Exercise 6

Problem 1: Bivariate Correspondence Analysis

Install the package **ca**. The data set SCIENCEDOCTORATES.txt contains the number of doctors graduated from different fields of science. The data is from USA between the years 1960-1975.

Apply correspondence analysis to the data set by using the function **ca**. Also, write an own code that applies correspondence analysis to the data set as presented in the lecture slides. Compare and interpret the results.

Problem 2: Association between row and column profiles

Let Z be the matrix defined as in the lecture slides. Then denote the PCA performed on the row profiles as V and on the column profiles as W. The matrices are defined the following way:

$$\begin{aligned} V =& Z^\top Z, \\ W =& ZZ^\top. \end{aligned}$$

Show that V and W have the same nonzero eigenvalues. Furthermore, show that the following relation holds for the normed eigenvectors that correspond to nonzero eigenvalues:

$$\begin{aligned} v_i &= \frac{1}{\sqrt{\lambda_i}} Z^\top w_i \\ w_i &= \frac{1}{\sqrt{\lambda_i}} Z v_i, \end{aligned}$$

where v_i is the i:th eigenvector of V and w_i is the i:th eigenvector of W.

Homework Assignment 6: Bivariate Correspondence Analysis

The data SMOKING1.txt contains a two-dimensional frequency table, where the employees of a company have been categorized according to their position (5 categories: SM = Senior Managers, JM= Junior Managers, SE = Senior Employees, JE = Junior Employees, SC = Secretaries). The smoking of the employees have 4 categories (None, Light, Medium, Heavy). Perform BCA using the function **ca**. Provide the requested answers, figures and tables in your report.

(a) Form the row and column profiles.

(b) How much of the variation is explained by the combination of components 1 and 3. Give the answer in percentages relative to the total variation.

(c) Produce the BCA graph with respect to the first two components.

(d) According to (c), which employees are more frequently heavy smokers? Justify!

(e) How much of the variation of the modality Heavy is explained by the first two components? How much of the variation of the modality Medium is explained by the first two components? Give the answers in percentages relative to the total variation. Hint: use the quality of representation.