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 13 Offprint

Correspondence Analysis

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

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

Fundación BBVA

CHAPTER 13

Correspondence Analysis

Correspondence analysis is one of the methods of choice for constructing ordinations of multivariate ecological data. Ecological data are often collected as counts, for example abundances, or other positive amounts such as biomasses, on a set of species at different sampling sites. Correspondence analysis is similar to PCA, but applies to data such as these rather than interval-scale data. It analyses differences between *relative* values: for example, if at a sampling site there is an overall abundance count of 320 individuals across all the species, and if a particular species is counted to be 55, then what is relevant for the analysis is the relative value of the abundance, 55/320 (17%). Furthermore, to measure inter-sample difference in such relative abundances across the species, the chi-square distance described in Chapter 4 is used to normalize species with different overall abundances. Classical MDS is applied to the chi-square distances to obtain an ordination of the sample points, with one important difference compared to PCA: each sample point is weighted proportionally to its total abundance (e.g., the value 320 mentioned above as an example), so that samples with higher overall abundance are weighted proportionally higher. These sample weights are also used in regressing the species on the dimensions to obtain a biplot. Finally, correspondence analysis has the special property that the analysis can be equivalently defined, and thought of, as the analysis of the rows (e.g., samples) or the analysis of the columns (e.g., species).

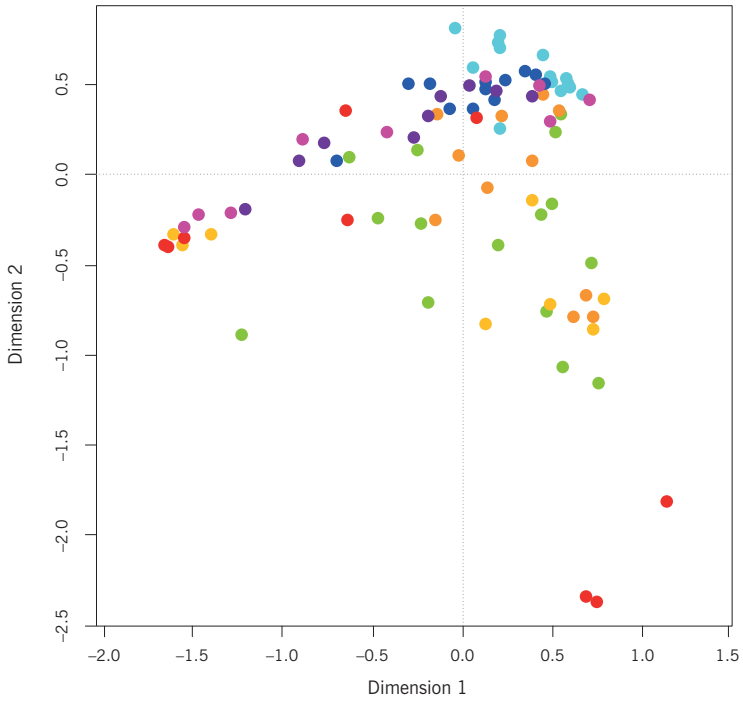
Contents

Weighted MDS of chi-square distances

In Chapter 11 we made a two-step analysis of the "Barents fish" abundance data: first, the Bray-Curtis dissimilarities between sampling sites were computed and a nonmetric MDS performed, and second, the species counts were regressed on the ordination dimensions using Poisson regression to obtain a biplot of the samples and species. These regressions were optimal conditional on the ordination obtained in the first step, so the question to consider now is what the ordination should be in two dimensions, say, in order that the regressions are the best that one can get using the two ordination axes as predictors. Like PCA, correspondence analysis, abbreviated as CA from now on, is going to be doubly optimal: the display of the sample points will be optimal and the biplot of the samples and species will be optimal in that the species regressions will explain maximum variance. One major difference in the CA approach is that it measures distance between the profiles of the abundances (i.e., vectors of relative abundances), described in Chapter 4 – see Exhibit 4.6 and the surrounding description. Then it uses the chi-square distance function between the profiles – see Exhibit 4.7 and its surrounding description. Furthermore, the sense of the optimality is by *weighted* least-squares in both the MDS of the sample profiles and in the regressions of species on the ordination axes – the sample weights are proportional to the abundance totals at the different sampling points. For example, the abundance totals at the 89 sites (see Exhibit 11.2) are 845, 1,740, 1,763, 767,..., 232, 36, with a grand total of 63,896. The weights, which are positive and add up to 1, will be 845/ 63,896 = 0.0132, 1,740/63,896 = 0.0272, and so on, until 232/63,896 = 0.0036 and 36/63,896 = 0.0006. Thus sites with higher abundances will be weighted more than those with lower abundances: for example, the profile of the second site will get a weight of 0.0272 (2.72%) whereas the last site, where overall abundance was low, will get a weight of 0.0006 (0.06%).

