of the multivariate normal sampling

distributions for each of the eigenvectors depend on all the other eigenvalues and eigenvectors in a somewhat complicated way. The sampling distribution for the kth eigenvector is

$$\hat{\boldsymbol{e}}_k \sim N_K(\boldsymbol{e}_k, [V_{\boldsymbol{e}_k}]), \tag{13.25}$$

where the covariance matrix for this distribution is

$$[V_{\boldsymbol{e}_k}] = \frac{\lambda_k}{n} \sum_{i \neq k}^{K} \frac{\lambda_i}{(\lambda_i - \lambda_k)^2} \boldsymbol{e}_i \boldsymbol{e}_i^T. \tag{13.26}$$

The summation in Equation 13.26 involves all K eigenvalue–eigenvector pairs, indexed here by i, *except* the kth pair, for which the covariance matrix is being calculated. It is a sum of weighted outer products of these eigenvectors, and so resembles the spectral decomposition of the true covariance matrix $[\Sigma]$ (cf. Equation 11.53). But rather than being weighted only by the corresponding eigenvalues, as in Equation 11.53, they are weighted also by the reciprocals of the squares of the differences between those eigenvalues, and the eigenvalue belonging to the eigenvector whose covariance matrix is being calculated. That is, the elements of the matrices in the summation of Equation 13.26 will be quite small, except for those that are paired with eigenvalues λ_i that are close in magnitude to the eigenvalue λ_k belonging to the eigenvector whose sampling distribution is being calculated.

13.4.2. Effective Multiplets

Equation 13.26, for the sampling uncertainty of the eigenvectors of a covariance matrix, has two important implications. First, the pattern of uncertainty in the estimated eigenvectors resembles a linear combination, or weighted sum, of all the *other* eigenvectors. Second, because the magnitudes of the weights in this weighted sum are inversely proportional to the squares of the differences between the corresponding eigenvalues, an eigenvector will be relatively precisely estimated (the sampling variances will be relatively small) if its eigenvalue is well separated from the other K–1 eigenvalues. Conversely, eigenvectors whose eigenvalues are similar in magnitude to one or more of the other eigenvalues will exhibit large sampling variations, and those variations will be larger for the eigenvector elements that are large in the eigenvectors with comparable eigenvalues.

The joint effect of these two considerations is that the sampling distributions of two (or more) eigenvectors having similar eigenvalues will be closely entangled. Their sampling variances will be large, and their patterns of sampling error will resemble the patterns of the eigenvector(s) with which they are entangled. The net effect will be that a realization of the corresponding sample eigenvectors will be a nearly arbitrary mixture of the true population counterparts. They will jointly represent the same amount of variance (within the sampling bounds approximated by Equation 13.21), but this joint variance will be arbitrarily mixed between (or among) them. Sets of such eigenvalue–eigenvector pairs are called effectively degenerate multiplets or *effective multiplets*. Attempts at physical interpretation of such sample eigenvectors will be frustrating if not hopeless.

The source of this problem can be appreciated in the context of a three-dimensional multivariate normal distribution, in which one of the eigenvectors is relatively large, and the two smaller ones are nearly equal. The resulting distribution has ellipsoidal probability contours resembling the cucumbers in Figure 12.3. The eigenvector associated with the single large eigenvalue will be aligned with the long axis of the ellipsoid. But this multivariate normal distribution has (essentially) no preferred direction in the plane perpendicular to the long axis (exposed face on the left-hand cucumber in Figure 12.3b). Any pair of perpendicular vectors that are also perpendicular to the long axis could as easily jointly represent variations in this plane. The leading

eigenvector calculated from a sample covariance matrix from this distribution would be closely aligned with the true leading eigenvector (long axis of the cucumber) because its sampling variations will be small. In terms of Equation 13.26, both of the two terms in the summation would be small because $\lambda_1 >> \lambda_2 \approx \lambda_3$. On the other hand, each of the other two eigenvectors would be subject to large sampling variations: the term in Equation 13.26 corresponding to one or the other of them will be large, because $(\lambda_2 - \lambda_3)^{-2}$ will be large. The pattern of sampling error for $\boldsymbol{e}_2$ will resemble the true generating-process $\boldsymbol{e}_3$, and vice versa. That is, the orientation of the two sample eigenvectors in this plane will be arbitrary, beyond the constraints that they will be perpendicular to each other, and to $\boldsymbol{e}_1$. The variations represented by each of these two sample eigenvectors will accordingly be an arbitrary mixture of the variations represented by their two population counterparts.

13.4.3. The North et al. Rule of Thumb

Equations 13.20 and 13.25, for the sampling distributions of the eigenvalues and eigenvectors, depend on the values of their true but unknown counterparts. Nevertheless, the sample estimates approximate the true values, so that large sampling errors are expected for those eigenvectors whose sample eigenvalues are close to other sample eigenvalues. The idea that it is possible to diagnose instances where sampling variations are expected to cause problems with eigenvector interpretation in PCA was expressed as a rule of thumb by North et al. (1982):

> *"The rule is simply that if the sampling error of a particular eigenvalue* λ *[*$\delta\lambda \sim \lambda(2/\mathrm{n})^{1/2}$*] is comparable to or larger than the spacing between* λ *and a neighboring eigenvalue, then the sampling errors for the EOF associated with* λ *will be comparable to the size of the neighboring EOF. The interpretation is that if a group of true eigenvalues lie within one or two* $\delta\lambda$ *of each other, then they form an 'effectively degenerate multiplet,' and sample eigenvectors are a random mixture of the true eigenvectors."*

However, caution is warranted in quantitatively interpreting the degree of overlap of the confidence intervals implied by the North et al. rule of thumb (see Section 5.2.2).

North et al. (1982) illustrated their rule of thumb with an instructive example. They constructed synthetic data from a set of known EOF patterns, the first four of which are shown in Figure 13.12a, together with their respective eigenvalues. Using a full set of such patterns, the covariance matrix $[\Sigma]$ from which they could be extracted was assembled using the spectral decomposition (Equation 11.53). Using $[\Sigma]^{1/2}$ (see Section 11.3.4), realizations of data vectors $\boldsymbol{x}$ from a distribution with covariance $[\Sigma]$ were generated as in Section 12.4. Figure 13.12b shows the first four eigenvalue–eigenvector pairs calculated from a sample of $n = 300$ such synthetic data vectors, and Figure 13.12c shows a realization of the leading eigenvalue–eigenvector pairs for $n = 1000$.

The leading four true eigenvector patterns in Figure 13.12a are visually distinct, but their eigenvalues are relatively close. Using Equation 13.21b and $n = 300$, 95% sampling intervals for the four eigenvalues are 14.02 ± 2.24, 12.61 ± 2.02, 10.67 ± 1.71, and 10.43 ± 1.67 (because $\Phi^{-1}(0.975) = 1.96$), all of which include the adjacent eigenvalues. Therefore it is expected according to the rule of thumb that the sample eigenvectors will be random mixtures of their population counterparts for this sample size, and Figure 13.12b bears out this expectation: the patterns in those four panels appear to be random mixtures of the four panels in Figure 13.12a. Even if the true eigenvectors were unknown, this conclusion would be expected from the North et al. rule of thumb, because adjacent sample eigenvectors in Figure 13.12b are within two estimated standard errors, or $2\,\hat{\delta\lambda} = 2\hat{\lambda}(2/n)^{1/2}$ of each other.

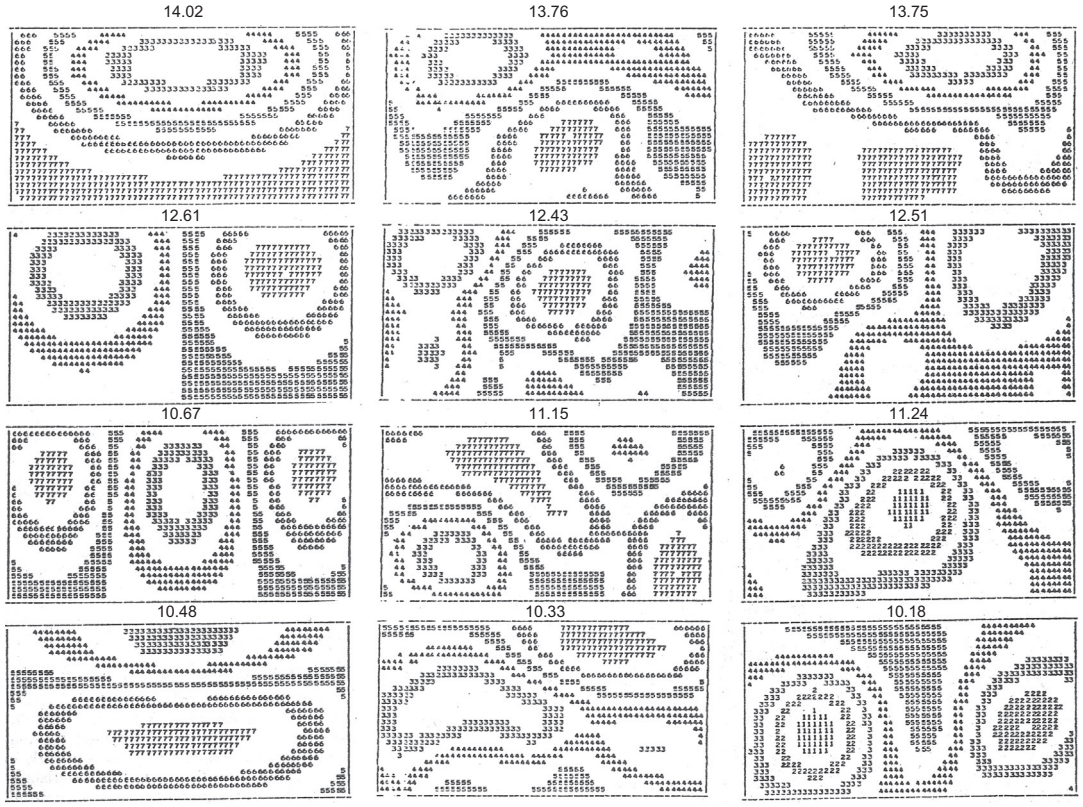

FIGURE 13.12 The North et al. (1982) example for effective degeneracy. (a) First four eigenvectors for the population from which synthetic data were drawn, with corresponding eigenvalues. (b) The first four eigenvectors calculated from a sample of $n = 300$, and the corresponding sample eigenvalues. (c) The first four eigenvectors calculated from a sample of $n = 1000$, and the corresponding sample eigenvalues. *From North et al. (1982). © American Meteorological Society. Used with permission.*

