Week 6: Exercises

Anikó Lovik

2024-03-15

Exercise 0. R Lab 12.5

If you have not yet done so, please start with the R Lab in the book. You can use the book or the HTML document which combines all relevant lab exercises for this week (you can find it on Brightspace under the 'Between Lecture + Workgroup' section).

Exercise 1. College dataset

In this exercise you will perform dimension reduction on the College dataset from the ISLR2 package. This dataset contains the number of applications, type of institution (private/public) and other characteristics of colleges in the US (for a description of the variables see ISLR2, pp. 54-55).

a) Check the sample size and variables of the dataset. Check for missing values, for this exercise, remove them if necessary. Remove the variables 'Private' and 'Apps'.

b) Run a PCA with and without scaling (normalisation) and check the explained variance per component. Do you expect any differences? Why yes/no?

c) Decide on how many components to keep using Kaiser's rule, the cumulative PVE and a scree plot for both (you can also add the average eigenvalue-rule, the MAP test, Revelle's VSS and parallel analysis, if you want). Do you see a difference between the methods? Do you see a difference between scaled and unscaled PCA results?

d) Which variables load high on each component? Can you give an interpretation of the components?

e) Create a bi-plot for the first two components.

f) Rotate the selected components using an orthogonal rotation. (There are many ways to do this, you could for example use the 'varimax' function in the EFA.dimensions package). Do you obtain a simple structure?

Exercise 2. Perform principal component analysis on the heptathlon dataset.

The heptathlon dataset contains data from 25 individuals on the seven disciplines that form the sport. Run a PCA with the normalised data. Decide on how many components to keep using at least three different methods. How much variance do the extracted components explain? Which variables load high on each component, can you give an interpretation of the components? Do the components become easier to interpret if you rotate the components?

Extra: do this without scaling the data (keep the centering). As depicted in the lecture slides and discussed, you will see some differences. What are these differences and how do they impact the interpretation of the results? Do you expect to see this in other datasets?

Exercise 3. Find an application.

Find an application of PCA (or PCR or PLS) related to something you are interested in. Answer the following questions:

a) What is the sample size? How many variables/feature were used for dimension reduction? (If applicable, what was the response variable?)

b) How was the method described? Would you be able to run the analysis based on what has been described? Why yes/no?

c) How many components were extracted? How was the number of components decided?

d) How are the results described? If the results are presented in a figure, describe it (are there any trade-offs?).

e) Were the components given an interpretation? If so, what do you think of it? Are the conclusions in line with the results?

f) If the code is available, have a look. Anything interesting to note?