12 Principal component analysis and redundancy analysis

In the previous chapter, Bray–Curtis ordination was explained, and more recently developed multivariate techniques were mentioned. Principal component analysis (PCA), correspondence analysis (CA), discriminant analysis (DA) and non-metric multidimensional scaling (NMDS) can be used to analyse data without explanatory variables, whereas canonical correspondence analysis (CCA) and redundancy analysis (RDA) use both response and explanatory variables. In this chapter, we present PCA and RDA, and in the next chapter CA and CCA are discussed.

The chapter is mainly based on Ter Braak and Prentice (1988), Ter Braak (1994), Legendre and Legendre (1998) and Jolliffe (2002). More easy readings are chapters in Kent and Coker (1992), McCune and Grace (2002), Quinn and Keough (2002), Everitt (2005) and especially Manly (2004)

12.1 The underlying principle of PCA

Let Y_{ij} be the value of variable j ($j = 1, ..,N$) for observation i ($i = 1, ..,M$). Most ordination techniques create linear combinations of the variables:

$$Z_{i1} = c_{11}Y_{i1} + c_{12}\,Y_{i2} + \ldots + c_{1N}\,Y_{iN} \tag{12.1}$$

Assume you have a spreadsheet where each column is a variable. The linear combination can be imagined as multiplying all elements in a column with a particular value, followed by a summation over the columns. The idea of calculating a linear combination of variables is perhaps difficult to grasp at first. However, think of a diversity index like the total abundance. This is the sum of all variables (all c_{ij}s are one), and summarise a large number of variables with a single diversity index.

The linear combination, $\mathbf{Z}_1 = (Z_{11},\ldots,Z_{M1})'$, is a vector of length M, and is called a principal component, gradient or axis. The underlying idea is that the most important features in the N variables are caught by the new variable $\mathbf{Z}_1$. Obviously one component cannot represent all features of the N variables and a second component may be extracted:

$$Z_{i2} = c_{21}Y_{i1} + c_{22}\,Y_{i2} + \ldots + c_{2N}\,Y_{iN} \tag{12.2}$$

Further axes can be extracted (in fact, there are as many axes as variables). Most ordination techniques are designed in such a way that the first axis is more important than the second, the second more important than the third, etc., and the axes represent different information. It should be noted that the variables in equations (12.1) and (12.2) are not the original variables. They are either centred (which means that the mean of each variable is subtracted) or normalised (centred and then divided by the standard deviation), and the implication of this is discussed later in this chapter.

The multiplication factors c_{ij} are called loadings. The difference among PCA, DA and CA is the way these loadings are calculated. In RDA and CCA, a second set of variables is taken into account in calculating the loadings. These latter variables are considered as *explanatory* variables, so we are modelling a cause-effect relationship.

12.2 PCA: Two easy explanations

PCA is one of the oldest and most commonly used ordination methods. The reason for its popularity is perhaps its simplicity. Before introducing PCA, we need to clear up some possible misconceptions. PCA cannot cope with missing values (but neither can most other statistical methods), it does not require normality, it is not a hypothesis test, there are no clear distinctions between response variables and explanatory variables, and it is not written as *principle* component analysis but as *principal* component analysis.

There are various ways to introduce PCA. Our first approach is based on Shaw (2003), who used the analogy with shadows to explain PCA. Imagine giving a presentation. If you put your hand in front of the overhead projector, your three-dimensional hand will be projected on a two-dimensional wall or screen. The challenge is to rotate your hand such that the projection on the screen resembles the original hand as much as possible. This idea of projection and rotation brings us to a more statistical approach of introducing PCA. Figure 12.1-A shows a scatterplot of the species richness and NAP variables of the RIKZ data (Chapter 27). There is a clear negative relationship between the two variables. Panel B shows the same scatterplot, except that both variables are now mean deleted and divided by the standard deviation (also called normalisation). We also used axes with the same range (from –3.5 to 3.5). Now suppose that we want to have two new axes such that the first axis represents most information, and the second axis the second most information. In PCA, 'most information' is defined as the largest variance. The diagonal line from the upper left to the lower right in Figure 12.1-B is this first new axis. Projecting the points on this new axis results in the axis with the largest variance; any line with another angle with the x-axis will have a smaller variance. The additional restriction we put on the new axes is that the axes should be perpendicular to each other. Hence, the other line (perpendicular to the first new axis) in the same panel is the second axis. Panel C shows the same graph as panel B except that the new axes are presented in a more natural direction. This graph is

called a PCA ordination plot. In the same way as the three-dimensional hand was projected on a two-dimensional screen, we can project all the observations onto the first PCA axis, and omit the second axis (Figure 12.1-D). PCA output (which is discussed later) shows that the first new axis represents 76% of the information in the original scatterplot.

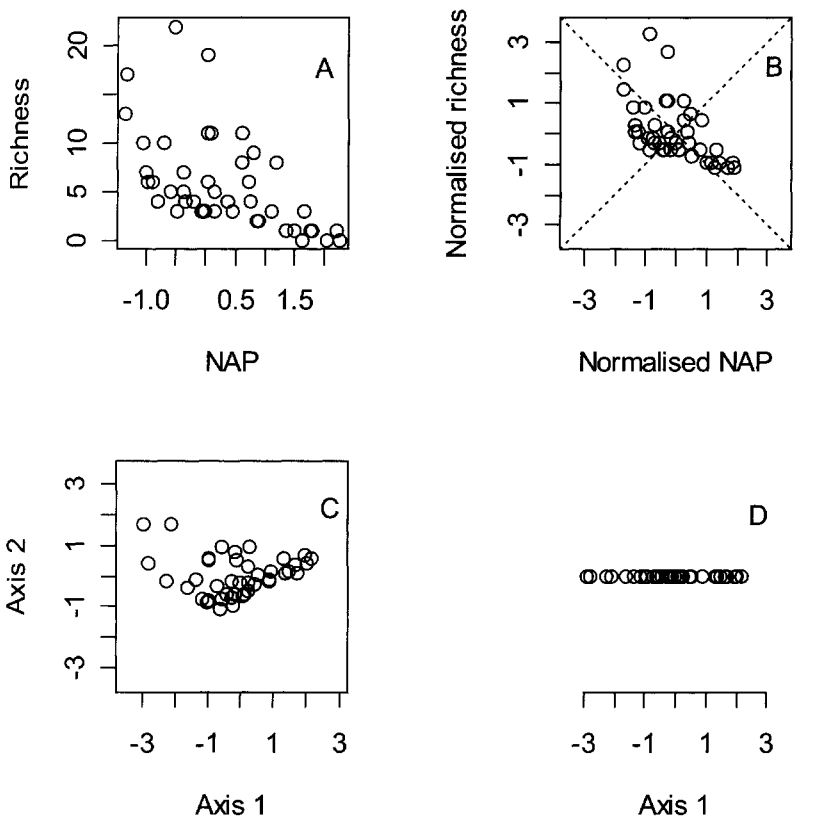

Figure 12.1. The underlying principle of PCA using the RIKZ data. A: Scatterplot of richness and NAP. B: Scatterplot of the same variables but now normalised. The two new axes have been added. C: The two PCA axes. D: Projection of all points on the first axis.

To go from Figure 12.1-A to Figure 12.1-D, we normalised and rotated the data. The two new axes are given by

$$Z_{i1} = -0.7 \times R_i + 0.7 \times NAP_i \quad \text{and} \quad Z_{i2} = 0.7 \times R_i + 0.7 \times NAP_i$$

In matrix notation, we have $\mathbf{Z} = \mathbf{XC}$, where $\mathbf{C}$ contains the multiplication factors, $\mathbf{X}$ the original variables and $\mathbf{Z}$ the principal components.

In applying PCA, one hopes that the variance of most components is negligible, and that the variation in the data can be described by a few (independent) principal components. Hence, instead of N original response variables, we end up with two or three principal components, which hopefully represent 70%-80% of the information in the data.

12.3 PCA: Two technical explanations

Having presented two relatively easy introductions to PCA, we feel it is necessary to present PCA in a mathematical context as well. Readers not familiar with matrix algebra may skip this section and continue with the illustrations of PCA in Section 12.4. We discuss two mathematical derivations of PCA.

PCA as an eigenvalue decomposition

More details on the technical aspects of PCA can be found in Jolliffe (2002), and only a short summary is presented below. The aim of PCA is to calculate an axis $\mathbf{Z}_1 = \mathbf{Yc}_1$ that has maximum variance. Because the mean of each variable is equal to zero, the variance of the first axis $\mathbf{Z}_1$ is given by $\mathbf{Z}_1'\mathbf{Z}_1 = \mathbf{c}_1'\mathbf{Y}'\mathbf{Yc}_1$. An aspect we have not mentioned so far is that the loadings are not unique. If a second axis $\mathbf{Z}_2 = \mathbf{Yc}_2$ is calculated, then both $\mathbf{c}_1$ and $\mathbf{c}_2$ can be multiplied with 10, resulting in axes that are also uncorrelated. Therefore, a restriction on the loadings is needed to make the solution unique: $\mathbf{c}_i'\mathbf{c}_i = 1$ for all axes $i = 1,..,N$. Although it seems rather arbitrary to set the sum of the squared loadings to 1, it does not have any important consequences.

