15-Multivariate Analysis of Ecological Community Data

2025-01-21

1 Introduction

The aim of the following exercise is to demonstrate some important multivariate methods by example of a simplified data set. In some cases, several alternatives are presented, but for a real analysis one does not need to include everything. On the other hand, the methods offer further possibilities that cannot all be presented, see the online help and corresponding books and tutorials.

The data set used originates from a field experiment to investigate the influence of fish predation on the species composition of the macrozoobenthos community in two streams. During the study period, however, an extreme flood occurred in August and caused major morphological changes to the two streams. This motivated the hypothesis that the species community has also changed.

2 Material and Methods

The data set originates from these two streams, "Gauernitzbach" and "Tännichtgrundbach") near Dresden. It contains a table of macrozoobenthos taxa and a description of the site (t = Tännichtgrund, g = Gauernitz, p = pool, r = riffle) and the time before (v) and after (n)) the flood. The data set is an unpublished teaching example with a selected number of time points and where the relatively large number of species was pragmatically aggregated to a small number of taxa. More about the experiment can be found in (Winkelmann et al., 2008, 2011).

The data set is available from https://tpetzoldt.github.io/datasets/data/gauernitz.csv, together with a file description.

After reading the data in we assing row names to the data frame from the `site` column that contains short codes for the observarions, e.g. `GP9` than means Gauernitzbach, pool, September. These rownames are very useful to indicate the observations in the plots.

In a second step, we split the original table in two separate tables `bio` with the species data only and `env` with the environmental factors. For the calculations, it is important that the `bio`-table contains only numbers. The data analysis is carried out with the **vegan** package (Oksanen et al., 2024).

```
library("dplyr")
library("vegan")
dat <- read.csv("gauernitz.csv")

## assign rownames to dataframe
row.names(dat) <- dat$site

## separate into two matrices, bio und env
env <- dat |> select(Habitat, Bach, Hochwasser)
bio <- dat |> select(Mollusca, Diptera, Baetis, Plecoptera, Coleoptera,
             Turbellaria, Heptageniidae, Ephemeroptera, Gammarus,
             Trichoptera, Acari, Nematoda, Oligochaeta)
```

After loading the data, it is always important to have a look at the data structures, for example with `str(bio)`, `str(env)` or in the "Global Environment" pane of RStudio.

3 Data Analysis

3.1 NMDS

We start with an NMDS (nonmetric multidimensional scaling), of the `bio`-data using the Bray-Curtis dissimilarity measure. It is the default, but we specify it here explicitly to make this transparent. Automatic transformation is switched off. This can be changed, depending on the properties of the data, for example to `wisconsin(sqrt(bio))`.

The function `metaMDS` then runs the NMDS several times with different starting values to avoid local minima. For difficult data sets, it may be necessary to increase the metaparameters `try` and `trymax`.

After that, we should have a look at the value of the stress and the stressplot.

```
mds <- metaMDS(bio, distance = "bray", autotransform = FALSE)
mds
stressplot(mds)
```

We can then plot the results of the NMDS-ordination.

```
plot(mds, type = "t")
```

In order to show the influence of environmental variables, we can fit vectors or factors to the ordination. In addition to this, we can show the significance of the fitted vectors. For getting reliable p-values, I recommend to increase `permu` to 3999 or 9999.

```
## Environmental fitting mit Permutations-Test
efit <- envfit(mds ~ Hochwasser + Bach + Habitat, env, permu = 999)
efit
```

Now, we can visualize the complete result.

```r
plot(mds, type = "t")
plot(efit, add = TRUE)
abline(h=0, col="grey", lty="dotted")
abline(v=0, col="grey", lty="dotted")
```

3.2 Hierarchical Clustering

The NMDS tries to project the distances as good as possible to a low number of dimensions, e.g. `k=2` that is the default. To see the full picture of distances in space, we may consider to apply hierarchical clustering. As agglomeration algorithm, we may consider `complete`, `ward.D2` or `ward.D`.

```r
hc <- hclust(vegdist(bio), method="ward.D")
plot(hc)
```

It is also possible, to colorize the clusters in the NMDS plot. Let's assume we have 4 clusters, we can first indicate it in the hierarchical cluster tree with `rect.hclust``, then cut the tree with`cutree‘.

```r
plot(hc)
rect.hclust(hc, 4)
```

The result is an assignment of the original observations to groups, that can be used to colorize the NMDS plot. It is possible to show the cluster tree directly in the nmds plot or to indicate it otherwise with, for example, `ordispider`, `ordihull` or `ordiellipse`

```r
grp <- cutree(hc, 4)  # assign observations to 4 groups
grp

plot(mds, type = "n")
text(mds$points, row.names(bio), col = grp)

## optional: show cluster tree
#ordicluster(mds, hc, col="blue")
```

Exercise: Include the fitted environmental variables.

