Multivariate Analysis of Ecological Data

MICHAEL GREENACRE
Professor of Statistics at the Pompeu Fabra University in Barcelona, Spain

RAUL PRIMICERIO
Associate Professor of Ecology, Evolutionary Biology and Epidemiology
at the University of Tromsø, Norway

Chapter 12 Offprint

Principal Component Analysis

First published: December 2013
ISBN: 978-84-92937-50-9

Supporting websites:
www.fbbva.es
www.multivariatestatistics.org

Fundación BBVA

CHAPTER 12

Principal Component Analysis

In the biplots shown so far, there has been a two-step process in their construction. First, a scatterplot of the samples is made, using as support axes either two observed variables or two dimensions from an MDS. In the latter case the display has been optimized according to the objective function in the MDS, either to come as close as possible to reproducing the proximities (metric MDS) or to come as close to reproducing their ordering (nonmetric MDS). Given the MDS ordination, we have shown how different variables can be added using regression coefficients from various types of linear models or as weighted averages of sample points. The visualization of these variables is optimized conditional on the ordination – that is, the ordination is not necessarily the best ordination for explaining the variance of the added variables. In this chapter we present the first method that simultaneously optimizes both the ordination of the samples and the explained variance of the variables. The method, principal component analysis, applies to matrices of interval-scale continuous measurements.

Contents

The "climate" data set

Climate data are important in many ecological research projects, since they form part of the body of environmental indicators that can help to explain biological patterns. In a marine research project in Kotzbehue, Alaska, a set of annual climate variables were gathered together in a table, part of which is shown in Exhibit 12.1. These are annual data over 23 years, from 1981 to 2003. The variables

Exhibit 12.1:
Annual climate data for years 1981-2003, consisting of 17 climate indices and meteorological variables. Part of the 23 × 17 data matrix is shown

Year	AO	AO_winter	AO_summer	NPI	NPI_spring	NPI_winter	Temp	⋯	IceCover	IceFreeDays
1981	–0.4346	–0.1683	–0.2410	–2.09	–0.15	–4.46	–3.9	⋯	–0.64	140
1982	0.2977	–0.3750	0.3083	0.75	0.13	1.70	–4.7	⋯	–1.65	144
1983	0.0319	0.1733	0.4653	–2.54	0.30	–5.44	–4.4	⋯	–0.34	116
1984	–0.1917	0.2627	0.0240	–1.20	–0.23	–2.62	–7.0	⋯	0.15	134
1985	–0.5192	–1.2667	0.2678	0.52	–0.43	1.11	–5.9	⋯	–0.21	120
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
2002	0.0717	0.4543	0.0187	0.13	–0.18	0.30	–3.3	⋯	0.78	203
2003	0.1521	–0.6453	0.0399	–1.67	–0.40	–3.84	–3.8	⋯	–1.60	179
mean	0.0466	0.0587	0.1652	–0.440	0.023	–0.950	–5.15		–0.317	151.8
variance	0.1699	1.1687	1.0505	1.166	0.491	5.603	1.08		0.888	398.5

form three sets: climate indices such as the Arctic Oscillation (AO) and North Pacific Index (NPI),[1] meteorological variables such as temperature and rainfall, and various measures relating to Arctic ice such as ice coverage and number of ice-free days in the year.

MDS of the sample points

Each year is described by a set of 17 variables – each variable is assumed to be on an interval scale: this means that to compare two values of any particular variable it is the difference between two values that is important rather than the ratio or percentage difference. There are many different scales amongst the variables, some are pure indices without any scale, others are in units of degrees centigrade or centimetres of precipitation, and another is a count of days. To measure overall differences between the years based on this disparate set of variables, we need to equalize their scales in some way so that large values do not count more just because they are on a different scale – notice, for example, in the last two columns of Exhibit 12.1 the large differences across years in the *IceFreeDays* column (i.e., high variance) compared to the small differences in the *IceCover* column (i.e., low variance). As explained in Chapters 3 and 4, the most common way of equalizing the scales is to express each variable relative to its standard deviation so that all variables have equal variance. But, depending on the nature of the data, some other way may be preferred – see the discussion in Chapter 3 for alternatives. In some other situations, when all the variables are measured on the same scale, standardization might not be necessary, so that

[1] The Atlantic Oscillation Index, based on sea-level pressure differences, is positive when there is low pressure over the North Pole, keeping the cold air there, while it is negative when cold air is released southwards. The North Pacific Index measures interannual to decadal variations in the atmospheric circulation, which anticipate changes in sea surface temperatures.

the natural variances of the variables come into play and are not equalized in any way.