The question is now how do we get the $\mathbf{c}_i$'s? Once we have the $\mathbf{c}_i$'s, we can easily calculate the $\mathbf{Z}_i$'s. It is relatively easy to show that the loadings can be obtained by solving the following constrained optimisation problem. For the first axis, we have:

$$\text{Maximise } \text{var}(\mathbf{Z}_1) = \mathbf{Z}_1'\mathbf{Z}_1 = \mathbf{c}_1'\mathbf{Y}'\mathbf{Yc}_1 \quad \text{subject to } \mathbf{c}_i'\mathbf{c}_i = 1.$$

The maximisation is with respect to the unknown parameters $\mathbf{c}_1$. Using matrix algebra, it is easy to show that the solution is given by $(\mathbf{S} - \mathbf{I}\lambda)\,\mathbf{c}_1 = 0$, see for example Jolliffe (2002). This may seem like magic to readers not used to matrix algebra, but this expression is fundamental in mathematics and is called the eigenvalue equation for $\mathbf{S}$, λ^* is the eigenvalue and $\mathbf{c}_1$ the eigenvector. $\mathbf{S}$ is either the covariance or correlation matrix, depending on whether the variables in $\mathbf{Y}$ are centred or normalised. This process can easily be extended to get the second or higher axes. In fact, the eigenvalue equation for all axes is given by $(\mathbf{S} - \mathbf{I}\Lambda)\mathbf{C} = \mathbf{0}$, where $\mathbf{C}$ contains the eigenvectors for all axes, $\mathbf{I}$ is the identity matrix and Λ the corre-

sponding eigenvalues. Statistical software can be used to obtain **C** (which can be used to obtain the axes).

The motivation to present the eigenvalue equation for PCA is that it justifies the iterative algorithm presented in the next paragraph, and this algorithm is used to explain RDA in the next chapter.

PCA as an iterative algorithm

The last approach to explain PCA is by using an iterative algorithm, which was presented in Jongman et al. (1995) and Legendre and Legendre (1998), and references in there. The algorithm has the following steps.

1. Normalise (or centre) the variables in **Y**.
2. Obtain initial scores **z** (e.g., by a random number generator).
3. Calculate new loadings: **c** = **Y**'**z**'.
4. Calculate new scores: **z** = **Yc**.
5. For second and higher axes: Make **z** uncorrelated with previous axes using a regression analysis.
6. Scale **z** to unit variance: $\mathbf{z}^* = \mathbf{z}/\lambda$, where λ is the standard deviation of the scores. Set **z** equal to **z***.
7. Repeat steps 2 to 6 until convergence. After convergence, divide λ by $M - 1$.

Once the algorithm is finished, the loadings **c** and principal components **z** (scores) are identical to those obtained from the PCA eigenvalue equation.

12.4 Example of PCA

To illustrate PCA, we use morphometric data from sparrows (Chris Elphick, University of Connecticut, USA), see also Chapter 14 for more details. This dataset consists of seven morphological variables taken from approximately 1000 sparrows. The morphological variables were wingcrd, flatwing, tarsus, head, culmen, nalopsi and weight (wt); see Section 14.3 for an explanation.

As the PCA will be based on the correlation matrix, it may be a good idea to inspect the correlation matrix before applying PCA. For the sparrow data, the correlation coefficients are relatively high (Table 12.1), and therefore we expect to find a PCA in which the first two axes (Z_1 and Z_2) explain a reasonable amount of information. The lengths of the wing are measured in two different ways, as the wing chord (wingcrd) and as the flattened wing (flatwing). Therefore the correlation between them is high, however the PCA can deal with this. Hence, for the moment we will use both variables.

Table 12.1. Correlations among the seven variables in the sparrow data.

	wingcrd	flatwing	tarsus	Head	culmen	nalopsi	Wt
wingcrd	1.00	0.99	0.55	0.53	0.42	0.38	0.66
flatwing		1.00	0.54	0.53	0.42	0.38	0.66
Tarsus			1.00	0.69	0.60	0.59	0.63
Head				1.00	0.72	0.73	0.62
culmen					1.00	0.70	0.49
nalopsi						1.00	0.50
Wt							1.00

The eigenvalue equation was solved and it gave the following values for the loadings of the first two axes (the variables were normalised):

$$Z_1 = -0.38\ \text{wingcrd} - 0.38\ \text{flatwing} - 0.39\ \text{tarsus} - 0.40\ \text{head} - 0.36\ \text{culmen} - 0.35\ \text{nalospi} - 0.38\ \text{wt}$$

$$Z_2 = 0.52\ \text{wingcrd} + 0.52\ \text{flatwing} - 0.12\ \text{tarsus} - 0.27\ \text{head} - 0.39\ \text{culmen} - 0.44\ \text{nalospi} + 0.16\ \text{wt}$$

The traditional way of presenting the graphical results of PCA is a scatterplot of $\mathbf{Z}_1$ and $\mathbf{Z}_2$; see Figure 12.2. However, the problem is the interpretation of this graph. The loadings for Z_1 and Z_2 indicate that the first axis is determined by all variables, and they all have approximately the same influence. This is typical for morphometric data; very often the first axis represents the overall shape of the animals. A more detailed discussion on morphometric data analysis is given in Chapter 30. The second axis seems to represent differences between wingcrd and flatwing versus culmen and nalospi.

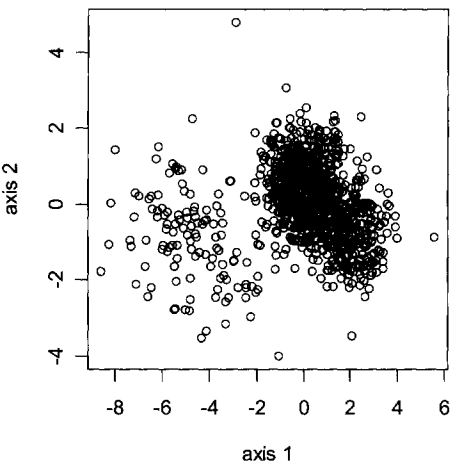

Figure 12.2. First two axes obtained by PCA for the sparrow data.

Eigenvalues in PCA represent the amount of variance explained by each axis. They can be expressed as numbers, percentage of the total variance, or as cumulative percentage of the total variance (Table 12.2). Some software packages present the eigenvalues that come out of the numerical routines, whereas others scale them so that the total sum is equal to one. The eigenvalues in the first column of Table 12.2 are scaled to have a sum of one. The first eigenvalue is 0.647, which means that it explains 64.7% of the variation in the data. The second eigenvalue is 0.163; hence the first two axes explain 81% of the variation.

Table 12.2. Eigenvalues and eigenvalues expressed as cumulative percentage. Some software packages rescale the eigenvalues so that the sum of all eigenvalues is equal to 1. These are given in the second column. Unscaled eigenvalues are in the third column. The sum of all eigenvalues is 7.

Axis	Eigenvalue (scaled)	Eigenvalue (unscaled)	Cumulative Eigenvalue as %
1	0.647	4.529	64.703
2	0.163	1.144	81.045
3	0.063	0.440	87.326
4	0.049	0.342	92.206

One problem with PCA is to decide how many components to present, and there are various rules of thumb. The first rule is the '80% rule'; present the first k axes that explain 80% (cumulative) of the total variation. In this case, two axes are sufficient. Another option is a scree plot of the eigenvalues (Figure 12.3). In such a graph, the eigenvalues are plotted as vertical lines next to each other. The aim is to detect an 'elbow-effect'. The justification is that the first k axes explain most information, whereas axes $k + 1$ and higher represent a considerably smaller amount of variation. A scree plot of the eigenvalues would then show which of the axes are important (long lines) and the change point (elbow) at which the axes become less important. In this case, one axis is sufficient. Some software packages use barplots or just lines in the scree plot.

Yet, another tool is the broken stick model approach (Jolliffe 2002; Legendre and Legendre 1998). If a stick of unit length is broken at random in p pieces, then the expected length of piece number j is given by

$$L_j = \frac{1}{p}\sum_{i=j}^{p}\frac{1}{i} \tag{12.3}$$

Comparing the eigenvalue of the j^{th} axis with L_j, gives an idea of the importance of the eigenvalue; if the eigenvalue for axis j is larger than L_j, then it can be considered as important. In this case, the broken stick values of the first three axes are 0.37, 0.23 and 0.16. So based on the broken stick model, only the first axis is of interest. Other tools to select the number of axes (e.g., using cross-validation) are discussed in Jolliffe (2002). However, in many scientific publications the

graphical output of PCA only consists of the first two axes, and occasionally the first four axes.

If the first few axes explain a low percentage, then it might be worthwhile to investigate whether there are outliers, or whether the relationships between variables are non-linear. If either occurs, then consider a transformation or accept the fact that ecological data are genuinely noisy.

The problem with the PCA ordination plot is that is does not tell us whether there are differences between groups of observations (e.g., two bird species), and neither does it give clear information on the influence of the original variables. Obviously, we could label species 1 as '1' and species 2 as '2' in the ordination diagram, and draw circles around groups of observations. Indeed, many authors do this. But with prior knowledge on a grouping structure in the observations available, this is not recommended as other statistical techniques can (and should) be applied for this purpose, for example discriminant analysis or classification models. In the next section, an extension of the ordination diagram is presented, which may give more visual information, namely the biplot.

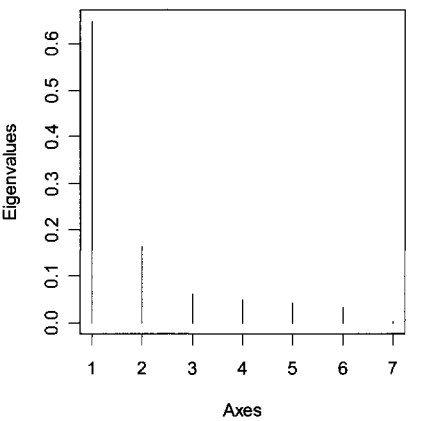

Figure 12.3. Eigenvalues for the sparrow data, obtained by PCA, are presented as vertical lines. The elbow rule suggests presenting only the first axis. Some programmes use barplots or a line plot.

