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 9 Offprint

Multidimensional Scaling

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

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

Fundación BBVA

CHAPTER 9

Multidimensional Scaling

In Chapter 7 a square matrix of distances or dissimilarities was visualized in the form of a dendrogram, trying to establish groups of points with relatively small within-group distances but relatively large between-group distances. Samples are not usually clustered naturally but are more often spread around continuously in the multidimensional space of the variables. Strongly correlated variables imply a certain amount of redundancy in the data, which means that less dimensions than the number of variables are required to describe the sample positions. Multidimensional scaling (MDS) is an alternative way of visualizing a distance or dissimilarity matrix, with the objective of representing the samples in a low-dimensional space, usually two- or three-dimensional, reproducing as closely as possible the inter-sample proximities (either distances or dissimilarities). The method is thus attempting to make a spatial map of the data to facilitate interpretation of the positions of the samples relative to one another. Since our intuitive understanding of a map is through the physical concept of a Euclidean distance, it will be an issue whether the sample proximities are Euclidean or not. Usually Euclidean-type distances will be mapped by so-called *metric* scaling methods, for example classical (metric) MDS, while non-Euclidean ones will be mapped by *nonmetric* methods.

Contents

A simple exact distance matrix

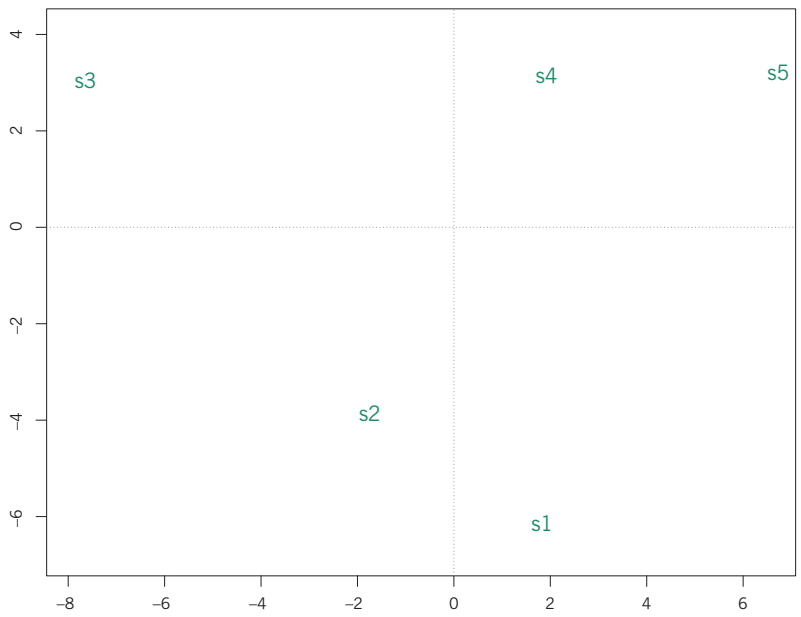
To introduce how MDS works in a simple way, consider the following matrix **D** of distances between five samples, numbered s1 to s5:

$$\mathbf{D} = \begin{array}{c} \\ \text{s1} \\ \text{s2} \\ \text{s3} \\ \text{s4} \\ \text{s5} \end{array} \begin{array}{c} \begin{array}{ccccc} \text{s1} & \text{s2} & \text{s3} & \text{s4} & \text{s5} \end{array} \\ \begin{bmatrix} 0 & 4.24 & 13.23 & 9.33 & 10.58 \\ 4.24 & 0 & 9.11 & 7.94 & 11.05 \\ 13.23 & 9.11 & 0 & 9.59 & 14.39 \\ 9.33 & 7.94 & 9.59 & 0 & 4.80 \\ 10.58 & 11.05 & 14.39 & 4.80 & 0 \end{bmatrix} \end{array} \tag{9.1}$$

First, we do not have an immediate way of telling whether this is a true distance matrix, but we can easily check the first two properties of a distance [see the properties in (5.1)], so at least we know that this is a dissimilarity matrix. A more laborious exercise would be to go through all triplets of samples (and there are 10 of these) to satisfy ourselves that the triangular inequality is indeed satisfied, for example, for samples s1, s2 and s3, with inter-sample values of 4.24; 13.23 and 9.11, the sum of any two is always greater than the third. In fact, we will see now that these five samples can be perfectly displayed in a two-dimensional map, reproducing exactly the above distance matrix.

Performing a so-called *classical MDS* (to be explained below) on the matrix **D**, the following map of the five samples in Exhibit 9.1 is obtained.

Exhibit 9.1: *Classical multidimensional scaling solution in two dimensions of the matrix* **D**, *using the* R *function* `cmdscale`



It can be verified that the distances between the samples are exactly those in the distance matrix **D**. Putting this another way, if we were given the coordinates of the five points in Exhibit 9.1 we could quite easily compute the distance matrix **D**, but MDS does the reverse, it starts with a distance matrix and constructs the map. The output of the MDS analysis is (1) the set of coordinates of the points i on each of the dimensions k of the solution, which we denote by f_{ik}, gathered in a matrix **F**; and (2) the parts of variance on each dimension, denoted by λ_k for the k-th dimension. For this example here are those results:

$$\mathbf{F} = \begin{array}{c} \\ s1 \\ s2 \\ s3 \\ s4 \\ s5 \end{array} \begin{array}{c} \begin{array}{cc} dim1 & dim2 \end{array} \\ \begin{bmatrix} 1.618 & -6.042 \\ -1.955 & -3.755 \\ -7.880 & 3.166 \\ 1.711 & 3.285 \\ 6.506 & 3.345 \end{bmatrix} \end{array} \qquad [\lambda_1 \quad \lambda_2] = \begin{array}{c} \begin{array}{cc} dim1 & dim2 \end{array} \\ [11.380 \quad 8.260] \end{array} \tag{9.2}$$

Computing the Euclidean distances between the rows of **F** will lead to exactly the same matrix of distances in **D**.

