Ecological ordination with Past

Øyvind Hammer, Natural History Museum, University of Oslo, 2011-06-26

Introduction

This text concerns *taxa-in-samples* data. Such a data set contains a number of samples, each sample occupying one row in the spreadsheet. Each sample contains counts, percentages or presence-absence of a number of taxa (in columns). The samples may come from different localities or different levels in a section or core. A basic requirement is to plot the samples as points in 2D or 3D, so that similar samples plot closely to each other, and more different samples are more distant. From such a plot, called an *ordination*, it may be possible to extract different types of information:

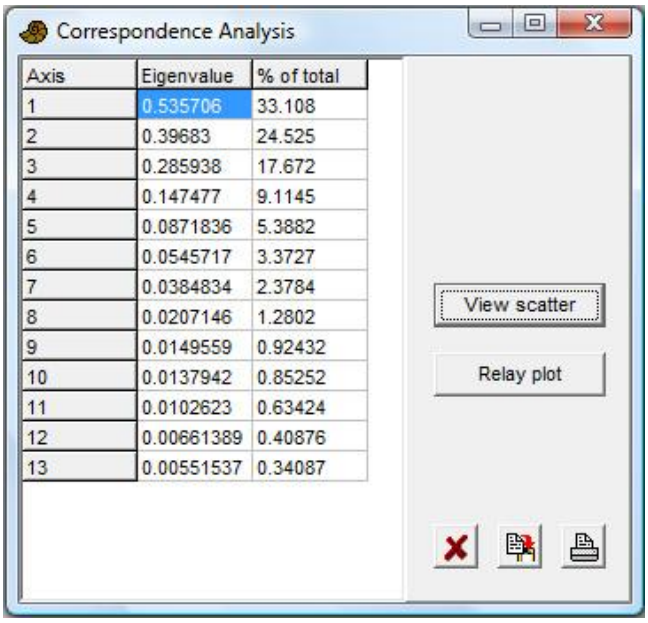
- Are there groups of points? How may such groups be interpreted, e.g. in terms of biogeography, biostratigraphy or environment?
- Are the points ordered according to geographical, stratigraphical or environmental gradients?

Ordination is a fundamental technique in modern ecology, and often one of the first things we try in order to get an overview of a complex taxa-in-samples data set.

Correspondence analysis

Correspondence analysis (CA) is one of the most popular ordination methods for taxa-in-samples data, especially for samples collected along one or several gradients along which taxa come and go in an overlapping sequence. Like other ordination methods, CA attempts to place similar samples in similar positions in the ordination plot. The measure of distance between samples is proportional to the chi-squared statistic.

We will use a dataset containing abundances (on a scale from 0 to 10) of Recent benthic foraminifera along a depth gradient in the Gulf of Mexico. Open the file `bentforams.dat`, select the whole table and run “Correspondence” from the Multivar menu.

