10. Correspondence Analysis (CA) and Detrended CA (DCA)

FISH 560: Applied Multivariate Statistics for Ecologists

Topics

- Perform CA and DCA, run diagnostics, generate biplots

R Packages: vegan, FactoMineR, factoextra,
R Source: biostats, evplot

BACKGROUND

Correspondence analysis (CA) is one of two names given to a method of ordination that was developed in the 1970s and continues to be heavily used by ecologists. Interestingly, CA was independently developed by Hill (1973,1974) in England (under the name reciprocal averaging) and by Benzecri (1969) in France (under the name analyse factorielle des corresponences). Legendre and Legendre (1998) trace its origins back farther than that outside of ecology.

Correspondence analysis aims to compare the correspondence between samples and species from a table of counted data (or any dimensionally homogenous table) and to represent it in reduced ordination space. PCA and CA are actually quite similar, both conceptually and statistically (Legendre and Legendre 1998). Whereas PCA operates in Euclidean data space to repartition the total variance as a series of optimized linear additive components (ordination axes), CA partitions the total contingency chi-square (or inertia) as a series of linear additive components within a chi-square data space (Digby and Kempton 1987). Noticeably, instead of maximizing the amount of variance explained by the ordination (PCA), CA maximizes the correspondence (or inertia) between species scores and sample scores. In other words, CA is essentially an eigenanalysis of a chi-square distance matrix. A chi-square distance matrix is defined by the deviation from expectation. The use of the chi-square metric results in a weighted linear mapping of samples (and species) into a reduced ordination space, where the particular weighting scheme (doubly weighting based on row and column totals) assumes a unimodal (rather than linear) species response to the underlying gradients. For this reason, CA is better able to capture ecologically meaningful gradients in species abundance data sets with many zeros and handles presence/absence data with greater ease.

CA is popular among ecologists for at least three reasons. First, this technique is recommended when species display unimodal (bell shaped or Gaussian) relationships with environmental gradients (ter Braak 1985), as it happens when a species favors specific values of a given environmental variable. Second, the reciprocal averaging algorithm disregards species double absences because the relationships between rows and columns of the table are quantified using the chi-square coefficient that excludes double absences (Legendre and Legendre 1998). Third, in contrast to the other ordination methods, CA simultaneously ordinates objects (e.g., sites) and descriptors (e.g., species) using a weighted averaging approach. That is, species centroids (the hypothetical mode of their distribution in ordination space) can be plotted as well as the locations of the samples. Given these advantages, CA does have one slight drawback (not a big deal!), which we will discuss later in relation to detrended correspondence analysis (DCA).

SET-UP

In this exercise you will be working with the MAHA species presence-absence dataset. But first remember to set-up your R work session by defining the current work directory to your folder of choice and loading the vegan library. Also, make sure to source the BIOSTATS file from the *File* pull-down menu. You can also do this using the functions **`setwd`**, **`library`** and **`source`**. There are two ways that we can import the species presence-absence data.

First, we can transform the species abundances into presence/absence (i.e., binary transformation) using the power method with an exponent equal to zero, by typing:

```
speabu <- read.csv('MAHA_speciesabu.csv',header=TRUE, row.names=1)
speocc <- data.trans(speabu,method='power',exp=0,plot=F)
```

Alternatively we can directly import the data, by typing:

```
speocc <- read.csv('MAHA_speciesocc.csv',header=TRUE, row.names=1)
```

CONDUCTING CORRESPONDENCE ANALYSIS

We will use the cca() function in the vegan library, which allows for both unconstrained correspondence analysis (CA) as well as the constrained version, canonical correspondence analysis (CCA), which we will address in a subsequent chapter. You can also use the decorana() function. The cca() function accepts matrices or data frames, computing "scores" for each sample and each species in the matrix or data frame. The scores are derived from the corresponding eigenvectors or, equivalently, from the weighted averages of the opposing scores. The cca() function expects the input matrix or data frame to contain only the samples and species of interest. Thus, make sure the data set has just the desired set of samples and species, and make sure that the data have already been transformed, if appropriate. Note, CA has a built in chi-square distance, so a row or column standardization may not be appropriate. The usage of cca() is:

cca(formula, data, ...)
cca(X, Y, Z, ...)

Where:

- **formula** Model formula, where the left hand side gives the community data matrix, right hand side gives the constraining variables, and conditioning variables can be given within a special function Condition.
- **data** Data frame containing the variables on the right hand side of the model formula.
- **X** Community data matrix.
- **Y** Constraining matrix, typically of environmental variables. Can be missing.
- **Z** Conditioning matrix, the effect of which is removed (`partialled out') before next step. Can be missing.
- **Scale** species to unit variance (like correlations).
- **...** Other arguments for print or plot functions (ignored in other functions).