In the previous description of MDS methods there was no question of weighting the points, in other words all were weighted equally. It is a fairly simple adaptation of the methodology to accommodate different weights, which means that points with higher weight will tend to be better displayed than points with lower weight. This reweighting can make a big difference to the final MDS solution, as illustrated for this particular data set in Exhibit 13.1. In the unweighted MDS there is a curve of points from the left across to the top and then down to the three red points. This curve is essentially reproduced on the right hand side of the weighted MDS, following the vertical axis from top to bottom, but two samples have separated out on the left. These latter samples have high abundances, and so have high weights and become much more prominent in the weighted analysis. The ordination map in Exhibit 13.1(b) is based on the CA solution.

(a)



(b)

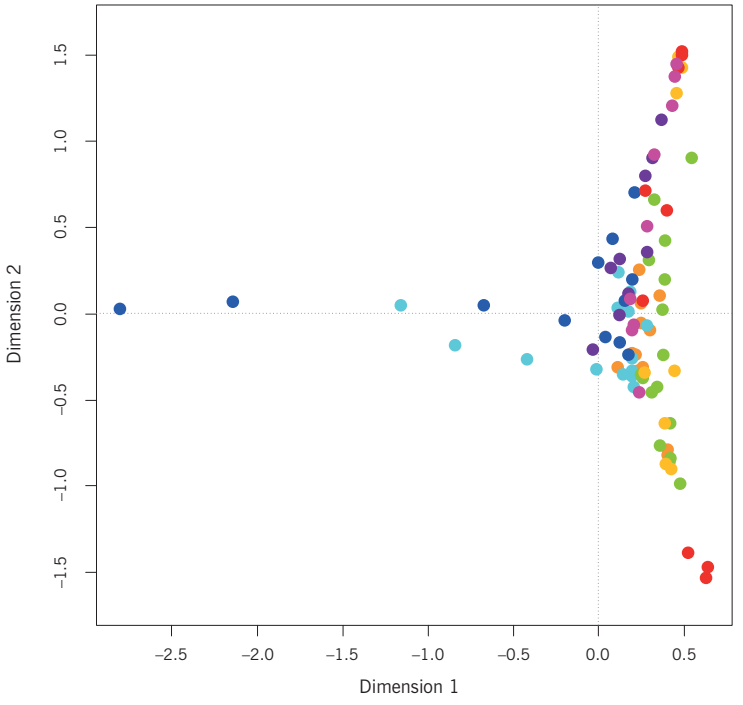
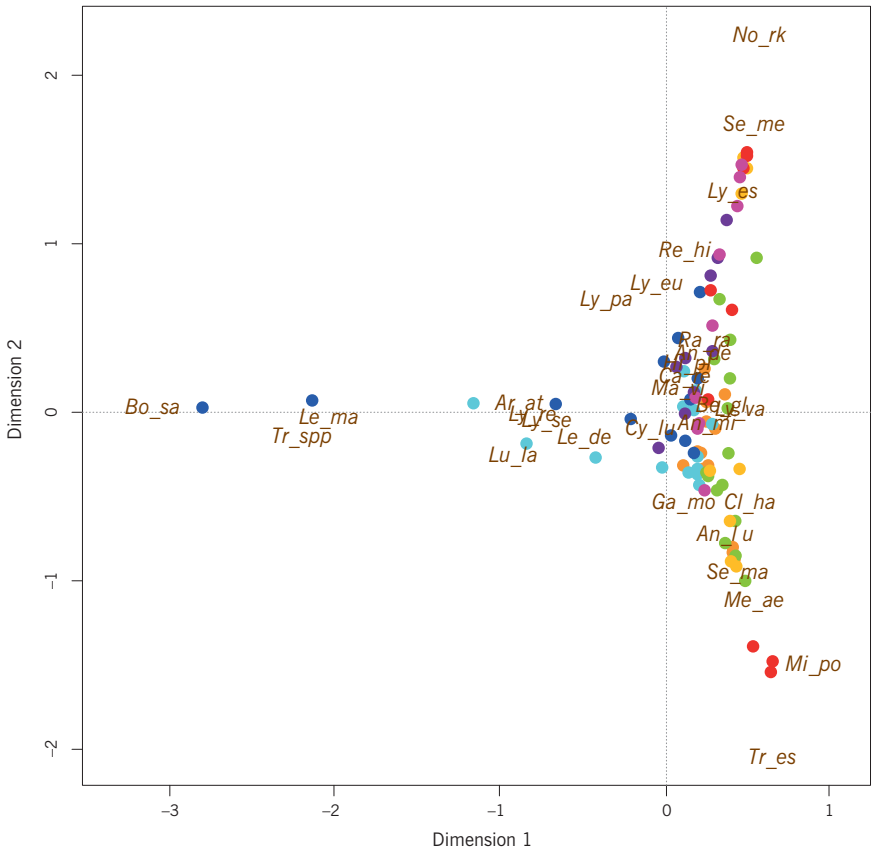