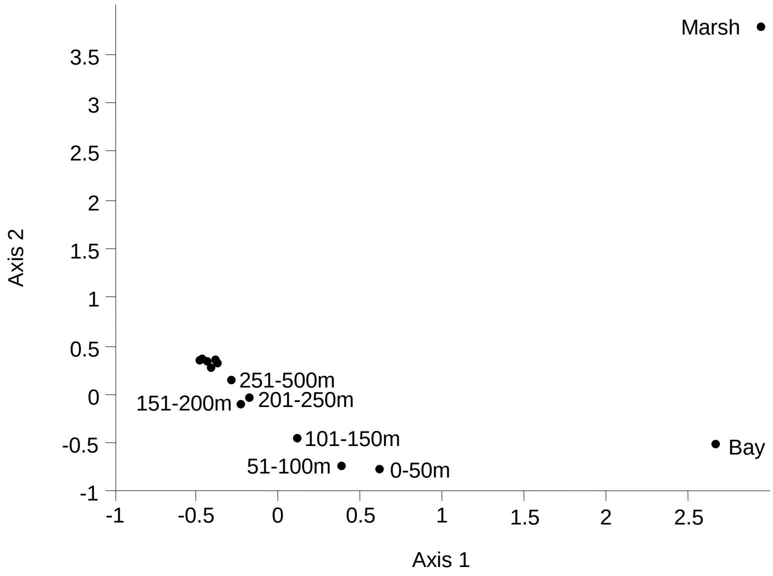


Axis	Eigenvalue	% of total
1	0.535706	33.108
2	0.39683	24.525
3	0.285938	17.672
4	0.147477	9.1145
5	0.0871836	5.3882
6	0.0545717	3.3727
7	0.0384834	2.3784
8	0.0207146	1.2802
9	0.0149559	0.92432
10	0.0137942	0.85252
11	0.0102623	0.63424
12	0.00661389	0.40876
13	0.00551537	0.34087

The first window shows the 13 *axes* constructed by the analysis. Consider that two points will always lie on a straight line (one dimension) and three points always in a plane (two dimensions). In this case we have 14 samples, and the corresponding points occupy a 13-dimensional space.

The axes are ordered according to their *eigenvalues*. The first axis, with the largest eigenvalue, contains 33.1% of the “information” in the data set, measured using the chi-squared criterion. The second axis contains 24.5%. This means that if we plot the points in two dimensions, using the two first CA axes, we retain 57.6% of the information, which is impressive considering that the dimensionality has been reduced from 14 to 2. This works because there is *structure* in the data, and the analysis has been successful at extracting this structure.

Click the “View scatter” button to see the result of the ordination.

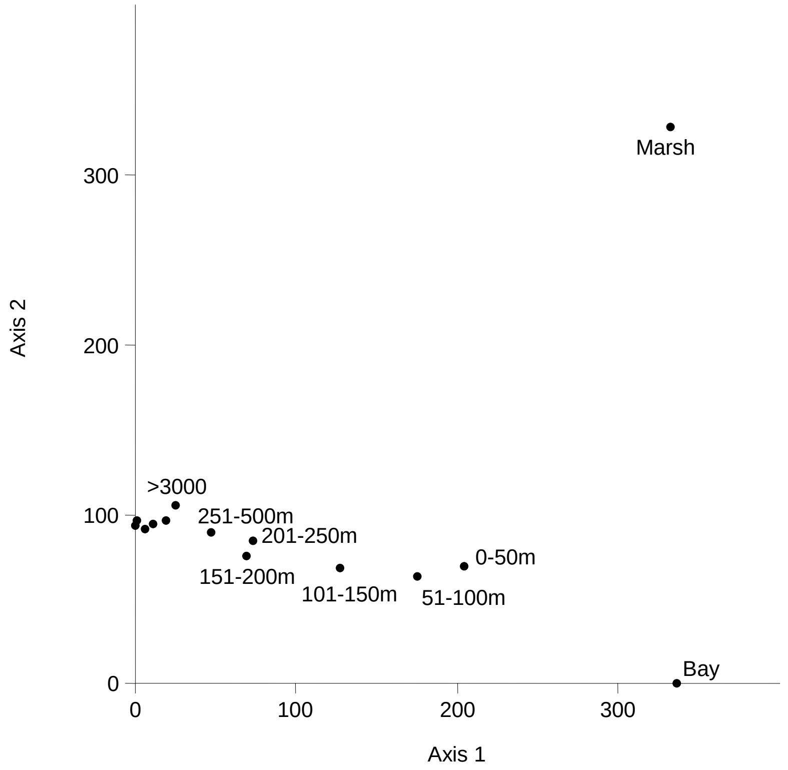


The important thing to note is that the shallowest sample (Marsh) is found at the right end of the plot. To the left of it is “Bay”, then “0-50m”, “51-100m” and “101-150m”. Going deeper than this, the points lie very close together, and partly in the wrong order considering their depths. Still, it is clear that the CA axis 1 can be interpreted as reflecting depth. Remember that the computer had no information about depth, but managed to place the samples in this order based only on their foram abundances. Firstly, this indicates that depth (or rather some other factor correlating with depth) is an important control on the foram fauna. Secondly, if this were a paleontological data set, with no *a priori* information on depth, such an analysis might provide clues about paleodepth.