For additional documentation on cca() type **`?cca`**.

Let's perform CA on the species presence-absence matrix, by typing:

```
spe.ca<-cca(speocc)
```

The results of a CA are complex and stored in a list, including:

- eigenvalues (inertia) associated with each axis
- eigenvectors (sample and species scores)

Type **`summary(spe.ca)`** to view the summary of the eigenvalues. Your results should be as below.

```
Call:
cca(X = speocc)
```

```
Partitioning of mean squared contingency coefficient:
              Inertia Proportion
Total           2.845          1
Unconstrained   2.845          1

Eigenvalues, and their contribution to the mean squared contingency
coefficient

Importance of components:
                          CA1    CA2     CA3     CA4    CA5     CA6
Eigenvalue            0.4013 0.3089 0.26673 0.20120 0.1934 0.17992
Proportion Explained  0.1410 0.1086 0.09375 0.07072 0.0680 0.06324
Cumulative Proportion 0.1410 0.2496 0.34338 0.41409 0.4821 0.54533
…

Scaling 2 for species and site scores
* Species are scaled proportional to eigenvalues
* Sites are unscaled: weighted dispersion equal on all dimensions

Species scores

              CA1      CA2       CA3       CA4      CA5      CA6
BANDDART  2.75402 -0.78122  0.005164 -0.424512  0.38277 -0.44415
BANDSCUL  0.70288  0.16340 -0.222250  0.281777 -0.80426  1.15851
BLACDACE -0.20830  0.44562 -0.214540 -0.342707 -0.31654  0.16223
…

Site scores (weighted averages of species scores)

          CA1      CA2      CA3      CA4      CA5      CA6
S1  -0.398438 -0.13939  1.17466  0.81799 -1.08594 -0.17203
S2  -1.110843 -0.94457  0.13356 -1.17710  1.08426  0.22256
S3  -0.415313  0.02718  0.69385 -0.57723 -0.66535 -0.32217
```

Eigenvalues and Inertia

Recall that in PCA eigenvalues represent 'variance' explained. In CA, eigenvalues represent 'inertia', where total inertia equals the chi-squared statistic of the data matrix standardized to unit total. Thus, the eigenvalues are something akin to variance, but they are not variances exactly. The chi-square statistic is a measure of association between samples and species. If species are not independently distributed among samples, so that species tend to co-occur in samples, the chi-square statistic will be large. The larger the chi-square statistic, the greater the correspondence of species' distributions among samples, and vice versa, the thus the greater the inertia.

To see the inertia (eigenvalue) of each axis, type:

spe.caCAeig

Note the syntax above. The results of the CA were stored in an object called 'spe.ca'. This object is a list containing several components, including one called 'CA'. This component is actually another list containing several components. Type **names(spe.ca)** to see the complete list, including the eigenvalues ('eig'). The call requests the list component named 'eig' in the list component named 'CA' in the object named 'spe.ca'.

The total inertia is equal to the sum of all the eigenvalues, and you can get this by typing either one of the commands below:

```
sum(spe.ca$CA$eig)
spe.ca$CA$tot.chi
```

Recall that you divide each eigenvalue by the sum of eigenvalues to calculate the proportion of variation accounted for by each axis. To calculate this as a percentage you can type:

```
spe.ca$CA$eig/sum(spe.ca$CA$eig)*100
```

This will return the following (just the first 6 axes are shown):

```
       CA1         CA2         CA3         CA4         CA5         CA6
14.10396566 10.85837238  9.37518000  7.07179790  6.79953686  6.32383406
```