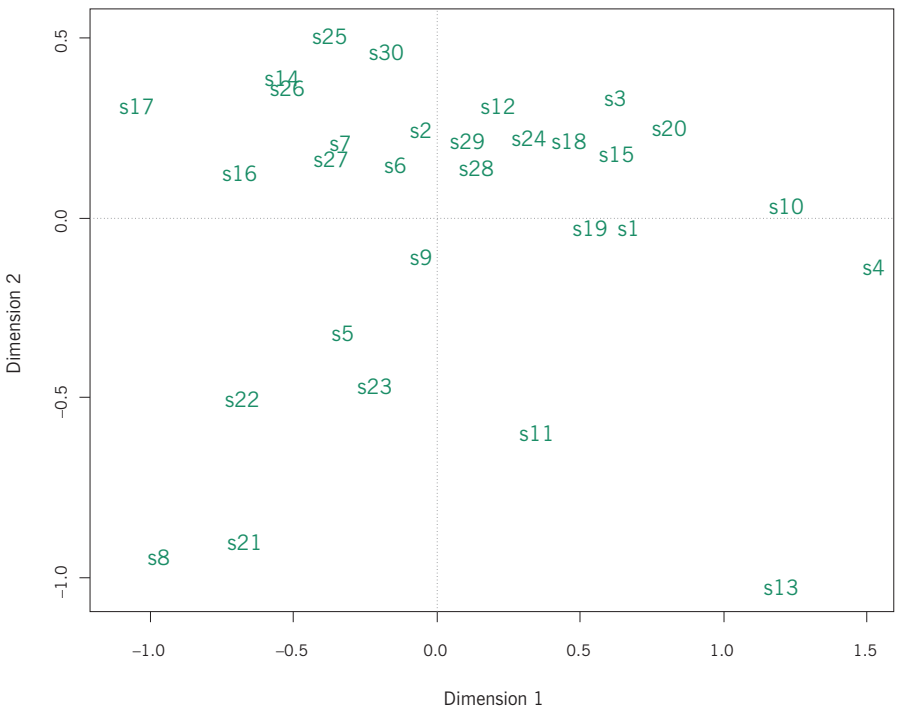
Classical MDS

The classical MDS procedure used to pass from the distance matrix **D** to the map in Exhibit 9.1 works as follows. For n points, the maximum dimensionality is one less, $n - 1$. For example, two points lie on a line (one-dimensional), three points lie in a plane (two-dimensional), but the three points could also lie on a line and be one-dimensional, and so on. In their $(n - 1)$–dimensional space, using an eigenvalue-eigenvector routine, we can identify the principal dimensions of the set of points, in descending order of importance. This order is determined by the eigenvalues, with the highest eigenvalue and associated eigenvector indicating dimension 1, the second dimension 2 and so on. In fact, the eigenvalues quantify the variance explained by each dimension. In the above example of five points, the dimensionality is at most 4, but it turns out that the third and fourth eigenvalues are zero, which means that the points are exactly two-dimensional. The total variance of the points, that is the quantification of their dispersion in the two-dimensional space, is the sum of the two eigenvalues, 11.38 + 8.26 = 19.64, with the first dimension accounting for 11.38/19.64 = 0.579, or 57.9%, of the total, and the second dimension the remaining 42.1%.

In practice: more complex proximity matrices

From now on we shall often call distance and dissimilarity matrices collectively as *proximity* matrices, and distinguish between them where necessary. In practice, you will never have a proximity matrix that is so simple as to be exactly two-dimensional. Let's take two examples of matrices we have encountered already, first the matrix of chi-square distances between the 30 samples in Chapter 4 – see Exhibit 4.7. Applying classical MDS to this matrix, the following eigenvalues are obtained: 12.37; 5.20; 3.83; 2.23, and all the remaining ones are zeros. This indicates that the distances are four-dimensional,[1] while in theory they could have any dimensionality up to 29. Exhibit 9.2 shows the MDS map of the chi-square distances between the 30 samples with respect to the first two dimensions. According to the eigenvalues, the total variance is 12.37 + 5.20 + 3.83 + 2.23 = 23.63, of which 12.37/23.37, or 52.4%, is accounted for by dimension 1 and 5.20/23.63, or 22.0%, is accounted for by dimension

Exhibit 9.2: *Classical multidimensional scaling solution in two dimensions of the matrix of chi-square distances of Exhibit 4.7. The percentages of variance on the horizontal and vertical axes are 52.4% and 22.0% respectively*



[1] The chi-square distances were computed on only five abundance values per sample, and Euclidean-type distances in this situation would usually be of dimensionality 5. So then why is the dimensionality equal to 4? This is because the chi-square distances are based on the relative abundances and since the relative abundances for each sample always add up to fixed value, 1, only four values are *free* and the fifth is one minus the sum of the other four. This is one of the properties inherent in correspondence analysis, treated in Chapter 13.

