7. Principal Coordinate Analysis

FISH 560: Applied Multivariate Statistics for Ecologists

Topics

- Principal coordinate analysis (PCoA)

R Packages: vegan
R Source: biostats

BACKGROUND

Principal coordinate analysis (PCoA), also known as metric multidimensional scaling, is a generalized variant of PCA in that it uses a linear mapping of the distances between objects onto the ordination space, and the algorithm attempts to explain most of the variance in the original data set (Gower 1966). As opposed to PCA, PCoA allows for the ordination of data based on any association coefficient (not just correlation/covariance). Flexibility in the choice of similarity/distance matrix translates to flexibility in the application of PCoA. This is particularly useful for species and mixed data types, where PCA is not appropriate. Although there are important differences between PCoA and PCA, the eigenanalysis underpinning both approaches are very similar. For example, a PCoA ordination based on Euclidean distances among objects is identical to that of a PCA based on a covariance matrix among variables. Despite its usefulness, PCoA is rarely used by ecologists.

One minor drawback of PCoA is that doesn't provide a direct link between the components and the original variables and so the interpretation of variable contribution may be more difficult. This is because PCoA components, instead of being linear combinations of the original variables as in PCA, are complex functions of the original variables depending on the selected dissimilarity measure. Besides, the non-Euclidean nature of some distance measures does not allow for a full representation of the extracted variation into a Euclidean ordination space. In that case, the non-Euclidean variation cannot be represented and the percent of total variance cannot be computed with exactness. The choice of the dissimilarity measure is thus of great importance, and subsequent transformation of the data to correct for negative eigenvalues is sometimes necessary (see Legendre & Legendre, 1998, section 9.2.4. for how to correct for such negative eigenvalues). Although there is no direct, linear relationship between the components and the original variables, it is still possible to correlate object scores on the main axis (or axes) with the original variables to assess their contribution to the ordination.

SET-UP

In this exercise you will be working with the MAHA species abundance 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`. Import the dataset by typing:

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

Let's transform the data before diving into the analysis. We will do this because the species abundance dataset is highly skewed and contains some rather large values which are valid but highly influential. Remember, the log of zero is undefined so we'll add 1 to each value in our data set.

```
speabu.log<-log(speabu+1)
```

PRINCIPAL COORDINATE ANALYSIS

Calculating the distance matrix

The first step in PCoA is the selection of a distance matrix to describe similarities among objects according to the descriptors. Again, any (dis) similarity coefficient can be selected. For today, let's generate a distance matrix using an appropriate dissimilarity coefficient for species abundance – Bray-Curtis coefficient. We'll use the vegdist() function. Its usage is:

vegdist (x, method, binary=FALSE)

where

- x is a data frame (here a sitexspecies abundance matrix)
- method is the desired similarity coefficient. Bray-Curtis ("bray"), Jaccard's ("jaccard"), Gower's ("gower") and several other similarity coefficients are available
- binary; logical. If equals TRUE, the original data set is converted to presence/absence (1, present; 0, absent) and then the desired similarity coefficient is applied. Default: FALSE.

For additional documentation type **?vegdist**.

Type,

```
speabu.d<-vegdist(speabu.log, "bray")
```

Perform the PCoA

Now we can perform PCoA. In vegan, the function is called cmdscale(),which refers to classic (metric) multidimensional scaling. Its usage is:

cmdscale(d, k, eig = FALSE, add = FALSE)

where

- d is a dissimilarity object (generated by dist or vegdist)
- k is the number of principal components (PC) that should be extracted from the distance matrix (max number = min(col, rows)-1)
- eig, logical. If TRUE eigenvalues for each PC are retuned. Default: FALSE.
- add, logical. If TRUE a constant is added to each value in the dissimilarity matrix so that the resulting eigenvalues are non-negative. Default: FALSE.

Let's perform PCoA:

```
spe.pcoa<-cmdscale(speabu.d, eig=TRUE, add=T)
```

Note that k is set to 35 (# species (columns) – 1). Type **`spe.pcoa`** at the prompt to view the returned principal scores (PCs) and eigenvalues.

The principal scores are contained in spe.pcoa$points and the eigenvalues are contained in spe.pcoa$eig.

