QMM: Exercise Sheet 4 - Principal Component Analysis and Factor Analysis

Fabien Baeriswyl, Jérôme Reboulleau, Tom Ruszkiewicz

Exercise 1. For this exercise, use the `Vehicles.csv` datafile. This dataset consists of 156 vehicles (trucks and automobiles) for which we have various details and characteristics, among which the horsepower, the miles-per-gallon, the price,... We want to perform an exploratory factor analysis on this dataset.

a) Analyse first the correlation matrix of the variables in the dataset. Then, run a principal component analysis for this dataset. Analyse the proportion of the total variance explained by this PCA.

b) According to the Kaiser-Guttman rule, how many components should you take into account?

c) Produce and analyse the variables factor map and identify which variables are connected to each other.

d) Run an unrotated factor analysis with 2 factors. Decide whether you need more or not. Then, run a factor analysis using a Varimax rotation. Compare the two.

e) Run a factor analysis relaxing the factors' correlation assumptions (Promax method). Compare it to the factor analysis only using a Varimax rotation from point d).

Exercise 2. For this exercise, use the `Service.csv` datafile. This dataset consists of a marketing study in which 106 clients graded different aspects (in total 28) on a scale from 1 to 7, 7 corresponding to the highest quality in the service. This kind of setting is typical and one wants to reduce the dimensionality, since using 28 variables in any model would be too much.

a) Produce a correlation matrix to determine the relations between the variables. Since you have many of them, it might be wise to find a way to visually assess the correlation matrix.

b) Run a PCA to determine how many factors are of interest.

c) Run an unrotated factor analysis with 2 factors to start off. Then, using a chi-square test, decide how many factors you should use. For your final model, use Kaiser-Guttman's rule to lead your choice.

d) Run a factor analysis, first using a Varimax method and then a Promax method. Decide which is best. Also, decide whether this chosen model is better than the one in c).

e) Indicate the top 3 variables that are sharing the most variance with the others. Indicate also the one that are sharing the less.

Exercise 3. For this exercise, use the `ship3.csv` datafile. Recall the shipping case that we already talked about.

a) Visualise the correlation matrix between the continuous variables of the dataset, using for example `corrplot()`. Then, run a PCA and determine how many principal components are necessary using Kaiser-Guttman's rule.

b) Plot an individual factor map with the first two principal components and colour the observations for the period they are observed in, i.e. either in the crisis period or in the "normal" period. Discuss your plot.

c) Try to run a factor analysis. Explain the main limitations.