12.5 The biplot

Figure 12.2 shows the general presentation of PCA as a plot of $\mathbf{Z}_1$ against $\mathbf{Z}_2$. Interpretation of the plot is difficult, and the biplot was developed to simplify this. We do not present the detailed mathematics behind the biplot, just the principle. However, if you are not familiar with matrix algebra, you may skip a few paragraphs. Useful references are Krzanowski (1988), Gabriel and Odoroff (1990),

Jongman et al. (1995), Gower and Hand (1996), Legendre and Legendre (1998) or Jolliffe (2002). Recall that the PCA loadings were obtained from an eigenvalue equation of the covariance or correlation matrix. Instead of this, we can also use the singular value decomposition (SVD): A mathematical technique closely related to the eigenvalue equation. It calculates matrices **U**, **L** and **V** such that **Y** = **U L V'**. The matrix **Y** contains all data in an observation-by-variable format. The variables are either centred or normalised. The matrices **U** and **V** are special in the sense that **U'U** = **I** and **V'V** = **I**, where **I** is an identity matrix. This property is also called orthonormal. The matrix **L** is diagonal and contains the square root of the eigenvalues. So, why do we need this? Well, if we take **YV**, we are back to our linear combination of the variables. So, as in the previous section we can write

$$\mathbf{Z} = \mathbf{YV} = \mathbf{UL}$$

The columns of **V** contain the loadings of each axis. Furthermore, each axis (which is a column in **Z**) is given by a column in **U** multiplied with a square root of an eigenvalue (the diagonal element on **L**). The reason we introduce the singular value decomposition is that it can be used for the following formula:

$$\mathbf{Y} = \mathbf{GH}' \quad \text{and} \quad \mathbf{G} = \mathbf{UL}^{\alpha} \text{ and } \mathbf{H}' = \mathbf{L}^{1-\alpha}\mathbf{V}'$$

This formula is the basis for the biplot. The idea is that we are going to plot columns of **G** and columns of **H** in the same graph, and this will allow us to make statements on the variables and observations. The matrix **H** contains information on the loadings and **G** on scores. The problem is that the interpretation of the biplot is based on the choice of α. There are two conventional choices, $\alpha = 0$ and α=1, and each of them is discussed below. This is also called the scaling of the biplot.

The biplot with α = 0 (correlation biplot)

When $\alpha = 0$, the biplot is also called a correlation biplot. In this case it is easy to show that: **Y'Y=HH'**. Except for a constant factor $N - 1$, the left part is the correlation or covariance matrix. This depends on whether Y was centred or normalised. Let us first assume that it is the covariance matrix. Denote the first two columns of **H** by $\mathbf{H}_2$. In the biplot, variable j is represented by $\mathbf{h}_j$, the j^{th} row of $\mathbf{H}_2$. It is conventional to do this by drawing a line from the origin to a point with coordinates given by $\mathbf{h}_j$. Using basic algebra it can be shown that the angle between $\mathbf{h}_i$ and $\mathbf{h}_j$ represents the similarity between two variables. To be more precise, the cosine of the angle between two lines *approximates* (as it is a two-dimensional representation) the correlation. This means that lines pointing in the same direction have a high correlation. Variables with an angle of 90 degrees have a small correlation (close to zero), and variables pointing in the opposite direction have a large but negative correlation. The length of each line is proportional to the variance of a particular variable.

Denote the first two columns of **G** by $\mathbf{G}_2$. In the same graph, each observation can be plotted with coordinates $\mathbf{g}_i$ (the i^{th} row of $\mathbf{G}_2$). This is typically done as a point (or label) and not as a line. Distances between two observations in the biplot

(represented by $\mathbf{g}_i$ and $\mathbf{g}_j$) are an approximation of the Mahalanobis distance. This distance measure works in a similar way as the Euclidean distance except that variables with large variance and groups of highly correlated observations are down-weighted.

The last rule makes use of geometry. As the i,j^{th} element of $\mathbf{Y}$ is approximated by $\mathbf{g}_i*\mathbf{h}_j$, the observations (represented by points in the biplot) can be projected perpendicularly on the lines (variables). The position of the point along the line gives an indication of the value of this observation. The biplot only shows a line from the origin to a point with coordinates $\mathbf{h}_i$ for each observation. The origin represents the average, and the visible part of the line reflects above average values of the variable. One has to imagine a line pointing in the opposite direction representing the values below average.

In case the variables are normalised (centred and divided by the standard deviation), the length of the line shows how well the variable is represented in the two-dimensional approximation. Ideally, all lines should have length 1.

Figure 12.4 shows the PCA correlation biplot for the sparrow data. The variables were normalised. All lines are pointing to the left indicating that all variables are highly correlated with each other. In more detail, wingcrd and flatwing are highly correlated with each other. The same holds for culmen and nalospi. These results are in line with the correlation matrix in Table 12.1. Points (or dots) in the graph (=scores) can be projected on lines. The observations close to the origin have average values for all variables. The observations towards the left have above average values for all variables. However, we cannot easily interpret the Euclidean distances between the observations. The eigenvalues are not influenced by the scaling process.

The biplot with α = 1 (distance biplot)

Let us now turn to the situation in which $\alpha = 1$. This is also called the distance biplot. The reason for this is that Euclidean distances in the biplot are now a two-dimensional approximation of the Euclidean distances between observations. This makes it easier to compare the position of observations with each other. We can still project the points (observations) on the lines. However, angles between lines do not have a direct interpretation in terms of the correlation anymore. The principle of projecting points on lines applies just as before. This means that if two lines point in the same direction we can only say that they both have high or low values for the same observations. So, this is a slightly more descriptive interpretation compared with the interpretation based on angles in the correlation biplot.

Figure 12.5 shows the distance biplot for the same data. We cannot interpret the angles between lines directly, but we can still project observations on lines. The advantage of this approach is that distances between points can now be more easily interpreted: These are approximations of Euclidean distances between the observations. Most of the points on the left are actually from species 2. The fact that they are all separated from species 1 means that bird species 2 has higher values for all morphometric values.

Summary of the scaling process

A biplot is a visualisation tool to present results of PCA. There are various options for the PCA biplot and this is called the scaling process. There are two main options, the correlation biplot, and the distance biplot. In the correlation biplot, we can use the following rules:

- If the data are centred, then the length of a line is proportional to the variance of the corresponding variable. If the data are normalised, then the length of a line is an indication of how well it is represented by the two-dimensional approximation. Long lines are represented well, but short lines should be interpreted with care.
- Angles between lines approximate the correlation.
- Points (observations) can be projected perpendicular on lines, and this gives information of the observed values (abundances).
- Distances between observations are two-dimensional approximations of Mahalanobis distances, but these are difficult to interpret.

For the distance biplot we have:

- If the data are centred, then the length of a line is proportional to the variance of the corresponding variable. If the data are normalised, then the length of a line is an indication of how well it is represented by the two-dimensional approximation.
- Angles between lines are not directly interpretable in terms of correlations.
- Points (observations) can be projected perpendicular on lines, and this gives information of the observed values (abundances).
- Distances between observations are two-dimensional approximations of Euclidean distances.

The choice of scaling depends on the underlying question: Are we interested in making a statement on the variables or observations? If it is the first question, then use the correlation biplot, else the distance biplot. The biplot interpretation should be done with care if the two-dimensional approximation only explains a small percentage of variance.

In Chapter 10, we discussed the pros and cons of the correlation and covariance coefficients as a measure of association. One of the things we mentioned was that if the variation between the variables is important, then use the covariance as the PCA is dominated by such variables. If the variables are in different units, or if only relative changes are important, use the correlation coefficient. The same arguments hold in the choice between the covariance and the correlation matrix for the PCA. Do not confuse this with the correlation biplot as this is a scaling option.

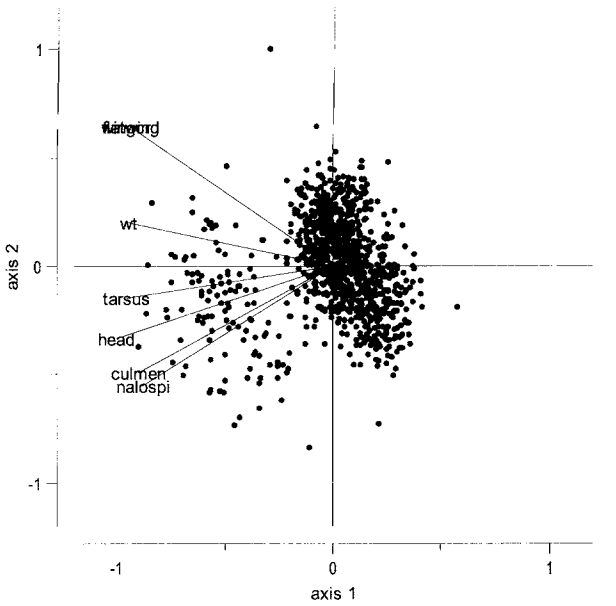

Figure 12.4. PCA biplot for sparrow data. The correlation biplot was used.