The principal scores for the first 5 observations should read:

```
            [,1]        [,2]         [,3]         [,4]         [,5]
S1  -0.379962052 -0.29157310  0.008993613 -0.002743771  0.191458393
S2  -0.301474146 -0.29455879  0.388392767  0.011907596 -0.180354539
S3  -0.117052609 -0.22344133  0.044101350 -0.072273726  0.253420678
S4   0.662126194 -0.17361377 -0.208290288  0.155837150 -0.064838489
S5  -0.115646174 -0.31700403 -0.323314608  0.162097678 -0.006921289
```

The eigenvalues for the first should read:

```
[1] 5.2402590 3.6946927 3.1018646 2.3737229 2.0522328 1.6856022 1.5614293
[8] 1.3452143 1.1556406 1.1417494 1.0759328 1.0231588 0.9406274 0.8869959
[15] 0.7093863 0.6563471 0.6343436 0.5251099 0.4788888 0.4456985 0.4217005
[22] 0.4165729 0.3806084 0.3529408 0.3108155 0.2931376 0.2767940 0.2492598
[29] 0.2369490 0.2134872 0.2104035 0.1760313 0.1647904 0.1488193 0.1395068
```

The percent of variation explained by any principal coordinate can be calculated by dividing its eigenvalue by the sum of the eigenvalues across all PCs. Recall, the same approach is used in PCA. Let’s calculate this for the first five PCs.

```
spe.pcoa$eig/sum(spe.pcoa$eig)*100
```

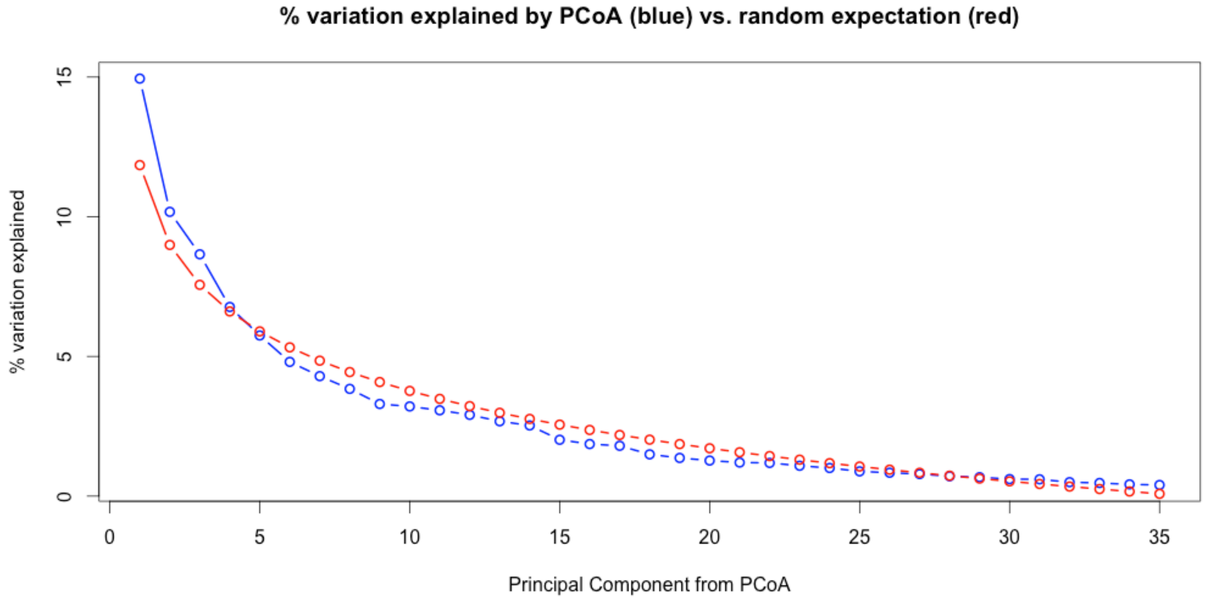
This should return the following values for the first 5 principal components:

```
[1]  1.494490e+01  1.017524e+01  8.655463e+00  6.769729e+00  5.749803e+00
```

Let’s compare the eigenvalues to expectations according to the broken stick model.