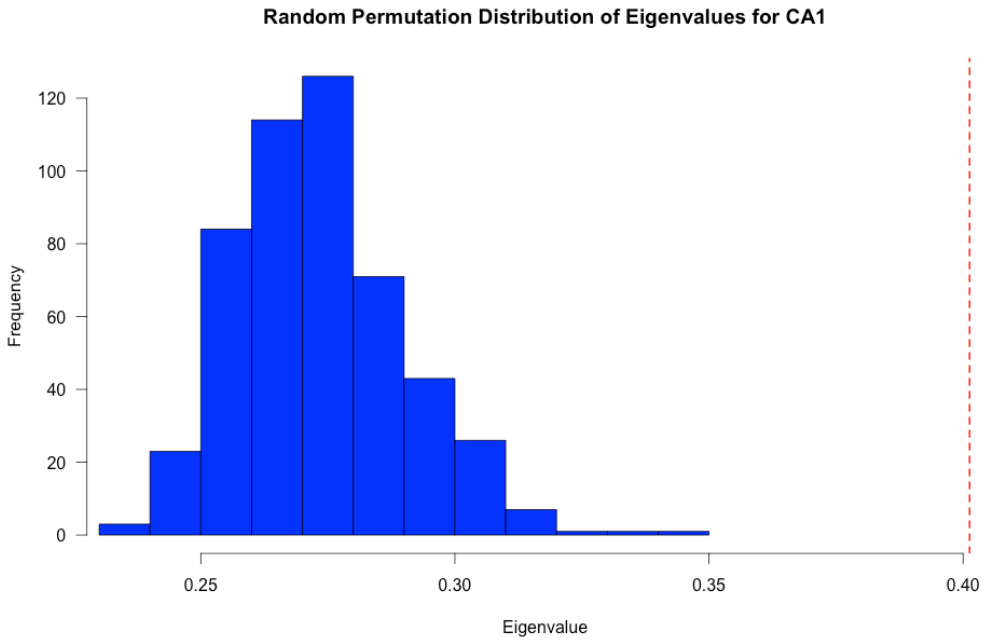
Note that all of this information is also presented in the summary output (see above). I just wanted to make sure you know where the values are coming from!

We may wish to test the statistical significance of the first several eigenvalues using a randomization test, as follows:

```
ordi.monte(speocc,ord='ca',dim=5,perm=500)
```

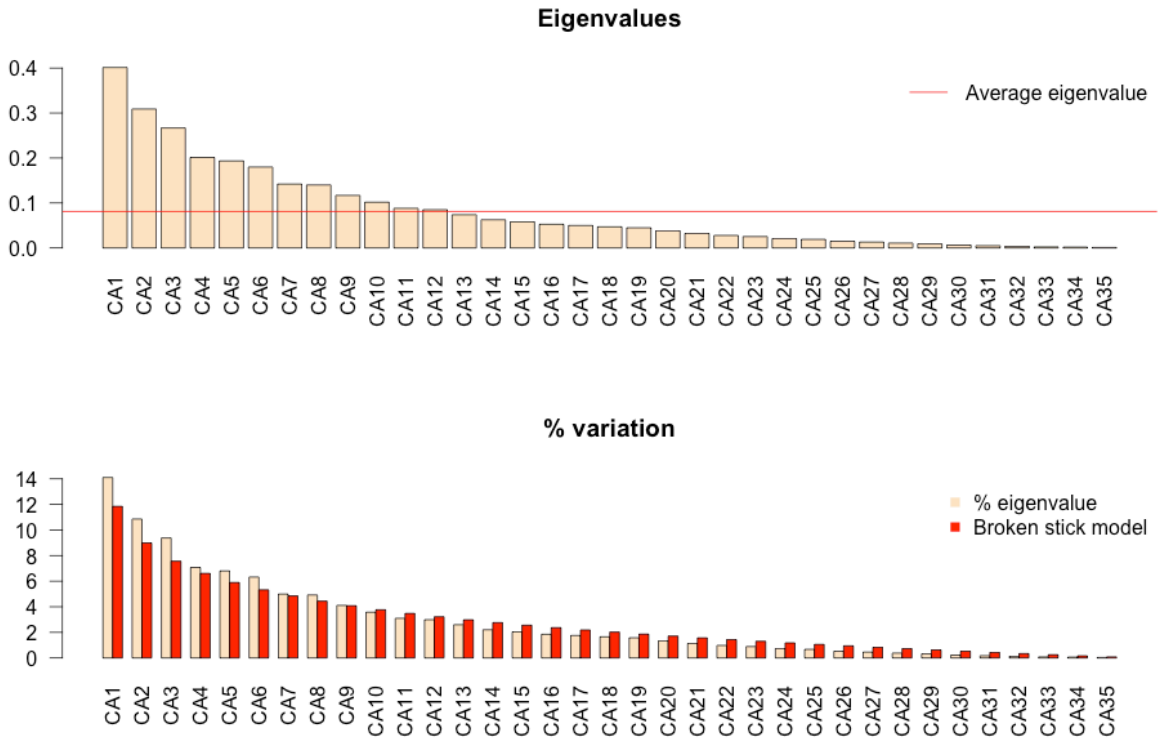
This will return the table below as well as histograms of null and observed eigenvalues for each axis. Axis 1 results are shown.

```
Randomization Test of Eigenvalues:
             CA1   CA2   CA3   CA4   CA5
Eigenvalue 0.401 0.309 0.267 0.201 0.193
P-value    0.000 0.000 0.000 0.810 0.442
```