Exhibit 13.1:
Unweighted MDS (a) and weighted MDS (b) of the chi-square distances between sampling sites, for the "Barents fish" data. Colour coding as in Chapter 11

Display of unit profiles

CA has some special features that are not present in PCA, chiefly because the displayed profiles consist of nonnegative values that add up to 1. These are also called *compositional data*, the proportions of the species in each sample. One of the classic ways of displaying the species in CA is to show where the so-called *unit profiles* are in the ordination space: these are vectors of zeros except for a 1 in the position corresponding to a species, as if there were a sample with only that species observed in it. The unit profiles for the species are shown as supplementary points in Exhibit 13.2. Thus the unit point of the species *Bo_sa* (*Boreogadus saida,* polar cod) is on the extreme left, and species points *Tr_spp* (*Triglops* species) and *Le_ma* (*Leptoclinus maculatus,* spotted snake blenny) are also separate from the others on the left hand side. It turns out that the sample on the extreme left has very high relative abundances of these three species, and this explains its outlying position in the direction of these species. These high distances are not apparent in Exhibit 13.1(a) because all the points are weighted equally, whereas CA gives prominence to the samples with higher weight, and

Exhibit 13.2: *Row-principal CA biplot (asymmetric map) of "Barents fish" data. The sample profiles are shown as well as unit profiles for the species. There is a barycentric (weighted average) relationship between the samples and species points. Explained variance is 47.4%*



Bo_sa is a species with high overall abundance. The CA solution in Exhibit 13.2 is sometimes called an *asymmetric map:* the sample (row) points are in principal coordinates and it turns out that the species (column) points are in standard coordinates in the CA sense. To make this more precise, each species also has a weight in CA, its total abundance relative to the grand total. In Exhibit 11.2 the last species *Tr_spp*, for example, has an abundance of 653, which gives it a weight of 653/63,896 (1.02%). With these weights the species points in Exhibit 13.2, representing unit profiles, have weighted sum of squares equal to 1 on each dimension, and thus have coordinates referred to as standard coordinates.

Barycentric (weighted average) relationship

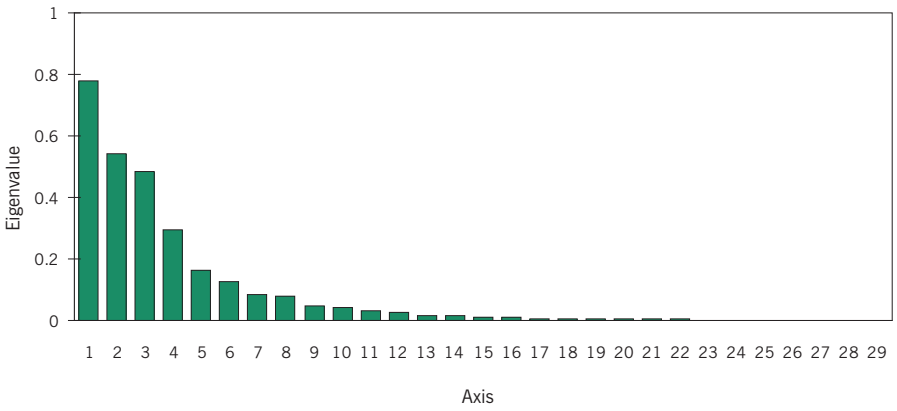
Before continuing let us start to call the sample and species weights *masses*, which is the preferred term in CA. This also distinguishes these masses from other sets of weights which we discuss now. For example, the masses of the 30 species in the "Barents fish" data are the relative abundances of the species in the whole data set. So the masses reflect the expected, or average, relative abundances in a sample if there were no differences in species distribution across the study region.