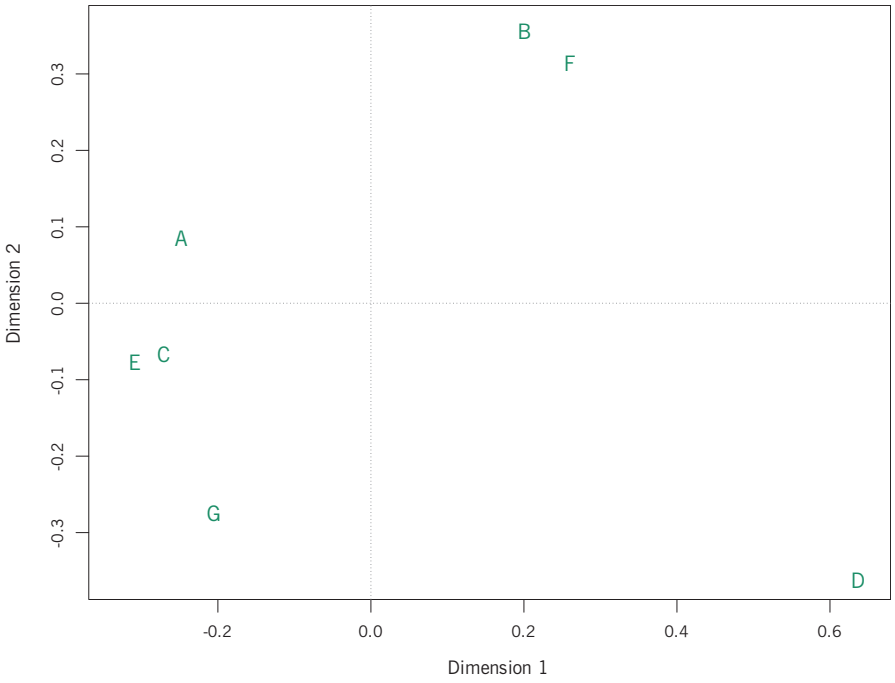
2, totalling 74.4% for the map as a whole. It can be seen in Exhibit 9.2 that the first dimension shows more variance than the second.

Exhibit 9.2 is an approximate display of the matrix of chi-square distances, but it is the best one can do in a two-dimensional display according to the optimization criterion inherent in classical scaling, namely to maximize explained distance variance. Each dimension of the display can be thought of as a variable trying to explain the variance in the distance matrix, like independent variables in a regression. These dimensions are uncorrelated, so their separate percentages of variance can be simply added to give the cumulative percentage of variance explained for by the set of dimensions. Hence, the percentage 74.4% of variance explained can be interpreted just like an R^2 in regression analysis.

Now as a second example we apply the same procedure to the matrix of Jaccard proximity values in Exhibit 7.1 between seven samples, which in theory has a dimensionality no higher than 6. The eigenvalues that emerge are: 0.786; 0.452; 0.148; 0.037; 0.000; –0.002; –0.030. Indeed, six of them are nonzero, but there are two negative eigenvalues, indicating that this proximity matrix is not Euclidean (hence the inclusion of the Jaccard index in Chapter 5 on non-Euclidean dissimilarity functions). So in this case it is impossible to represent these proximities in any Euclidean space, and the negative eigenvalues give an idea of how much of the variance cannot be displayed. The part that is Euclidean is the sum of the positive eigenvalues: 0.786 + 0.452 + 0.148 + 0.037 = 1.423, while the part that cannot be displayed is the sum of the absolute values of the negative eigenvalues: 0.002 + 0.030 = 0.032, which is quite small compared to 1.423. Exhibit 9.3 shows the classical MDS display in two dimensions of the samples.

Before interpreting this display, how do we quantify the variance explained in this case? There are two ways to do it, depending on whether the non-Euclidean part is included in the total variance or not. The first two eigenvalues, 0.786 and 0.452, can be expressed relative to the sum of the positive eigenvalues, 1.423, or the sum of the absolute values of all the eigenvalues, 1.423 + 0.032 = 1.455. In the former case the percentages would be 56.5% and 32.5%, totalling 89.0%, while in the latter case they would be slightly lower, 55.3% and 31.8%, totalling 87.1%. The non-Euclidean part is quite small in this case, hence the small differences between the two options. An acceptable way of reporting the results would be to say that 2.1% (i.e., 0.032/1.455 = 0.021) of the total variance, is non-Euclidean, and that, of the Euclidean part of the variance, 89.0% (i.e., (0.786 + 0.452)/1.423 = 0.890) is displayed in Exhibit 9.3.

Exhibit 9.3:
Classical multidimensional scaling solution in two dimensions of the matrix of Jaccard dissimilarities of Exhibit 7.1.The percentages of variance on the horizontal and vertical axes are 56.5% and 32.5% respectively (expressed relative to the four-dimensional Euclidean part of the variance)



Comparing the MDS map in Exhibit 9.3 with the dendrogram in Exhibit 7.8, the separation of the groups {B,F}, {A,C,E,G} and {D} is verified here. The only noticeable difference between the two results is that in the clustering A was joined with E and C with G before they all joined together in a cluster of four, whereas in Exhibit 9.3 C and E look like they should have been clustered first. This is because C and E are, in fact, quite far apart in the third dimension, with coordinates of −0.24 and 0.18 respectively. This large difference in the third dimension can not be seen in Exhibit 9.3 but the original data in Exhibit 7.1 show C indeed further from E compared to the dissimilarity between A and E or C and G.