The interpretation of Axis 2 is more obscure, but it is clearly dominated by the difference between the Marsh and the Bay samples.

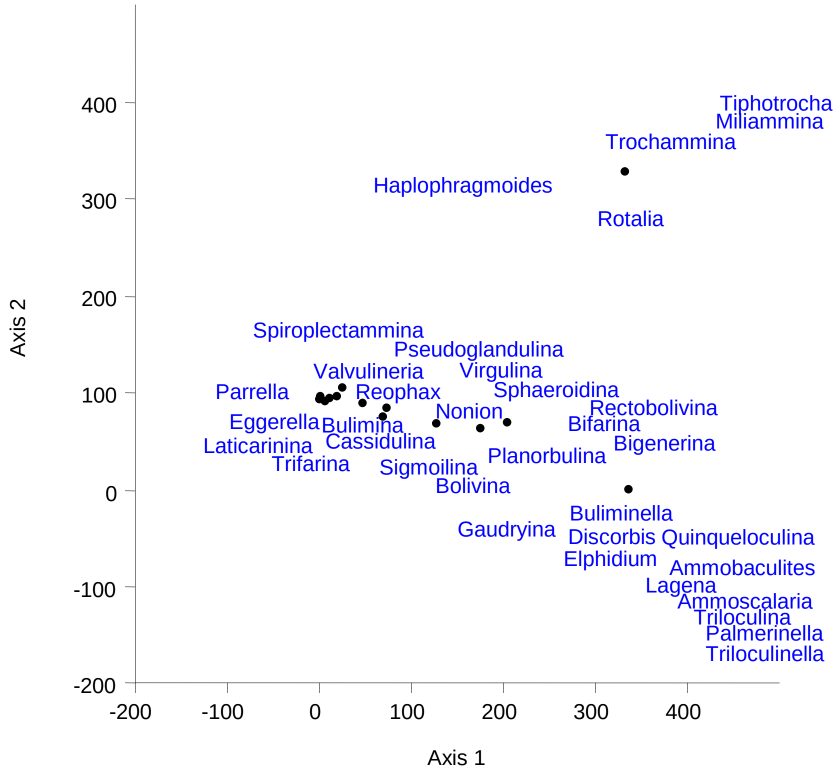
Detrended correspondence analysis

Sometimes, nonlinear relationships can cause the main gradient, reflected by CA axis 1, to “spill over” into axis 2, producing an arch rather than a linear trend in the CA plot. The points can also get compressed near the ends of the gradient. To reduce these perhaps annoying effects, one can attempt to “straighten out” the arch. Detrended correspondence analysis (DCA) is one method with this purpose. Select all, and run “Detrended correspondence” from the Multivar menu. In this case the effect is not dramatic compared with the usual correspondence analysis, but the depth gradient is more parallel with axis 1 (tick and untick the “Detrending” box to compare the two methods).



An interesting feature of CA (and DCA) is that is can show both the samples and the taxa in the same plot, illustrating which taxa are moreimportant in different regions of the diagram. Select the “Column labels” option. In the figure below I have moved and removed names to improve readability. E.g. *Rotalia* and *Miliammina* are typical of the Marsh environment, e.g. *Lagena* and

Elphidium are found mainly in the Bay, *Virgulina* and *Bolivina* on the inner shelf, *Bulimina* and *Cassidulina* in deeper water.



Principal coordinates analysis

The idea of placing the samples in the ordination plot so that similar samples are close, can be generalized to any measure of sample distance. This leads to principal coordinates analysis (PCoA), which attempts to make the (Euclidean) distance between any pair of points proportional to sample distance. We have large flexibility in the choice of distance (or similarity) measure, and different people have different favorites.

