

Multivariate Statistics: Exercise 5

November 13, 2014

Principal component analysis:

1. We consider the data set *world.R*, see Exercise 5. Our aim is to represent the main multivariate data structure by means of few principal components. The first two variables in the data frame should not be used for computing the principal components.
 - (a) Compute the principal components from the standardized data (`princomp()` using the option `cor=TRUE`). Apply `plot` and `summary` on the result object and interpret the results.

Interpret the directions of the principal components (stored as `$loadings` in the result object).

The `$scores` in the result object are the data values projected on the principal components. Show pairwise plots of these scores of the most important principal components, and visualize the variable *Continent* by different choices of symbols or colors in the plots. Which multivariate relations are visible?
 - (b) Perform the above analysis with a robustly estimated covariance matrix using `covMcd()` from the `library(robustbase)`. The MCD result object can be used in `princomp()` via the argument `covmat`. How do the results change compared to (a)?
2. Use the data set *data(yarn)* from the `library(pls)`, and from this list only the element `yarn$NIR`.
 - (a) Visualize the data with `matplot()` such that each observation forms a line in the plot.
 - (b) Compute the first 5 principal components of the centered (not scaled) data. Use the algorithm `PCAgrid()` from the `library(pcaPP)` for this task. Use as a scale measure the classical estimator (`method="sd"`) as well as robust estimators (`mad`, `qn`).
 - (c) Visualize the resulting directions (`$loadings`) from (b) in terms of lines, in analogy to (a). Interpret these results.

Save your (successful) R code together with short documentations and interpretations of results in a text file, named as *Familyname5.R*. Send the file as an email attachment to mehmet.mert@tuwien.ac.at, at latest Tuesday (11.11).