Multivariate Statistics: Exercise 3

Eduard Szöcs (szoecs@uni-landau.de)

January 20, 2016

1 Part I - Principal component analysis

In this exercise we will analyse a dataset from Australia collected by R. Schäfer and colleagues [1]. The focus of this study was to analyse the effects of pesticide toxicity and salinisation on macroinvertebrate communities [2].
Macroinvertebrates, pesticides and other environmental variables were sampled at 24 sites situated in a 120km radius around Melbourne on three sampling occasions. These sites covered a gradient of both, pesticide exposure and salinisation. Pesticide toxicity was expressed in terms of Toxic Units (TU) with respect to *Daphnia magna* and salinity in terms of electrical conductivity ($\mu S/cm$ at 25°C).
You can find the data in the *data* folder of this exercise. There are two data files:

envdata.csv 22 measured environmental variables at the sites. Some of the variables have already been transformed. See Table 1 for details.

abudata.csv Counts of 75 taxa collected during the study - mostly on family level.

The first three columns are the same in both files: `Site`, `Month` and `Site_Month` are ID variables - they code uniquely each sample. These should not be included in your analysis, but are useful to join both tables.

Table 1: Overview of environmental variables.

Column	Variable	Unit	Transformation
`T`	Temperature	°C	-
`pH`	pH	-	-
`oxygen`	Dissolved oxygen	% sat.	-
`Depth`	stream depth	m	-
`max_width`	maximum stream width	m	-
`min_width`	minimum stream width	m	-
`rif_prec`	Pool	%	-
`pool_perc`	Riffle	%	-
`Bedrock`	Bedrock	%	-
`Boulder`	Boulder (>25.6 cm)	%	-
`Cobble`	Cobble (6.4 - 26.5 cm)	%	-
`Pebble`	Pebble (1.6 - 6.4 cm)	%	-
`Gravel`	Gravel (0.2 - 1.6 cm)	%	-
`Sand`	Sand (0.06 - 0.2 cm)	%	-
`Clay.silt`	Clay (<0.06 cm)	%	-
`log_Cond`	Conductivity	uS / cm	log10
`log_Nh4`	Ammonia	mg / L	log10
`log_NO2`	Nitrite	mg / L	log10
`log_NO3`	Nitrate	mg / L	log10
`log_PO4`	Phosphate	mg / L	log10
`log_Turb`	Turbidity	NTU	log10
`log_maxTU`	Maximum TU	$TU_{D.manga}$	log10

1.1 Tasks

q1_1 Read both data files into R and name them according to their file name! Check that all variables (except the three ID variables) are either numeric or integer.

The first step of an analysis is to get an impression about the data (or the conditions at the sampling sites), the relationships between variables and the main gradients in the data set.
Since we know that there is a gradient of salinity and pesticide exposure (due to experimental design), we are interested which other gradients may be present in the data set.

q1_2 If you run a PCA on a data set with 20 variables: how many axes will the resulting PCA have? Give your answer as integer.

```
e.g. 2
```

q1_3 Conduct a PCA on the environmental data set, excluding the variables `log_Cond` and `log_maxTU`. Note, that we are interested in the correlation between variables. Therefore, we set `scale = TRUE` to scale the variables (they are measured on different scales!) which gives us correlations.

Create a temporary data.frame, excluding the ID variables, `log_Cond` and `log_maxTU`. Run a PCA on this temporary data.frame.

q1_5 Create a correlation biplot of the PCA from **q1_3**!

q1_6 Figure 1 shows a correlation biplot of the environmental variables.