Once the variables are standardized, there are several possibilities for computing an overall measure of difference, i.e. distance or dissimilarity, between any two years. The most obvious choices are the sum (or average) of the absolute differences between the 18 variables (city-block distance) or the Euclidean distance. In principal component analysis the Euclidean distance is used, followed by classical MDS. We will use a very slight adaptation of the Euclidean distance, averaging the squared distances between the variables rather than summing them, so that the measure of distance is unaffected by the number of variables included. For the first two years in Exhibit 12.1, the distance between them is computed as follows:

$$d_{1981,\,1982} = \sqrt{\frac{1}{18}\left[\frac{(-0.4346-0.2977)^2}{0.1699} + \frac{(-0.1683-(-0.3750)^2}{1.1687} + \cdots + \frac{(140-144)^2}{398.5}\right]} = 1.277$$

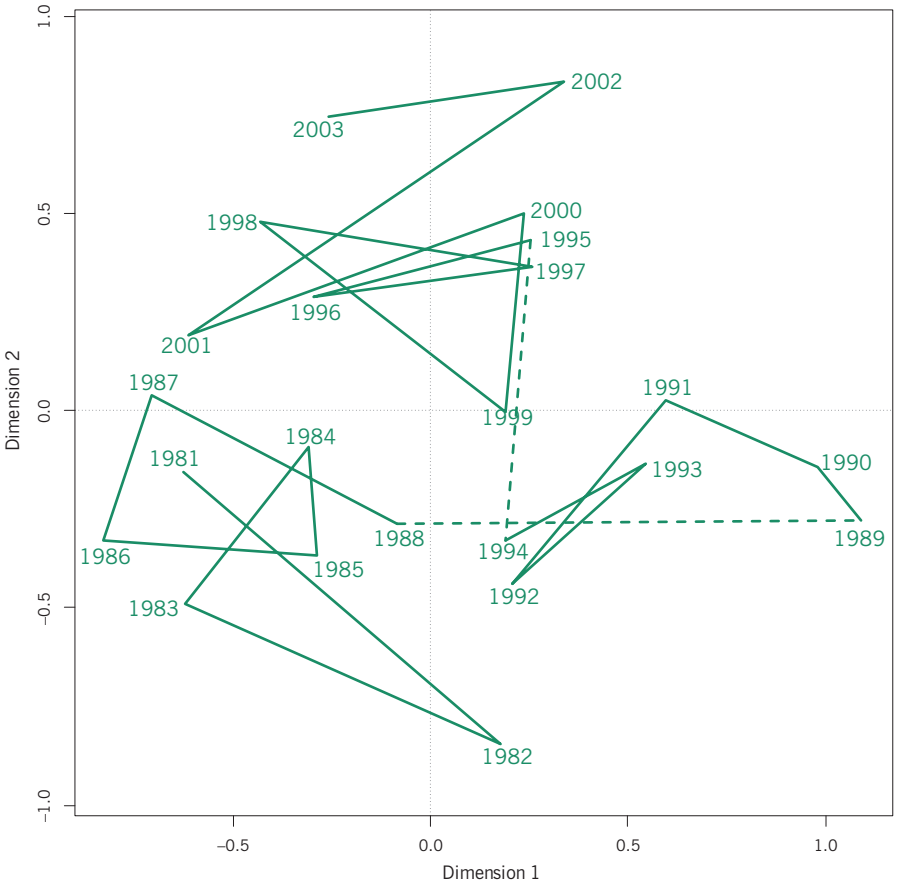
The denominators 0.1699, 1.1687, …, 398.5 are the variances of the variables *AO*, *AO_Winter*, …, *IceFreeDays*, so that the values inside the square of each numerator are divided by the corresponding variable's standard deviation. As described in Chapter 4, this distance function is called the *standardized Euclidean distance.*

