15 Principal coordinate analysis and non-metric multidimensional scaling

15.1 Principal coordinate analysis

In Chapter 12, principal component analysis (PCA) was introduced. The visual presentation of the PCA results is by plotting the axes (scores) in a graph. Some books use the phrase 'scores are plotted in a Euclidian space'. What this means is that the scores can be plotted in a Cartesian axes system, another notation is $|^2$, and the Pythagoras theorem can be used to calculate distances between scores. The problem is that PCA is based on the correlation or covariance coefficient, and this may not always be the most appropriate measure of association. Principal co-ordinate analysis (PCoA) is a method that, just like PCA, is based on an eigen-value equation, but it can use any measure of association (Chapter 10). Just like PCA, the axes are plotted against each other in a Euclidean space, but the PCoA does not produce a biplot (a joint plot of the variables and observations).

The aim of PCoA is to calculate a distance matrix and produce a graphical configuration in a low-dimensional (typically two or three) Euclidean space, such that the distances (as measured by the Pythagoras theorem) between the points in the configuration reflect the original distances as good as possible. The PCoA can be applied either on the variables or on the observations. Other names for PCoA are metric multidimensional scaling and classical scaling.

How does PCoA work?

To illustrate how PCoA works, we will use data from a plant study in Mexico. A detailed analysis of these data is presented in Chapter 32. The data consist of 200 observations (percentage cover) on a large number of plant families (32). In Chapter 32, totals per pastures (20 in total) are used, but here we will use the original 200 observations. Further details are given in Chapter 32. There are many observations equal to zero, and various species have a patchy distribution, which makes the correlation, covariance and Chi-square functions less appropriate tools to define association. Hence, we have a data matrix of dimension 200-by-32 with many zeros, and the question we focus on here is which families co-occur. The motivation for this question is that one of the plant families was introduced, and it may cause stress for the indigenous families. So, we need to calculate a distance

matrix of dimension 32-by-32, which will tell us how dissimilar the families are, and we want a graphical representation of this to help with the interpretation. Because PCA and correspondence analysis are not suitable for these data (they are based on covariance, correlation or the Chi-square function), we use PCoA.

The first choice we have to make is how to define association between the families. In Chapter 10, and in Legendre and Legendre (1998, p. 294), it is argued that the Jaccard index can be used to measure 'co-occurrence of species'. We used the Jaccard index to measure the association between the families. There are 32 families and 200 sites; hence the similarity matrix **S** is of dimension 32-by-32. It is converted into a distance matrix **D** by **D** = 1 − **S** (Thorrington-Smith 1971). In the next step of the algorithm, the ij^{th} element of **D** is transformed as follows:

$$A_{ij} = -\frac{1}{2} D_{ij}^2$$

This gives a matrix **A** where the ij^{th} element is calculated as above, and it has the same dimension as **D**. A second transformation is then applied on the elements of **A**:

$$E_{ij} = A_{ij} - \overline{A}_i - \overline{A}_j + \overline{A}$$

The notations $\overline{A}_i$, $\overline{A}_j$ and $\overline{A}$ stand for row, column and overall average. It can be shown that this transformation preserves the distances relationship between family i and j (Legendre and Legendre 1998, pg 430). In the last step, an eigenvalue equation is solved for **E**, and after applying the appropriate scaling on the eigenvectors, the eigenvector (or axes) can be plotted against each other in a two- or three-dimensional graph, just like in PCA. Distances between the points on this graph approximate the distances in **D**. An example is given in Figure 15.1-A. The first two axes explain 30% of the variation in the distance matrix (just as in PCA, the eigenvalues are used for this). Families close to each other co-occur at the same sites. Our special interest is on the family grcyn, and it co-occurs with ac, co and the group of families at the right side of the graph (ma, ru, eu, la, le, cy and grresto).

To assess how good the PCoA approximates the original distances in **D**, these distances can be plotted versus the ones obtained by PCoA. Such a graph is called a Shepard plot, and ideally, the points should lie on a straight line. If that is indeed the case, the PCoA distances are identical to the original ones. The Shepard plot in Figure 15.1-B shows there is a considerable mismatch for the larger distances. Increasing the number of axes (e.g., three) seems to be an appropriate step. Indeed this gives better results, but the axes are not presented here.

Occasionally, the PCoA may produce negative eigenvalues and solutions exist to solve this problem (Legendre and Legendre 1998). In most occasions this problem will not affect the first few axes used for plotting.

The results from the PCoA in Figure 15.1 indicate that more axes are needed; alternatively the method discussed in the next section can be applied, as it is more capable in reducing a high-dimensional space into a small number of axes.