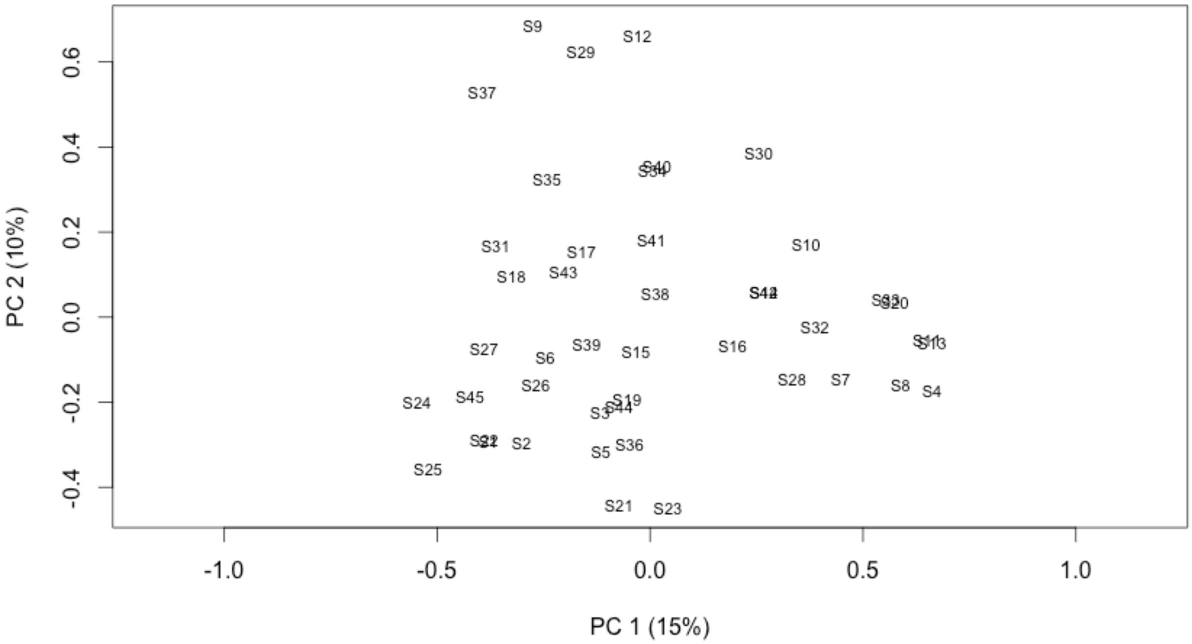
```
plot(spe.pcoa$eig[1:35]/sum(spe.pcoa$eig)*100,type="b",lwd=2,col="blue",xlab=
"Principal Component from PCoA", ylab="% variation explained", main="%
variation explained by PCoA (blue) vs. random expectation (red)")

lines(bstick(35)*100,type="b",lwd=2,col="red")
```



Now, let’s view the ordination plot:

```
ordiplot(spe.pcoa, choices = c(1, 2), type="text", display=”sites”, xlab=”PC
1 (15%)”, ylab=”PC 2 (11%)”)
```



This plot represents each of the sites in 2-D ordination space (x-axis = principal component 1, y-axis = principal component 2).

Calculate the PC loadings (i.e., variable weights)

Now, to calculate and depict species loadings (i.e., principal weights in the eigenvectors) on each principal coordinate we’ll use the function envfit() along with the PC scores from our PCoA object. For brevity, please see additional documentation and usage information in the vegan manual (type ?envfit, ?scores). The function envfit() simply performs a linear correlation analysis based on standardized data (in other words, a simple linear regression) between each of the original descriptors (i.e., species) and the scores from each principal component. A permutation test is used to assess statistical significance, rather than using the F distribution.

```
vec.sp<-envfit(spe.pcoa$points, k=45, speabu.log, perm=1000)
```

Another way of doing this is to extract the first two PCs from our PCoA object using the scores() function. To do this, type:

```
vec.sp<-envfit(scores(spe.pcoa), speabu.log, perm=1000)
```

This should return the values (note yours will be slightly different because it is based on random permutations) listed below for the first 7 species:

```
vec.sp
```

```
***VECTORS
```

```
                Dim1       Dim2     r2  Pr(>r)
BANDDART -0.183732  0.982976 0.3384 <0.001 ***
BANDSCUL -0.056155  0.998422 0.0598  0.262
BLACDACE  0.774004 -0.633181 0.6104 <0.001 ***
BLUECHUB -0.690920 -0.722931 0.7034 <0.001 ***
BLUEGILL -0.853874 -0.520480 0.1510  0.029 *
BLUNMINN -0.422624  0.906305 0.1471  0.025 *
BROOTROU  0.856519 -0.516115 0.0115  0.807
CCHUBSUC -0.740762 -0.671768 0.0412  0.451
COMMSHIN -0.971232  0.238134 0.1038  0.098 .
CREECHUB  0.932674  0.360719 0.0047  0.901
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
P values based on 1000 permutations.
```