Applying classical MDS to these distances, Exhibit 12.2 is obtained, explaining 45.6% of the variance. The years have been connected in sequence and there seem to be big changes in the climate variables from 1988 to 1989 and from 1994 to 1995, shown by dashed lines. Thus three groups of climate "regimes" are apparent, from 1982 to 1988 bottom left, then 1989 to 1994 on the right and finally from 1995 to 2003 in the upper section of the map. Next, adding the variables to the map will give the interpretation for these groupings.

Adding the variables to make a biplot

The 17 standardized variables are now regressed on the MDS dimensions, and their gradient vectors of regression coefficients are shown in Exhibit 12.3. The reason why the three groups of years separate is now apparent. The first period from 1982 to 1988 is characterized by high ice at all times of the year and to a lesser extent low winter temperatures (remember that the longer vectors here, corresponding to higher regression coefficients, will be the more important variables to interpret). In this period the climate indices, which are pointing to the right, will generally be below average. The period from 1989 to 1994, especially 1989, show a relatively sharp increase in all the climate indices. Then for the period 1995 to 2003, the climate indices move generally to average values

Exhibit 12.2:
MDS map of the 23 years according to the standardized Euclidean distances between them, across 17 climate variables. Variance explained by the two dimensions is 27.8% and 17.8%, totalling 45.6%



again but the annual temperature is generally higher, total ice is lower and the number of ice-free days higher.

Principal component analysis

The MDS solution for the years and the addition of the variables by regression once again looks like a two-step process, but what we have done is in fact the principal component analysis (PCA) solution of the climate data set, which is a one-step process. The difference between this analysis and all the other two-step analyses described before in this book is that here both the display of the cases and the display of the variables are simultaneously optimized. If one computes the overall variance explained of the 17 variables by the two MDS dimensions in this case, one gets exactly the same percentage of variance, 45.6%: 27.8% by the first dimension and 17.8% by the second. To summarize, PCA of a cases-by-variables data matrix can be thought of as an MDS of the Euclidean distances between the cases plus the regressions of the variables on the dimensions of the MDS solution.

