Multivariate Statistics: Exercise 4

Eduard Szöcs (szoecs@uni-landau.de)

January 20, 2016

1 Part I - Constrained Ordination

This exercise is a follow up to exercise 3.
Condensing all the community data into one number (like the Shannon index) comes along with loss of information, but simplicity. However, we have abundance data at hand and we should use this in the most efficient way.
In this exercise we'll look how community composition changes with salinity and pesticide load.
"RDA is an extremely powerful tool in the hand of ecologists [...]" [1]. Unfortunately RDA should not directly be used with abundance data because of the double-zero problem (euclidean distance) and non-linear responses along long gradients.
Legendre and Gallagher (2001) [2] solved this problem : "[..] especially since the introduction of the Legendre and Gallagher (2001) transformations that open RDA to the analysis of community composition data tables (transformation-based RDA or tb-RDA)."
This exercise is about transformation-based RDA (tbRDA) and its application to the Australia data.

1.1 Tasks

q1_1 Read both data files (abudata and envdata) into R.

q1_2 Transform the abundance data using a Hellinger transformation!

Do not forget to exclude the three ID columns.

q1_3 Fit a RDA model to this hellinger tranformed data.

Use conductivity (`log_Cond`), pesticide load (`log_maxTU`), as well as the first two PCA axes from exercise 3 as predictors!

Here is the code to get you started:

We run a PCA on the environmental data, extract the site scores and create a new data.frame with the predictor variables:

```
# PCA from exercise 3
PCA <- rda(envdata[ , -c(1:3, 25, 19)], scale = TRUE)
# extract site scores
sc <- scores(PCA, scaling = 1, display = 'sites')
# create new predictor data.frame
pred_data <- data.frame(log_Cond = envdata$log_Cond, log_maxTU = envdata$log_maxTU, sc)
str(pred_data)

## 'data.frame': 69 obs. of  4 variables:
##  $ log_Cond : num  2.06 1.95 1.98 1.86 1.88 ...
##  $ log_maxTU: num  -2.15 -5.14 -2.63 -5.14 -2.61 ...
##  $ PC1      : num  -0.601 -0.652 -0.498 -0.423 -0.573 ...
##  $ PC2      : num  -0.3664 -0.3518 -0.171 -0.2149 -0.0238 ...
```

Now compute with this the RDA!

q1_4 Create a triplot of the resulting PCA. Use a symmetric scaling (scaling = 3).

q1_5 Figure 1 shows the resulting triplot.

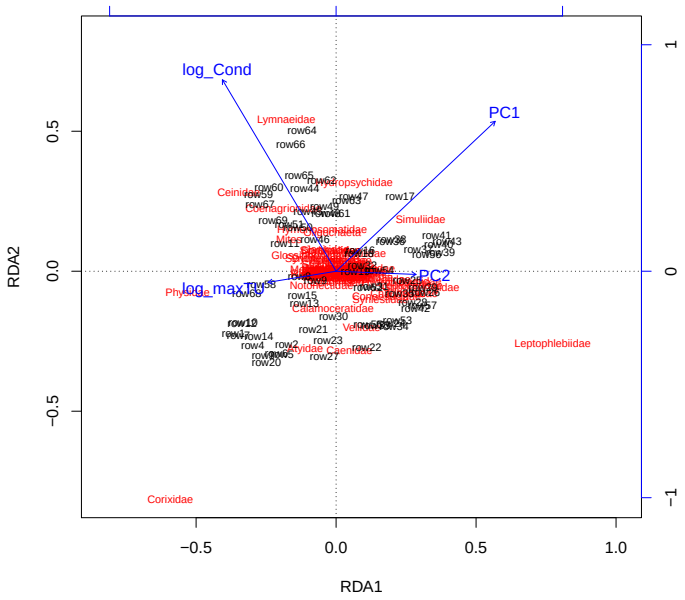

Figure 1: Triplot of the Australia data

Describing the plot, which of the following statements is correct?

1. Conductivity and PC1 are the two most influential predictors.
2. Lymnaeidea are mainly found at sites with high electrical conductivity.
3. The abundance of Simulidae is affected by hydrology.
4. The sample in row 64 has higher salinity then the sample in row 22.
5. Physidae can tolerate a relatively high amount of toxicity.

One or more statements are correct.

1_6 What proportion of total variance can be explained by the predictor variables? Give your answer in percent and round to integer (0.3916 = 39%)!

```
e.g. 39
```

1_7 What proportion of (total) variance that can be explained by the predictors is displayed on the first axis? Give your answer in percent and round to integer (0.3916 = 39%)!

```
e.g. 39
```

1_8 Perform a permutational significance test of the RDA axes. Which axes displays a significant amount of variation?

1. None.
2. Axis 1.
3. Axes 1 and 2.
4. Ax2s 1, 2 and 3.
5. All RDA axes.

One statement is correct.

1_9 Perform a permutational significance test of the predictor variables (use a marginal test (`by = 'margin'`)). Which predictor explains most of the variation in the data?

1. None.
2. log_Cond

3. log_maxTU
4. PC1
5. PC2

One statement is correct.

References

[1] Daniel Borcard, Francois Gillet, and Pierre Legendre. *Numerical Ecology with R (Use R)*. Springer, 1st edition. edition, 2011.

[2] P. Legendre and E. D. Gallagher. Ecologically meaningful transformations for ordination of species data. *Oecologia*, 129(2):271–280, 2001.