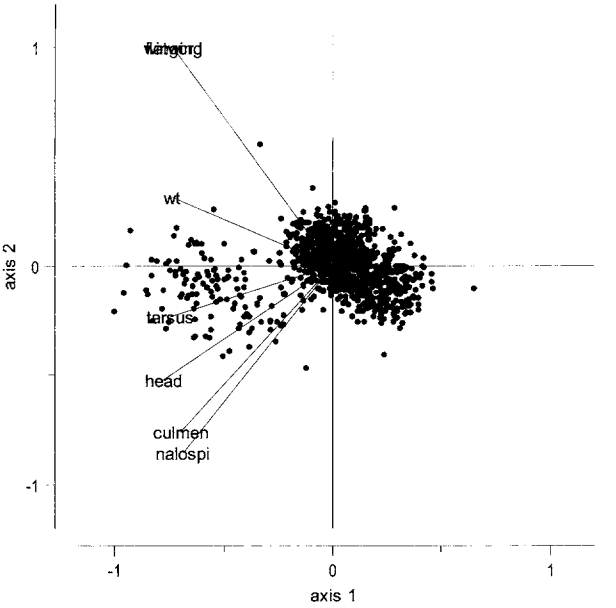

Figure 12.5. PCA distance biplot for the sparrow data.

12.6 General remarks

PCA, as with most other statistical methods, cannot cope with missing values. Missing values must be replaced by a sensible estimate (e.g., mean value of the response variable). Alternatively, consider omitting the variable or the entire observation from the data.

Jolliffe (2002) discusses how PCA can be used for outlier detection. He distinguishes between two situations: A variable with one (or more) large values that should have been detected using dotplots or boxplots, or observations that are outliers but cannot be detected with univariate plotting tools like the Cleveland dotplot (Chapter 4). We discussed this type of outliers in Chapter 4. Outliers of the first type tend to dominate one of the first few axes in PCA. The original sparrow data contained an observation with a large tarsus value, and this mainly determined the third PCA axis. According to Jolliffe (2002) the second type of outlier tends to dominate the last few axes.

Figure 12.6 shows the correlation biplot for the CSL bird radar data that we used in Chapter 10. The covariance matrix was used. The first two eigenvalues are 0.28 and 0.10, representing 38% of the variation. The original data matrix was of dimension 70-by-400. This means that there are fewer observations than variables and as a result there will be 330 axes with zero eigenvalues (these cannot be interpreted). In principle, one should have more observations than variables. However, this does not influence the first few axes and you can still use PCA, but avoid interpreting the higher axes (numbers 71 to 400).

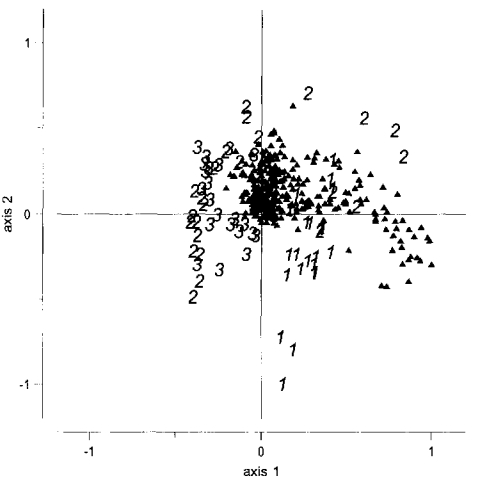

Figure 12.6. PCA correlation biplot for the CSL bird radar data. The triangles represent the variables (counts in grid cells), and the numbers are the observations. The labels 1, 2 and 3 indicate on which day an observation was taken. The covariance matrix was used. Grid cells with less than 10 birds were omitted from the analysis.

The biplot in Figure 12.6 is rather difficult to interpret. The 400 grid cells were used as variables. We have to omit the lines and labels to avoid the graph becoming cluttered with detail to the extent of becoming unreadable. Instead, we used triangles for the variables and the observations are labelled with values 1, 2 and 3, depending on which day the bird was measured. We decided to use the covariance matrix as some of the grid cells at the edge had considerably lower values than in the centre. The additional information we have are the spatial coordinates for each 'variable'. This allows us to make a contour plot of the loadings for an axis; see Figure 12.7. The contour plot for the loadings of the first axis shows that the grid cells with high loadings are all centred at a particular part of the study area, which is located on the right-hand side of the figure. Based on the position of the labels 1, 2 and 3 in Figure 12.6, the high loadings are mainly associated with days one and two. This means that 28% of the variation in the data is related to a geographical hotspot in the study area.

If the underlying question is whether there is a difference between the days, a distance biplot should be made. The resulting biplot is not shown here, but it is similar to Figure 12.6 except that the groups with the labels '1' are more isolated.

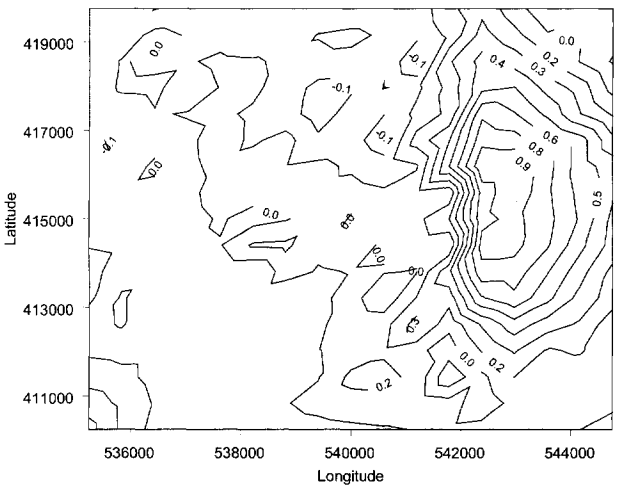

Figure 12.7. Contour graph of the loadings of the first PCA axis for the CSL radar data. Note the hotspot towards the midle right.

12.7 Chord and Hellinger transformations

Three potential problems with PCA are as follows: It measures linear relationships (because it is based on the correlation or covariance coefficient), the presence of double zeros (Chapter 10) and the arch effect. All three problems are

mainly associated with ecological community data; counts of multiple species at multiple sites. For such data, relationships are typically non-linear, and there are many zeros in the data matrix. A data transformation might solve the first problem, but double zeros will stay a problem even after a transformation.

If two species contain many zeros, then both the correlation and the covariance coefficients will indicate that they are similar. Figure 12.8 shows species abundance along an artificial gradient. Species one and two are at different sides of the gradient; they were not sampled at the same sites. Therefore, these species should be labelled as dissimilar. A PCA should place such species and sites at different ends of the first axis. However, the correlation coefficient will indicate that species one and two are highly correlated because of the double zeros. As a result, the second PCA axis will bend the scores of both sides of the gradient slightly inwards, showing the so-called horseshoe or arch effect.

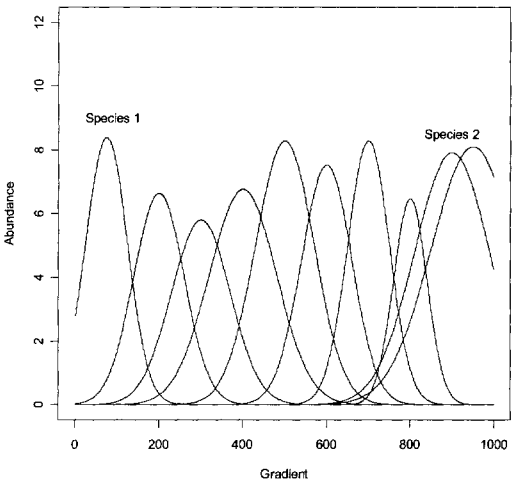

Figure 12.8. Simulated species abundances along a gradient.

Legendre and Gallagher (2001) showed that various other measures of association can be combined with PCA: The Chord distance, Hellinger distance, and two Chi-square related transformations. For the Chord distance, this process works as follows. Let y_{ij} be the abundance of species j at site i, and let y_{ij}' be the transformed abundance according to

$$y'_{ij} = \frac{y_{ij}}{\sqrt{\sum_{j=1}^{p} y_{ij}^2}}$$

Then the Euclidean distance between two (transformed) observations is equal to the Chord distance between them, and the biplot shows a two-dimensional representation of these distances. The Hellinger and Chi-square distances can be ob-

tained with similar data transformations but these are not given here. Legendre and Gallagher (2001) showed that the data transformation using the Chord and Hellinger distance function was less influenced by the arch effect than ordinary PCA (and correspondence analysis). The data transformation is applied before the PCA algorithm is started.

If a species is truly rare but locally abundant (patchy), and if it is felt that it should be included in the analysis, then one of these transformations can be applied. However, if such species are only rare because of the sampling design or if they are considered unimportant, then they should not be used.

The data used in Figure 12.6 contains a large percentage of zeros and a Chord transformation might be appropriate. The resulting biplot after the Chord transformation was applied is shown in Figure 12.9. Distances between observations in this graph are now two-dimensional approximations of the Chord distances.

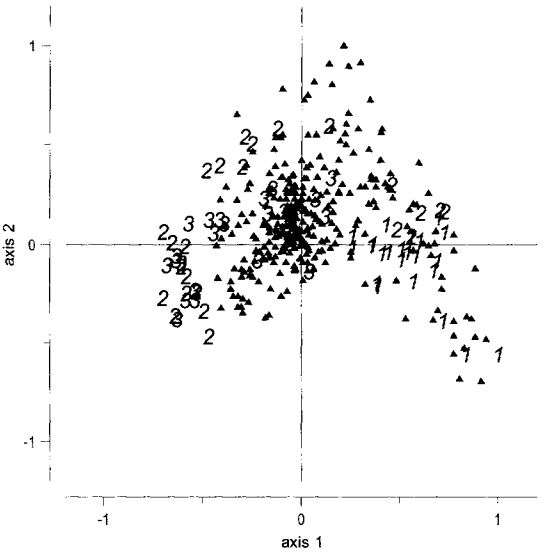

Figure 12.9. PCA distance biplot for the bird radar data. The covariance matrix and a Chord transformation was used.