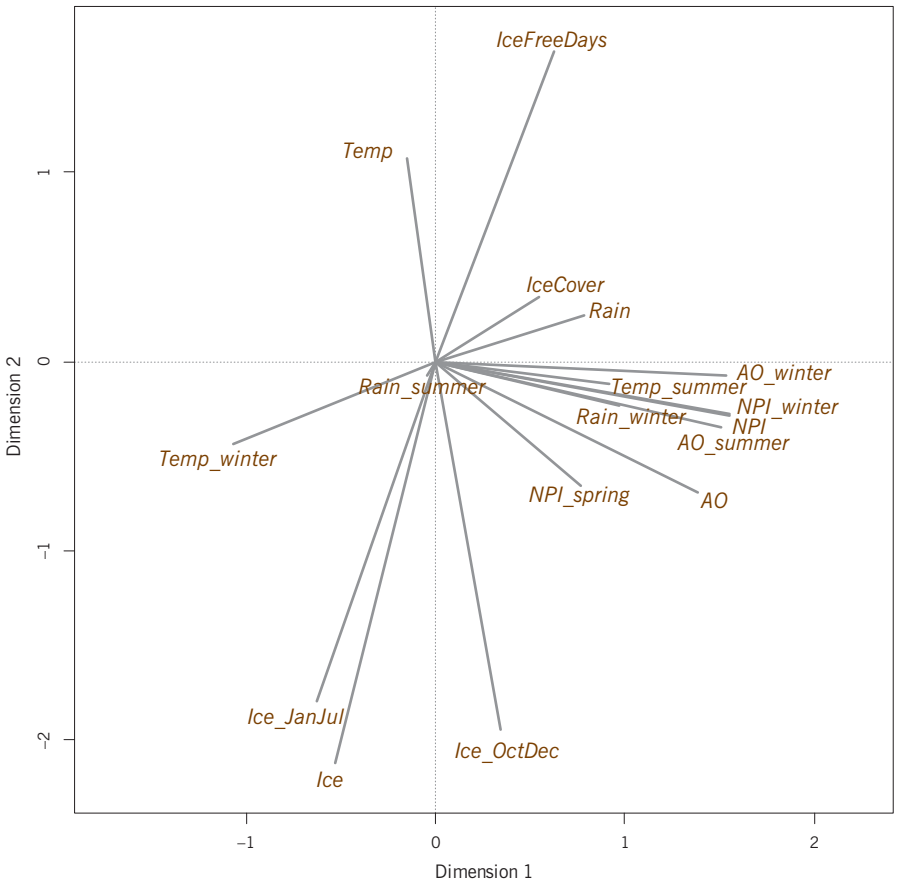


Exhibit 12.3:
Regression relationships of the variables with the two dimensions of the MDS map in Exhibit 12.2. Superimposing this configuration on Exhibit 12.2 would give a biplot of the years and the variables. This would be the so-called row-principal *biplot, explained on the following page*

To compute a PCA it is not necessary to do these two consecutive steps: they can be done in a single step using a famous theorem in mathematics called the *singular value decomposition,* or SVD. This result is similar to the eigenvalue-eigenvector theorem for square matrices but applies to rectangular matrices. Applying the SVD to a matrix results in a least-squares approximation of the matrix of a lower rank, where rank is the algebraic equivalent of dimensionality. In the application to the climate data, the SVD provides a rank 2 approximation to the 17-dimensional standardized data matrix, and it is this two-dimensional approximation that is represented in Exhibits 12.2 and 12.3. The approximation explains 45.6% of the variance in the original matrix, and this is the same if one thinks of the explanation of the row points (i.e., the years as displayed in Exhibit 12.2) or the variables (i.e., the climate variables as displayed in Exhibit 12.3). Thus, Exhibit 12.2 explains 45.6% of the (squared) Euclidean distances between the rows, and at the same time the two dimen-

sions of the solution are predictors of the 17 variables, also explaining 45.6% of their total variance.

Scaling of the solution

There is one subtle but important difference, not discussed before, between the regression biplots of Chapter 10 and the MDS and PCA biplots in Chapter 11 and in Exhibits 12.2 and 12.3. In the regression biplots of Chapter 10 the support space of the sample points was constructed using standardized variables, with variance one on both axes, whereas in Chapter 11 and this chapter so far, the samples had different variances on the axes. In Exhibit 12.2, for example, the variances along dimensions 1 and 2 were 6.12 and 3.90 respectively, explaining 27.8% and 17.8% of the total variance of the sampled years. This means that the year points are more spread out horizontally than vertically, although this may not appear obvious in Exhibit 12.2. In other examples there may be a much greater disparity between the horizontal and vertical spread of points in the distance map, in which case the discussion in this section is more of an issue.

To mimic the regression biplots of Chapter 10, we could rescale the coordinates of the sample points (i.e., year points here) in the MDS map to have unit variance on both dimensions, and then add the variables by regression. This will not affect the variance explained in the regression analyses but has some advantage in that the gradient vectors are then standardized regression coefficients and more easily compared.

Let us introduce some terminology that will be essential in future descriptions of different types of biplot. If a set of points, row or column points, has equal sum of squares on each dimension (usually equal to 1, but not necessarily), we call their coordinates on the dimensions *standard coordinates*. If they have sum of squares equal to (or proportional to) the variance explained by the dimensions, then their coordinates are called *principal coordinates*. In the case of PCA, the Exhibit 12.2 displays the year points (rows) in principal coordinates, and Exhibit 12.3 displays the variables (columns) in standard coordinates. The superimposition of these two displays is a true biplot, called the *row-principal biplot*. The other possibility, given in Exhibit 12.4, is the *column-principal biplot*, where the years are in standard coordinates, and the variables in principal coordinates. In this case it may seem hardly different to the combination in Exhibits 12.2 and 12.3, apart from the scale on the axes, so we should clarify the difference in interpretations of the two alternatives. In the row-principal biplot obtained by superimposing Exhibits 12.2 and 12.3, the year points are a spatial approximation of the inter-year Euclidean distances. In the column principal biplot of Exhibit 12.4 where the year coordinates have been standardized, this distance approximation property is not true any more. In Exhibit 12.4 the focus is on the climate variables and their spatial properties, in particular the angles