Nonmetric MDS

The type of MDS described above is included in the family of *metric* MDS methods, since the observed proximities are accepted as quantitative measures of difference between the samples, and the error in the solution is quantified by calculating actual differences between the observed values and those that are displayed in the solution map. Especially when it is known that the proximity matrix is non-Euclidean, an alternative form of MDS may be used, called *nonmetric* MDS, which has a more relaxed way of measuring the quality of the solution. In nonmetric MDS we are not interested in reproducing the proximities themselves, but rather their ordering, that is if we sorted all the observed proximities from

smallest to largest, and we did the same for all the interpoint distances in the solution, then a perfect solution would be if the two orderings were identical. In the matrix of Jaccard dissimilarities of Exhibit 7.1 there are ½ × 7 × 6 = 21 values, ordered in the first set of columns of Exhibit 9.4 (notice the tied ranks). In the second set of columns, the distances between points in Exhibit 9.3 are ordered (there are usually no tied ranks in such fitted distances). Corresponding pairs of samples are linked: if all these links were horizontal, then the distances would be perfectly in order.

The objective of nonmetric MDS would be to get a better correspondence in the orderings of the points. The result for this small data set is surprising, because of the inherent clustering of the samples, shown in Exhibit 9.5.

A comparison of the observed and fitted distances in Exhibits 9.3 and 9.5 clarifies what has happened – see the two plots in Exhibit 9.6. The objective

(a) Observed Jaccard indices

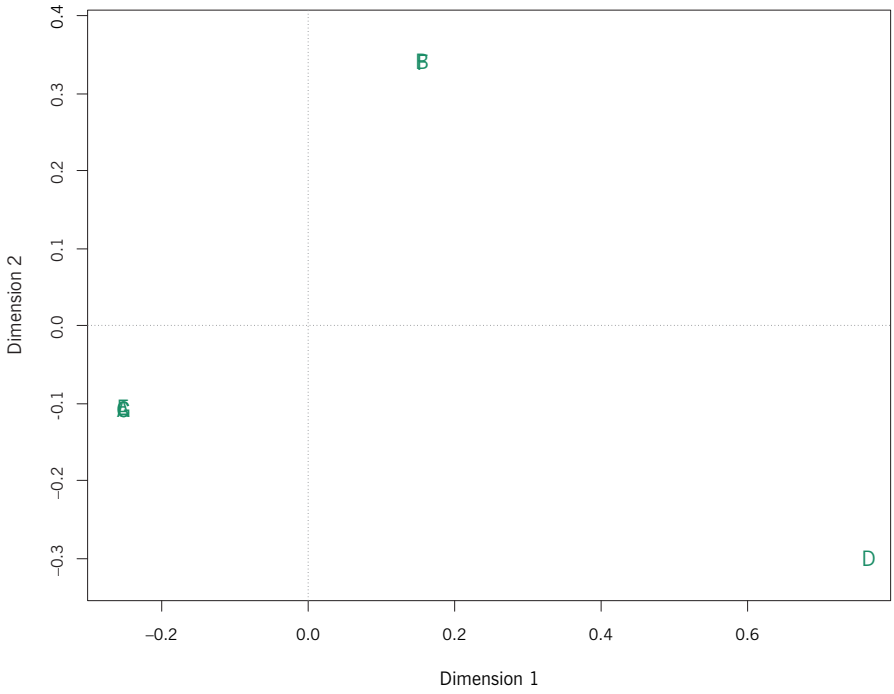
Rank	Jaccard	Pair
1	0.200	(B,F)
2	0.250	(A,E)
3	0.333	(C,G)
4.5	0.375	(E,G)
4.5	0.375	(A,G)
6.5	0.429	(A,C)
6.5	0.429	(C,E)
8	0.500	(A,B)
9	0.625	(A,F)
10.5	0.667	(B,E)
10.5	0.667	(C,F)
12	0.714	(B,C)
13	0.750	(F,G)
14.5	0.778	(B,G)
14.5	0.778	(E,F)
16	0.800	(D,F)
17	0.833	(B,D)
18	0.857	(D,G)
20	1.000	(A,D)
20	1.000	(C,D)
20	1.000	(D,E)

(b) Distances in metric MDS

Rank	Distance	Pair
1	0.040	(C,E)
2	0.074	(B,F)
3	0.154	(A,C)
4	0.175	(A,E)
5	0.219	(C,G)
6	0.223	(E,G)
7	0.365	(A,G)
8	0.524	(A,B)
9	0.559	(A,F)
10	0.633	(B,C)
11	0.655	(C,F)
12	0.669	(B,E)
13	0.692	(E,F)
14	0.752	(B,G)
15	0.754	(F,G)
16	0.775	(D,F)
17	0.842	(B,D)
18	0.848	(D,G)
19	0.955	(C,D)
20	0.988	(D,E)
21	0.993	(A,D)

Exhibit 9.4:
Ordering of the original Jaccard dissimilarities, from lowest to highest, and ordering of the interpoint distances in the metric MDS of Exhibit 9.3

Exhibit 9.5: *Nonmetric MDS of the Jaccard dissimilarities of Exhibit 7.1. The samples agglomerate into three groups, identical to the clustering in Exhibit 7.8*



of the nonmetric MDS is to find a configuration of the points such that the interpoint distances are close to the ordering of the original distances. In each plot a monotonically increasing function is shown (i.e., a function that never decreases) which best fits the interpoint distances in the map – this function is obtained by a procedure called *monotonic regression*. The error is quantified by the sum of squared deviations between the fitted distances (green circles) and the monotonic regression line. If the sequence of fitted points were always ascending then the monotonic regression line would simply join the points and the error would be zero. Clearly, the upper plot of Exhibit 9.6 shows that there are relatively large deviations of the points from the best-fitting monotonic regression line compared to the near zero deviations in the lower plot. In the lower plot it looks like a perfect fit, but the enlargement of the first seven points shows that there are indeed very small deviations). To explain what has happened there, notice that the interpoint distances between B and F and among all pairs of points in the set {A,C,E,G} are the smallest in the original dissimilarity matrix. Hence, the nonmetric approach puts them all at near-zero distance from one another, and all their values can thereby be reduced to near zero. This maintains their ordering, with very little error from a monotonically increasing relationship, as shown in the enlargement.