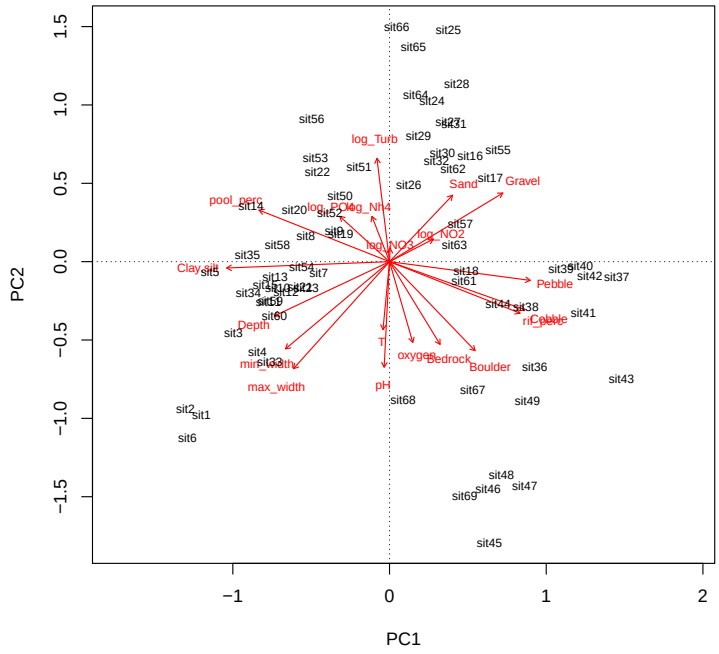

Figure 1: Correlation biplot of environmental variables

Which of the following statements is correct?

1. Temperature (`T`) and `pH` are correlated.
2. `Clay.silt` and `log_Turb` are negatively correlated.
3. The proportion of riffles (`rif_perc`) and pools (`pool_perc`) are negatively correlated.
4. `Clay.silt` and Temperature (`T`) are not correlated.
5. Sand and Gravel are not correlated.

One or more statements are correct.

q1_7 Figure 1 shows a correlation biplot of the environmental variables. Which of the following statements is correct?

1. The first axis could be interpreted as a gradient in hydrological conditions.
2. The first and second axis are not correlated.

3. The second axis is a gradient of chemical conditions.
4. `Clay.silt` is the most important variable for the first axis.

One or more statements are correct. G

q1_8 What proportion of variance can be explained by the first two axis? Give your answer in percent (0.39 = 39%) and round to one digit!

```
e.g. 38.1
```

2 Part II - PCA regression

We are interested which variables drive the diversity of macroinvertebrates. Our main hypotheses are that salinity and pesticide toxicity may affect diversity. Nevertheless, also other variables may have an impact on diversity.
The data set is very small and consists of only 69 observations [1]. Therefore, we cannot fit a model containing all variables (Note, that you should have at least 10 observations per variable).
One way to deal with this issue, is to reduce the dataset into fewer variables. If PCA-axes describe interpretable gradients, we could use the axes as surrogates for the variables that load on these. PCA axis are orthogonal to each other and therefore, we also have no issue with collinearity.

2.1 Tasks

q2_1 First need to quantify macroinvertebrate diversity in our samples. There are many possibilities - an easy measure is the number of species, but there are also other diversity indices. Use the `diversity()` function from the vegan packages to compute the Shannon diversity of the samples.

@ Don't forget to exclude the ID variables, before computing diversity!

q2_2 Using the broken stick criterion - how many axes would you extract from the PCA from **q1_3**? Give your answer as integer.

```
q2_2 <- 7
```

q2_3 Extract the site scores of the first two axes (You should use scaling = 1, as we are interested in the relationship between samples). These will be used as predictors in the next step.

q2_4 Create a new data.frame with the columns "shannon" (=shannon diversity from q2_1), "PC1" (=sites scores on first axis (from q2_3)), "PC2" (site scores on second axis), `"log_Cond"` (from envdata) and `"log_maxTU"` (from envdata).