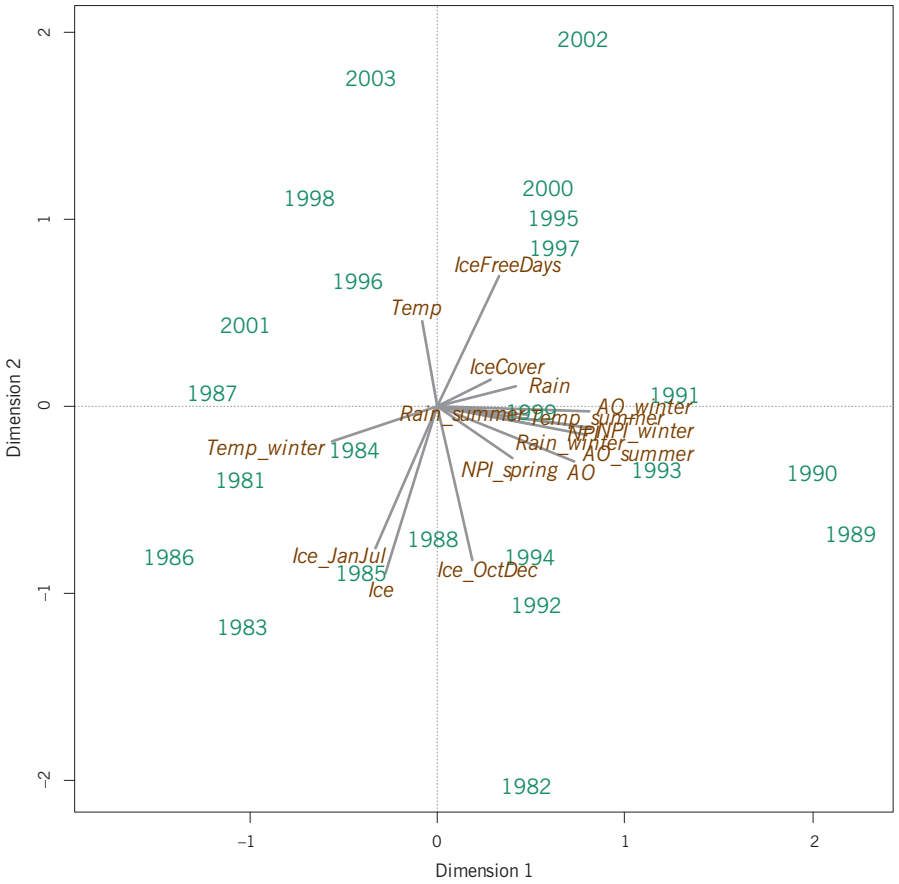


Exhibit 12.4:
Column-principal biplot of the climate data. Here the year points have coordinates that are standardized, while the sum of squares of the variable points on each dimension is proportional to the variance explained

between them, which have cosines that are approximately equal to the pairwise inter-correlations (see Chapter 6).

However, in Exhibit 12.4 the coordinates of the columns are standardized regression coefficients. In addition, because the two support dimensions are uncorrelated, the standardized regression coefficients are identical to the correlation coefficients of the column variables with the dimensions. For example, in Exhibit 12.4, the variable *IceFreeDays* can be seen to have a correlation with the first and second dimensions of approximately 0.3 and 0.7, respectively. In the same display sample points lie more or less within plus/minus two units, that is two standard deviations (because they are standardized), whereas the column points all have absolute values less than one (because their coordinates are correlations). One final remark: the sample points have means equal to zero, but the variable points are not centred. This is why, for

the biplot in Exhibit 12.4, the legend says that the sample coordinates are standardized (mean zero, variance one), whereas the sum of squares of the variable points on each dimension is proportional to the respective variance explained.