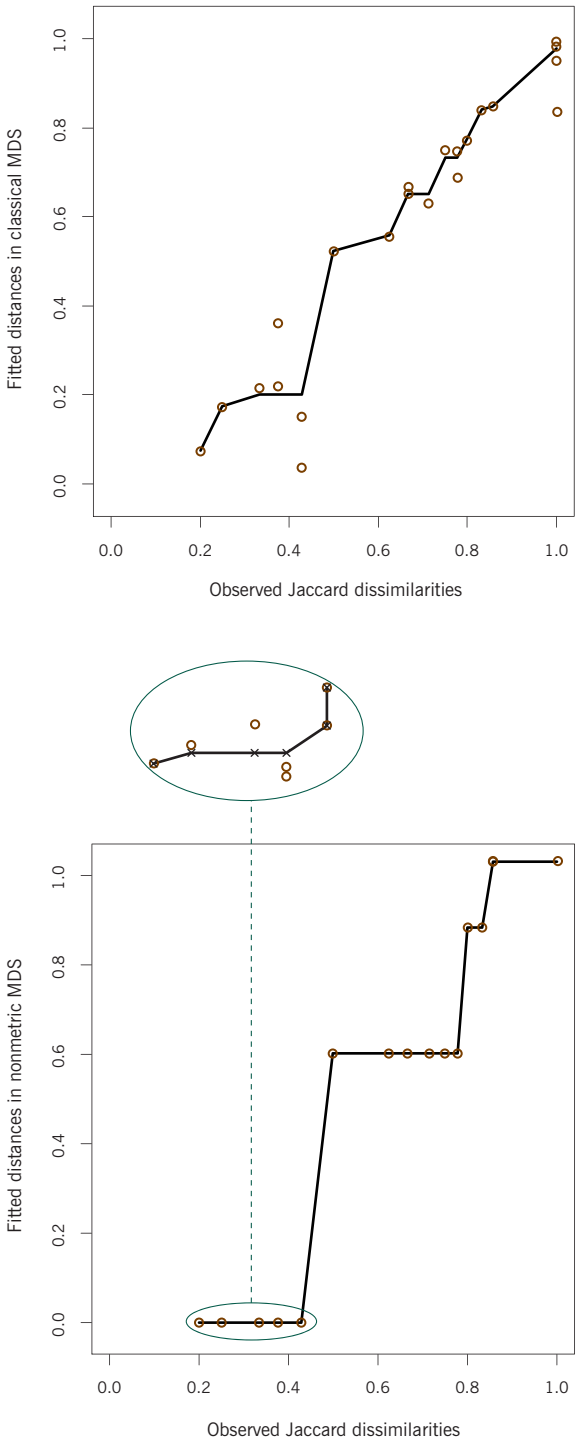


Exhibit 9.6:
The horizontal axis shows the observed dissimilarities from Exhibit 7.1, and the vertical axes show the fitted interpoint distances from Exhibits 9.3 and 9.5 respectively. In both plots the closest fitting monotonically increasing function is shown. The vertical scale of the first seven points in the nonmetric MDS (see lower plot) is expanded considerably to show the small lack of fit for those points

The actual measure of error in a nonmetric MDS is a normalized version of the sum of squared errors, called *stress*. The most popular one is known as *Kruskal's stress formula 1:*

$$\text{stress} = \left(\frac{\sum_{i<j} (\hat{d}_{ij} - d_{ij})^2}{\sum_{i<j} d_{ij}^2} \right)^{1/2} \qquad (9.1)$$

where d_{ij} is the distance between points i and j in the MDS map and $\hat{d}_{ij}$ is the corresponding value on the monotonic regression line (hence the values $|\hat{d}_{ij} - d_{ij}|$ are the vertical discrepancies between points and the line in Exhibit 9.6). This stress measure is often multiplied by 100 and considered as a percentage error: using this convention the metric MDS in Exhibit 9.3 would have a stress of 8.3% while the nonmetric MDS in Exhibit 9.5 would have a stress near zero equal to 0.008% (very low stress values are typical for small data sets like this one – later we will show the result for a larger data set).

Adding variables to MDS maps

The MDS maps in Exhibits 9.3 and 9.5 show the samples only, but the Jaccard dissimilarity matrix was constructed on a samples-by-species data matrix (Exhibit 5.6). Since this is simple presence-absence data, the species can be shown in the MDS maps at positions near the samples that contain them. Usually, they would be situated at the average spatial position of the corresponding samples, shown in Exhibit 9.7 for the two MDS maps. For example, species sp6 is present in samples D and G, so is at an average position halfway between them, while species sp4 is present in samples B, D and F, and is thus positioned at an average position of these three samples.