The situation is somewhat different for the larger sample size (Figure 13.12c). Again using Equation 13.21b but with $n = 1000$, the 95% sampling intervals for the four generating-process eigenvalues are 14.02 ± 1.22, 12.61 ± 1.10, 10.67 ± 0.93, and 10.43 ± 0.91. These intervals indicate that the first two sample EOFs should be reasonably distinct from each other and from the other EOFs, but that the third and fourth eigenvectors will probably still be entangled. Applying the rule of thumb to the sample eigenvalues in Figure 13.12c indicates that the separation between all adjacent pairs is close to $2\,\delta\hat{\lambda}$. The additional sampling precision provided by the larger sample size allows an approximation to the first two true EOF patterns to emerge, although an even larger sample still would be required before the sample eigenvectors would correspond well to their population counterparts.

The synthetic data realizations $\boldsymbol{x}$ in this artificial example were chosen independently of each other. If the data being analyzed are serially correlated, the unadjusted rule of thumb will imply better eigenvalue separation than is actually the case, because the variance of the sampling distribution of the sample eigenvalues will be larger than $2\,\lambda_k^2/n$ (as given in Equation 13.21). The cause of this discrepancy is that the sample eigenvalues are less consistent from batch to batch when calculated from

autocorrelated data, so the qualitative effect is the same as was described for the sampling distribution of sample means, in Section 5.2.4. However, the effective sample size adjustment in Equation 13.15 would be appropriate in this case, which implies a much less extreme effect on the effective sample size than does Equation 5.12. Here r_1 would correspond to the lag-1 autocorrelation for the corresponding principal component time series when using Equation 13.21 or 13.24; and to the geometric mean of the autocorrelation coefficients for the two corresponding principal component series, when using Equation 13.26.

13.4.4. Bootstrap Approximations to the Sampling Distributions

The conditions specified in Section 13.4.1, of large sample size and/or underlying multivariate normal data, may be too unrealistic to be practical in some situations. In such cases it is possible to build good approximations to the sampling distributions of sample statistics using the bootstrap (see Section 5.3.5). Beran and Srivastava (1985) and Efron and Tibshirani (1993) specifically describe bootstrapping sample covariance matrices to produce sampling distributions for their eigenvalues and eigenvectors. The basic procedure is to repeatedly resample the underlying data vectors $\boldsymbol{x}$ with replacement; to produce some large number, n_B, of bootstrap samples, each of size n. Each of the n_B bootstrap samples yields a bootstrap realization of $[S]$, whose eigenvalues and eigenvectors can be computed. Jointly these bootstrap realizations of eigenvalues and eigenvectors form reasonable approximations to the respective sampling distributions, which will reflect properties of the underlying data that may not conform to those assumed in Section 13.4.1.

Be careful in interpreting these bootstrap distributions. A (correctable) difficulty arises from the fact that the eigenvectors are determined up to sign only, so that in some bootstrap samples the counterpart of $\boldsymbol{e}_k$ may very well be $-\boldsymbol{e}_k$. Failure to rectify such arbitrary sign switches will lead to large and unwarranted inflation of the computed sampling distributions for the eigenvector elements. Difficulties can also arise when resampling effective multiplets, because the random distribution of variance within a multiplet may be different from resample to resample, so the resampled eigenvectors may not bear one-to-one correspondences with their original sample counterparts. Finally, the bootstrap procedure destroys any serial correlation that may be present in the underlying data, which would lead to unrealistically narrow bootstrap sampling distributions. The moving-blocks bootstrap can be used for serially correlated data vectors (Wilks, 1997b) as well as scalars. Wang et al. (2014) provide an example using monthly surface pressure data.

13.5. ROTATION OF THE EIGENVECTORS

13.5.1. Why Rotate the Eigenvectors?

There is a strong tendency to try to ascribe physical interpretations to PCA eigenvectors and the corresponding principal components. The results shown in Figures 13.4 and 13.6 indicate that it can be both appropriate and informative to do so. However, the orthogonality constraint on the eigenvectors (Equation 11.48) can lead to problems with these interpretations, especially for the second and subsequent principal components. Although the orientation of the first eigenvector is determined solely by the direction of the maximum variation in the data, subsequent vectors must be orthogonal to each higher-variance eigenvector, regardless of the nature of the physical processes that may have given rise to the data. To the extent that those underlying physical processes are not independent, interpretation of the corresponding principal components as being independent modes of variability will not be justified (North, 1984). The first principal component may represent an important mode of variability or physical

process, but 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 chosen using a truncation criterion such as those discussed in Section 13.3. Rotated eigenvectors can be less prone to the artificial features resulting from the orthogonality constraint on the unrotated eigenvectors, such as Buell patterns (Richman, 1986). They also appear to exhibit better sampling properties (Richman, 1986; Cheng et al., 1995) than their unrotated counterparts. A large fraction of the review of PCA by Hannachi et al. (2007) is devoted to rotation.

Several procedures for rotating the original eigenvectors exist, but all seek to produce what is known as *simple structure* in the resulting analysis. Simple structure generally is understood to have been achieved if a large fraction of the elements of the resulting rotated vectors are near zero, and few of the remaining elements correspond to elements that are not near zero in the other rotated vectors. The desired result is that each rotated vector represents mainly the few original variables corresponding to 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. Simple structure aids interpretation of a rotated PCA to the extent that it allows association of each rotated eigenvector with a small number of the original K variables whose corresponding eigenvector elements are not near zero.

Following rotation of the eigenvectors, a second set of new variables is defined, called *rotated principal components*. The rotated principal components are obtained from the original data analogously to Equation 13.1 and 13.2, as the dot products of data vectors and the rotated eigenvectors. They can be interpreted as single-number summaries of the similarity between their corresponding rotated eigenvector and a data vector $\boldsymbol{x}$. Depending on the method used to rotate the eigenvectors, the resulting rotated principal components may or may not be mutually uncorrelated.

A price is paid for the improved interpretability and better sampling stability of the rotated eigenvectors. One cost is that the dominant-variance property of PCA is lost. The first rotated principal component is no longer that linear combination of the original data with the largest variance. The variance represented by the original unrotated eigenvectors is spread more uniformly among the rotated eigenvectors, so that the corresponding eigenvalue spectrum is flatter. Also lost is either the orthogonality of the eigenvectors, or the uncorrelatedness of the resulting principal components, or both.

13.5.2. Rotation Mechanics

Rotated eigenvectors are produced as a linear transformation of a subset of M of the original K eigenvectors,

$$\underset{(\mathrm{KxM})}{\left[\widetilde{E}\right]} = \underset{(\mathrm{KxM})}{[E]}\ \underset{(\mathrm{MxM})}{[T]}\ , \tag{13.27}$$

where $[T]$ is the rotation matrix, and the matrix of rotated eigenvectors is denoted by the tilde. If $[T]$ is orthogonal, that is, if $[T][T]^{\mathrm{T}} = [I]$, then the transformation Equation 13.27 is called an *orthogonal rotation*. Otherwise the rotation is called *oblique*.

Richman (1986) lists 19 approaches to defining the rotation matrix $[T]$ in order to achieve simple structure, although his list is not exhaustive. However, by far the most commonly used approach is

the orthogonal rotation called the *varimax* (Kaiser, 1958). A varimax rotation is determined by choosing the elements of [*T*] to maximize

$$\sum_{m=1}^{M}\left[\sum_{k=1}^{K} {e_{k,m}^{*}}^{4} - \frac{1}{K}\left(\sum_{k=1}^{K} {e_{k,m}^{*}}^{2}\right)^{2}\right], \tag{13.28a}$$

where

$$e_{k,m}^{*} = \frac{\widetilde{e}_{k,m}}{\left(\sum_{m=1}^{M} \widetilde{e}_{k,m}^{2}\right)^{1/2}}, \tag{13.28b}$$

are scaled versions of the rotated eigenvector elements. Together Equations 13.28a and 13.28b define the "normal varimax," whereas Equation 13.28a alone, using the unscaled eigenvector elements $\widetilde{e}_{k,m}$, is known as the "raw varimax." In either case the transformation is sought that maximizes the sum of the variances of the (either scaled or raw) squared rotated eigenvector elements, which tends to move them toward either their maximum or minimum (absolute) values (which are 0 and 1), and thus tends toward simple structure. The solution is iterative and is a standard feature of many statistical software packages.