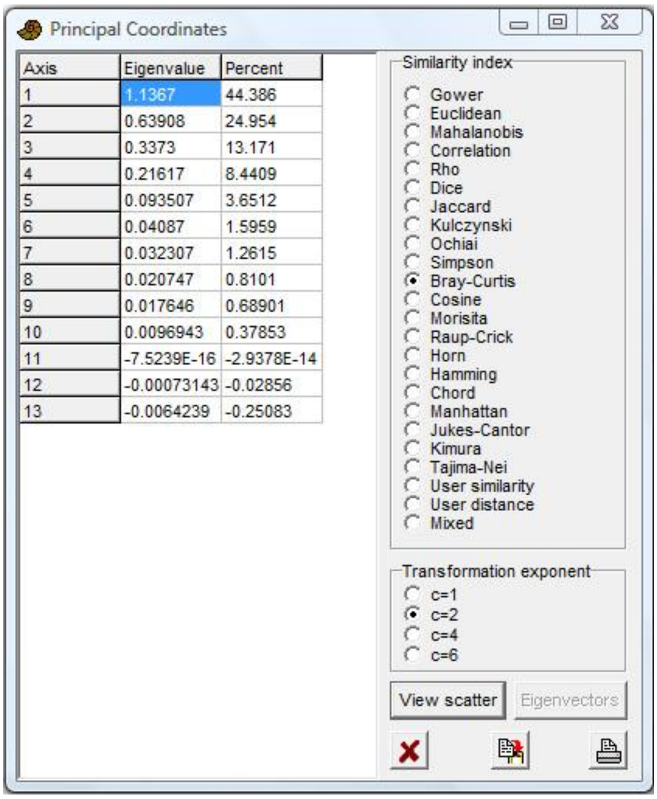
In marine ecology, the Bray-Curtis distance is now the “default” choice. The Bray-Curtis distance between samples *j* and *k* is defined as follows (the sums indexed by *i* go over all taxa):

$$d_{jk} = \frac{\sum_i \left| x_{ji} - x_{ki} \right|}{\sum_i \left(x_{ji} + x_{ki} \right)}.$$

For binary (presence-absence) data, the Dice similarity is a good start. The Dice similarity puts more weight on joint occurences than on mismatches. When comparing two samples, a match is counted for all taxa with presences in both samples. Using *M* for the number of matches and *N* for the the total number of taxa with presence in just one row, we have

$$d_{jk} = 2M / (2M+N).$$

Select all, and run ”Principal coordinates” in the Multivar menu. Then select the Bray-Curtis similarity index. Similarly to correspondence analysis, each ordination axis has an associated eigenvalue. The first two axes “explain” 69% of the total variation, which is again impressive considering the dimensionality reduction.



Axis	Eigenvalue	Percent
1	1.1367	44.386
2	0.63908	24.954
3	0.3373	13.171
4	0.21617	8.4409
5	0.093507	3.6512
6	0.04087	1.5959
7	0.032307	1.2615
8	0.020747	0.8101
9	0.017646	0.68901
10	0.0096943	0.37853
11	-7.5239E-16	-2.9378E-14
12	-0.00073143	-0.02856
13	-0.0064239	-0.25083