3.3 Canonical Correspondence Analysis

As an alternative to NMDS, we can also use CCA, that is a "constraied ordination method" and allows a more detailed numerical analysis (e.g. separatation of inertia), but is limited to χ^2-distance, while NMDS allows arbitrary distance measures, including Bray-Curtis.

```r
cc <- cca(bio ~ Habitat + Bach + Hochwasser, data = env)
#cc <- cca(bio ~ ., data = env) # same. The . means all from env
cc # print Eigenvalues
plot(cc)
ordihull(cc, env$Habitat, col = "blue")   # or: ordispider, ordiellipse ...
```

3.4 Test of significance

The CCA supports also significance tests and model selection.

```
## Resampling-ANOVAs of the CCA
anova(cc)
anova(cc, by = "terms") # most useful
anova(cc, by = "axis")

## Model selection to find the optimal model
step(cc)
```

Several other multivariate significance tests exist. The Adonis-Test is in particular popular, because it considers also interaction terms. It does not rely on an NMDS or CCA and works directly with a distance matrix. In order to increase its power, it may optionally consider strata. The following shows some examples.

Exercise: Try different model formulae and decide which one is most appropriate for the data set and the original hypothesis.

```
dist <- vegdist(bio, method = "bray")

adonis2(dist ~ Hochwasser * Habitat * Bach, data=dat)

## Comparison with and without strata
adonis2(dist ~ Hochwasser * Bach, strata = env$Habitat, data=dat)
adonis2(dist ~ Hochwasser * Bach, data=dat)
```

3.5 dbRDA and elimination of covariates

The following examples show further possibilities. Instead of a CCA (that uses χ^2) we can also use a so-called distance-based redundancy analysis (dbRDA), that supports arbitrary distance measures, e.g. Bray-Curtis.

Another option is a partial CCA or partial dbRDA where we can eliminate covariates (`condtion = ...`) that we are not much interested in, so that the ordination focuses on the variables we are interested in. This is the called a partial analysis (pCCA, p-dbRDA). We will then also get three kinds of eigenvalues and eigenvectors (components of the inertia).

```
## ============================================================================
## dbRDA, supports arbitrary dissimilarity measures
## ============================================================================

dbr <- dbrda(dist ~ Habitat * Bach * Hochwasser, data = env, distance="bray")
dbr
anova(dbr, by="terms", permutations=3999)
#summary(dbr)
plot(dbr)

## ============================================================================
## partial CCA: elimination of covariates
## ============================================================================
```

```r
pcc <- cca(bio ~ Hochwasser + Bach + Condition(Habitat), data = env)
pcc
plot(pcc)


## ================================================================================
## partial dbRDA
## ================================================================================
dbrc <- dbrda(bio ~ Hochwasser + Bach + Condition(Habitat),
              data = env, distance = "bray")
dbrc
plot(dbrc)
```

Exercise: Apply a method that eliminates the differences between the streams and investigate whether pools and riffles behave differently.

3.6 Procrustes test

To compare the ordinations, that we get with a different set of environmental variable and conditions, we can use the so-called Procrustes test.

```r
dbr  <- dbrda(bio ~ Hochwasser + Bach + Habitat,
              data = env, distance="bray")
pdbr <- dbrda(bio ~ Hochwasser + Bach + Condition(Habitat),
              data = env, distance="bray")

proc <- procrustes(dbr, pdbr)
plot(proc, type = "t")
protest(dbr, pdbr)
```

References

Oksanen, J., Simpson, G. L., Blanchet, F. G., Kindt, R., Legendre, P., Minchin, P. R., O'Hara, R. B., Solymos, P., Stevens, M. H. H., Szoecs, E., Wagner, H., Barbour, M., Bedward, M., Bolker, B., Borcard, D., Carvalho, G., Chirico, M., De Caceres, M., Durand, S., … Weedon, J. (2024). *Vegan: Community ecology package*. https://CRAN.R-project.org/package=vegan

Winkelmann, C., Hellmann, C., Worischka, S., Petzoldt, T., & Benndorf, J. (2011). Fish predation affects the structure of a benthic community. *Freshwater Biology*, *56*(6), 1030–1046. https://doi.org/https://doi.org/10.1111/j.1365-2427.2010.02543.x

Winkelmann, C., Petzoldt, T., Koop, J. H., Matthaei, C. D., & Benndorf, J. (2008). Benthivorous fish reduce stream invertebrate drift in a large-scale field experiment. *Aquatic Ecology*, *42*, 483–493. https://doi.org/https://doi.org/10.1007/s10452-007-9101-7