The circle of correlations

In the classical MDS of the Euclidean distances of this 17 variable problem, there are a maximum of 17 dimensions in the MDS solution, of which Exhibit 12.2 shows the best two corresponding to the two largest eigenvalues. (If there were 17 or fewer samples the maximum dimensionality of the solution would be one less than the number of samples.) Because the 17 "response" variables are standardized, a property of the standardized regression coefficients computed on standardized dimensions (where the coefficients are, we repeat, the correlations of the variables with the ordination axes) is that their sum of squares over all the dimensions for a particular variable is equal to 1, in other words the variable is fully explained by the complete set of MDS dimensions. In Exhibit 12.4, which shows the standardized regression coefficients with respect to the first two dimensions only, the sum of squares of the two coordinates for each variable, in other words the squared length of each vector shown, is equal to the proportion of variance explained for the respective variable. So we can draw a unit circle around the variable vectors and the variables that are better explained by the dimensions will be longer and closer to the unit circle. Exhibit 12.5 shows this circle. For example, the variable *Ice*, which lies close to the unit circle, has a very large part of its variance explained by the dimensions – the percentage is actually 88% – whereas *Temp* has only 21.2% (the length of the *Temp* vector is just under 0.5, and the length squared is the part of variance explained).

Having explained this relationship between the squared coordinates and the parts of variance explained for each variable, it follows that the average of the squared coordinates on each axis is equal to the part of variance (for all variables) explained by the axis: these averages are computed to be 0.278 and 0.178 respectively – see the caption of Exhibit 12.2.

In addition, as mentioned before, the cosines of the angles between the variables in Exhibit 12.5 are approximations of the correlations between the variables, and the approximation is improved when the variables are well explained, that is close to the unit circle. Thus we can be pretty sure that *IceFreeDays* and *Ice* are negatively correlated – the actual correlation is –0.57, the second most negative correlation amongst the variables. However, look at Exhibit 6.2 again – in order to see the correlation exactly as the angle cosine, we would need to see the two vectors *IceFreeDays* and *Ice* in their actual positions, not projected down onto this approximate MDS map.