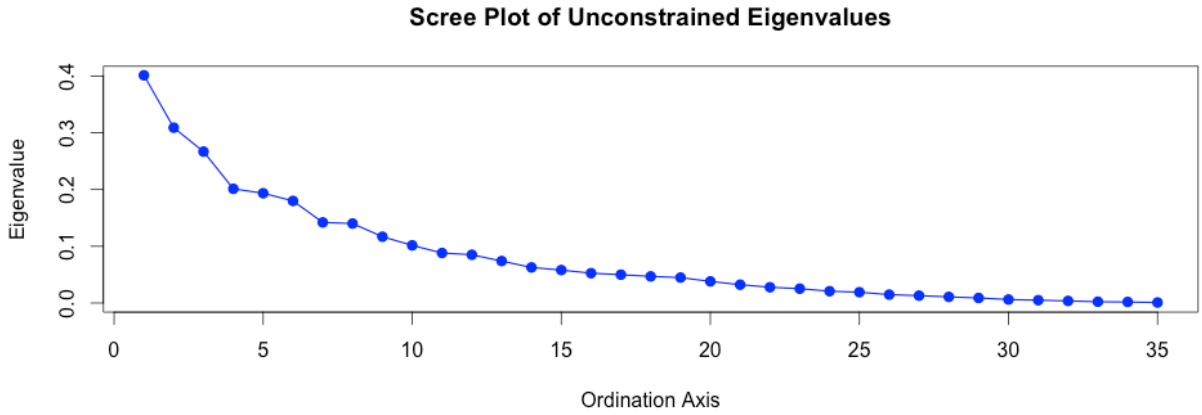
We can look at this same information graphically with the evplot() function (DON'T FORGET TO SOURCE THIS FILE):

```
evplot(spe.ca$CA$eig)
```



We can look at this same information graphically with the ordi.scree() function:

```
ordi.scree(spe.ca,ord='ca')
```



Site and species scores

To see the sample and species scores for the first two axes, type:

```
spe.ca$CA$u[,1:2]          # sample scores ('u')
spe.ca$CA$v[,1:2]          # species scores ('v')
```

Here are the scores for the first 5 sites:

```
              CA1          CA2
S1  -0.398437934 -0.13939364
S2  -1.110842589 -0.94456616
S3  -0.415313410  0.02718452
S4  -0.711761834  2.08266581
S5  -0.134982565  0.90637210
…
```

Note, in contrast to PCA, the structure correlations (i.e., linear correlations between the species, in this case, and the ordination axes) are not meaningful because we are assuming a unimodal response function. Therefore, there is little point on trying to interpret the axes on the basis of which species are loaded highly. Instead, we are interested in the relative positions of species in the ordination space and, reciprocally, the relative positions of objects in the ordination space.

Ordination plots

Recall that there are lots of interesting options for displaying relationships in ordination plots, including:

- simple plots of samples and species
- displaying related samples (groups) on ordination plots
- fitting and displaying predictor variables on ordination plots

All of the plotting functions used in the PCA lab are applicable here and operate identically. Just substitute the CA object (spe.ca) for the PCA object.

Issues in scaling: CA has several conventions for scaling the sample and species scores. Recall that CA is a weighted averaging method, therefore both samples and species are often jointly depicted in the ordination space (i.e. joint plot), where the center of inertia (centroid) of their scores corresponds to the zero for all axes. Depending on the choice of the scaling type, either the ordination of rows (samples) or the columns (species) is more meaningful, and can be interpreted as an approximation of the chi-square distances between samples or species, respectively.

In scaling type 1, sample scores are calculated as weighted averages of species scores; i.e., the means of the species scores for species that occur across the sites. Sample points that are close to each other in ordination space are similar with regard to the pattern of relative frequencies across species. This scaling is the most appropriate if one is primarily interested in the ordination of objects.

In scaling type 2, species scores are calculated as weighted averages of samples scores; i.e., the means of the sample scores in which the species occurs. Species points that are close to each other in ordination space are similar with regard to the pattern of relative frequencies across sites. This scaling is the most appropriate if one is primarily interested in the ordination of descriptors.

Note, when we take a weighted average, the range of averages shrinks from the original values. The shrinkage factor is equal to the eigenvalue of CA, which has a theoretical maximum of 1. Thus, in scaling type 1, the sample scores are scaled by the eigenvalues, whereas in scaling type 2, the species scores are scaled by the eigenvalues.