MDS of Bray-Curtis dissimilarities

In Chapter 5 we computed the Bray-Curtis dissimilarities between 30 samples, s1 to s30, based on the abundances of five species, *a* to *e* – see Exhibit 5.2, where we also pointed out that this measure violated the triangle inequality and was therefore not a metric. For this reason, nonmetric MDS is usually used to map Bray-Curtis indices, but first let us see how metric MDS would handle the display of Exhibit 5.2. The maximum dimensionality of this set of 30 samples is 29, and, as expected, we obtain several negative eigenvalues in the classical MDS: in fact, 14 eigenvalues are positive, with a sum of 57,729, while 15 are negative, with a sum of absolute values equal to 8,176 (Exhibit 9.8). This latter amount quantifies how much variance is impossible to display in a Euclidean space. The first two eigenvalues are 19,102 and 14,825, so the variance explained by the two-dimensional solution, relative to the Euclidean part, is (19,102 + 14,825)/57,729, or 58.8%, that is an error of 41.2%. Computing the stress on this solution, however, gives a value of 16.3%, showing again that the stress criterion always appears more optimistic than the explained variance one.

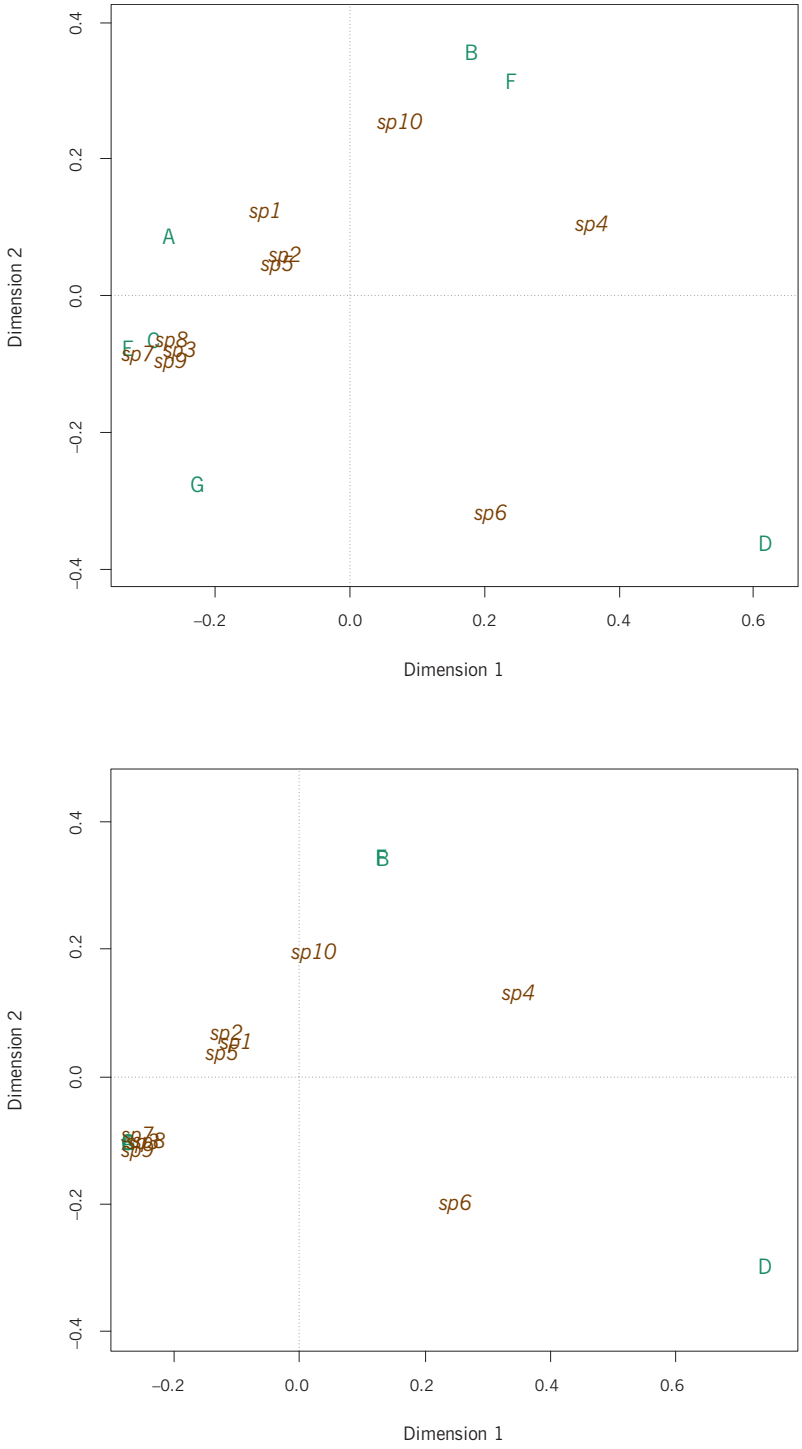
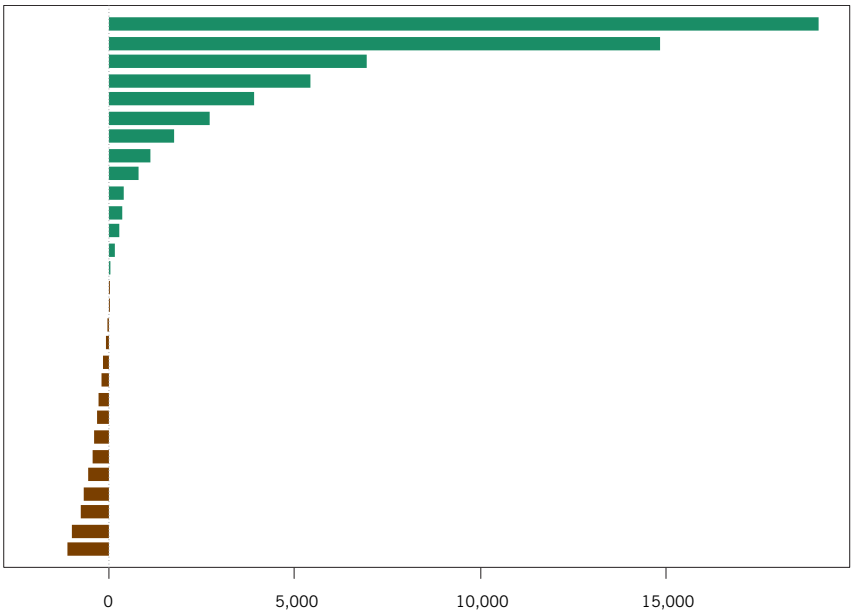