The results of eigenvector rotation can depend on how many of the original eigenvectors are selected for rotation. That is, some or all of the leading rotated eigenvectors may be different if, say, $M+1$ rather than M eigenvectors are rotated (e.g., O'Lenic and Livezey, 1988). Unfortunately there is often not a clear answer to the question of what the best choice for M might be, and typically an essentially subjective choice is made. Some guidance is available from the various truncation criteria in Section 13.3, although these may not yield a unique answer. Sometimes a trial-and-error procedure is used, where M is increased slowly until the leading rotated eigenvectors are stable, that is, insensitive to further increases in M. In any case, however, it makes sense to include either all, or none, of the eigenvectors making up an effective multiplet, since jointly they carry information that has been arbitrarily mixed. Jolliffe (1987, 1989) suggests that it may be helpful to separately rotate groups of eigenvectors within effective multiplets in order to more easily interpret the information that they jointly represent.

Figure 13.13 compares unrotated and varimax-rotated PCAs for reconstructing spatial patterns that are independent (Figure 13.13a) and overlapping (Figure 13.13b). Both synthetic examples pertain to a 30×30-gridpoint square domain ($K = 900$), with $n = 256$. The leftmost columns in each panel show the three true generating-process eigenvectors, the nonzero features of which in Figure 13.13a are spatially disjoint. In this case both the unrotated (middle column) and rotated (rightmost column) recover the true patterns well. Because the underlying spatial patterns of variability in the leftmost column of Figure 13.13a already exhibit "simple structure," the unrotated and rotated solutions are equivalent because the rotation matrix [*T*] in Equation13.27, implicitly resulting from Equation 13.28, very nearly equals the identity.

The features in the underlying "truth" eigenvectors in Figure 13.13b have overlapping spatial extents, and so cannot vary independently. The unrotated PCA in the middle column of Figure 13.13b is not able to separate the three. Here the first and third eigenvectors include portions of the underlying variability of the second mode, and the second eigenvector includes influences from the first and third underlying modes. In contrast, the varimax-rotated solution in the rightmost column of Figure 13.13b recovers the three true underlying modes well.

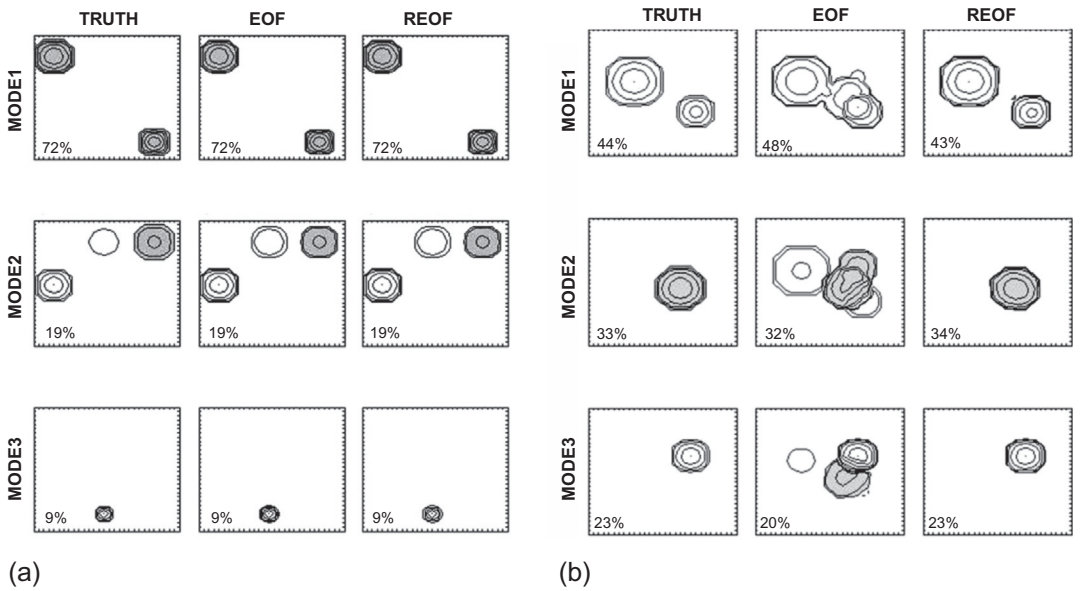

FIGURE 13.13 Synthetic example comparing unrotated and rotated PCAs when three underlying modes of variability are (a) spatially independent, and (b) spatially overlapping and thus nonindependent. The square spatial domain consists of 30 gridpoints in each direction, the contour interval is 0.25, and areas with negative eigenvector loadings are shaded. *Modified from Lian and Chen (2012). © American Meteorological Society. Used with permission.*

Figure 13.14 shows spatial displays of the first two rotated eigenvectors of monthly averaged hemispheric winter 500 mb heights. Using the truncation criterion of Equation 13.13 with $T = 1$, the first 19 eigenvectors of the correlation matrix for these data were rotated. The two patterns in Figure 13.14 are similar to the first two unrotated eigenvectors derived from the same data (see Figure 13.4a and 13.4b), 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 Figure 13.14, and each panel emphasizes more uniquely a particular feature of the variability of the 500 mb heights corresponding to the teleconnection patterns in Figure 3.33. The rotated vector in Figure 13.14a 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 Figure 14.14b, where the rotation has separated it from the eastern hemisphere pattern evident in Figure 13.4b.

Figure 13.15 shows schematic representations of eigenvector rotation in two dimensions. The upper diagrams in each section represent the eigenvectors in the two-dimensional plane defined by the underlying variables x_1 and x_2, and the corresponding lower diagrams represent "maps" of the eigenvector elements plotted at the two "locations" x_1 and x_2 (these are meant to correspond to real-world maps such as those shown in Figures 13.4 and 13.14). Figure 13.15a illustrates the case of the original unrotated eigenvectors. The leading eigenvector $\boldsymbol{e}_1$ is defined as the direction onto which a projection of the data

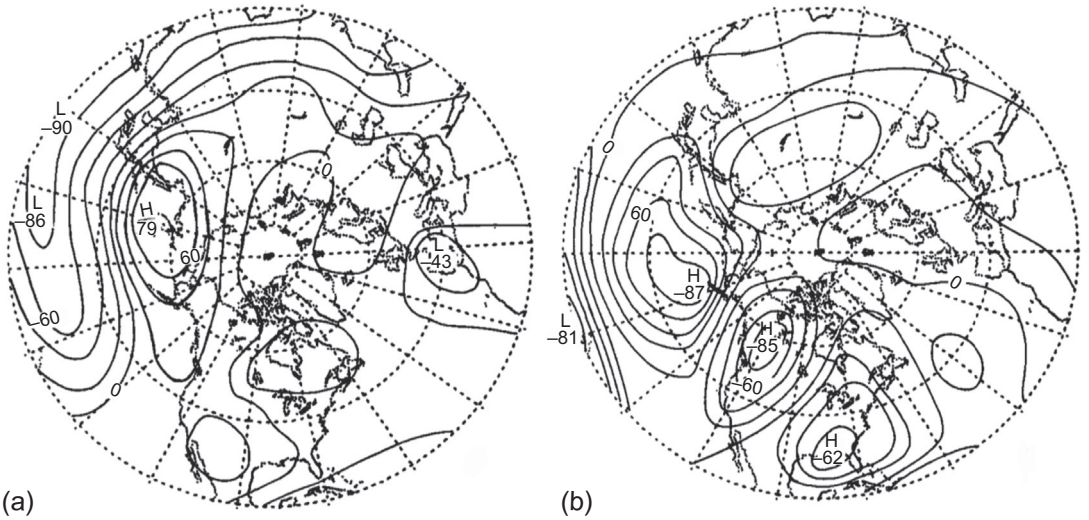

FIGURE 13.14 Spatial displays of the first two rotated eigenvectors of monthly averaged hemispheric winter 500 mb heights. The data are the same as those underlying Figure 13.4, but the rotation has better isolated the patterns of variability, allowing a clearer interpretation in terms of the teleconnection patterns in Figure 3.33. *From Horel (1981). © American Meteorological Society. Used with permission.*

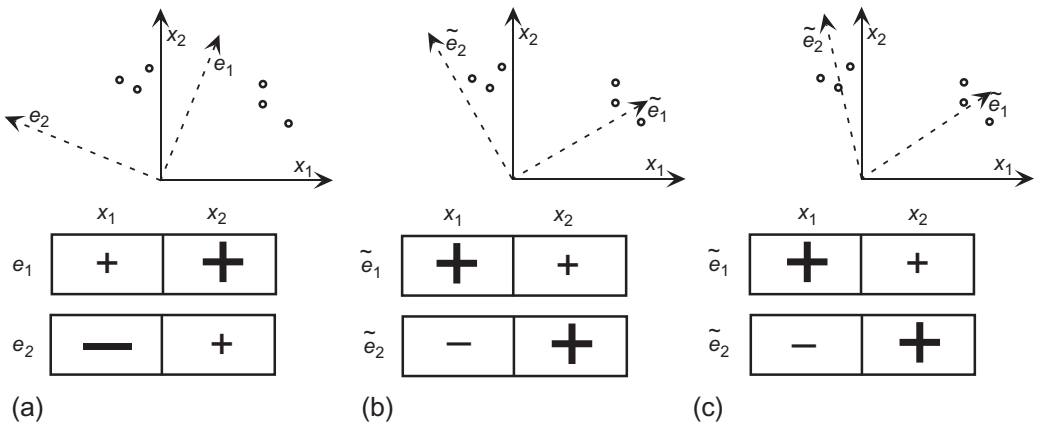

FIGURE 13.15 Schematic comparison of (a) unrotated, (b) orthogonally rotated, and (c) obliquely rotated unit-length eigenvectors in $K = 2$ dimensions. Left panels show eigenvectors in relation to scatterplots of the data, which exhibit two groups or modes. Right panels show schematic two-point maps of the two eigenvectors in each case. *After Karl and Koscielny (1982).*

points (i.e., the principal components) has the largest variance, which locates a compromise between the two clusters of points (modes). That is, it locates much of the variance of both groups, without really characterizing either. The leading eigenvector $\boldsymbol{e}_1$ points in the positive direction for both x_1 and x_2, but is more strongly aligned toward x_2, so the corresponding $\boldsymbol{e}_1$ map below shows a large positive "+" for x_2, and a smaller "+" for x_1. The second eigenvector is constrained to be orthogonal to the first, and so corresponds to large negative x_1, and mildly positive x_2, as indicated in the corresponding "map."

