Multivariate analysis,Home assignment

Exercise 1. The U.S. crime data set consists of the reported number of crimes in the 50 states in 1985 (the file 'uscrime.rds'). The crimes were classified according to 7 categories: murder, rape, robbery, assault, burglary, larceny and auto theft. The data set also contains identification of the region: Northeast, Midwest, South and West. Perform principal component analysis (PCA) on the crime variables.

1) Start by studying the covariance matrix of the crime variables. Should we perform PCA with the sample covariance or correlation matrix?

2) How large part of the total variation is explained by the first two principal components? By the first 3 PCs? Write down the formulas for calculating the first two PCs. Can we interpret the first two PCs in this example?

3) Calculate the correlations between the first two PCs and the input variables. Plot the correlations for the 7 variables in a scatter plot with a circle of radius 1. How can you summarize/interpret these correlations?

4) Make a scatter plot of the first two PC scores, label the four regions with different colour or symbols. Does any region stick out?

Exercise 2. Perform a factor analysis on the U.S. crime data set.

1) How many factors can we consider at most in this example? Check the degrees of freedom.

2) Estimate the factor model using the following methods: principal component method, principal factor method and maximum likelihood method. Use both $k = 2$ and $k = 3$ factors. Interpret the factors after rotation (use VARIMAX rotation). What is the main difference between the factors from a 2-factor and 3-factor model, respectively?

3) Choose the model that you think is most suitable and summarize the main characteristics for the chosen model.

4) Estimate the factor scores for the chosen model, these can be used to describe the differences between different states. Which states have the highest factor scores?