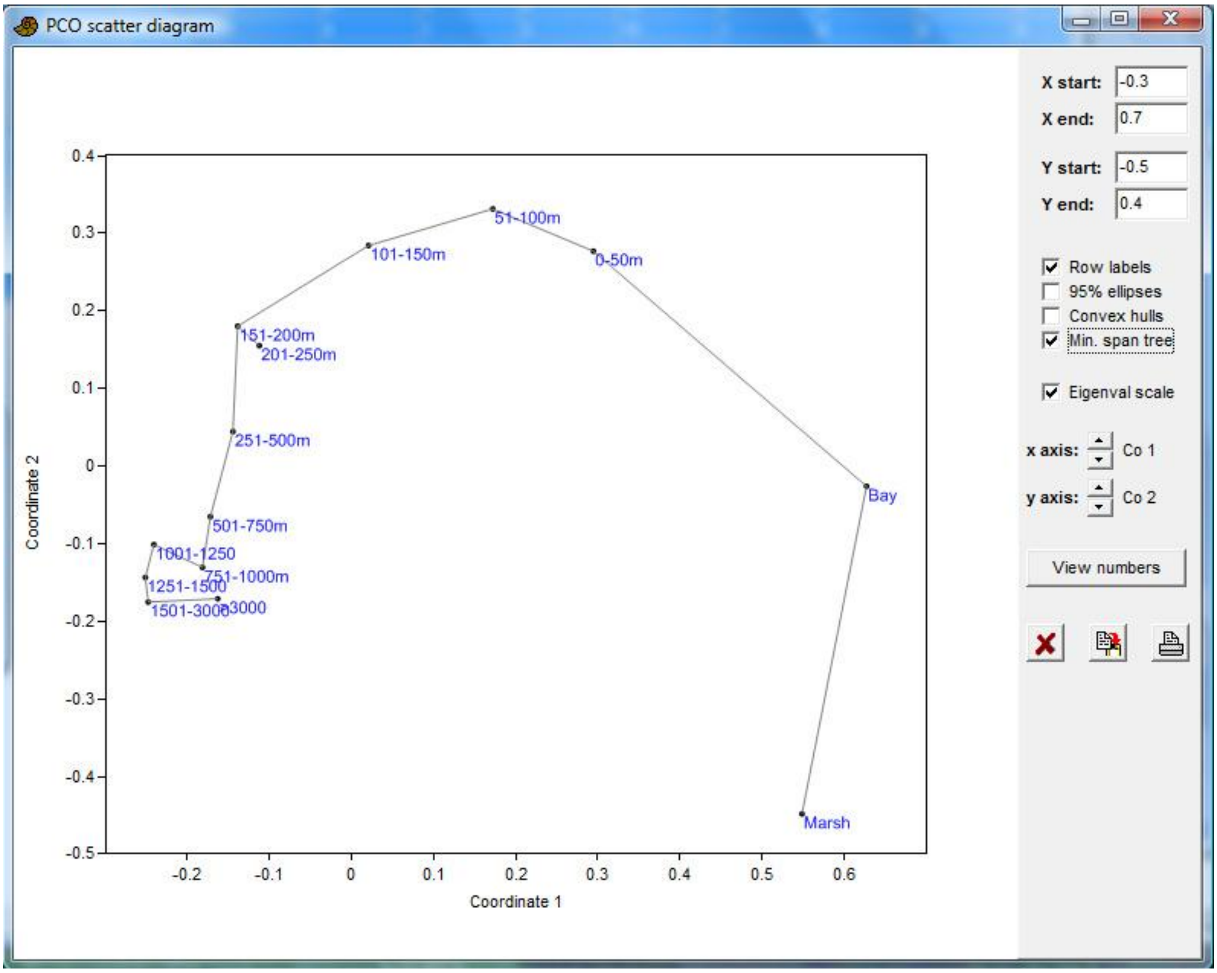
Similarity index

- ○ Gower
- ○ Euclidean
- ○ Mahalanobis
- ○ Correlation
- ○ Rho
- ○ Dice
- ○ Jaccard
- ○ Kulczynski
- ○ Ochiai
- ○ Simpson
- ● Bray-Curtis
- ○ Cosine
- ○ Morisita
- ○ Raup-Crick
- ○ Horn
- ○ Hamming
- ○ Chord
- ○ Manhattan
- ○ Jukes-Cantor
- ○ Kimura
- ○ Tajima-Nei
- ○ User similarity
- ○ User distance
- ○ Mixed

Transformation exponent

- ○ c=1
- ● c=2
- ○ c=4
- ○ c=6

Click the “View scatter” button to see the ordination results. The samples are clearly placed along a depth gradient, but forming a large arch instead of a straight line. The order of samples along the gradient is further emphasized by plotting the “minimal spanni ng tree”, which is a set of lines connecting all the dots so that the total length is as small as possible, measured with the selected index (Bray-Curtis) in the fully dimensional data set. We see that except for the 201-250m sample, the depth gradient is captured perfectly.