Figure 13.15b represents orthogonally rotated eigenvectors. Within the constraint of orthogonality they approximately locate the two point clusters, although the variance of the first rotated principal

component is no longer maximum since the projections onto $\widetilde{\boldsymbol{e}}_1$ of the three points with $x_1 < 0$ are quite small. However, the interpretation of the two features is enhanced in the maps of the two eigenvectors below, with $\widetilde{\boldsymbol{e}}_1$ indicating large positive x_1 together with modest but positive x_2, whereas $\widetilde{\boldsymbol{e}}_2$ shows large positive x_2 together with modestly negative x_1.

Finally, Figure 13.15c illustrates an oblique rotation, where the resulting rotated eigenvectors are no longer constrained to be orthogonal. Accordingly they have more flexibility in their orientations and can better accommodate features in the data that are not orthogonal.

13.5.3. Sensitivity of Orthogonal Rotation to Initial Eigenvector Scaling

An underappreciated aspect of orthogonal eigenvector rotation is that the orthogonality of the result depends strongly on the scaling of the original eigenvectors before rotation (Jolliffe, 1995, 2002; Mestas-Nuñez, 2000). This dependence is usually surprising because of the name "orthogonal rotation," which derives from the orthogonality of the transformation matrix $[T]$ in Equation 13.27, that is, $[T]^{\mathrm{T}}\,[T] = [T]\,[T]^{\mathrm{T}} = [I]$. The confusion is compounded because of the incorrect assertion in a number of papers that an orthogonal rotation produces both orthogonal rotated eigenvectors and uncorrelated rotated principal components. At most one of these two results can be obtained by an orthogonal rotation, but neither will occur unless the eigenvectors are scaled correctly before the rotation matrix is applied. Because of the confusion about the issue, an explicit analysis of this counterintuitive phenomenon is worthwhile.

Denote as $[E]$ the possibly truncated $(K \times M)$ matrix of eigenvectors of $[S]$. Because these eigenvectors are orthogonal (Equation 11.50) and are originally scaled to unit length, the matrix $[E]$ is orthogonal, and so satisfies Equation 11.44b. The resulting principal components can be arranged in the matrix

$$\underset{(n\,x\,M)}{[U]} = \underset{(n\,x\,K)}{[X]}\ \underset{(K\,x\,M)}{[E]}\,, \tag{13.29}$$

each of the n rows of which contain values for the M retained principal components, $\boldsymbol{u}_m{}^T$. As before, $[X]$ is the original data matrix whose K columns correspond to the n observations on each of the original K variables. The uncorrelatedness of the unrotated principal components can be diagnosed by calculating their covariance matrix,

$$\begin{aligned}(n-1)^{-1}\underset{(M\,x\,M)}{[U]^T[U]} &= (n-1)^{-1}([X][E])^T[X][E] \\ &= (n-1)^{-1}[E]^T[X]^T[X][E] \\ &= [E]^T\Big([E][\Lambda][E]^T\Big)[E] = [I][\Lambda][I] \\ &= [\Lambda].\end{aligned} \tag{13.30}$$

The u_m are uncorrelated because their covariance matrix $[\Lambda]$ is diagonal, and the variance for each u_m is λ_m. The steps on the third line of Equation 13.30 follow from the diagonalization of $[S] = (n\text{–}1)^{-1}[X]^{\mathrm{T}}[X]$ (Equation 11.52a) and the orthogonality of the matrix $[E]$.

Consider now the effects of the three eigenvector scalings listed in Table 13.3 on the results of an orthogonal rotation. In the first case, the original eigenvectors are not rescaled from unit length, so the matrix of rotated eigenvectors is simply

$$\underset{(K\,x\,M)}{\left[\widetilde{E}\right]} = \underset{(K\,x\,M)}{[E]}\,\underset{(M\,x\,M)}{[T]}\,. \tag{13.31}$$

That these rotated eigenvectors are still orthogonal, as expected, can be shown by

$$\begin{aligned}\left[\widetilde{E}\right]^T\left[\widetilde{E}\right] &= ([E][T])^T[E][T] = [T]^T[E]^T[E][T] \\ &= [T]^T[I][T] = [T]^T[T] = [I].\end{aligned} \tag{13.32}$$

That is, the resulting rotated eigenvectors are still mutually perpendicular and of unit length. The corresponding rotated principal components are

$$\left[\widetilde{U}\right] = [X]\left[\widetilde{E}\right] = [X][E][T], \tag{13.33}$$

and their covariance matrix is

$$\begin{aligned}(n-1)^{-1}\underset{(M\,x\,M)}{\left[\widetilde{U}\right]^T}\left[\widetilde{U}\right] &= (n-1)^{-1}([X][E][T])^T[X][E][T] \\ &= (n-1)^{-1}[T]^T[E]^T[X]^T[X][E][T] \\ &= [T]^T[E]^T\Big([E][\Lambda][E]^T\Big)[E][T] \\ &= [T]^T[\Lambda][T].\end{aligned} \tag{13.34}$$

This matrix is not diagonal, reflecting the fact that the rotated principal components are no longer uncorrelated. This result is easy to appreciate geometrically, by looking at scatterplots such as Figure 13.1 or Figure 13.3. In each of these cases the point cloud is inclined relative to the original (x_1, x_2) axes, and the angle of inclination of the long axis of the cloud is located by the first eigenvector. The point cloud is not inclined in the $(\boldsymbol{e}_1, \boldsymbol{e}_2)$ coordinate system defined by the two eigenvectors, reflecting the uncorrelatedness of the unrotated principal components (Equation 13.30). But relative to any other pair of mutually orthogonal axes in the plane, the points would exhibit some inclination, and therefore the projections of the data onto these axes would exhibit some nonzero correlation.

The second eigenvector scaling in Table 13.3, $||\mathbf{e}_m|| = (\lambda_m)^{1/2}$, is commonly employed, and indeed is the default scaling in many statistical software packages for rotated principal components. In the notation of this section, employing this scaling is equivalent to rotating the scaled eigenvector matrix $[E][\Lambda]^{1/2}$, yielding the matrix of rotated eigenvectors

$$\left[\widetilde{E}\right] = \Big([E][\Lambda]^{1/2}\Big)[T]. \tag{13.35}$$

The orthogonality of the rotated eigenvectors in this matrix can be checked with

$$\begin{aligned}\left[\widetilde{E}\right]^T\left[\widetilde{E}\right] &= \Big([E][\Lambda]^{1/2}[T]\Big)^T[E][\Lambda]^{1/2}[T] \\ &= [T]^T[\Lambda]^{1/2}[E]^T[E][\Lambda]^{1/2}[T] \\ &= [T]^T[\Lambda]^{1/2}[I][\Lambda]^{1/2}[T] = [T]^T[\Lambda][T].\end{aligned} \tag{13.36}$$

Here the equality on the second line is valid because the diagonal matrix $[\Lambda]^{1/2}$ is symmetric, so that $[\Lambda]^{1/2} = ([\Lambda]^{1/2})^{\mathrm{T}}$. The rotated eigenvectors corresponding to the second, and frequently used, scaling

in Table 13.3 are *not* orthogonal, because the result of Equation 13.36 is not a diagonal matrix. Neither are the corresponding rotated principal components independent. This can be seen by manipulating their covariance matrix, which is also not diagonal, that is,

$$\begin{aligned}(n-1)^{-1}\underset{(M\,x\,M)}{\left[\widetilde{U}\right]^T}\left[\widetilde{U}\right] &= (n-1)^{-1}\Big([X][E][\Lambda]^{1/2}[T]\Big)^T[X][E][\Lambda]^{1/2}[T] \\ &= (n-1)^{-1}[T]^T[\Lambda]^{1/2}[E]^T[X]^T[X][E][\Lambda]^{1/2}[T] \\ &= [T]^T[\Lambda]^{1/2}[E]^T\Big([E][\Lambda][E]^T\Big)[E][\Lambda]^{1/2}[T] \\ &= [T]^T[\Lambda]^{1/2}[I][\Lambda][I][\Lambda]^{1/2}[T] \\ &= [T]^T[\Lambda]^{1/2}[\Lambda][\Lambda]^{1/2}[T] \\ &= [T]^T[\Lambda]^2[T].\end{aligned} \tag{13.37}$$