12.8 Explanatory variables

In this paragraph, we set the scene for redundancy analysis. Figure 12.10-A shows the correlation biplot for the classes of the RIKZ data (Chapter 27) and indicates a clear zonation. Insecta are negatively correlated with Crustacea and Mollusca, and the later two are correlated with each other. The first two eigenvalues are 0.35 and 0.29, which means that they explain 64% of the variation.

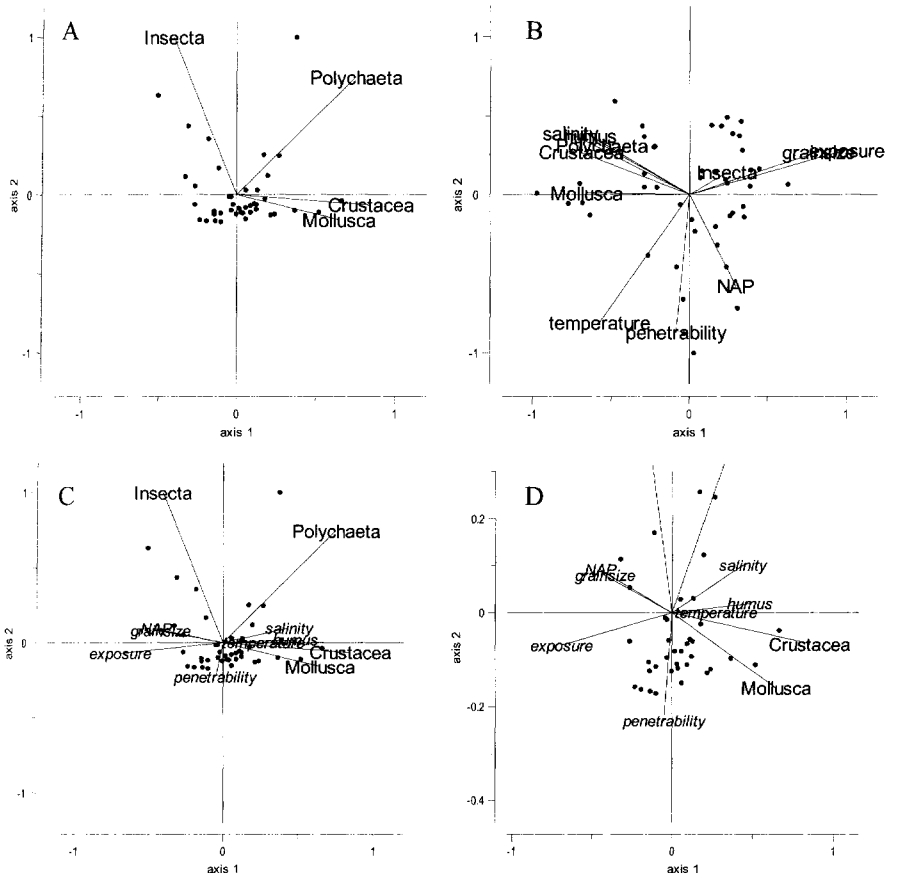

Figure 12.10. A: PCA correlation biplot for RIKZ class data. B: PCA correlation biplot for a combined (species classes and explanatory variables) data matrix. C: PCA correlation biplot applied on the species class data. Explanatory variables are superimposed using cross-correlations. D: As panel C but we zoomed in.

So far, the multivariate analysis has not taken into account any explanatory variables. The first option is to apply the PCA on a data matrix that contains both the class variables and the explanatory variables. The resulting correlation biplot is given in Figure 12.10-B. The first axis seems to be determined by grain size and exposure versus salinity, humus and three of the species classes. Insecta has a rather short line indicating that it is not represented well by the first two axes. The second axis is determined by the variables, NAP, penetrability and temperature. The first two eigenvalues are 0.31 and 0.17, corresponding to 49% of the variation. The problem with this approach is that the PCA will focus on the variables that have the highest correlations (or covariances). If the explanatory variables

have higher correlations with each other, than with the response variables, the main gradients will only contain information on the explanatory variables. This situation is common in ecology. In this particular case, perhaps three species classes might be related to environmental variables, but there is no guarantee that this is the case with other datasets. Insecta is clearly not related to the main gradients in the combined data.

Another option is to apply PCA on the class data and calculate correlations between the PCA axes $\mathbf{Z}_1$ and $\mathbf{Z}_2$ and each of the explanatory variables. This information can then be superimposed on the biplot; see Figure 12.10-C. The line for an explanatory variable is obtained by drawing a line from the origin to a point (c_1,c_2), where c_1 is the correlation between the first axis and the explanatory variable, and c_2 with the second axes. A long line indicates that the explanatory variable is highly correlated with at least one of the axes. In this case, correlations between axes and explanatory variables are relatively low. We zoomed in to obtain a better view of the correlations (Figure 12.10-D). The first axis is mainly related to Crustacea and Mollusca. Using the biplot rule, observations that have positive $\mathbf{Z}_1$ values have above average values for these two species classes. These samples have below average values for exposure, grain size and NAP. These correlations are approximately 0.5. Correlations between the second axis and the explanatory variables are small. The problem with this approach is that the major gradients in the species data might not be related to the environmental variables. It might well be that the third and fourth PCA axes are related to the environmental variables. In such cases we would not detect it in Figure 12.10-C. In the next section, we will use a more appropriate technique to relate multiple species with multiple explanatory variables.

12.9 Redundancy analysis

In previous sections, it was shown that PCA is a useful tool to visualise correlations or covariances between variables, but the explanation of the results in terms of explanatory variables was cumbersome. Redundancy analysis (RDA) is an interesting extension of PCA that explicitly models response variables as a function of explanatory variables. We start by looking at RDA, by ignoring all the formulae and only discuss how to interpret the graphical and numerical output. We then look at the mathematics, which is an extension of the iterative algorithm discussed in the PCA section.

The graphical output of RDA consists of two biplots on top of each other, and is called a triplot. There are three components in such a graph:

1. The quantitative explanatory variables are represented by lines, and the qualitative (nominal) explanatory variables by squares (one for each level).
2. The response variables (often the species) by lines or labels. If there are many response variables, it may help visual interpretation to use labels.
3. The samples by points or labels. If there are many samples, using points instead of labels improves the visual quality of the graph.

The choices and interpretation are similar to the PCA biplot. First, we have to decide the scaling: The correlation ($\alpha = 0$) or distance triplot ($\alpha = 1$). Second, whether we want to use the covariance or correlation matrix for the response variables. In the triplot, the response variables (typically species) are represented by lines, and the observations by dots. The new thing is the lines for the quantitative explanatory variables and square for the levels of the nominal explanatory variables. The rules for the RDA *correlation triplot* interpretation are as follows (Ter Braak 1994; Legendre and Legendre 1998; Lepš and Šmilauer 2003):

1. Angles between species lines represent correlation between the species.
2. Points for observations can be projected perpendicularly on the species lines and indicate the values.
3. Points for observations cannot be compared directly with each other.
4. Angles between lines of quantitative explanatory variables represent correlations between them. But these are not as accurate as the ones obtained by a PCA applied on only the explanatory variables.
5. Angles between lines of species and explanatory variables represent correlations.
6. Observations can be projected perpendicular on the lines for explanatory variables indicating the values of the explanatory variables at those sites.
7. The nominal explanatory variables (coded as 0–1) can be represented as a square. Its position is determined by the centroid of the observations that have the value 1 (for this variable). Distances between centroids and between observations and centroids are *not* approximations of their Euclidean distances. A square can be projected perpendicular on a line for a species and represents the mean species abundances for that class. Squares cannot be compared with the qualitative explanatory variables

The first three rules are similar to the PCA correlation biplot. For the *distance triplot*, the following rules hold:

1. Angles between lines for species do not represent correlations.
2. Points for observations can be projected perpendicularly on the species lines and indicate the (fitted) values.
3. Distances between observations represent a two-dimensional approximation of their Euclidean distances.
4. Angles between lines of species and qualitative explanatory variables represent a two-dimensional approximation of correlations.
5. The nominal explanatory variables (coded as 0–1) can be represented as a square. Its position is determined by the centroid of the observations that have the value 1. Distances between centroids and between observations and centroids are approximations of their Euclidean distances. A square can be projected perpendicular on a line for a species and represents the mean species abundances for that class. Squares cannot be compared with the qualitative explanatory variables

The correlation and distance triplots are also called species-conditional triplot and the site conditional triplot, or scaling 2 and scaling 1. If the response variables are not species by sites, then perhaps the nomenclature Y-conditional and sample

conditional scaling is more appropriate. Legendre and Legendre (1998) used the names correlation biplot and distance biplot. Mathematically, the difference between these two scalings is merely a few simple multiplications involving eigenvalues (Section 12.5, Legendre and Legendre 1998), but the ecological interpretation of the two types of triplots is different, as discussed above. If the interest is on observations, then the distance triplot should be used.

An example is presented next. We have used the same RIKZ data that we used earlier in Section 12.8. The response variables are the four class variables: Insecta, Polychaeta, Mollusca and Crustacea. There are 10 explanatory variables. One of them is nominal (week, with four levels) and one (exposure) could be considered as nominal or continuous. In first instance, we will omit week and consider exposure as continuous. The resulting triplot is presented in Figure 12.11. The rules to read this graph are similar to the PCA biplot. The triplot (Figure 12.11) is based on the correlation biplot, and because we want to make all 'species' equally important, the correlation matrix was used, and not the covariance matrix. The triplot for the RIKZ data shows that NAP and grain size are highly correlated, and exposure is negatively correlated to humus and temperature. Mollusca, and Crustacea are correlated with each other, and also with Polychaeta, but not with Insecta. In terms of species-environmental relations, there seems to be a negative effect of exposure, NAP and grain size on Crustacea, Mollusca and Polychaeta. Insecta seem to be positively related with chalk and sorting, and negatively to penetrability. We can also infer at which sites this is happening.