X start: -0.3
X end: 0.7
Y start: -0.5
Y end: 0.4

☑ Row labels
□ 95% ellipses
□ Convex hulls
☑ Min. span tree

☑ Eigenval scale

x axis: Co 1
y axis: Co 2

Non-metric multidimensional scaling

PCoA is now in relatively little use compared with a conceptually similar method called non-metric multidimensional scaling (NMDS). This method attempts to place the points in a two- or three-dimensional coordinate system such that the *ranked differences* are preserved. For example, if the original distance between points 4 and 7 is the ninth largest of all distances between any two points, points 4 and 7 will ideally be placed such that their Euclidean distance in the ordinated 2D plane or 3D space is still the ninth largest. NMDS intentionally does not take absolute distances into account. It usually performs better than PCoA.

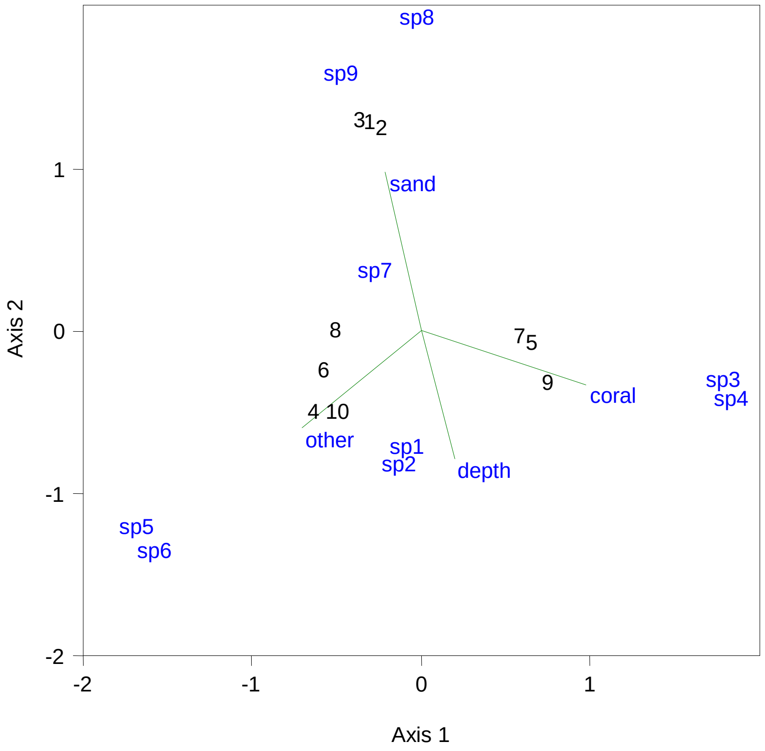
Because there is no closed algebraic solution to this problem, the computer must proceed by trial and error. The program may converge on a different solution in each run, depending upon the random initial conditions. Each run is actually a sequence of 11 trials, from which the best one is chosen. One of these trials uses PCoA as the initial condition.

Select all, and run “Non-metric MDS” from the Multivar menu. You are asked to select a similarity measure – try Bray-Curtis.

The “Shepard plot” of obtained versus observed (target) ranks indicates the quality of the result. Ideally, all points should be placed on a straight ascending line (*x*=*y*). The “stress” value should be small, at least less than 0.20 and ideally less than 0.10. In this case the result is excellent (stress 0.05), showing that the reduction to two dimensions implies very little loss of information.

Canonical correspondence analysis

If we have independent measurement of environmental variables (temperature, pH, substrate type etc.) it is possible to *constrain* the analysis so that the ordination axes are linear combinations of these variables. This can give precise visualization of how the environment controls the faunal gradients. Canonical correspondence analysis (CCA) is one such method.



In the example above, three different sets of items are shown in the same plot (a “triplot”). The taxa are sp1 to sp9. The samples are numbered 1-10. The environmental variables (depth and substrate type) are shown as lines (vectors) from the origin. For example, samples 1-3 are sandy, shallow, and characterized by taxa sp8 and sp9.

CCA is now very popular in ecology, but less often used in paleontology because we do not have independent environmental information (in contrast, the task is often to reconstruct environment based on fossil data).

References

Legendre, P. & Legendre, L. 1998. Numerical Ecology. Elsevier.