The third eigenvector scaling in Table 13.3, $||\mathbf{e}_m|| = (\lambda_m)^{-1/2}$, is used relatively rarely, although it can be convenient in that it yields unit variance for all the principal components u_m. The resulting rotated eigenvectors are not orthogonal, so that the matrix product

$$\begin{aligned}\left[\widetilde{E}\right]^T\left[\widetilde{E}\right] &= \Big([E][\Lambda]^{-1/2}[T]\Big)^T[E][\Lambda]^{-1/2}[T] \\ &= [T]^T[\Lambda]^{-1/2}[E]^T[E][\Lambda]^{-1/2}[T] \\ &= [T]^T[\Lambda]^{-1/2}[I][\Lambda]^{-1/2}[T] = [T]^T[\Lambda]^{-1}[T],\end{aligned} \tag{13.38}$$

is not diagonal. However, the resulting rotated principal components are uncorrelated since their covariance matrix,

$$\begin{aligned}(n-1)^{-1}\underset{(M\,x\,M)}{\left[\widetilde{U}\right]^T}\left[\widetilde{U}\right] &= (n-1)^{-1}\Big([X][E][\Lambda]^{-1/2}[T]\Big)^T[X][E][\Lambda]^{-1/2}[T] \\ &= (n-1)^{-1}[T]^T[\Lambda]^{-1/2}[E]^T[X]^T[X][E][\Lambda]^{-1/2}[T] \\ &= [T]^T[\Lambda]^{-1/2}[E]^T\Big([E][\Lambda][E]^T\Big)[E][\Lambda]^{-1/2}[T] \\ &= [T]^T[I][I][T] = [T]^T[T] = [I],\end{aligned} \tag{13.39}$$

is diagonal, and also reflects unit variances for all the rotated principal components.

Most frequently in meteorology and climatology, the eigenvectors in a PCA describe spatial patterns, and the principal components are time series reflecting the importance of the corresponding spatial patterns in the original data. When calculating orthogonally rotated principal components in this context, we can choose to have either orthogonal rotated spatial patterns but correlated rotated principal component time series (by using $||\mathbf{e}_m|| = 1$), or nonorthogonal rotated spatial patterns whose time sequences are mutually uncorrelated (by using $||\mathbf{e}_m|| = (\lambda_m)^{-1/2}$), but not both. It is not clear what the advantage of having neither property (using $||\mathbf{e}_m|| = (\lambda_m)^{1/2}$, as is often done) might be. Differences in the results for the different scalings will be small if sets of effective multiplets are rotated separately, because their eigenvalues will necessarily be similar in magnitude, resulting in similar lengths for the scaled eigenvectors.

13.5.4. Simple Structure Through Regularization

A quite different approach to achieving simple structure, so that few of the eigenvector loadings in PCA are appreciably different from zero, is through Lasso regularization (Tibshirani, 1996). Use of the Lasso for regularization of least-squares regression, which imposes a budget on the sum of absolute values for the regression coefficients, was discussed in Section 7.5.2. The result is that, as the ceiling c on the sum of absolute values of the regression coefficients is decreased, more of them are progressively driven to zero.

A similar approach can be taken in PCA (Jolliffe et al., 2003), in which case the constraint can be expressed as

$$\sum_{k=1}^{K} |e_{k,m}| \le c, c > 1, \tag{13.40}$$

for each eigenvector $\boldsymbol{e}_m$. Jolliffe et al. (2003) named the approach Simplified Component Technique—LASSO, or SCoTLASS. For $c = 1$ the method yields exactly one nonzero loading in $\boldsymbol{e}_m$, and so produces regularized eigenvectors that are exactly aligned with the original coordinate axes. Unconstrained ordinary PCA is produced when $c \ge \sqrt{K}$. For $1 < c \le \sqrt{K}$ the resulting eigenvectors are orthogonal, and so still define a rigid rotation of the coordinate axes, but the data projections onto them (the regularized principal components) are not uncorrelated.

Of course the results will depend on the regularization parameter c, the choice of which is more ambiguous for PCA than in the regression setting because predictive cross-validation will in general not be available. Smaller values of the regularization parameter produce more loadings that are exactly zero, but simultaneously the fraction of variance represented decreases and the correlations among the regularized principal components increase. Jolliffe et al. (2003) suggest recomputation using multiple values of the regularization parameter, and then choosing c subjectively, perhaps on the basis of the problem-specific interpretability of the resulting regularized eigenvectors. Computation of regularized PCA is more difficult than its conventional counterpart and may involve multiple local minima in the numerical optimization. More details can be found in Hastie et al. (2015) and Jolliffe et al. (2003).

13.6. COMPUTATIONAL CONSIDERATIONS

13.6.1. Direct Extraction of Eigenvalues and Eigenvectors from [*S*]

The sample covariance matrix [S] is real and symmetric, and so will always have real-valued and nonnegative eigenvalues. Standard and stable algorithms are available to extract the eigenvalues and eigenvectors from real, symmetric matrices (e.g., Press et al., 1986), and this approach can be a very good one for computing a PCA. As noted earlier, it is sometimes preferable to calculate the PCA using the correlation matrix [R], which is also the covariance matrix for the standardized variables. The computational considerations presented in this section are equally appropriate to PCA based on the correlation matrix.

One practical difficulty that can arise is that the required computational time increases very quickly as the dimension of the covariance matrix increases. A typical application of PCA in meteorology or climatology involves a field observed at K grid- or other space-points, at a sequence of n times, where $K >> n$. The typical conceptualization is in terms of the $(K \times K)$ covariance matrix, which is very large—it is not unusual for K to include thousands of gridpoints. Hours of computation may be required to extract this many eigenvalue–eigenvector pairs. Yet since $K > n$–1 the sample covariance matrix is

singular, implying that the last $K–n+1$ of its eigenvalues are exactly zero. It is pointless to calculate numerical approximations to these zero eigenvalues and their associated arbitrary eigenvectors.

In this situation fortunately it is possible to focus the computational effort on the n nonzero eigenvalues and their associated eigenvectors, using a computational trick (Von Storch and Hannoschöck, 1984). Recall that the $(K \times K)$ covariance matrix $[S]$ can be computed from the centered data matrix $[X']$ using Equation 11.30. Reversing the roles of the time and space points (although it is still the columns of the anomaly matrix $[X']$ that have zero mean), we also can compute the $(n \times n)$ covariance matrix

$$\underset{(n\,x\,n)}{[S^*]} = \frac{1}{n-1} \underset{(n\,x\,K)}{[X']} \; \underset{(K\,x\,n)}{[X']^T}. \tag{13.41}$$

Both $[S]$ and $[S^*]$ have the same min$(n–1, K)$ nonzero eigenvalues, $\lambda_k = \lambda^*_k$, so the required computational time may be much shorter if they are extracted from the smaller $(n \times n)$ matrix $[S^*]$, and this latter computation will be much faster in the usual situation where $K >> n$.

The eigenvectors of $[S]$ and $[S^*]$ are different, but the leading $n–1$ (i.e., the meaningful) eigenvectors of $[S]$ can be computed from the eigenvectors $e_k{}^*$ of $[S^*]$ using

$$\boldsymbol{e}_k = \frac{[X']^T \boldsymbol{e}_k^*}{\left\| [X']^T \boldsymbol{e}_k^* \right\|}, k = 1, \ldots, n-1. \tag{13.42}$$

The dimensions of the multiplications in both numerator and denominator are $(K \times n)\,(n \times 1) = (K \times 1)$, and the role of the denominator is to ensure that the resulting $\boldsymbol{e}_k$ have unit length.

13.6.2. PCA via SVD

The eigenvalues and eigenvectors in a PCA can also be computed using the SVD (singular value decomposition) algorithm (Section 11.3.5), in two ways. First, as illustrated in Example 11.5, the eigenvalues and eigenvectors of a covariance matrix $[S]$ can be computed through SVD of the matrix $(n–1)^{-1/2}[X']$, where the centered $(n \times K)$ data matrix $[X']$ is related to the covariance matrix $[S]$ through Equation 11.30. In this case, the eigenvalues of $[S]$ are the squares of the singular values of $(n–1)^{-1/2}[X']$—that is, $\lambda_k = \omega_k{}^2$—and the eigenvectors of $[S]$ are the same as the right singular vectors of $(n–1)^{-1/2}[X']$—that is, $[E] = [R]$, or $\boldsymbol{e}_k = \boldsymbol{r}_k$.

An advantage of using SVD to compute a PCA in this way is that the left singular vectors (the columns of the $(n \times K)$ matrix $[L]$ in Equation 11.72) are proportional to the principal components (i.e., to the projections of the centered data vectors $\boldsymbol{x}'_i$ onto the eigenvectors $\boldsymbol{e}_k$). In particular,

$$u_{i,k} = \boldsymbol{e}_k^T \boldsymbol{x}'_i = \sqrt{n-1}\, \ell_{i,k} \sqrt{\lambda_k}, \quad i = 1, \ldots, n, \; k = 1, \ldots, K; \tag{13.43a}$$

or

$$\underset{(n\,x\,K)}{[U]} = \sqrt{n-1} \; \underset{(n\,x\,K)}{[L]} \; \underset{(K\,x\,K)}{[\Lambda]}^{1/2}. \tag{13.43b}$$

Here the matrix $[U]$ is used in the same sense as in Section 13.5.3, that is, each of its K columns contains the principal component series u_k corresponding to the sequence of n data values x_i, $i = 1, \ldots, n$.