The numerical output of RDA for the RIKZ data is given in Table 12.3 and shows that all the explanatory variables explain 38% of the variation in the species data. This is the sum of all so-called canonical eigenvalues. From this 38%, the first two axes explain 87%. This is the column labelled 'eigenvalue as cumulative percentage of the sum of all eigenvalues'. Hence, the first two axes are a good representation of what can be explained with all the explanatory variables. This value tends to be high for most datasets due to high correlations between explanatory variables. The fourth column is more interesting. It shows that the first two axes explain 33% of the variation in the species data. This value is obtained by multiplying 0.38 (the sum of all canonical eigenvalues) with 0.87 (the variation explained by the first two axes). The RDA algorithm rescales the response variables to have a total sum of squares of 1.

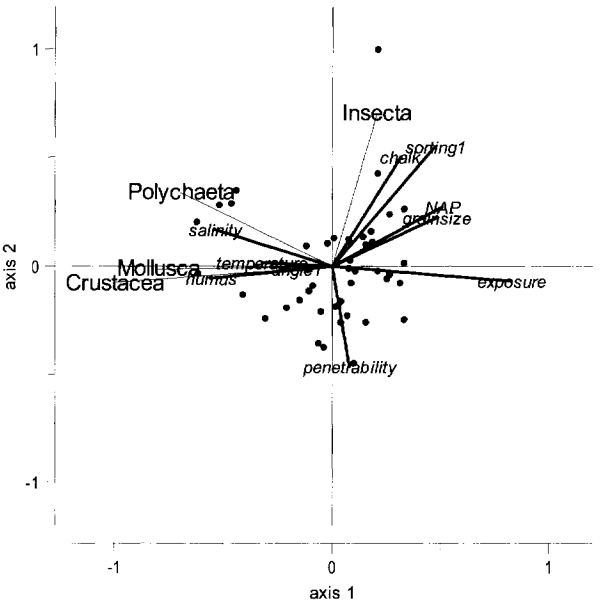

Figure 12.11. RDA correlation triplot for RIKZ class data.

Table 12.3.Numerical output of PCA applied to the RIKZ data. The sum of all canonical eigenvalues is 0.38, and the total variance is 1.

Axis	λ	λ as %	λ as cumulative %	λ as % of sum of all canonical eigenvalues	λ as cumulative % of sum of all canonical eigenvalues
1	0.26	26	26	68	68
2	0.07	7	33	19	87

A second approach to explain RDA is as follows. PCA calculates the first principal component as

$$Z_{i1} = c_{11}Y_{i1} + c_{12}\,Y_{i2} + \ldots + c_{1N}\,Y_{iN}$$

Redundancy analysis is a sort of PCA that requires that the components are linear functions of the explanatory variables:

$$Z_{i1} = a_{11}X_{i1} + a_{12}X_{i2} \ldots\ldots + a_{1q}\,X_{iQ}$$

Hence, the axes in RDA are not only a linear combination of the response variables, but also of the explanatory variables. It is as if you tell the computer to apply a PCA, but only show the information in the biplots, which can be (linearly) related to the explanatory variables. Further axes are obtained in the same way. RDA can be applied if there are N response variables measured at M sites and Q

explanatory variables measured at the same sites. Note that this technique requires an explicit division of the variables into response and explanatory variables. The maximum number of RDA axes is the minimum of N and M.

The more mathematical explanation of RDA requires the iterative algorithm for PCA, which was presented in the previous section. The algorithm has the following steps:

1. Normalise (or centre) the variables **Y**, and normalise the explanatory variables **X**.
2. Obtain initial scores **z** (e.g., by a random number generator).
3. Calculate new loadings: **c** = **Y'z'**.
4. Calculate new scores: **z** = **Yc**.
5. For second and higher axes: Make **z** uncorrelated with previous axes using a regression analysis.
6. Apply a linear regression of **z** on **X,** and set **z** equal to the fitted values.
7. Scale **z** to unit variance: $\mathbf{z}^* = \mathbf{z}/\lambda$, where λ is the standard deviation of the scores. Set **z** equal to **z***.
8. Repeat steps 2 to 7 until convergence.
9. After convergence, divide λ by $M - 1$.

The only extra steps are the normalisation of **X** and the regression in step 6. The fitted values obtained by a linear regression model represent the information in the response variable that is linearly related with the explanatory variables. So the effect of this regression is that only the information in **z** (linear combination of **Y**'s) that is related to the **X** is presented. This ensures that the scores **z** are a linear combination of the explanatory variables.

We also discussed the PCA eigenvalue equation. Indeed, RDA can also be solved with an eigenvalue equation and a couple of matrix multiplications, see Legendre and Legendre (1998) for details.

Because RDA is basically a series of multiple linear regression steps, we need more observations than explanatory variables.

Nominal variables

Different statistics programmes treat nominal variables slightly differently, and this is discussed next using an artificial example. Suppose that abundances of two species were measured at five sites and that sampling took place over a period of three months by two observers (Table12.4). Sites 1 and 4 were measured in April, site 2 in May and sites 3 and 5 in June. The first observer sampled sites 1 and 2, whereas the second observer sampled the remaining months. The explanatory variables Month and Observer are nominal variables. The last variable is easily dealt with. Define $\text{Observer}_i = 0$ if the i^{th} site was sampled by observer A and 1 if observer B measured the data. Because Month has three classes (April, May and June), three new columns are defined: April, May and June. Where the value is 1 if sampling took place in the corresponding month and 0 elsewhere. However, there is one little problem: The variables April, May and June are linearly related and therefore one of the columns should be omitted. If this is not done, the soft-

ware will give an error message. For the RDA and CCA, the nominal variables must be coded with values 0 and 1.

We re-applied the RDA model to the same RIKZ data, except that the explanatory variable 'week' was taken into account. This variable contains four classes; week 1, 2, 3 and 4, and four new variables W1, W2, W3 and W4 were created. If an observation was from week 1, W1 was set to 1, and W2, W3 and W4 to 0. The same was done for observations from other weeks. To avoid collinearity, the variable W4 was not used in the analysis. The RDA triplot is shown in Figure 12.12. Nominal variables are represented by squares. Adding the three extra variables did not change the main patterns in the triplot. The sum of all canonical eigenvalues is now 44%. Although difficult to see, the square for week 1 indicates that Polychatea, Mollusca, Crustacea and temperature were high in this week.

Table 12.4. Set up of an artificial dataset. Measurements were made in April, May and June (indicated by a '1') and by two observers.

	Species 1	Species 2	Temp	Wind	April	May	June	Observer
Site 1	...	...	...	...	1	0	0	0
Site 2	...	...	...	...	0	1	0	0
Site 3	...	...	...	...	0	0	1	1
Site 4	...	...	...	...	1	0	0	1
Site 5	...	...	...	...	0	0	1	1

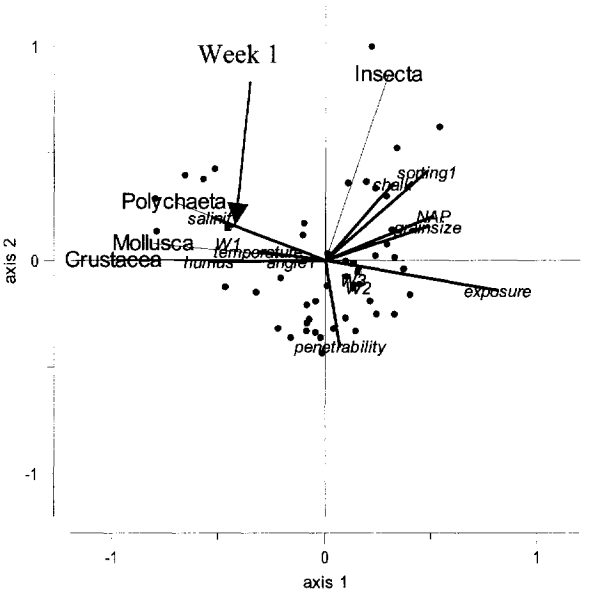

Figure 12.12. RDA triplot for the RIKZ data. The nominal variable week was added.

The order of importance

An interesting question is which of the explanatory variables is the most important, which are the least important, and which are irrelevant. Just as in linear regression, this question can be answered with a forward selection. However, in linear regression, we used the AIC to choose the optimal model. Here, we have a slightly different strategy. The sum of all canonical eigenvalues is used as a tool to assess how well a specific selection of explanatory variables explains the variance in the species data. But if there is only one explanatory variable, the total sum of all canonical eigenvalues is equal to the eigenvalue of the first and only axis. We call these the marginal effects. Table 12.5 shows the marginal effects for the same RIKZ data. It shows the eigenvalue and percentage of explained variance if only one explanatory variable is used in RDA. Results indicate that exposure is the single best explanatory variable, followed by week 1.