Exhibit 9.7:
The MDS maps of Exhibits 9.3 and 9.5 with the species added at the average positions of the samples that contain them

Exhibit 9.8:
The eigenvalues in the classical MDS of the Bray-Curtis dissimilarity indices of Exhibit 5.2, showing positive eigenvalues in green and negative ones in brown



Performing nonmetric MDS on the same data gives a stress value of 13.5%, which is not a big improvement on the 16.3%, suggesting that the two resulting maps will not be as different as we found for the smaller data set of Jaccard indices. This is indeed the case, as shown by the quite similar maps in Exhibit 9.9.

In our experience, when there is a large number of samples (and by "large" we mean, as most statisticians do, 30 or more, as in this example), the metric and nonmetric approaches generally agree in their solutions. Where they disagree is in the quantification of the success of their results, with the stress measure always giving a more optimistic value because it does not measure the recovery of the proximities themselves, but their ordering in the map.

Adding count variables to MDS maps

The maps in Exhibit 9.9 emanate originally from abundance data on five species, so the question now is how to include these species on the map. We shall consider alternative ways of doing this in future chapters, but for the moment let us use the same approach as in Exhibit 9.7 when the species were positioned at the averages of the samples that contained them. The difference here is that we have abundance counts for the species across the samples, so what we can do is to position each species at their weighted average across the samples. For example, species *a* has abundances of 0, 26, 0, 0, 13, etc., and a total abundance of 0 + 26 + 0 + 0 + 13 + ... = 404, so the position of *a* is at a weighted average position of the 30 species, with weights 26/404 = 0.064 on sample s2, 13/404 = 0.032 on sample s5, and