The SVD algorithm can also be used to compute a PCA by operating on the covariance matrix directly. Comparing the spectral decomposition of a square, symmetric matrix (Equation 11.52a) with its SVD (Equation 11.72), it is clear that these unique decompositions are the same. In particular, since a

covariance matrix [S] is square and symmetric, both the left and right matrices of its SVD are equal, and contain the eigenvectors, that is, [E] = [L] = [R]. In addition, the diagonal matrix of singular values is exactly the diagonal matrix of eigenvalues, [Λ] = [Ω].

Computation of PCA using the SVD algorithm is comparatively fast. However, be aware that particular software implementations of the SVD may not sort the eigenvalues in descending order, although each eigenvector will still be associated with the correct eigenvalue.

13.6.3. The Power Method

In some applications (e.g., Wilks, 2016c) only the leading eigenvalue and eigenvector are needed, so that computation of a full PCA is unnecessarily slow and wasteful of computing resources. In such cases it can be advantageous to find the leading eigenvalue–eigenvector pair using the *power method* (e.g., Golub and van Loan, 1996).

Beginning with an arbitrary initial guess for the leading eigenvector, $\boldsymbol{e}_1$, with $||\boldsymbol{e}_1|| = 1$, the power method algorithm proceeds by iterating

$$\boldsymbol{v} = [S]\,\boldsymbol{e} \tag{13.44a}$$

$$\lambda_1 = \|\boldsymbol{v}\| \tag{13.44b}$$

$$\boldsymbol{e}_1 = \boldsymbol{v}/\lambda_1 \tag{13.44c}$$

until convergence. Here $\boldsymbol{v}$ is an intermediate storage vector, and $||\boldsymbol{v}||$ denotes its Euclidean length.

13.6.4. PCA and Missing Data

Computation of a sample covariance or correlation matrix using Equation 11.30 as input to a PCA, or alternatively computation of a PCA through SVD on the data matrix as in Example 11.5, both require that the input data anomaly matrix [X'] contains no missing values. Of course there are often missing value in real data sets, requiring that some accommodation be made before computation of a PCA. On the other hand, once the PCA has been computed, it can be used to estimate the missing values.

If there are very few missing values in a data set it may be reasonable to simply delete any incomplete data vectors before proceeding, but usually this approach will lead to an excessive portion of the data being lost. Alternatively, there are two approaches to dealing with the missing data before computation of a PCA. The first is to substitute the appropriate sample mean for any missing data, yielding a zero anomaly for the corresponding missing value $x'_{i,k}$, which is equally applicable when computing a covariance or correlation matrix, and when computing the SVD of a data matrix. The second approach is to estimate the elements of the covariance or correlation matrix using the numerator of Equation 3.28, including only terms for which both elements are nonmissing.

The two methods will differ only with respect to the divisor in the covariance calculation in the numerator of Equation 3.28, because the first approach will yield zero contribution to the sum when either of the two anomalies are zero, even though the divisor is n–1. Accordingly imputing zero anomaly for any missing values leads to negative bias in the absolute values of the resulting variance and covariance estimates. On the other hand, using the second approach yields different sample sizes for the estimated covariances or correlations, inconsistencies among which may result in a singular matrix. In that case some of the trailing eigenvalues may be negative which, as a consequence of Equation 11.54, will lead to an upward bias in the leading estimated eigenvalues.

Once a PCA involving missing data values has been computed, its eigenvectors can be used to estimate those missing data. First, the principal components in the matrix $[U]$ (Equation 13.29) are estimated using

$$u_{i,m} = \frac{\sum_{\substack{k=1 \\ x_{i,k} \text{ valid}}}^{K} x'_{i,k} e_{k,m}}{\sum_{\substack{k=1 \\ x_{i,k} \text{ valid}}}^{K} e_{k,m}^2}, \tag{13.45}$$

where the summations include only terms for which data values $x_{i,k}$ are nonmissing. When a data vector $\boldsymbol{x}_i$ contains no missing elements, the denominator of Equation 13.45 is 1 (Equation 11.49), in which case Equation 13.45 is equivalent to Equation 13.1. The missing anomaly values can then be estimated through application of a synthesis,

$$x'_{i,k} = \sum_{m=1}^{M} u_{i,m} e_{m,k}. \tag{13.46}$$

Equation 13.45 is the result of a least-squares approach that minimizes the error $([X]-[U][E^{\mathrm{T}}])^{\mathrm{T}}([X]-[U][E^{\mathrm{T}}])$, which is derived from Equation 13.29. It yields reasonable results when the proportion of missing data is not excessive. For larger (than perhaps 10% to 20%) fractions of missing data, more elaborate iterative approaches have been found to be more accurate (Taylor et al., 2013).

13.7. SOME ADDITIONAL USES OF PCA

13.7.1. Singular Spectrum Analysis (SSA): Time-Series PCA

Principal component analysis can also be applied to scalar or multivariate time series. This approach to time-series analysis is known both as *singular spectrum analysis* and *singular systems analysis* (SSA, in either case). Fuller developments of SSA than is presented here can be found in Elsner and Tsonis (1996), Ghil et al. (2002), and Vautard et al. (1992).

SSA is easiest to understand in terms of a scalar time series x_t, $t = 1, \ldots, n$; although the generalization to multivariate time series of a vector $\boldsymbol{x}_t$ is reasonably straightforward. As a variant of PCA, SSA involves extraction of eigenvalues and eigenvectors from a covariance matrix. This covariance matrix is calculated from a scalar time series by passing a *delay window*, or imposing an *embedding dimension*, of length K on the time series. The process is illustrated in Figure 13.16. For $K = 3$, the first K-dimensional data vector, $\boldsymbol{x}_{(1)}$ is composed of the first three members of the scalar time series, $\boldsymbol{x}_{(2)}$ is composed of the second three members of the scalar time series, and so on, yielding a total of $n - K + 1$ overlapping lagged data vectors.

If the time series x_t is covariance stationary, that is, if its mean, variance, and lagged correlations do not change through time, the $(K \times K)$ population covariance matrix of the lagged time-series vectors $\boldsymbol{x}_{()}$ takes on a special banded structure known as *Toeplitz*, in which the elements $\sigma_{i,\,j} = \gamma_{|\,i-j|} = \mathrm{E}[x'_t \; x'_{t+|i-j|}]$ are arranged in diagonal parallel bands. That is, the elements of the resulting covariance matrix are taken from the autocovariance function (Equation 3.39), with lags arranged in increasing order away from the main diagonal. All the elements of the main diagonal are $\sigma_{i,i} = \gamma_0$, that is, the variance. The elements on

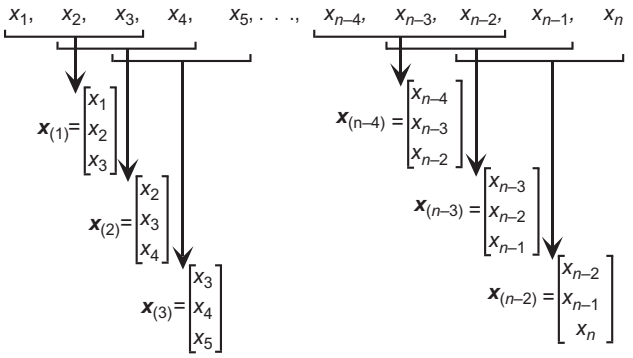

FIGURE 13.16 Illustration of the construction of the vector time series $\boldsymbol{x}_{(t)}$, $t = 1, \ldots, n–M+1$, by passing a delay window of embedding dimension $M = 3$ over consecutive members of the scalar time series x_t.

the diagonal bands adjacent to the main diagonal are all equal to γ_1, reflecting the fact that, for example, the covariance between the first and second elements of the vectors $\boldsymbol{x}_{(t)}$ in Figure 13.16 is the same as the covariance between the second and third elements. The elements separated from the main diagonal by one position are all equal to γ_2, and so on. Because of edge effects at the beginnings and ends of sample time series, the sample covariance matrix may be only approximately Toeplitz, although the diagonally banded Toeplitz structure is sometimes enforced before calculation of the SSA (Allen and Smith, 1996; Elsner and Tsonis, 1996; Groth and Ghil, 2015).

Since SSA is a PCA, the same mathematical considerations apply. In particular, the principal components are linear combinations of the data according to the eigenvectors (Equations 13.1 and 13.2). The analysis operation can be reversed to synthesize, or approximate, the data from all (Equation 13.20) or some (Equation 13.21) of the principal components. What makes SSA different follows from the different nature of the underlying data, and the implications of that different nature on interpretation of the eigenvectors and principal components. In particular, the data vectors are fragments of time series rather than the more usual spatial distribution of values at a single time, so that the eigenvectors in SSA represent characteristic time patterns exhibited by the data, rather than characteristic spatial patterns. Accordingly, the eigenvectors in SSA are sometimes called T-EOFs. Since the overlapping time series fragments $\boldsymbol{x}_t$ themselves occur in a time sequence, the principal components also have a time ordering, as in Equation 13.11. These temporal principal components u_k, or T-PCs, index the degree to which the corresponding time-series fragment $\boldsymbol{x}_t$ resembles the corresponding T-EOF, $\boldsymbol{e}_k$. Because the data are consecutive fragments of the original time series, the principal components are linear combinations of these time-series segments, with the weights given by the T-EOF elements. The T-PCs are mutually uncorrelated, but in general an individual T-PC will exhibit temporal autocorrelations.