The joint display of samples and species in Exhibit 13.2 has an additional property that is particular to CA and is, in fact, one of the reasons for its relevance in ecology. Each sample point, originally the profile of the sample across the species, is at the weighted average of the species points, where weights are defined here as the elements of the profile. Let us take the sample on the extreme left of Exhibit 13.2 as an example. This sample has a total abundance of 4,399 and its profile across the 30 species consists of 19 zeros and 11 positive values, of which a few are extremely high compared to the species masses. For example, 82.9% is in the species *Bo_sa* (3,647 out of 4,399), whereas the relative abundance (i.e., mass) of *Bo_sa* in the whole data set is only 8.3% (5,297 out of 63,896). The sample is situated at the weighted average of the species points, and 82.9% of its weight is on *Bo_sa*, hence its position close to it. It has also much higher than average relative abundances of *Tr_spp* and *Le_ma*. For the same reason, the three sample points at bottom right must have high values in their profiles on the species *Tr_es* and *Mi_po* in order to be situated so close to them in that direction. Weighted averages are also called *barycentres* and this relationship between sample and species points in this version of the CA solution is called the *barycentric relationship*.

Dimensionality of CA solution

CA also leads to an eigenvalue measure of the part of variance on each dimension, as in PCA, and these eigenvalues can be viewed in a scree plot, shown in Exhibit 13.3. Here we introduce some terminology particular to CA: the total variance is called the total *inertia* of the data set, and is equal to 2.781 in this case. The eigenvalues, or *principal inertias*, decompose this total along the prin-

Exhibit 13.3:
Scree plot of eigenvalues in the CA of the "Barents fish" data



cipal axes. Notice that there are only 29 eigenvalues – the dimensionality of the full space is not 30, the number of species, but one less because the profile matrix analysed has constant row sums of 1. To decide on how many dimensions are worth interpreting we proceed as in PCA: it looks like there may be at most four dimensions distinguishing themselves from the others that tend to fall off in a pattern typical of random data. Later in Chapter 17 we will show more formally by a permutation test that in fact there are only three highly significant dimensions. So we should be looking at the third dimension as well. This poses a technological challenge, but it is now fairly easy to observe three-dimensional displays. In Exhibit 13.4 is a snapshot of the three-dimensional view of the points, and if you click on the image in the electronic version of this book it will revolve around the vertical axis.

Contribution biplots

The caption of Exhibit 13.2 refers to the display as a row-principal biplot, but this is not exactly the same as the regression biplots discussed before. In Chapters 11 and 12 the standardized variables were regressed onto the axes using ordinary least squares. Here there are two differences: firstly, the fact that the chi-square distances between profiles are being displayed, and secondly, the fact that each sample is weighted differently according to its corresponding mass. Thus it should be the columns of the standardized profile matrix that are regressed on the axes, standardized by centring with respect to the average profile (in this case, the set of species masses) and dividing columns by the square root of the corresponding masses, i.e. the standardization inherent in the chi-square distance. This gives another version of the biplot which we call the *contribution biplot*, shown in Exhibit 13.5 – just the species vectors are shown, the sample points are identical to those of Exhibit 13.2. With this scaling the species that are the most outlying on the axes are the ones contributing mostly to the CA solution, and thus the important ones for interpretation.