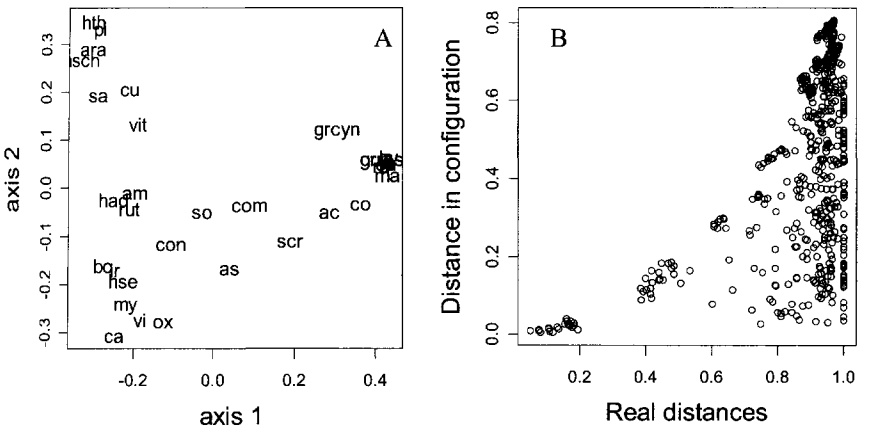

Figure 15.1. A: Results obtained by PCoA. Two axes were estimated. The first two eigenvalues are 2.68 and 0.97, which corresponds to 22% and 8% of the variation in the distance matrix, respectively. B: Shepard plot for the PCoA. The horizontal axis represents the original Jaccard distances between the families and the vertical axis contains the distances in the two-dimensional PCoA configuration. Ideally, the points should be on a straight line.

15.2 Non-metric multidimensional scaling

PCA, CA and PCoA are all methods that solve an eigenvalue equation. The advantage of PCoA above PCA is that any measure of association can be used; PCA is limited to the correlation and covariance coefficients. We will now discuss a method that has the same aim as PCoA, it can also use any measure of association, but it is better in preserving the high-dimensional structure with a few axes. It is called non-metric multidimensional scaling (NMDS). Its disadvantage is that it is not based on an eigenvalue solution but on numerical optimisation methods and for larger datasets the calculations tend to become time consuming, even on fast computers.

The aim of NMDS is to calculate a distance matrix **D** and visualise this matrix in a low (typically two- or three-) dimensional configuration. The difference between PCoA and NMDS is that in PCoA the distances in the configuration should match the original distances as closely as possible. In NMDS it is the order, or ranking, of the distances in **D** that we try to represent as closely as possible. As an example, Jaccard indices are given for four families in Table 15.1. The indices were converted to dissimilarities. PCoA uses this type of matrix to obtain the configuration such that Euclidian distances match the numbers in the table as closely as possible. In Table 15.2 we have converted the distances into ranks. As and Ac had the lowest dissimilarity (they were the most similar) and therefore have rank

1, etc. NMDS will produce a configuration that matches the data in Table 15.2 as closely as possible. Hence, points close to each other in the NMDS ordination diagram represent families that are more similar than others.

Table 15.1. Matrix with Jaccard dissimilarities among four families.

	Ac	Aam	Ara	As
Ac				
Am	0.949			
Ara	0.963	0.929		
As	0.642	0.952	0.962	

Table 15.2. Matrix with the dissimilarities from Table 15.1 transformed to ranks.

	Ac	am	ara	as
Ac				
Am	3			
Ara	6	2		
As	1	4	5	

Most books, including the text above, tend to say that 'NMDS is better in preserving relationships in a low dimensional space compared to PCoA'. In fact this is not a fair comment as we are comparing two different things: absolute distances and ranks. It is just that in most ecological studies, one is still content with the information that families A and B are more similar than C and D (NMDS), as to knowing that A and B are five times more similar than C and D (PCoA).
The NMDS ordination diagram for the Mexican plant species is given in Figure 15.2-A and the Shepard diagram in Figure 15.2-B. We used only two axes. The interpretation of the ordination diagram is simple; families close to each other are more similar than points far away from each other, but we do not know by how much. The only problem is then, what does similar mean? As we used the Jaccard index, it means that families close to each other in the graphical configuration coexist at the same sites. As the method is not based on an eigenvalue decomposition, the ordination diagram can be rotated and scaled. This does not affect its interpretation.

To give an impression of how NMDS works, the underlying mathematical algorithm is summarised next. Readers not interested in the underlying maths may skip this paragraph.

1. Choose a measure of association and calculate the distance matrix **D**.
2. Specify m, the number of axes.
3. Construct a starting configuration **E**. This can be done with PCoA.
4. Regress the configuration on **D**: $D_{ij} = \alpha + \beta E_{ij} + \varepsilon_{ij}$.
5. Measure the relationship between the m dimensional configuration and the real distances by fitting a non-parametric (monotonic) regression curve in

the Shepard diagram. A monotonic regression is constrained to increase. If a parametric regression line is used, we obtain PCoA.