Conditional effects (Table 12.6) show the increase in the total sum of eigenvalues after including a new variable during the forward selection. The first variable is exposure, used because it was the best single explanatory variable (Table 12.5). To test the null hypothesis that the explained variation is larger than a random contribution, a Monte Carlo permutation test (see below) is applied. The F-statistic and p-value indicate that the null-hypothesis can be rejected. The second variable to enter the model is NAP, and the total sum of all eigenvalues increases with 0.06. The Monte-Carlo test indicates that it is significant. The next two variables to enter the model are sorting and week 1. After these four variables, the increase in the total sum of eigenvalues is only 0.02 and the Monte Carlo test shows that it is not significant.

Table 12.5. Marginal effects for the RIKZ. Group classes were square root transformed. The total sum of all eigenvalues is 0.44, and the total variance is 1. The second column shows the eigenvalue using only one explanatory variable, and the third column is the eigenvalue as percentage of the sum all eigenvalues (using all variables).

Explanatory Variable	Eigenvalue Using Only One Explanatory Variable	Eigenvalue as %
angle1	0.01	1.83
exposure	**0.18**	**40.52**
salinity	0.08	18.23
temperature	0.04	8.58
NAP	0.07	16.53
penetrability	0.02	4.10
Grain size	0.06	14.57
Humus	0.08	18.41
Chalk	0.05	11.14
sorting1	0.08	18.40
W1	0.17	39.64
W2	0.04	8.91
W3	0.03	6.02

Table 12.6. Conditional effects for the RIKZ data. The classes were square root transformed. The total sum of all eigenvalues is 0.44, and the total inertia is 1. The second column shows the increase in explained variation due to adding an extra explanatory variable. If one level of week is important, then all its levels should be selected in the final model (or none).

Order	Explanatory Variable	Increase Total Sum Eigenvalues After Including New Variable	F-statistic	p-value
1	**exposure**	**0.18**	**9.340**	**0.000**
2	**NAP**	**0.06**	**3.466**	**0.011**
3	**sorting1**	**0.04**	**2.532**	**0.068**
4	**W1**	**0.05**	**3.214**	**0.017**
5	salinity	0.02	1.099	0.352
6	W3	0.02	1.328	0.257
7	W2	0.02	0.951	0.450
8	chalk	0.01	0.835	0.456
9	penetrability	0.01	0.490	0.726
10	temperature	0.01	0.602	0.621
11	humus	0.01	0.389	0.780
12	angle1	0.01	0.210	0.878
13	grain size	0.00	0.228	0.906

A bit more info on the permutation tests

Details of the permutation test can be found in Legendre and Legendre (1998) or Lepš and Šmilauer (2003). Let us start simply. We have one explanatory variable and multiple response variables. The data are in the following format:

$$\begin{matrix} Y_{11} & \cdots & Y_{1N} \\ Y_{21} & \cdots & Y_{2N} \\ \vdots & & \vdots \\ Y_{M1} & \cdots & Y_{MN} \end{matrix} \qquad \begin{matrix} X_1 \\ X_2 \\ \vdots \\ X_M \end{matrix}$$

We have two data matrices, the N response variables Y and one explanatory variables $\mathbf{X}$, measured both at the same M sites. In linear regression, we used an F-test to compare nested models and this gave information on the importance of explanatory variables. Here, we do something similar. First, we apply the model with the explanatory variable and obtain an F-statistic, which is defined differently to the F-statistic in linear regression. Because it is for the original data, we call it F^*. Then we assume that there is no relationship between the rows in the $\mathbf{Y}$s and the $\mathbf{X}$. This is our null hypothesis. Under the null hypothesis, we can randomly change the order of the rows in the $\mathbf{X}$ (or indeed $\mathbf{Y}$, but let's do the $\mathbf{X}$). This is called a permutation. Each time we do a permutation, we can also calculate our F-statistic, which is a measure how much the Ys are related to the (permutated) $\mathbf{X}$. We could do this process a large number of times, say 9999 times. We will explain shortly why 9999 and not 10000. If the F^* from the original data has very similar values

compared with permutated 9999 F values, then there is no reason to doubt the validity of the null hypothesis. However, if the F^* is considerably larger than the majority of the 9999 F values, then perhaps our assumption of no relationship between the rows of **Y** and **X** is incorrect. In fact, the number of times that F is larger than F^* is closely related to the p-value. To be more precise, the p-value is given by

$$p = \frac{\text{Number of times that } F \geq F^* + 1}{\text{Number of permutations} + 1}$$

The '+1' is used because F^* itself is also used as evidence against the null hypothesis. Hence, if we choose 9999 permutations in a software package, we basically have to write in the paper or report that 10000 permutations were carried out.

Instead of one explanatory variable, it is more likely that we have multiple (say Q) explanatory variables:

$$\begin{matrix} Y_{11} & \cdots & Y_{1N} \\ Y_{21} & \cdots & Y_{2N} \\ \vdots & & \vdots \\ Y_{M1} & \cdots & Y_{MN} \end{matrix} \qquad \begin{matrix} X_{11} & \cdots & X_{1Q} \\ X_{21} & \cdots & X_{2Q} \\ \vdots & & \vdots \\ X_{M1} & \cdots & X_{MQ} \end{matrix}$$

For multiple explanatory variables, exactly the same procedure can be carried out: The rows of **X** are permutated a large number of times, and each time an F-statistic is calculated. The next question is now, how to define the F-statistic. And here is where things get a bit complicated. There is an F-statistic for testing only the first axis, for the overall effect of the Q explanatory variables, for data involving covariables (these are discussed in the next section), and for doing a forward selection. They are all defined in terms of eigenvalues, information of explained variance and number of axes and explanatory variables. The interested reader is referred to Legendre and Legendre (1998) for more detail.

Just as in linear regression, the use of forward selection is criticised by some scientists (and routinely applied by others). If a large number of forward selection steps are made, it might be an option to apply a Bonferroni correction. In such a correction, the significance level α is divided by the maximum number of steps in the forward selection. Alternatively, the p-value of the F-statistic can be multiplied by the number of steps. The main message is probably, as always, be careful with explanatory variables that have a p-value close to significance.

The last point to discuss is the permutation process itself. We explained the permutation process as changing the order of the rows in the **X** matrix. This is called raw data permutation. Just as in bootstrapping techniques (Efron and Tibshirani 1993), you can also modify the order of the residuals. Within RDA, this is called permutation of residuals under a full model. In this case, the RDA is applied using the **Y**s and the **X**s, and residuals from the **Y**s are permuted. Legendre and Legendre (1998) mentioned that simulation studies have shown that both approaches are adequate. In case of permutation tests with covariables (next section),

permutation of residuals is preferred if the covariables contain outliers. However, these should already have been identified during the data exploration.

When the rows in the data matrices correspond to a time series, a spatial gradient, or are blocks of samples, permuting the rows as discussed above might not be appropriate as it increases the type I error. For time series and spatial gradient data, permuting the rows with blocks might be an option. In this approach, blocks of samples are selected, and the blocks are permuted, but not the data within the blocks. If the time span of the data is not long enough to generate a sufficient number of blocks (> 20), it might be an option to consider the data as circular and connect the last observation (row) with the first one. However, this process does involve removing trends from data (Lepš and Šmilauer 2003), and this might be the main point of interest. If the data consist of different transects, and there is a large difference between transects, permuting the rows only within transects, and not between them, is better. This is also relevant if the data are obtained from an experiment that would normally (within a univariate context) be analysed with an ANOVA (different treatments).

Chord and Hellinger transformations

As with PCA, RDA is based on the correlation (or covariance) coefficient. It therefore measures linear relationships and is influenced by double zeros. The same special transformations as in PCA can be applied. This means that RDA can be used to visualise Chord or Hellinger distances. Examples can be found in Chapters 27 and 28.

12.10 Partial RDA and variance partitioning

If you are not interested in the influence of specific explanatory variables, it is possible to partial out their effect just as we did in partial linear regression (Chapter 5). For example, in Table 12.4 you might be interested in the relationship between species abundances and the two explanatory variables temperature and wind speed, with the effects of month and observer of less interest. Such variables are called covariables. Another example is the RIKZ data in Table 12.6 where you could consider the explanatory variables Weeks 1–3 as covariables and investigate the role of the remaining explanatory variables. Or possibly you are only interested in the pure exposure effect.

In partial RDA, the explanatory variables are divided into two groups by the researcher, denoted by **X** and **W**. The effects of **X** are analysed while removing the effects of **W**. This is done by regressing the covariables **W** on the explanatory variables **X** in step 1 of the algorithm for RDA, and continuing with the residuals as new explanatory variables. Additionally, the covariables are regressed on the canonical axes in step 6, and the algorithm continues with the residuals as new scores.

Suppose we want to know the relationships between all the variables after filtering out the effects of exposure. Simply leaving it out would not answer the question as various variables may be collinear with exposure, and would therefore take over its role. In the partial RDA, the effect of exposure is removed from the response variables, the explanatory variables and the axes. The resulting triplot is presented in Figure 12.13. Note that after filtering out the effects of exposure, Crustacea is positively related to humus and negatively to temperature and NAP. Insecta and Mollusca are not related to any of the remaining variables.