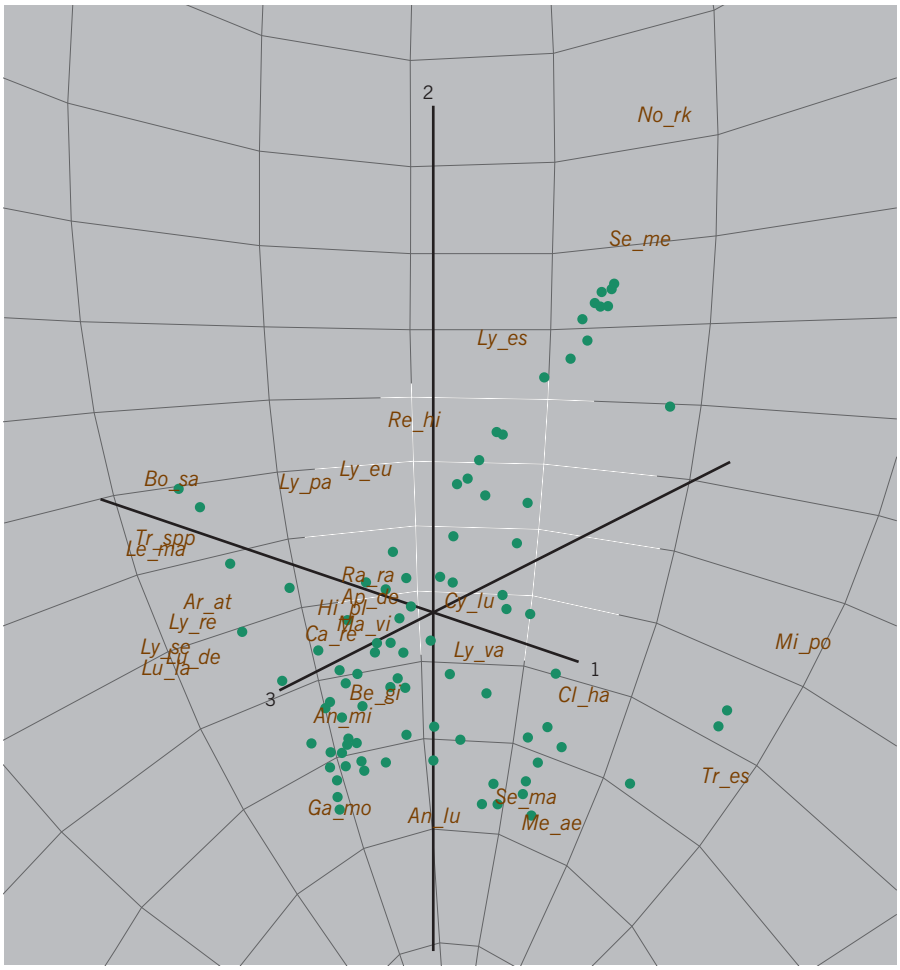


Exhibit 13.4:
Three-dimensional view of the samples and species, row principal biplot scaling. For readers of the electronic version: To see the rotation of these points around the vertical (second) axis, click on the display

Notice first the technical difference between the scalings of the species in Exhibits 13.2 and 13.5. In Exhibit 13.2 the standard coordinates have weighted average sum of squares equal to 1 on each ordination axis, using the species masses. In Exhibit 13.5 the contribution coordinates have unweighted sum of squared coordinates equal to 1 on each axis. These squared coordinates are the part contributions to the respective axes and are thus called *contribution coordinates*. In Exhibit 13.5 the species are shown as gradient vectors, and are oriented in the exact same directions as the unit profiles in Exhibit 13.2, but each species point has been pulled in by different amounts, with the rarer species being pulled in more than the more abundant ones. The exact relationship between the two types of species coordinates is that the contribution coordinates in Exhibit 13.5 are the standard coordinates in Exhibit 13.2 multiplied by the square roots of the respective species masses.