6. The discrepancy from the fitted curve is called STRESS.
7. Using non-linear optimisation routines, obtain a new estimation of **E** and go to step 4 until convergence.

There are different ways to quantify STRESS in step 6, but all use quadratic sums of the original distances and those in the reduced space; see Legendre and Legendre (1998) for the exact formulations.

The NMDS algorithm is iterative, and for large datasets, different starting values might give different results. It may be an option to run the algorithm a couple of times with different starting values and see how the STRESS is changing.

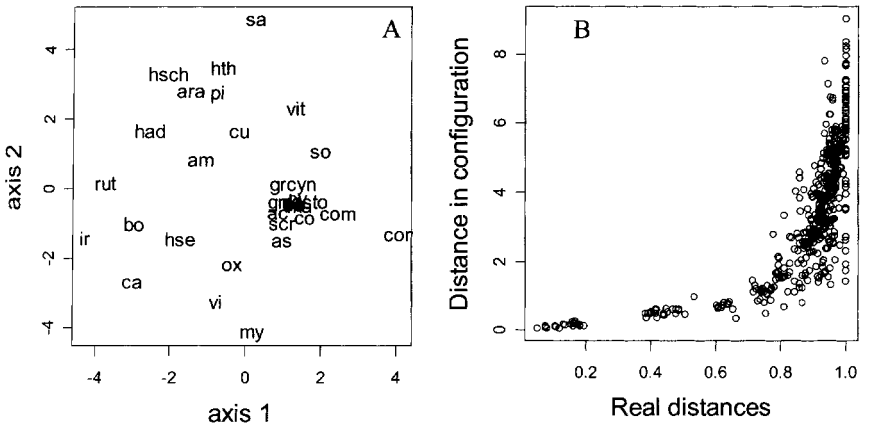

Figure 15.2. A: Results obtained by NMDS for the Mexican plant data. The STRESS is equal to 0.2336, and two axes were extracted. B: Shepard plot for the NMDS. The horizontal axis represents the Jaccard real distances between the families, and the vertical axis contains the distances obtained by the two NMDS configurations. The STRESS is 0.2336. If three axes are estimated, the STRESS is 0.1432.

Assessing the quality of the display

In PCA, CA and PCoA we have eigenvalues that can be used to assess the quality of the display. In NMDS we do not have eigenvalues and instead the STRESS is used to judge how good is the m-dimensional configuration. There are different ways to do this. One option is to calculate STRESS for different values of m (number of axes) and make a scree diagram, just as we did for the eigenvalues in PCA (Chapter 12). Along the x-axis we plot the number of axis m, and along the y-axis the STRESS. A clear change in stress (elbow effect) would indicate the optimal value of m. An alternative approach is to use the following rule of thumb (PRIMER manual):

- STRESS smaller than 0.05. The configuration is excellent and allows for a detailed inspection.
- STRESS between 0.05 and 0.1. Good configuration and no need to increase m.
- STRESS between 0.1 and 0.2. Be careful with the interpretation.
- STRESS between 0.2 and 0.3. Problems start, especially in the upper range of this interval.
- STRESS larger than 0.3. Poor presentation and consider increasing m.

Example

In Chapter 29, PCA is applied on fatty acid data measured in stranded dolphins in Scotland. There are 31 fatty acids (variables) and 89 dolphins (observations). Here, PCoA and NMDS are applied on the same data. In the first step of both methods, we calculate an appropriate distance matrix between the 89 dolphins. Both methods give a low-dimensional configuration of this matrix. Because the variation between the fatty acids differed (Chapter 29) we decided to normalise the fatty acid variables and use Euclidian distances between the dolphins. A PCoA gave exactly the same ordination diagram (not presented here) as the PCA (Chapter 29). Indeed, PCoA with the Euclidean distance function gives the same results as PCA. The NMDS ordination is given in Figure 15.3. The graph is similar (though not identical) to the PCoA and PCA ordination plots. The STRESS is 0.129, which means that the two-dimensional configuration is reasonable. Note the four observations in the upper left corner. An explanation of their separation is given in Chapter 29.

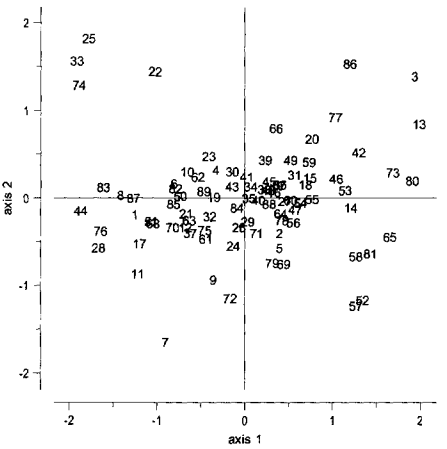

Figure 15.3. NMDS ordination diagram for the fatty acid dolphin data obtained with $k = 2$. The STRESS is 0.129. The numbers identify individual dolphins.