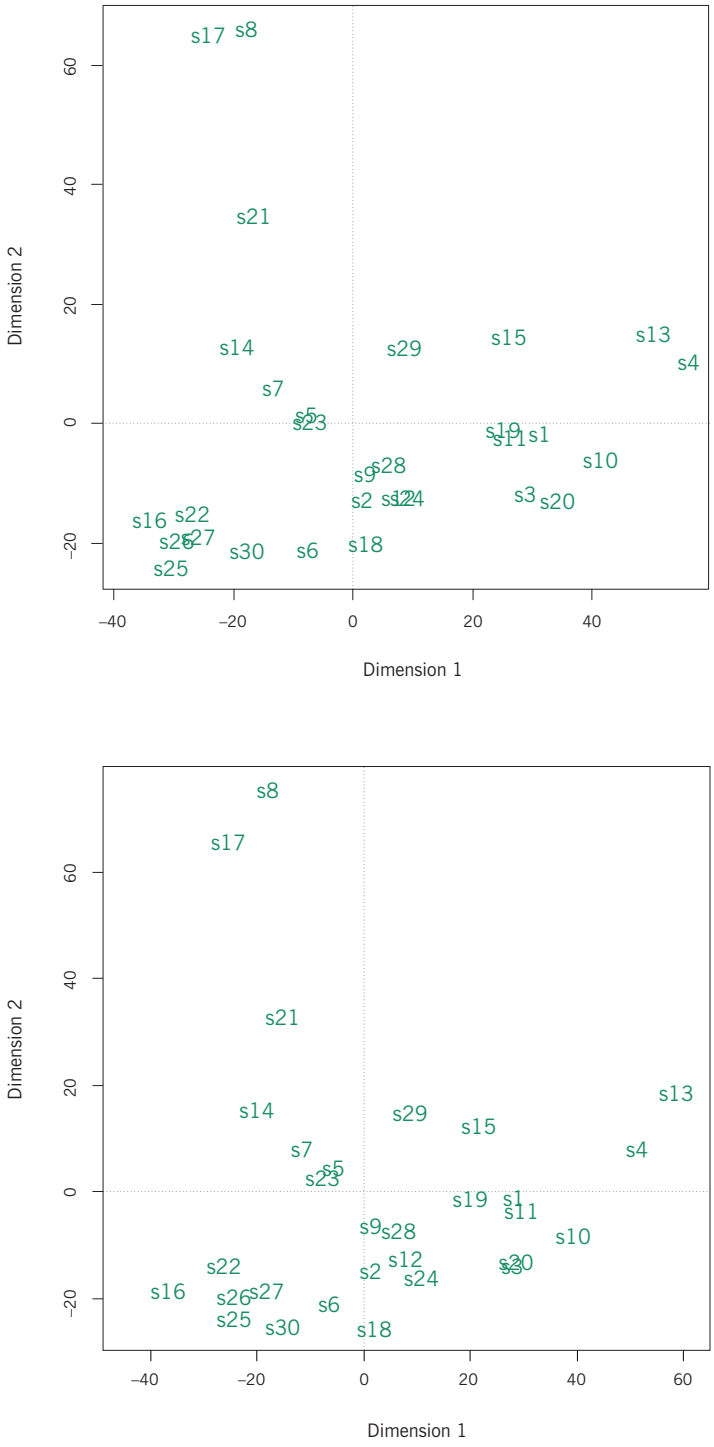
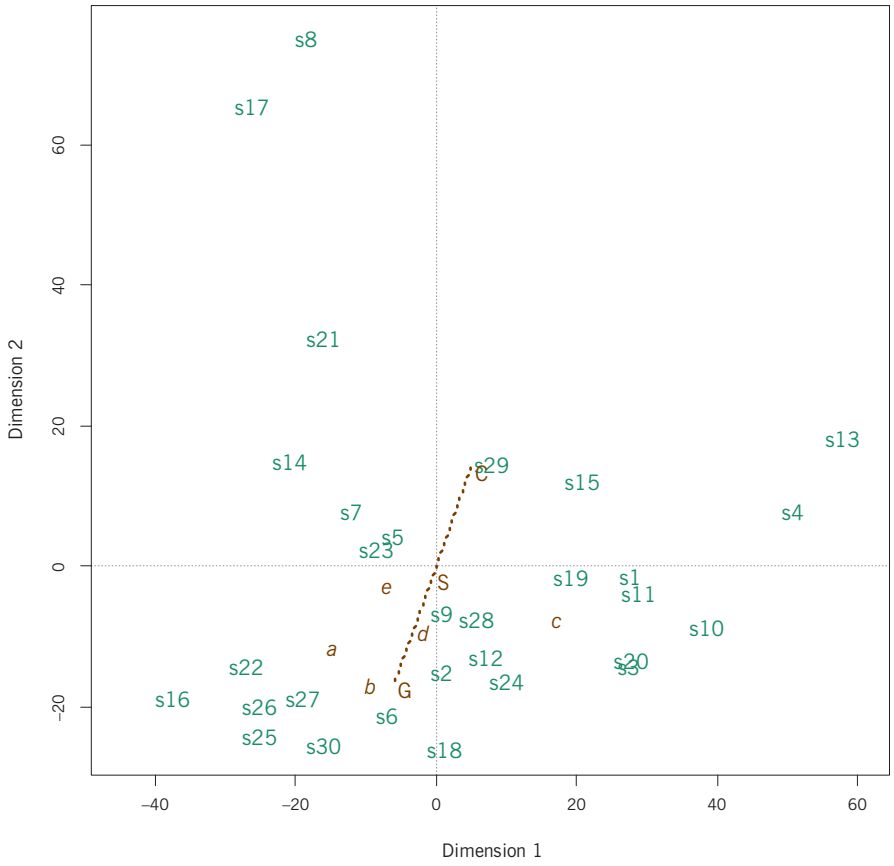


Exhibit 9.9:
Classical MDS map (upper) and nonmetric MDS map (lower) of the Bray-Curtis dissimilarities of Exhibit 5.2

Exhibit 9.10:
Nonmetric MDS solution (right hand map in Exhibit 9.9) with species a to e added by weighted averaging of sample points, and sediment types C, S *and* G *by averaging*



so on. Exhibit 9.10 shows the species positions on the nonmetric MDS solution, showing, for example, that species *a* and *b* are relatively more abundant in the samples at lower left, while *c* is more associated with samples on the right. Similarly, even though the ordinal sediment types C (clay), S (sand) and G (gravel) have not been used in the mapping, they can be depicted at the averages of the subsets of samples corresponding to them. The samples thus appear to follow a trend from top right (more clay) to bottom left (more gravel).

SUMMARY: Multidimensional scaling

1. Multidimensional scaling (MDS) is a method that attempts to make a spatial map of a matrix of proximities, either distances or dissimilarities defined between sample units, so that the interpoint distances in the map come as close as possible to the given proximities according to the chosen fit criterion.

2. The fit criterion in metric MDS involves approximating the actual proximity values by the mapped distances, for example by least-squares.

3. Classical MDS is a particular form of metric MDS that relies on the eigenvalue-eigenvector decomposition of a square matrix. The eigenvalues give convenient measures of variance explained on each axis, and the dimensions of the solution are uncorrelated.

4. Nonmetric MDS has a more relaxed fit criterion in that it strives to match only the ordering of the proximities to the ordering of the mapped distances.

5. The error in classical MDS is quantified by the percentage of unexplained variance, while in nonmetric MDS the error is quantified by the stress.

6. The stress measure always gives a more optimistic result, because of the relaxation of approximating the proximity values in the map in favour of their rank ordering.

7. In most cases, however, when the size of the proximity data matrix is quite large, say for at least 30 sample units, the results of the two approaches will be essentially the same.

8. When the proximities are of a Euclidean type, it will be more useful to use the metric scaling approach because of the connection with methods such as principal component analysis (Chapter 12) and correspondence analysis (Chapter 13). There would be little advantage, for example, in applying nonmetric scaling to a matrix of chi-square distances.

9. When the proximities are non-Euclidean, the nonmetric approach avoids the dilemma that the triangle inequality is violated by concentrating on ordering of proximities rather than their actual values.

List of Exhibits