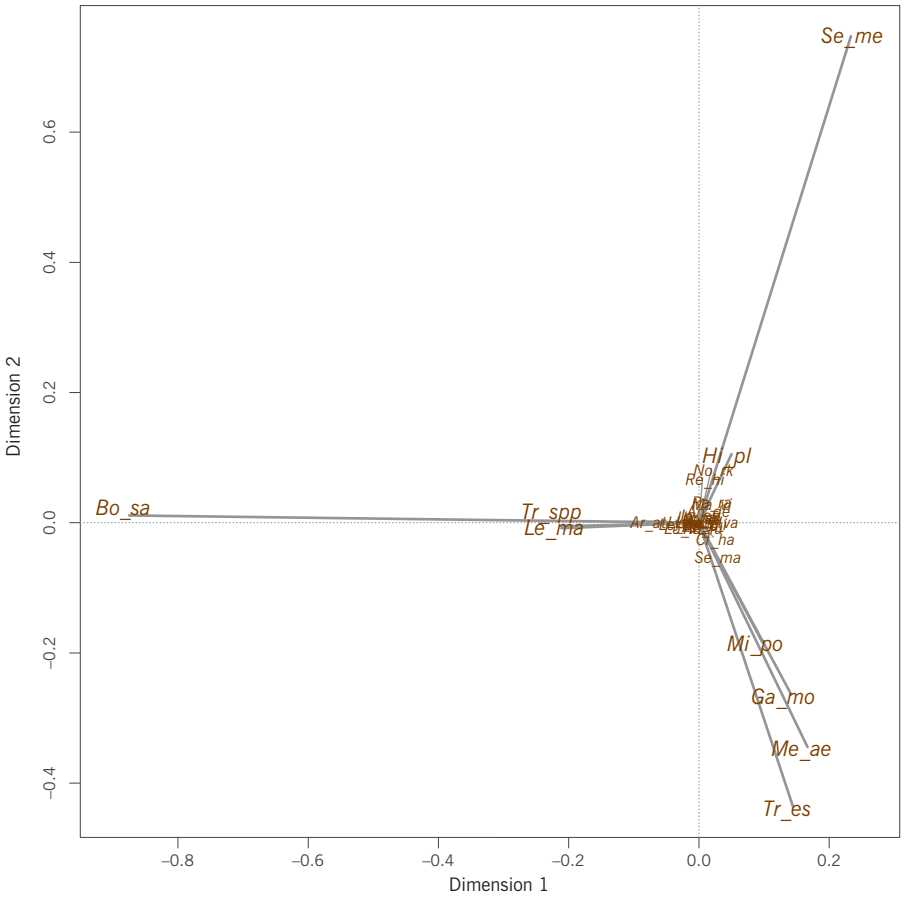


Exhibit 13.5: *Species in contribution coordinates. Combining this configuration with the sample points in Exhibit 13.2 would give the two-dimensional contribution biplot. The species that contribute more than average to an axis are shown in larger font (contributions to all three significant dimensions are taken into account here – the species Hi_pl contributes highly to the third dimension). Those near the origin in tiny font are very low contributors to the CA solution*

When it comes to the interpretation, the species *No_rk* and *Se_me* are ones that exemplify the difference between Exhibits 13.2 and 13.5. *No_rk* is a quite rare species in the data set, only 83 counted out of the total of 63,896, whereas the overall count of *Se_me* is 12,103. Thus *No_rk* is pulled in very strongly from its unit profile position in Exhibit 13.2 to its inlying position in Exhibit 13.5 – whereas it looked like it was the most important point before, it is now one of the species near the origin that are shown with tiny labels. By contrast, *Se_me* is not pulled in so strongly because of its high mass and in Exhibit 13.5 is confirmed to be the most important contributor to that spread of the samples upwards on the second axis. Both versions of the CA ordination are useful: from Exhibit 13.5 we know that *Se_me* is a strong contributor while Exhibit 13.2 tells us that the much sparser data for *No_rk* still correlates with that of *Se_me*. Another way of thinking about the high contributors, nine species in all in Exhibit 13.5, is that we could remove the other 21 species from the data set and get more or less the same result. To illustrate this, Exhibit 13.6