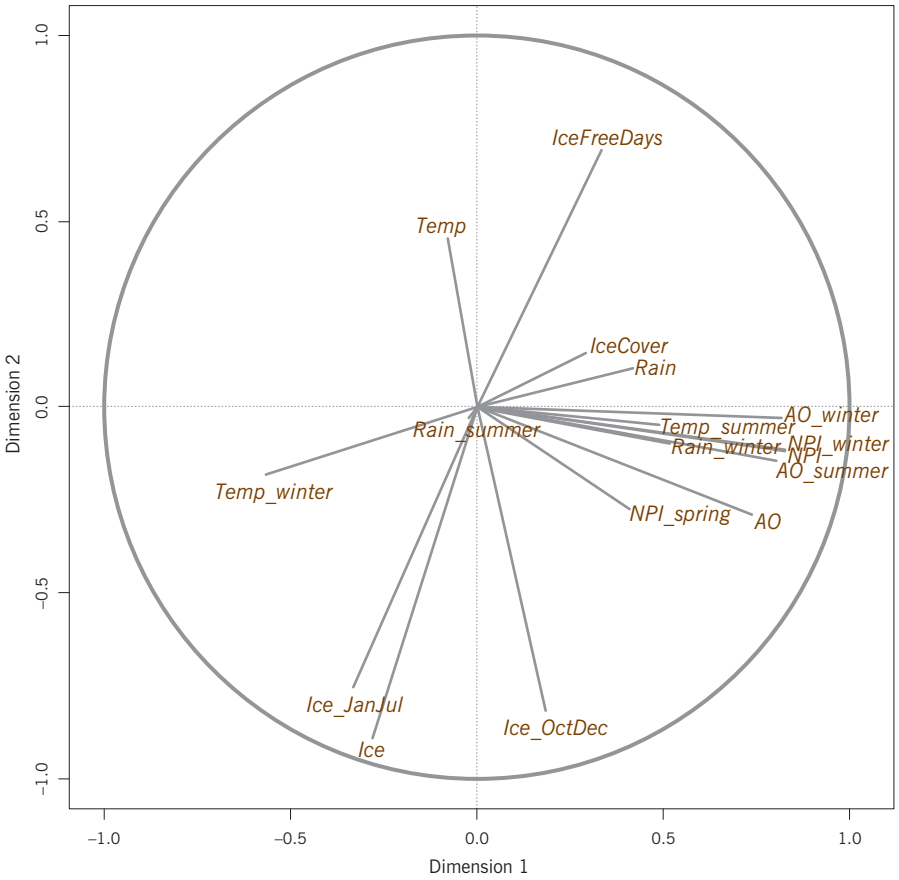


Exhibit 12.5:
Plot of the variables as in Exhibit 12.4, that is as standardized regression coefficients (i.e., principal coordinates in this PCA, which are the correlations between the variables and the dimensions), all lying within the unit circle. The closer the variable vector is to the unit circle, the better it is explained by the dimensions. The angle cosines between the vectors also approximate the correlations between the variables

Deciding on the dimensionality of the solution

Mostly for convenience of plotting, the biplots from PCA and MDS are shown with respect to the best two *principal axes*. But are these axes "significant" in the statistical sense? And what about other dimensions? If the third dimension of the solution were also "important", we could try to use three-dimensional graphics to visualize the biplots. But before going to these lengths we need a way of deciding how many dimensions are worth interpreting. One of the simplest ways of judging this, albeit quite informal, is to make a so-called *scree plot*, that is a bar graph of all the eigenvalues, or parts of variance, of the solution and then look for the so-called *elbow* in the plot. In Exhibit 12.6 on the left we show the scree plot of the eigenvalues of the present 23 × 17 example of climate data, whereas the scree plot on the right is of a 23 × 17 matrix of normally distributed random variables. Both PCAs have the same total variance of 17, equal to the number of variables, since each standardized variable has variance 1. Notice how for the random data the eigenvalues fall off gradually in value, whereas for

the climate data the first two stand out from the rest. From eigenvalue 3 onwards the values in Exhibit 12.6(b) are similar or greater than those in Exhibit 12.6(a), so it does seem that the first two dimensions of the climate data are nonrandom. Later in Chapter 17 we will make formal tests for dimensionality, based on the same method of comparing eigenvalues obtained in a PCA with those from PCAs of randomly generated data.