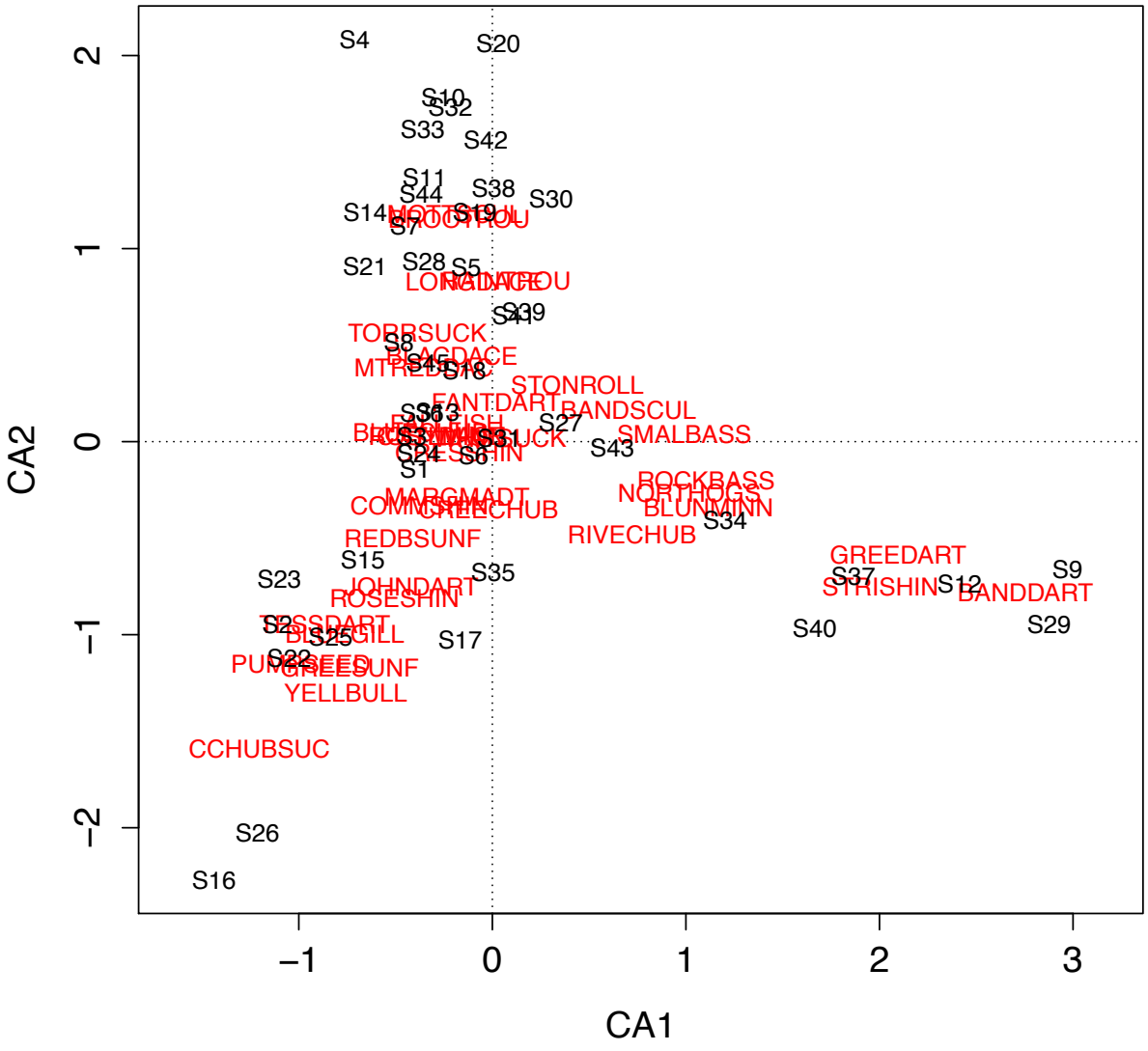
In scaling type 3, the sample and species scores are scaled symmetrically; both are scaled by the square root of eigenvalues.

Here, let's create a CA joint-plot using scaling type 2 by typing:

```
ordiplot(spe.ca,choices=c(1,2),scaling=2)
```

You might be interested in labeling the plot above with object (site) names to help with interpretation.