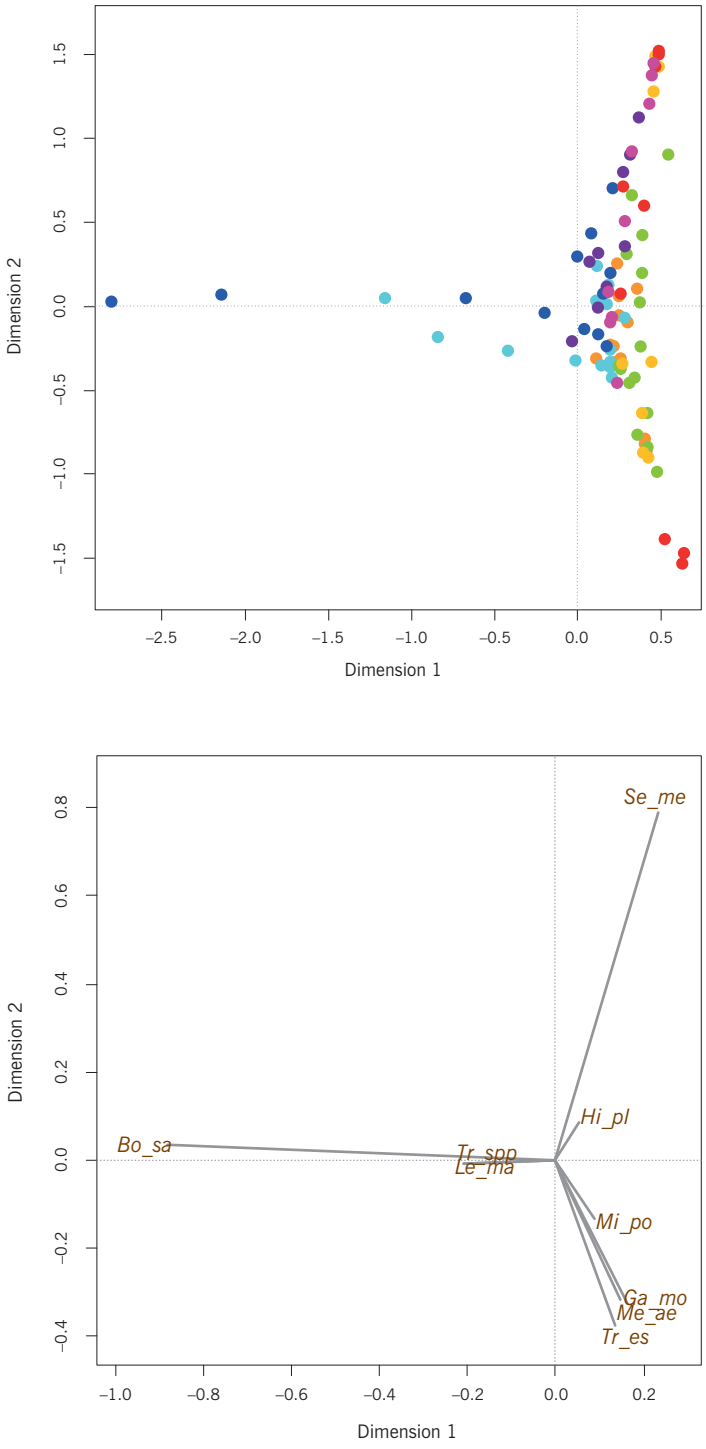


Exhibit 13.6:
Contribution biplot of the "Barents fish" data, retaining only the nine species with high contributions to the three-dimensional solution. The sample and species points are shown separately. The Procrustes correlations with the configurations obtained in Exhibits 13.2 (sample points) and 13.5 (species points), using all 30 species, are 0.993 and 0.997 respectively

shows the contribution biplot of this reduced data set of nine species. The result is almost identical – the Procrustes correlations with the previous results are almost 1.

Symmetric analysis of rows and columns

In all the above we have considered the case of the row profiles, the relative abundances of the species in each sample, with chi-square distances between them, mapped into a space using (weighted) classical MDS, with columns (i.e., species) displayed either as unit profiles or in contribution coordinates. We could turn this problem around by interchanging rows and columns and repeating everything as before. The matrix of column profiles is thus considered – these are the relative abundances across the samples of each species (i.e., the columns of Exhibit 11.1 divided by the column totals). Chi-square distances between these species profiles would be visualized (i.e., columns in principal coordinates), and the sample points added either as unit points (i.e., rows in standard coordinates) or as standardized regression coefficients (i.e., rows in contribution coordinates). In CA the row and column profile matrices, analysed in this similar and symmetric way, lead to exactly the same final solution, and all the sets of coordinates are related by simple scalar multipliers. The following are the basic results to remember, for both row and column points:

$$\text{principal coordinates} = \text{standard coordinates} \times (\text{principal inertias})^{1/2} \qquad (13.1)$$

$$\text{contribution coordinates} = (\text{masses})^{1/2} \times \text{standard coordinates} \qquad (13.2)$$

For example, suppose we had all the results from the analysis of the sample profiles, as discussed up to now and as shown in Exhibits 13.2, 13.4 and 13.5, and we wanted the equivalent results for the analysis of the species profiles. The species principal coordinates would be the species standard coordinates (shown in Exhibit 13.2) multiplied by the square roots of the principal inertias (eigenvalues) on respective axes: $(0.777)^{1/2}$ on first axis, $(0.541)^{1/2}$ on the second, $(0.485)^{1/2}$ on the third – notice that the principal inertias in CA are always less than one, so the principal coordinates are always contracted towards the centre compared to the standard coordinates. To obtain the sample standard coordinates we have to do the reverse operation by taking the sample principal coordinates (also in Exhibit 13.2) and divide by the corresponding square roots of the principal inertias, given above. Finally, to obtain sample contribution coordinates in order to see which are the highly contributing samples to the solution, the standard coordinates for each sample are multiplied by the corresponding square root of the sample mass.

A popular way of showing the results of a CA is to show the simultaneous display of the row and column profiles, that is both in principal coordinates. For the

"Barents fish" data set, this so-called *symmetric map* of the points, where both rows and columns are visualizing their inter-point chi-square distances, is shown in Exhibit 13.7. Because contributions are not directly visualized in the points' coordinates, we can introduce larger and smaller symbols or labels to give an indication of the important points to concentrate on in the interpretation. An advantage of this display is that the row and column points have the same inertias (parts of variance) along the dimensions, so they are spread out the same amount horizontally and vertically, which uses the plotting space better. Strictly speaking, however, the symmetric map is not a biplot as described before. However, when the square roots of the principal inertias along axes are not too different, so that principal and standard coordinates are approximately proportional to one another in the two-dimensional solution (see formula (13.1)), then the map comes close to a true biplot. In this example, $(0.777)^{½} = 0.881$ and $(0.541)^{½} = 0.736$, which are indeed quite close, so Exhibit 13.7 can be interpreted as an approximate biplot.