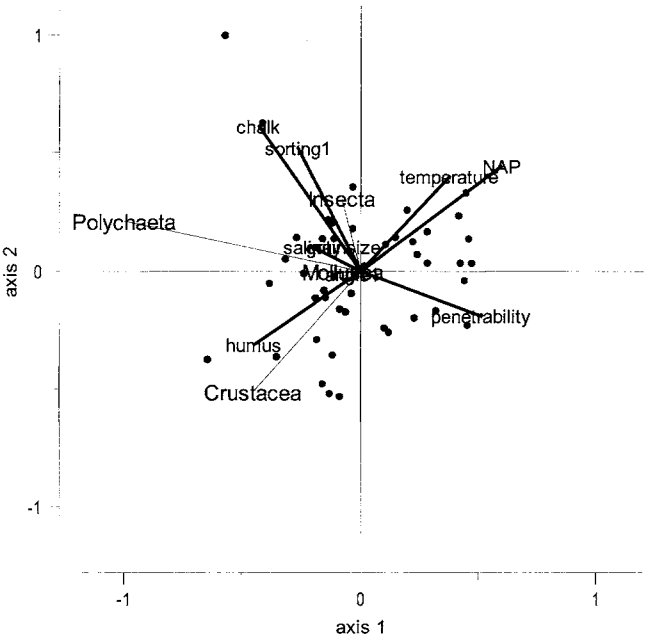

Figure 12.13. Triplot obtained by a partial RDA. The effect of exposure was removed. A correlation triplot using the covariance matrix is presented.

Variance partitioning for linear regression was explained in Chapter 5. Using the R^2 of each regression analysis, the pure **X** effect, the pure **W** effect, the shared effect and the amount of residual variation was determined. Borcard et al. (1992) applied a similar algorithm to multivariate data and used CCA instead of linear regression. Their approach results to the following sequence of steps for variance partitioning in RDA:

1. Apply a RDA on **Y** against **X** and **W** together.
2. Apply a RDA on **Y** against **X**.
3. Apply a RDA on **Y** against **W**.
4. Apply a RDA on **Y** against **X**, using **W** as covariates (partial RDA).

5. Apply a RDA on **Y** against **W**, using **X** as covariates (partial RDA).

Using the total sum of canonical eigenvalues of each RDA analysis (equivalent of R^2 in regression), the pure **X** effect, the pure **W** effect, the shared information and the residual variation can all be explained as a percentage of the total inertia (variation).

An example of variance partitioning for the RIKZ data is presented next. The question is: What is the pure exposure effect? As we suspected that week was strongly related to exposure, we excluded week from the analyses. The results from the five different RDA analyses are given in Table 12.7. Using this information, the pure exposure effect can easily be determined (Table 12.8) and is equal to 7% of the variation. The shared amount of variation is 10%. It is not possible to distinguish this information due to collinearity between exposure and some of the other variables.

Table 12.7. Results of various RDA and partial RDA analysis for the RIKZ data. Total variation is 1. Percentages are obtained by dividing the explained variance by total variance. The 'other' variables are the remaining nine explanatory variables.

Step	Explanatory variables	Explained Variance	%
1	Exposure and others	0.38	38
2	Exposure	0.18	18
3	Others	0.30	30
4	Exposure with others as covariable	0.07	7
5	Others with exposure as covariable	0.20	20

Table 12.8. Variance decomposition table showing the effects of exposure, and other variables for the RIKZ data. Components A and B are equal to the explained variances in steps 5 and 4, respectively. C is equal to variance step 3 minus variance step 5, and D is calculated as Total variance – the explained variance in step 1. In RDA, total variance is equal to 1. *: To avoid confusion, note that due to rounding error, the percentages do not add up to 100.

Component	Source	Calculation	Variance	%
A	Pure others		0.20	20
B	Pure exposure		0.07	7
C	Shared (3–5)	0.30 – 0.20	0.10	10
D	Residual	1.00 – 0.38	0.62	62
Total*				100

12.11 PCA regression to deal with collinearity

An interesting extension of PCA is PCA–regression. A detailed explanation can be found in Jolliffe (2002). In this method, the PCA components are extracted and

used as explanatory variables in a multiple linear regression. Using estimated regression parameters and loadings, it can be inferred which of the original variables are important.

As an example, we use the same RIKZ data that we used in the previous section. The question we address is which of the explanatory variables is related to the class Crustacea using a linear regression model. In the previous section, we used a selection of the explanatory variables, but here we use all 10. Hence, the linear regression model is of the form:

$$\text{Crustacea} = \alpha + \beta_1 \text{Angle1} + \beta_2 \text{Exposure} + \ldots + \beta_{10} \text{Sorting} + \varepsilon$$

To find the optimal regression model, we need to apply a backward or forward selection to find the optimal selection of explanatory variables. However, due to high collinearity this may be a hazardous exercise. To visualise this potential problem, we applied a PCA on all explanatory variables and the resulting correlation biplot is presented in Figure 12.14. Note that several lines are pointing in the same direction, which is an indication of collinearity. However, before we can confirm this, we need to check the quality of the two-dimensional biplot. The first four eigenvalues are 0.37, 0.18, 0.12 and 0.11 and therefore all four axes explain 77% of the variation in all ten variables, and the first two axes explain 55%. This is enough to get nervous about (collinearity), so we need to do something. There seems to be three groups of variables. The quick solution is to select one explanatory variable from each group in the biplot in Figure 12.14, for example grain size, salinity and temperature, and use these in the linear regression model as explanatory variables. This is a valid approach, and is used in some of the case study chapters. The higher the explained amount of variation along the first two axes, the more confident you can be with the selected variables.

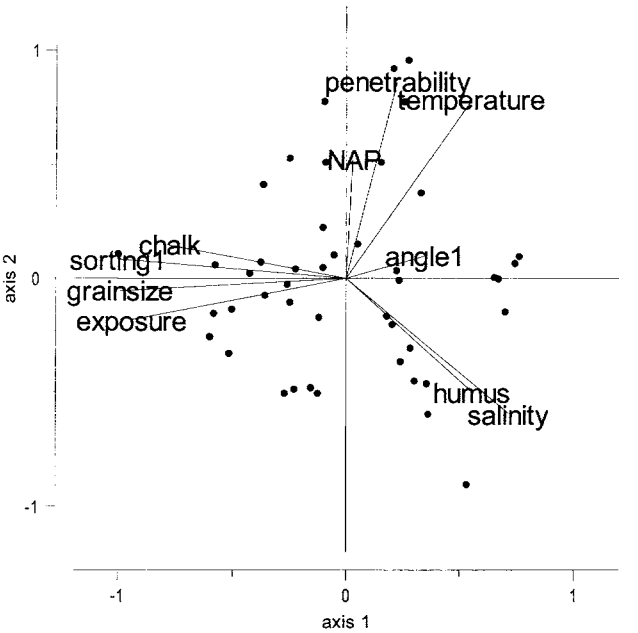

Figure 12.14. PCA biplot for the RIKZ data using all explanatory variables.

Alternative options to identify collinear explanatory variables are variance inflation factors (Chapter 26), pairplots or correlation coefficients. A more advanced method is PCA regression. All the PCA axes, which are uncorrelated, are used as explanatory variables in the linear regression model. The following linear regression model is applied:

$$\text{Crustacea} = \alpha + \beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_3 + \beta_4 Z_4 + \ldots \beta_{10} Z_{10} + \varepsilon$$

Z_1 is the first axes (representing the major environmental gradient), Z_2 the second, etc. The estimated model was as follows:

$$\text{Crustacea} = 20.44 + 9.31\text{x}\, Z_1 + 3.30\, Z_2 - 6.07\, Z_3 + 0.97 \times Z_4 - 1.91\, Z_5 - 1.16 \times Z_6 + 12.08 \times Z_7 + 19.78 \times Z_8 - 32.51 \times Z_9 + 7.54 \times Z_{10} + \varepsilon$$

The t-values of the regression parameters (not presented here) indicate that only Z_1 and Z_9 are significantly related to Crustacea. A backward selection indicated the same. All axes explained 41% of the variation in Crustacea, Z_1 and Z_9 28%, Z_1 alone 20% and Z9 alone 8%. The optimal model can be written as

$$\text{Crustacea} = 20.44 + 9.31 \times Z_1 - 32.51 \times Z_9 + \varepsilon$$

However, the loadings of Z_1 and Z_9 tell us how the axes are composed in terms of the 10 original explanatory variables. So, we can multiply the loadings of each

axis with the corresponding regression coefficient and then add up the loadings of the same explanatory variable (Table 12.9). The multiplication factors are the estimated regression coefficients 9.31 and –32.51. So, all the loadings for Z_1 are multiplied with 9.31, and those of Z_9 with –32,51. If we add up all these multiplied loadings (see the last column in Table 12.9), we obtain a regression model of the form

Crustacea = 20.44 + 4.17 × Angle − **21.82 × Exposure** + 8.14 × Salinity − **15.89 × Temperature** + 6.60 × NAP + **11.23 × Penetrability** − 5.49 × Grain size + 0.45 × Humus + 6.95 × Chalk − **9.12 × Sorting**

Because all explanatory variables were standardised, we can directly compare these loadings. The results suggest that high values of exposure, temperature and sorting are associated with low values of Crustacea, and high values of salinity and with high values of Crustacea. Other explanatory variables have less influence. Humus has a very small influence. The PCA regression approach allows you to assess which of the explanatory variables are important while avoiding problems with collinearity. It is also possible to obtain confidence bands and *p*-values (Jolliffe 2002, Chapter 8).

Jolliffe (2002) states that the low-variance axes can sometimes be better predictors than the high-variance ones. So you could also try applying the PCA regression method on the low-variance axes.

Table 12.9. Loadings for each axis and the total sum of loadings per variable.

Variable	Loadings Z_1	Loadings Z_9	Sum
angle1	0.16	–0.08	4.17
exposure	–0.43	0.55	–21.82
salinity	0.33	–0.16	8.14
temperature	0.25	0.56	–15.89
NAP	0.01	–0.20	6.60
penetrability	0.10	–0.32	11.23
grain size	–0.45	–0.30	5.49
humus	0.25	0.06	0.45
chalk	–0.36	–0.32	6.95
sorting1	–0.46	0.15	–9.12