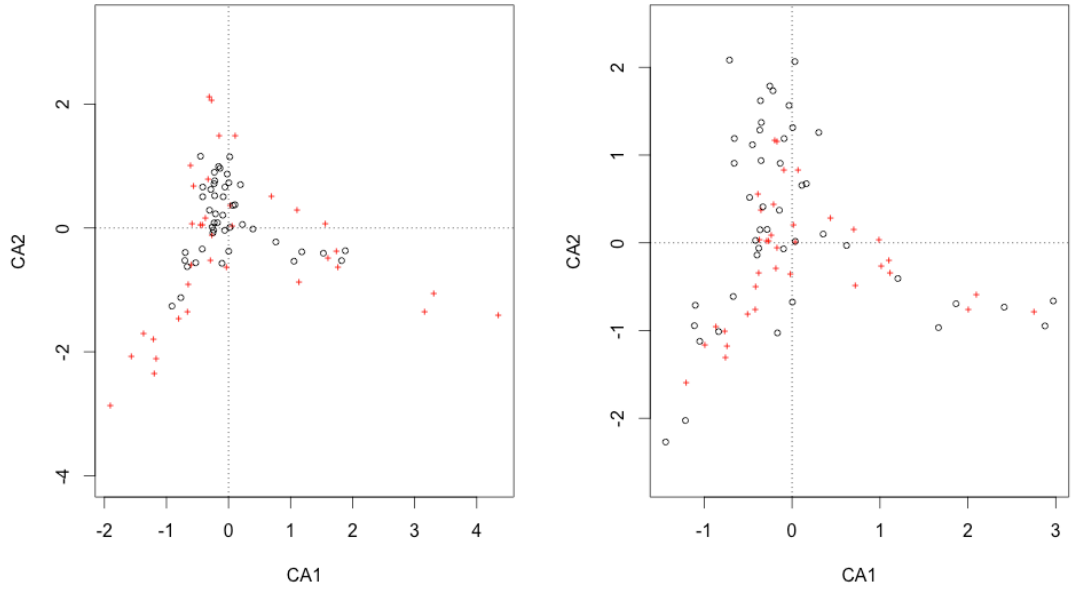
```
ordiplot(spe.ca,type='t', scaling=2)
```



You might also interested in comparing the two scaling approaches. To do this you can type:

```
par(mfrow=c(1,2))
```

```
ordiplot(spe.ca, scaling=1)
ordiplot(spe.ca, scaling=2)
```

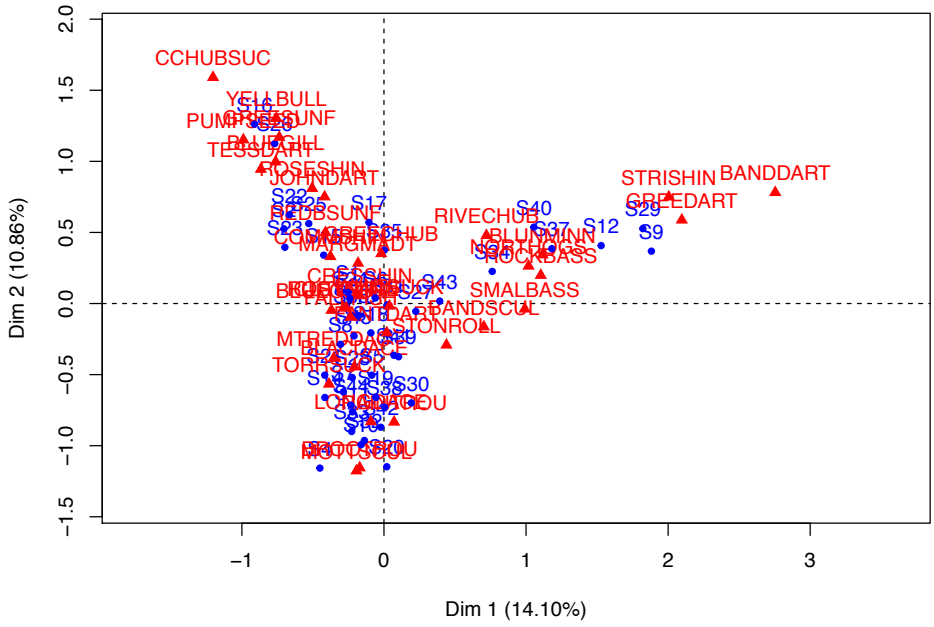


Additional visualization options

Let's use the FactoMineR and factoextra packages to produce some nicer looking outputs from CA. In a nutshell, factoextra is an **R** package to easily extract and visualize the output of multivariate data analyses conducted in FactoMineR.