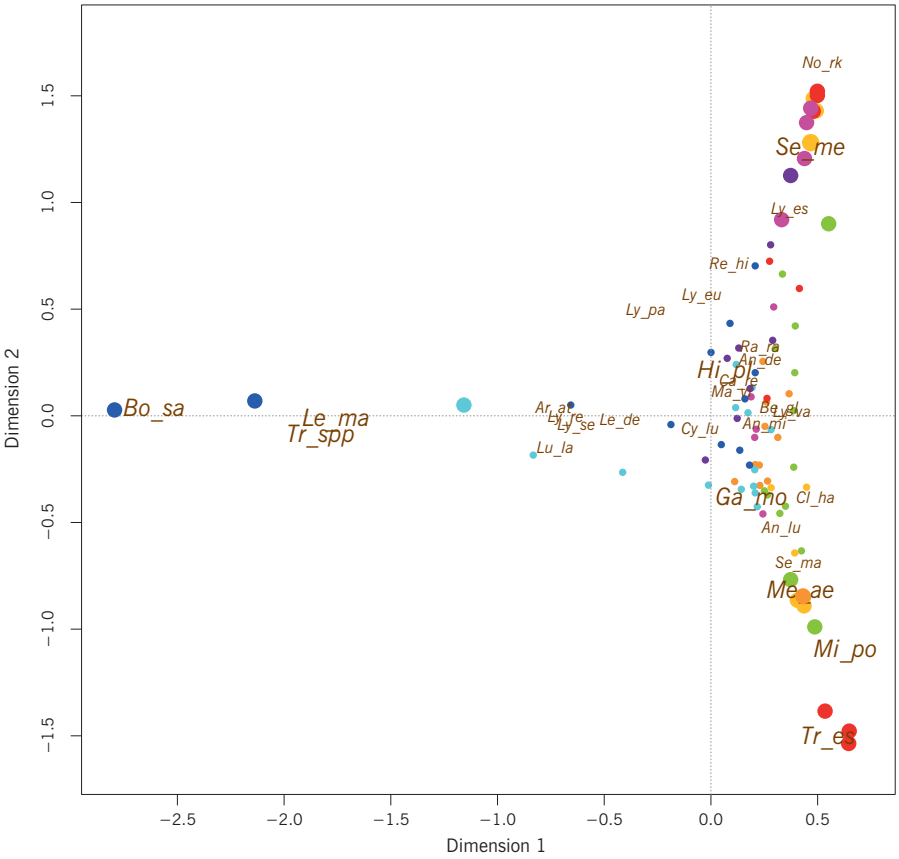


Exhibit 13.7:
Symmetric map of "Barents fish" data set, both samples and species in principal coordinates, with higher than average contributing samples and species in larger symbols and font sizes

SUMMARY: Correspondence analysis

1. Correspondence analysis (CA) is the analogue of principal component analysis (PCA) for data that are nonnegative such as abundance counts, biomasses and percentages. All the data must be measured on the same scale, so that it makes sense to compute row sums and column sums.
2. CA analyses the row profiles and/or the column profiles of the data matrix: these are the rows of data divided by their respective row sums or the columns divided by their respective column sums.
3. Each row and each column is weighted by its respective mass: the masses are the row and columns sums relative to the grand total of the data.
4. Distances between row profiles or between column profiles are defined by the chi-square distance.
5. For a samples-by-species data matrix, CA is generally thought of asymmetrically as an analysis of the sample (row) profiles, visualizing the inter-profile chi-square distances in a low-dimensional map (i.e., samples displayed in principal coordinates).
6. The species can then be visualized in two alternative ways as a biplot: as unit profiles, showing fictitious samples consisting of just one species (i.e., species in standard coordinates), or as gradient vectors showing the regression relationships between the species and the principal axes (i.e., species in contribution coordinates). These alternatives indicate identical orientations of biplot axes, but the latter alternative has the advantage that the more outlying species are the higher contributors to the solution.
7. Thinking of the analysis from the column profile point of view gives another way of interpreting the CA solution, as distances between species. The solution of this problem is identical to the row profile problem, with simple scaling factors linking the two solutions.
8. A popular way of showing the CA result is the symmetric map, where both row and column profiles are visualized simultaneously, that is both in principal coordinates. The row and column points have the same spread along the dimensions, and they can each be interpreted in terms of approximate chi-square distances.

List of Exhibits