Multivariate analysis

December 5, 2018

Answers for this lab must be submitted before **December 12th at 5pm on Moodle**. In your answer for each question, please include a copy of the R code used (if applicable) and the results obtained.

1. Composition of mineral springs

The `springs.csv` dataset contains data from a study by Tanaskovic et al. (2012) on the physicochemical properties (temperature, pH, electrical conductivity) and ion concentrations of mineral springs in Serbia.

```
sources <- read.csv("springs.csv")
str(sources)
```

```
## 'data.frame':    30 obs. of  14 variables:
##  $ waterSource: Factor w/ 30 levels "Banjska","Bogutovacka",..: 9 8 10 19 24 26 2 14 21 16 ...
##  $ tempCels   : num  45 25 25 48 14.4 33 25 39.1 43 35 ...
##  $ pH         : num  7.7 7.3 6.6 7 7.5 7.4 7.2 7.6 7.5 7 ...
##  $ elecCond   : int  3770 5100 677 2219 1100 722 440 265 248 324 ...
##  $ totSolid   : num  4.173 6.2 0.641 2.119 1.1 ...
##  $ Ca         : num  5.6 24 120 60 56 60 50 62 74 102 ...
##  $ Mg         : num  3.8 24 36.6 18 48.8 67 61 14 17 32 ...
##  $ Na         : num  1157 476 200 2280 226 ...
##  $ K          : num  7.9 4 22.2 30 3.5 7.1 1.5 1.2 8 2.5 ...
##  $ Cl         : num  56.8 369 89 1560 198.8 ...
##  $ SO4        : num  2 2 55 2 82 90 15 40 20 18 ...
##  $ HCO3       : num  2904 829 890 3360 573 ...
##  $ SiO2       : num  36.7 25 18 20 22 10 20 18 18 25 ...
##  $ geoStruct  : int  1 1 1 1 1 2 2 2 2 2 ...
```

We first look at the 8 columns ranging from *Ca* to *SiO2*, which represent the concentration of calcium, magnesium, potassium, chloride, sulfate, bicarbonate and silicate ions (in this order), all measured in mg / L.

a) Since these 8 variables have the same units, it is not necessary to standardize them before performing multivariate analyzes. However, we will apply a logarithmic transformation to these data. Inspect the data and describe why this transformation is useful here.

Note: If a data frame consists only of numeric variables, you can apply the `log` function to the array to take the logarithm of all variables at once.

b) Conduct a principal component analysis of the transformed concentrations. What do you notice about the contributions of the variables on the first axis (PC1)? Is there a general property (not related to a particular variable) that distinguishes samples with negative and positive values on that axis?

c) Create a biplot for the 2nd and 3rd principal components. Which variables have the greatest effect on each of these axes? From this plot, give an example of a pair of ions whose concentrations are highly positively correlated, and of another pair whose concentrations are negatively correlated.

d) The `geoStruct` variable in the original data frame is a categorical variable representing the geological type of the site. How could you check if a principal component varies significantly between sites of different geological types? Perform this test separately for the PC1 and PC2 components and determine the nature of the significant differences, if any.

e) Create a scatterplot of the components PC1 and PC2 that also indicates the geological type for each site.

2. Composition of Arctic sediments

The `arctic.csv` dataset, from a study by Martinez et al. (2009), contains data on the composition of samples at different depths (`depth`, unspecified units) of a sediment core from the Arctic Ocean. For each sample, the columns *Al* to *P* indicate the percentage of the mass of the sample corresponding to that element.

```
arctic <- read.csv("arctic.csv")
head(arctic)
```

```
##    sampleID depth   Al    Ti   Fe    Mn   Ca   Mg   Na    K     P
## 1 4C-01H-1W  0.20 8.34 0.520 5.34 0.286 4.04 1.71 1.86 2.15 0.085
## 2 4C-01H-2W  2.36 7.56 0.539 5.75 0.307 0.50 1.48 1.85 2.07 0.084
## 3 4C-02H-2W  5.62 7.84 0.468 4.67 0.149 0.46 1.13 1.70 1.90 0.075
## 4 3A-01H-4W  8.52 8.58 0.503 5.06 0.288 0.50 1.17 1.85 2.08 0.093
## 5 4C-03H-3W 12.41 7.46 0.494 5.06 0.089 0.42 1.15 1.84 2.10 0.093
## 6 2A-05X-2W 20.83 8.49 0.520 6.21 0.097 0.40 1.13 1.74 1.99 0.127
```

a) Again, since the variables are on the same scale (% mass), there is no need to standardize them. Conduct a PCA with the original variables and another one with the standardized variables, then compare the biplots. Explain how the choice to standardize or not affects the importance of different elements on the results of the PCA.

b) From the standardized variables, create a hierarchical clustering using the complete linkage method and another using Ward's method. Overlay the classification into (i) 2 groups and (ii) 4 groups obtained by each method to the results of the PCA for these same data. Did you obtain different groups between the two methods? Do the resulting classifications look good relative to the principal component coordinates?

Note: For the 4-group classification, inspect up to four principal components to evaluate the classification. (This is not a general rule, only a suggestion for this problem.)

c) Graph the variation of the first two principal components as a function of depth, then graph the group membership of the 4-group classification (according to either method) as a function of depth. Do these two methods detect significant changes in sediment composition along the core?