First type the following to conduct the CA:

```
spe.ca2<-CA(speocc)
```



Type the following to obtain the other results

```
summary(spe.ca2)
```

DETRENDED CORRESPONDENCE ANALYSIS

When the species composition of the sites progressively changes along the environmental gradient, sample positions may appear in the ordination plot as nonlinear configurations called 'arch' (Gauch 1982) (or 'horseshoe' in the case of PCA), which may impair further ecological interpretation. In CA, the arch effect may be mathematically produced as a side-effect of the CA procedure that tries to obtain axes that both maximally separate species and that are uncorrelated to each other (ter Braak 1987): when the first axis suffices to correctly order the sites and species, a second axis (uncorrelated with the former) can be obtained by folding the first axis in the middle and bringing its extremities together, thus resulting in an arch configuration.

To remove the arch effect in CA, a mathematical procedure, detrending, is used to flatten the distribution of the sites along the first CA axis without changing their ordination on that axis. The approach is then designated as detrended correspondence analysis (DCA).

Specifically, DCA attempts to correct each of these drawbacks with the following solutions:

- Single long gradients appear as curves or arcs in ordination: the solution is to detrend the later axes by making their means equal along segments of previous axes.

- Sites are packed more closely at gradient extremes than at the center: the solution is to rescale the axes to equal variances of species scores.

- Rare species seem to have an unduly high influence on the results: the solution is to downweight rare species.

The execution of DCA and the examination of the results are almost identical to those described already for CA, and we will not repeat the procedures again here. Briefly, we will use the function decorana() to conduct DCA. Its usage is:

decorana (veg, iweigh=0, ira=0, mk=26)

Where:

6. **veg** is the community data set
7. **iweigh** indicates whether rare species should be downweighted (0, no; 1, yes). Default is 0.
8. **ira** indicates the type of analysis (0, DCA; 1, CA). Default is 0.
9. **mk** indicates the number of segments in rescaling. Default is 26. This option is only invoked when ira=0.

For additional documentation on decorana type **`?decorana`**.

Now, let's conduct a DCA by typing:

```
spe.dca<-decorana(speocc,ira=0)

spe.dca

Call:
decorana(veg = speocc, ira = 0)

Detrended correspondence analysis with 26 segments.
Rescaling of axes with 4 iterations.

                 DCA1   DCA2   DCA3   DCA4
Eigenvalues     0.3674 0.2450 0.2239 0.1656
```

```
Decorana values 0.4013 0.2320 0.1961 0.1197
Axis lengths    3.1419 3.1044 3.7364 2.8201
```

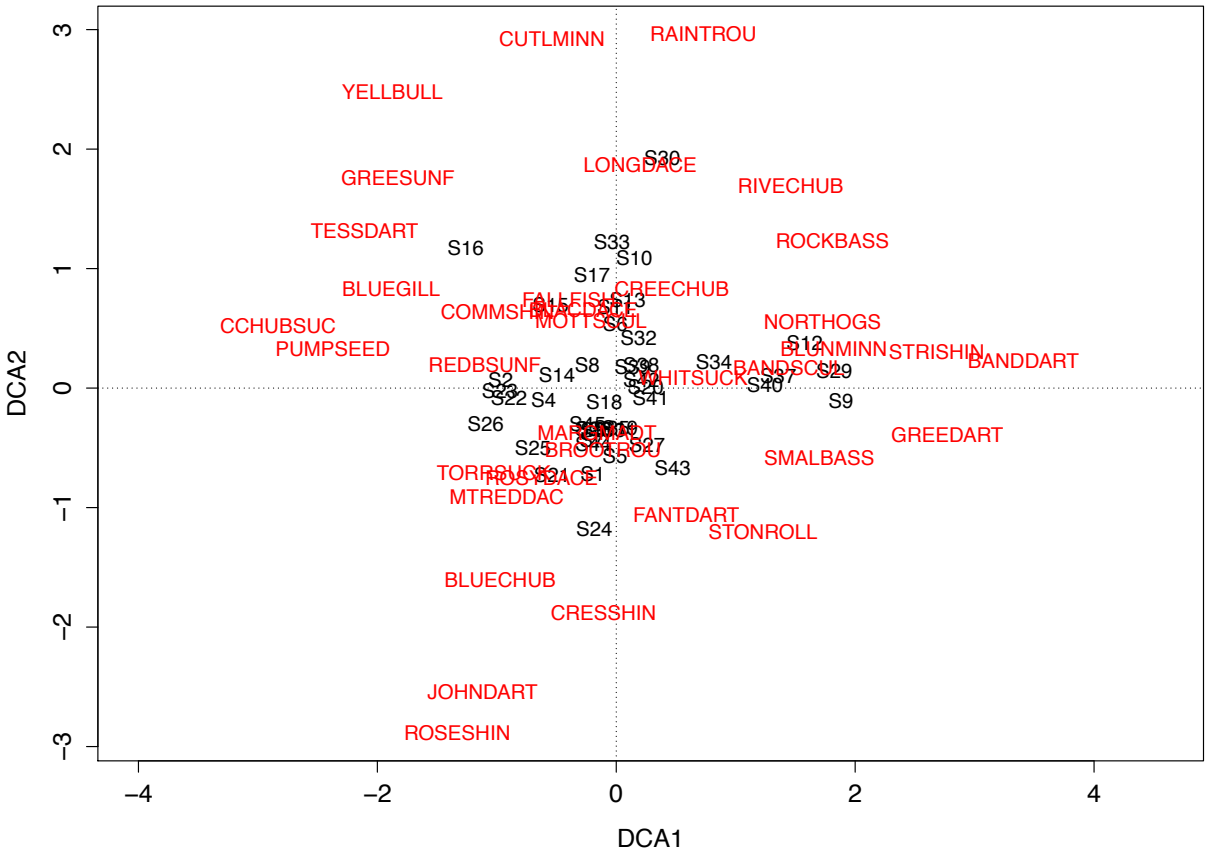
From the above summary we can see that DCA returns eigenvalues and decorana values. Unlike CA, these results are not straightforward to interpret because they do not equate to the proportion of variance explained (because of the detrending and rescaling of axes).