The analogy between SSA and Fourier analysis of time series is especially strong, with the T-EOFs corresponding to the sine and cosine functions, and the T-PCs corresponding to the amplitudes. However, there are two major differences. First, the orthogonal basis functions in a Fourier decomposition are the fixed harmonic functions, whereas the basis functions in SSA are the data-adaptive T-EOFs. Therefore an SSA may be more efficient than a Fourier analysis, in the sense of requiring fewer basis functions to represent a given fraction of the variance of a time series. Similarly, the Fourier amplitudes are time-independent constants, but their counterparts, the T-PCs, are themselves functions of

time. Therefore similarly to wavelet analysis (Section 10.6), SSA can represent time variations that may be localized in time, and so not necessarily recurring throughout the time series.

Also in common with Fourier analysis, SSA can detect and represent oscillatory or quasi-oscillatory features in the underlying time series. A periodic or quasi-periodic feature in a time series is represented in SSA by pairs of T-PCs and their corresponding eigenvectors. These pairs have eigenvalues that are equal or nearly equal. The characteristic time patterns represented by these pairs of eigenvectors have the same (or very similar) shape, but are offset in time by a quarter cycle (as are a pair of sine and cosine functions). Unlike the sine and cosine functions these pairs of T-EOFs take on shapes that are determined by the time patterns in the underlying data. A common motivation for using SSA is to search, on an exploratory basis, for possible periodicities in time series, which periodicities may be intermittent and/or nonsinusoidal in form. Features of this kind are indeed identified by a SSA, but false periodicities arising only from sampling variations may also easily occur in the analysis (Allen and Robertson, 1996; Allen and Smith, 1996).

An important consideration in SSA is choice of the window length or embedding dimension, K. Obviously the analysis cannot represent variations longer than this length, although choosing too large a value results in a small sample size, $n–K+1$, from which to estimate the covariance matrix. Also, the computational effort increases quickly as K increases. Usual rules of thumb are that an adequate sample size may be achieved for $K < n/3$, and that the analysis will be successful in representing time variations with periods between $K/5$ and K.

Example 13.5 SSA for an AR(2) Series

Figure 13.17 shows an $n = 100$-point realization from the AR(2) process (Equation 10.27) with parameters $\phi_1 = 0.9, \phi_2 = –0.6, \mu = 0$, and $\sigma_\varepsilon = 1$. This is a purely random series, but the parameters ϕ_1 and ϕ_2 have been chosen in a way that allows the process to exhibit pseudoperiodicities. That is, there is a tendency for the series to oscillate, although the oscillations are irregular with respect to their frequency and phase. The spectral density function for this AR(2) process, included in Figure 10.21, shows a maximum centered near $f = 0.15$, corresponding to a typical period near $\tau = 1/f \approx 6.7$ time steps.

Analyzing the series using SSA requires choosing a delay window length, K, that should be long enough to capture the feature of interest yet short enough for reasonably stable covariance estimates to be calculated. Combining the rules of thumb for the window length, $K/5 < \tau < K < n/3$, a plausible choice is $K = 10$. This choice yields $n–K+1 = 91$ overlapping time series fragments $\boldsymbol{x}_{(t)}$ of length $K = 10$.

Calculating the covariances for this sample of 91 data vectors $\boldsymbol{x}_{(t)}$ in the conventional way yields the (10×10) matrix

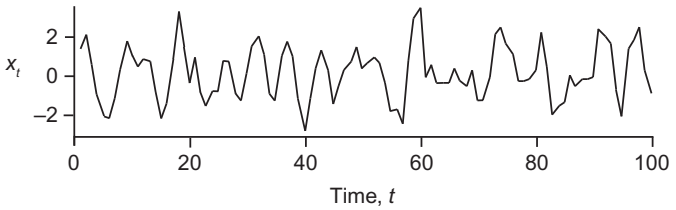

FIGURE 13.17 An $n = 100$-point realization from an AR(2) process with $\phi_1 = 0.9$ and $\phi_2 = –0.6$.

$$[S] = \begin{bmatrix} 1.792 & & & & & & & & & \\ .955 & 1.813 & & & & & & & & \\ -.184 & .958 & 1.795 & & & & & & & \\ -.819 & -.207 & .935 & 1.800 & & & & & & \\ -.716 & -.851 & -.222 & .959 & 1.843 & & & & & \\ -.149 & -.657 & -.780 & -.222 & .903 & 1.805 & & & & \\ .079 & -.079 & -.575 & -.783 & -.291 & .867 & 1.773 & & & \\ .008 & .146 & -.011 & -.588 & -.854 & -.293 & .873 & 1.809 & & \\ -.199 & .010 & .146 & -.013 & -.590 & -.850 & -.289 & .877 & 1.809 & \\ -.149 & -.245 & -.044 & .148 & .033 & -.566 & -.828 & -.292 & .874 & 1.794 \end{bmatrix}. \quad (13.47)$$

For clarity, only the elements in the lower triangle of this symmetric matrix have been printed. Because of edge effects in the finite sample, this covariance matrix is approximately, but not exactly, Toeplitz. The 10 elements on the main diagonal are only approximately equal, and each is estimating the true lag-0 autocovariance $\gamma_0 = \sigma^2{}_x \approx 2.29$. Similarly, the nine elements on the second diagonal are approximately equal, with each estimating the lag-1 autocovariance $\gamma_1 \approx 1.29$, the eight elements on the third diagonal estimate the lag-2 autocovariance $\gamma_2 \approx -0.21$, and so on. The pseudoperiodicity in the data is reflected in the large negative autocovariance at three lags, and the subsequent damped oscillation in the autocovariance function, which can be seen easily by reading down the first column, or reading the bottom row from right to left.

Figure 13.18 shows the leading four eigenvectors of the covariance matrix in Equation 13.47 and their associated eigenvalues. The first two of these eigenvectors (Figure 13.18a), which are associated with nearly equal eigenvalues, are very similar in shape and are separated by approximately a quarter of the period τ corresponding to the middle of the spectral peak in Figure 10.21. Jointly they represent the dominant feature of the data series in Figure 13.17, namely the pseudoperiodic behavior, with successive peaks and crests tending to be separated by six or seven time units.

The third and fourth T-EOFs in Figure 13.18b represent other, nonperiodic aspects of the time series in Figure 12.17. Unlike the leading T-EOFs in Figure 13.18a, they are not offset images of each other and

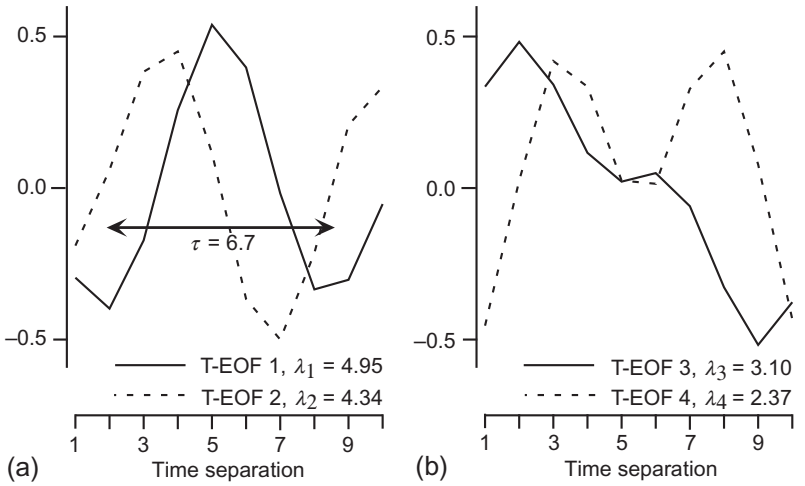

FIGURE 13.18 (a) First two eigenvectors of the covariance matrix in Equation 12.38, and (b) the third and fourth eigenvectors.

do not have nearly equal eigenvalues. Jointly the four patterns in Figure 13.18 represent 83.5% of the variance within the 10-element time series fragments (but not including variance associated with longer timescales).

Ghil et al. (2002) present a similar extended example of SSA, using a time series of the Southern Oscillation Index (Figure 3.16). ◇

It is conceptually straightforward to extend SSA to simultaneous analysis of multiple (i.e., vector) time series, which is called *multichannel SSA*, or MSSA (Ghil et al., 2002; Plaut and Vautard, 1994). The relationship between SSA and MSSA parallels that between an ordinary PCA for a single field and simultaneous PCA for multiple fields as described in Section 13.2.2. The multiple channels in a MSSA might be the L gridpoints representing a spatial field at time t, in which case the time series fragments corresponding to the delay window length K would be coded into a $(LK \times 1)$ vector $\boldsymbol{x}_{(t)}$, yielding a $(LK \times LK)$ covariance matrix from which to extract space-time eigenvalues and eigenvectors (ST-EOFs). The dimension of such a matrix may become unmanageable, and one solution (Plaut and Vautard, 1994) can be to first calculate an ordinary PCA for the spatial fields, and then subject the first few principal components to the MSSA. In this case each channel corresponds to one of the spatial principal components calculated in the initial data compression step. Vautard et al. (1996, 1999) describe MSSA-based forecasts of fields constructed by forecasting the space-time principal components, and then reconstituting the forecast fields through a truncated synthesis.