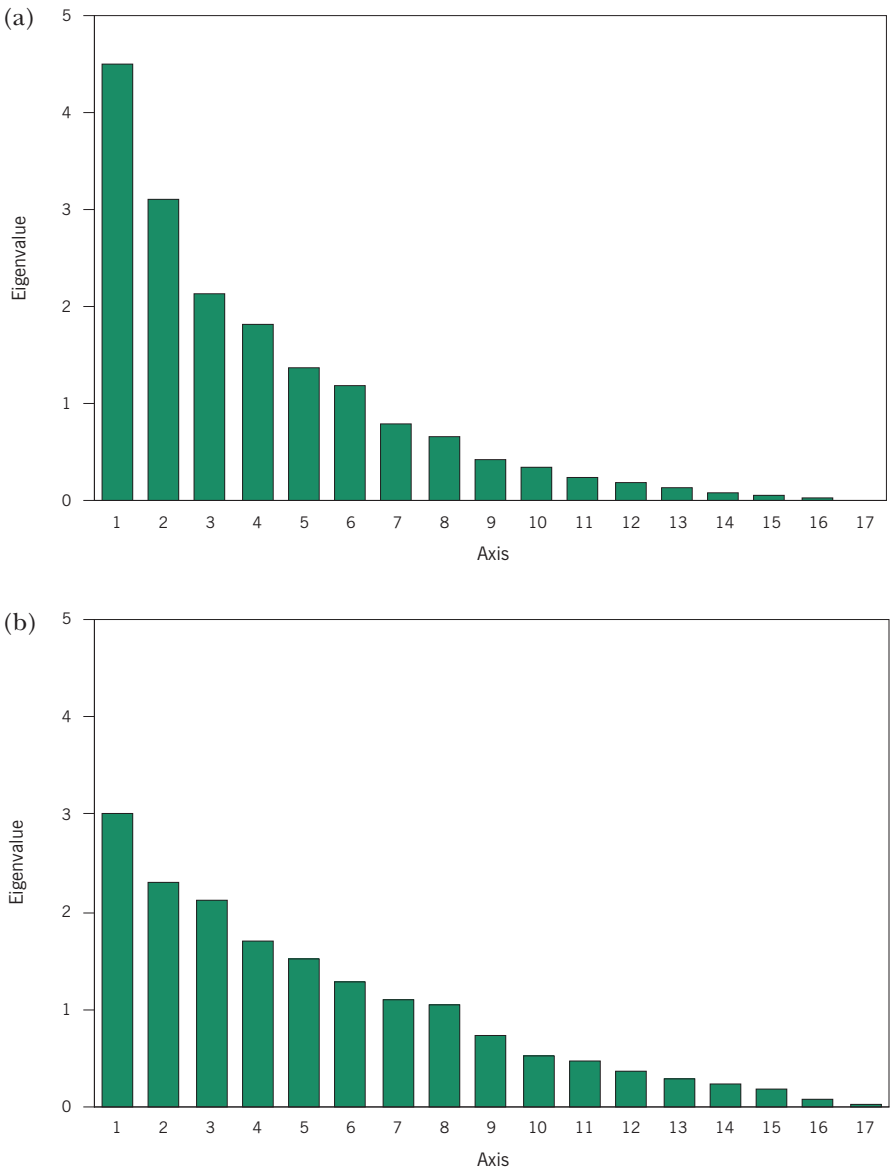


Exhibit 12.6: *Scree plots of the eigenvalues for (a) the climate data matrix; (b) a random data matrix*

SUMMARY: Principal component analysis

1. Principal component analysis (PCA) can be thought of as a multidimensional scaling (MDS) of a sample of multivariate observations, followed by the addition of the variables to the MDS map by linear regression on the dimensions, to give a biplot. Proximities between the multivariate sample points are defined using Euclidean or weighted Euclidean distance.

2. The special feature of PCA is that the visualization of both the sample points and the variables is optimized simultaneously: that is, the MDS optimally displays the sample points by least-squares and the dimensions are at the same time the best ones for predicting the variables by least-squares regression. In fact, the PCA solution is obtained in a single computational step, rather than a two-step MDS and regression approach.

3. There are two ways of displaying the results of PCA in the form of a biplot, differing only by scale factors of the sample (row) and variable (column) coordinates: the row-principal biplot and the column-principal biplot.

4. In the row-principal biplot the sample points are scaled in principal coordinates, as they would be from the MDS solution: they have mean 0 on each principal axis and their variances on each axis are the parts of variance explained. The variables added by regression will have equal variance on the axes (usually equal to 1) and their coordinates are thus called *standard coordinates*. Visually, the sample points will be spread out more on the first (horizontal) axis than on the second (vertical) axis, whereas the variables will be equally spread out on the two axes.

5. In the column-principal biplot the sample points are standardized on each principal axis, usually to have variance 1, in which case the regressions of the (standardized) variables on the principal axes are standardized regression coefficients, identical to the correlations between the variables and the axes. Now the variables, which are in principal coordinates, will be more spread out on the first axis than the second, whereas the sample points are equally spread out on the axes.

6. In the column-principal biplot the variables can be depicted as vectors inside a unit circle: the closer they lie to the circle the better they are explained by the principal dimensions. Also the angle cosines between the vectors are approximations of the correlations between them.

7. PCA is a dimension-reduction technique that attempts to separate "signal" (i.e., true structure) in the data, from "noise" (i.e., random variation), concentrating the signal in the first principal axes. Choosing how many axes are nonrandom can be performed informally by inspection of the scree plot of the eigenvalues in descending order, observing which eigenvalues stand out from the rest.

List of Exhibits