Site and species scores can be obtained by typing:

```
summary(spe.dca)
```

The DCA ordination plot can be obtained by typing:

```
plot(spe.dca)
```



MULTIPLE CORRESPONDENCE ANALYSIS

Multiple correspondence analysis is an extension of the simple correspondence analysis for summarizing and visualizing a data table containing more than two *categorical variables*. We will not explore this approach, but I recommend the FactoMineR library for the analysis and the factoextra library for ggplot2-based elegant visualization.

OPTIONAL READINGS (* recommended)

Benzecri, J.P. 1969. Statistical analysis as a tool to make patterns emerge from data. Pages 35-60 in S. Watanabe [ed.] Methodologies of Pattern Recognition. Academic Press, New York.

Bradfield, G. E. and Kenkel, N. C. 1987. Nonlinear ordination using flexible shortest path adjustment of ecological distances. Ecology 68: 750–753.

Digby, P. G. N. and Kempton. R. A. 1987. Multivariate analysis of ecological communities. Chapman and Hall, London, UK

Greenacre, M. J. 1984. Theory and applications of correspondence analysis. Wiley, New York, NY.

Greenacre, M. J. and Hastie, T. 1987. The geometric interpretation of correspondence analysis. J. Am. Statist. Assoc. 82: 437–447.

Hill, M. O. 1973. Reciprocal averaging: an eigenvector method of ordination. J. Ecol. 61: 237–249.*****

Hill, M.O. 1974. Correspondence analysis: a neglected multivariate method. Appl Statis 23: 340–354.

Hill, M.O. and H.G. Gauch. 1980. Detrended Correspondence Analysis: an improved ordination technique. Vegetatio 42: 47-58. *****

Jackson, D. A., and K. M. Somers. 1991. Putting things in order: the ups and downs of detrended correspondence analysis. American Naturalist 137: 704-12. *****

Legendre, P. and Legendre, L. 1998. Numerical Ecology. Elsevier Science BV, Amsterdam, 853pp.

Palmer, M. W. 1993. Putting things in even better order: the advantages of canonical correspondence analysis. Ecology 74: 2215-30.

Peet, R. K., R. G. Knox, J. S. Case, and R. B. Allen. 1988. Putting things in order: the advantages of detrended correspondence analysis. American Naturalist 131: 924-34.

ter Braak, C.J.F. 1985. Correspondence analysis of incidence and abundance data: properties in terms of a unimodal response model. Biometrics 41: 859–873.

Wartenberg, D., S. Ferson, and F. J. Rohlf. 1987. Putting things in order: a critique of detrended correspondence analysis. American Naturalist 129: 434-48.

EXERCISE

Purpose

Upon completion of this chapter, you should be able to do the following:

(1) Carry out a CA and DCA
(2) Assess the most powerful solution with respect to the number of dimensions; and
(3) Interpret sample and species scores.

Tasks

Perform a CA using the fish species abundance dataset, and address the following:

- How much variation is explained by the first two components? Are these statistical significant?
- Interpret the joint plot (i.e., ordination with object and species scores)
- Compare joint plots using different scaling? How different are they? Does it affect your interpretation of similarities between sites and between species?
- Is there any evidence for arching in the ordination? If so, does a DCA remedy this problem?