13.7.2. Principal-Component Regression

A pathology that may occur in multiple linear regression (see Section 7.3.1) is that a set of predictor variables having strong mutual correlations can result in the calculation of an unstable regression relationship, in the sense that the sampling distributions of the estimated regression parameters may have very high variances. The problem can be appreciated in the context of Equation 11.40, for the covariance matrix of the joint sampling distribution of the estimated regression parameters. This equation depends on the inverse of the matrix $[X]^{\mathrm{T}}[X]$, which is proportional to the covariance matrix $[S_x]$ of the predictors. Very strong intercorrelations among the predictors leads to their covariance matrix (and thus also $[X]^{\mathrm{T}}[X]$) being nearly singular, or small in the sense that its determinant is near zero. The inverse, $([X]^{\mathrm{T}}[X])^{-1}$ is then large, and inflates the covariance matrix $[S_b]$ in Equation 11.40. The result is that any specific set of estimated regression parameters may be very far from their correct values as a consequence of sampling variations, leading the fitted regression equation to perform poorly on independent data. The prediction intervals (based upon Equation 11.42) are also inflated.

An approach to remedying this problem is to first transform the predictors to their principal components, the correlations among which are zero. The resulting *principal-component regression* is convenient to work with, because the uncorrelated predictors can be added to or taken out of a tentative regression equation at will without affecting the contributions and parameter estimates of the other principal-component predictors. If all the principal components are retained in a principal-component regression, then nothing is gained over the conventional least-squares fit to the full predictor set. However, Jolliffe (2002) shows that multicollinearities, if present, are associated with the principal components having the smallest eigenvalues. As a consequence, the effects of the multicollinearities, and in particular the inflated covariance matrix for the estimated parameters, can in principle be removed by truncating the trailing principal components associated with the very small eigenvalues.

There are problems that may be associated with principal-component regression. Unless the principal components that are retained as predictors are interpretable in the context of the problem being analyzed, the insight to be gained from the regression may be limited. It is possible to reexpress the principal-component regression in terms of the original predictors using the synthesis equation (Equation 13.6), but the result will in general involve all the original predictor variables even if only one or a few principal component predictors have been used. This reconstituted regression will be biased, although often the variance is much smaller than for the least-squares alternative, resulting in a smaller MSE overall.

13.7.3. The Biplot

It was noted in Section 3.6 that graphical EDA for high-dimensional data is especially difficult. Since principal component analysis excels at data compression using the minimum number of dimensions, it is natural to think about applying PCA to EDA. The *biplot*, originated by Gabriel (1971), is such a tool. The "bi-" in biplot refers to the simultaneous representation of the n rows (the observations) and the K columns (the variables) of a data matrix $[X]$.

The biplot is a two-dimensional graph, whose axes are the first two eigenvectors of $[S_x]$. The biplot represents the n observations as their projections onto the plane defined by these two eigenvectors, that is, as the scatterplot of the first two principal components. To the extent that $(\lambda_1+\lambda_2)/\Sigma_k\lambda_k \approx 1$, this scatterplot will be a close approximation to their higher-dimensional relationships, in a graphable two-dimensional space. Exploratory inspection of the data plotted in this way may reveal such aspects of the data as the points clustering into natural groups, or time sequences of points that are organized into coherent trajectories in the plane of the plot.

The other element of the biplot is the simultaneous representation of the K variables. Each of the coordinate axes of the K-dimensional data space defined by the variables can be thought of as a unit basis vector indicating the direction of the corresponding variable, that is, $\boldsymbol{b}_1{}^T = [1, 0, 0, \ldots, 0], \boldsymbol{b}_2{}^T = [0, 1, 0, \ldots, 0], \ldots, \boldsymbol{b}_K{}^T = [0, 0, 0, \ldots, 1]$. These basis vectors can also be projected onto the two leading eigenvectors defining the plane of the biplot, that is,

$$\boldsymbol{e}_1^T \boldsymbol{b}_k = \sum_{k=1}^{K} e_{1,k}\, b_k \tag{13.48a}$$

and

$$\boldsymbol{e}_2^T \boldsymbol{b}_k = \sum_{k=1}^{K} e_{2,k}\, b_k. \tag{13.48b}$$

Since each of the elements of each of the basis vectors $\boldsymbol{b}_k$ is zero except for the kth, these dot products are simply the kth elements of the two eigenvectors. Therefore each of the K basis vectors $\boldsymbol{b}_k$ is located on the biplot by coordinates given by the corresponding eigenvector elements. Because the data values and their original coordinate axes are both projected in the same way, the biplot amounts to a projection of the full K-dimensional scatterplot of the data, including the coordinate axes, onto the plane defined by the two leading eigenvectors.

Figure 13.19 shows a biplot for the $K = 6$-dimensional January 1987 data in Table A.1, after standardization to zero mean and unit variance, so that the PCA pertains to their correlation matrix, $[R]$. The PCA for these data is given in Table 13.1b. The numbers indicate the calendar date for each plotted data point. The projections of the six original basis vectors (plotted longer than the actual projections in

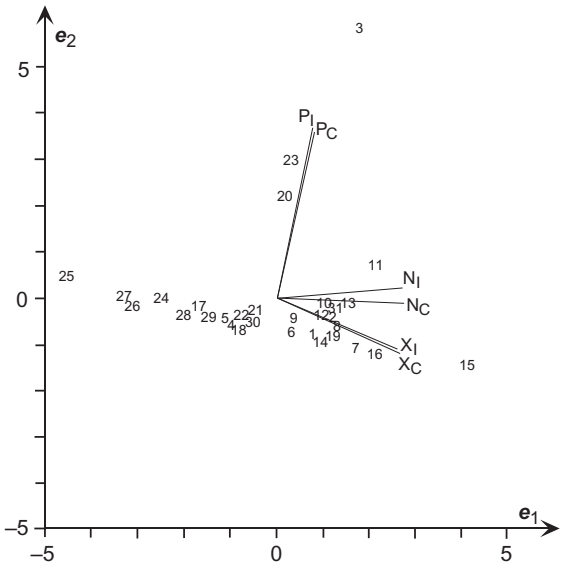

FIGURE 13.19 Biplot of the January 1987 data in Table A.1, after standardization. P = precipitation, X = maximum temperature, and N = minimum temperature. Numbered points refer to the corresponding calendar dates. The plot is a projection of the full six-dimensional scatterplot onto the plane defined by the first two principal components.

Equation 13.48 for clarity, but with the correct relative magnitudes) are indicated by the line segments diverging from the origin. "P," "N," and "X" indicate precipitation, minimum temperature, and maximum temperature, respectively, and the subscripts "I" and "C" indicate Ithaca and Canandaigua. It is immediately evident that the pairs of lines corresponding to like variables at the two locations are oriented nearly in the same directions, and that the temperature variables are oriented nearly perpendicularly to the precipitation variables. Approximately (because the variance described by the first two principal components is 92% rather than 100%), the correlations among these six variables are equal to the cosines of the angles between the corresponding lines in the biplot (compare Table 3.5), so the variables oriented in very similar directions form natural groupings.

The scatter of the n data points not only portrays their K-dimensional behavior in a potentially understandable way, but their interpretation is informed further by their relationship to the orientations of the variables. In Figure 13.19 most of the points are oriented nearly horizontally, with a slight inclination that is about midway between the angles of the minimum and maximum temperature variables, and perpendicular to the precipitation variables. These are the days corresponding to small or zero precipitation, whose primary variability characteristics relate to temperature differences. They are mainly located below the origin, because the mean precipitation is a bit above zero, and the precipitation variables are oriented nearly vertically (i.e., correspond closely to the second principal component). Points toward the right of the diagram, that are oriented similarly to the temperature variables, represent relatively warm days (with little or no precipitation), whereas points to the left are the cold days. Focusing on the dates for the coldest days, we can see that these occurred in a single run, toward the end of the month. Finally, the scatter of data points indicates that the few values in the upper portion of the biplot are different from the remaining observations, but it is the simultaneous display of the variables that allows us to see that these result from large positive values for precipitation.

13.8. EXERCISES

13.1. Using information from Exercise 11.8,
   a. Calculate the values of the first principal components for 1 January and for 2 January.
   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?

13.2. a. Compute the first two principal components for 1 January in the PCA in Table 13.1b.

   b. Reconstruct the six (standardized) weather variable values for 1 January, using the first 2 PCs only.

13.3. A principal component analysis of the data in Table A.3 yields the three eigenvectors $\boldsymbol{e}_1{}^T = [.593, .552, -.587]$, $\boldsymbol{e}_2{}^T = [.332, -.831, -.446]$, and $\boldsymbol{e}_3{}^T = [.734, -.069, .676]$, 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, Jolliffe's modification, and the broken stick model?
   c. Estimate the missing precipitation value for 1956 using the first PC only.

13.4. Use the information in Exercise 13.3 to
   a. Compute 95% confidence intervals for the eigenvalues, assuming large samples and multinormal data.

   b. Examine the eigenvalue separation using the North et al. rule of thumb.

13.5. Using the information in Exercise 13.3, calculate the eigenvector matrix $[E]$ to be orthogonally rotated if
   a. The resulting rotated eigenvectors are to be orthogonal.
   b. The resulting principal components are to be uncorrelated.

13.6. Use the SVD in Equation 11.74 to find the first three values of the first principal component of the minimum temperature data in Table A.1.

13.7. Table 13.1b is the summary of a PCA on (the standardized) daily weather data at Ithaca and Canandaigua for January 1987. Using these results, the principal-component regression equation predicting the corresponding daily maximum temperatures at Central Park, NY city is:

$$\text{Max}_{\text{NYC}} = 37.5 + 7.15\,u_1 - 2.81\,u_3 \quad \left(R^2 = 75.9\%\right)$$

   where u_1 and u_3 are the first and third principal components, respectively. Determine the corresponding regression equation in terms of the six original (standardized) predictor variables.

13.8. Construct a biplot for the data in Table A.3, using the information in Exercise 13.3.