Based on these results you will notice that banded darter (BANDDART), black dace (BLACDACE), blue chub (BLUECHUB), bluegill (BLUEGILL) and bluntnose minnow (BLUMINN) show statistically significant loadings on the first two principal components. These species could be used to interpret the position of the stream sites (objects) in ordination space. You can drill down into the sub-objects contained in vec.sp, by typing:

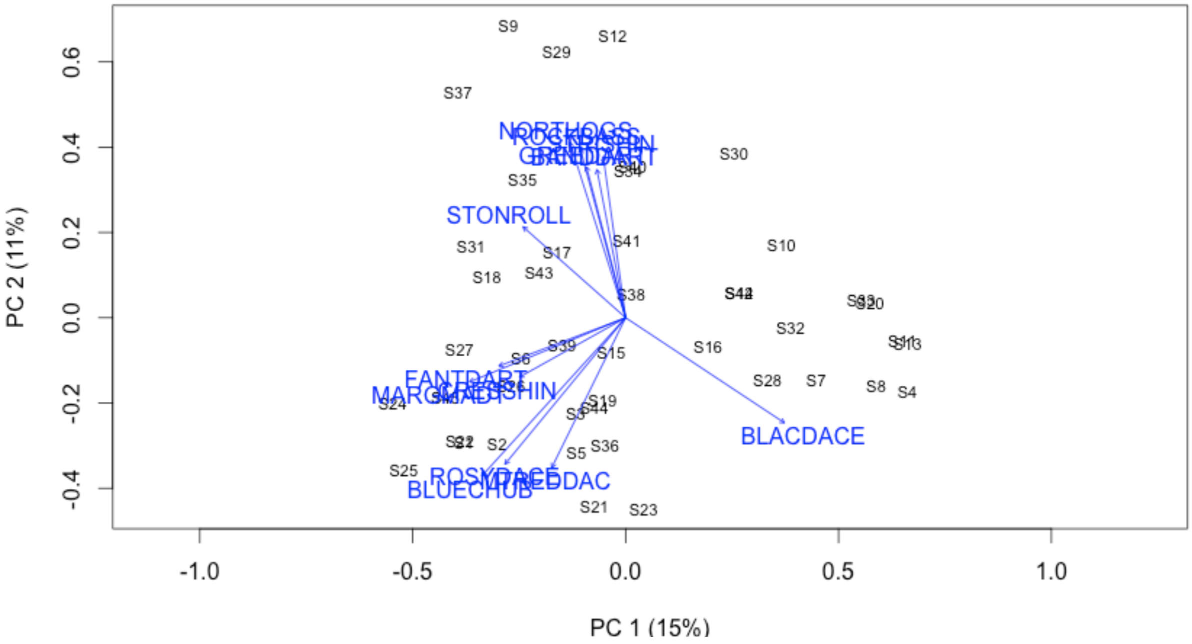
```
names(vec.sp)
names(vec.sp$vectors)
```

You will see that the normalized eigenvectors (scaled to unit length) are presented in the matrix vec.sp$vectors$arrows. These values should look familiar!

Now, plot the eigenvectors on the ordination plot.

```
ordiplot(spe.pcoa, choices = c(1, 2), type="text", display="sites", xlab="PC
1 (15%)", ylab="PC 2 (11%)")
plot(vec.sp, p.max=.01, col="blue")
```

where p.max is the significance level that the species occurrence data must have with either PC in order to be depicted (these p-values were presented in vec.sp).



Other options for conducting PCoA

It is common that multiple functions exist to perform the same multivariate analysis, and this is the case for PCoA. You can use the pcoa() function in the ape package. Be sure to install the library first.

Here is some example code:

```
spe.pcoa<-pcoa(speabu.d,correction="cailliez")
```

OPTIONAL READINGS

Gower JC (1966) Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika 53: 325-338.

Podani, J. 2005. Multivariate exploratory analysis of ordinal data in ecology: pitfalls, problems and solutions. J. Veg. Sci. 16: 497–510.

Podani, J. and Miklos, I. 2002. Resemblance coefficients and the horseshoe effect in principal coordinates analysis. Ecology 83: 3331–3343.