Your data.frame should have exactly this structure:

```
## 'data.frame': 69 obs. of  5 variables:
##  $ shannon  : num  1.17 1.13 1.09 1.39 1.38 ...
##  $ PC1      : num  -0.601 -0.652 -0.498 -0.423 -0.573 ...
##  $ PC2      : num  -0.3664 -0.3518 -0.171 -0.2149 -0.0238 ...
##  $ 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 ...
```

q2_5 Build a linear model explaining shannon diversity with conductivity (`log_Cond`), salinity (`log_maxTU`), as well as the two extracted PCA axes. Use the formula interface together with the data.frame from **q2_4**. Use the order `log_Cond`, `log_maxTU`, PC1, PC2 in your formula.

2_6 Figure 2 shows the residuals vs. fitted values of the model from q2_5.

[1] which additionally may not be independent (temporal autocorrelation), but we will ignore this at the moment.

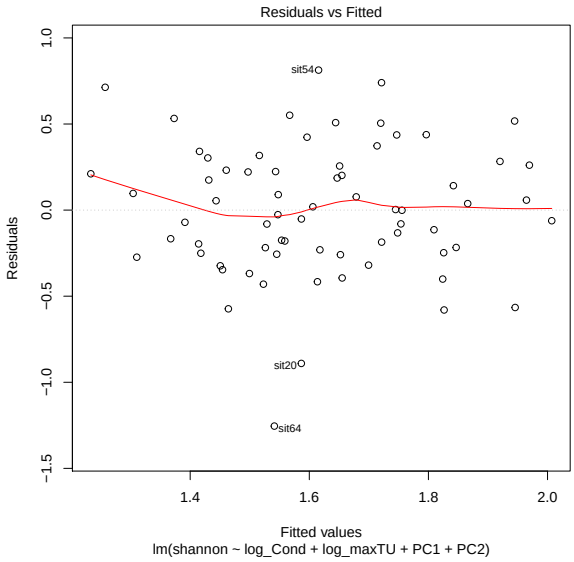

Figure 2: Residuals vs. fitted values of the model from q2_5.

Is the assumption of constant variances met? (yes = TRUE, no = FALSE)

```
TRUE or FALSE
```

2_7 Below is the summary-output of this model.

```
##
## Call:
## lm(formula = shannon ~ log_Cond + log_maxTU + PC1 + PC2, data = q2_4)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.25445 -0.24753 -0.00092  0.25610  0.81304
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.08297    0.24591   8.470 4.79e-12 ***
## log_Cond    -0.22128    0.07711  -2.870  0.00556 **
## log_maxTU   -0.02322    0.03650  -0.636  0.52689
## PC1          0.28597    0.13354   2.142  0.03605 *
## PC2          0.23844    0.17313   1.377  0.17323
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3923 on 64 degrees of freedom
## Multiple R-squared:  0.1881,Adjusted R-squared:  0.1374
## F-statistic: 3.708 on 4 and 64 DF,  p-value: 0.008905
```

Can the hypothesis that pesticides decrease macroinvertebrate diversity confirmed? (TRUE = yes, FALSE = no)

```
TRUE or FALSE
```

2_8 Which of the following statements is correct?

1. `Clay.silt` has a positive impact on diversity.
2. The model explains only 13.7% of the variance in the data.
3. Increasing conductivity decreases diversity.

4. There are 64 degrees of freedom, because there are 64 data points.
5. Hydrology influences diversity.

One or more statements are correct.

References

[1] Ralf B. Schäfer, Mirco Bundschuh, Duncan A. Rouch, Eduard Szöcs, Peter C. von der Ohe, Vincent Pettigrove, Ralf Schulz, Dayanthi Nugegoda, and Ben J. Kefford. Effects of pesticide toxicity, salinity and other environmental variables on selected ecosystem functions in streams and the relevance for ecosystem services. *Science of the Total Environment*, 415(1):69–78, 2012.

[2] E. Szöcs, B. J. Kefford, and Ralf B. Schäfer. Is there an interaction of the effects of salinity and pesticides on the community structure of macroinvertebrates? *Science of the Total Environment*, 437(1):121–126, 2012.