Statistical Methods 2

Portfolio 2 – Factor analysis and independent component analysis

Complete the following tasks and submit your work on Blackboard by 4pm on Friday 03/02/2023

Task 1 (70/100)

Perform factor analysis in `R` using a dataset of your choice. More precisely, you need to

- Choose the number k of factors. Justify your choice (i.e. show that the model fits well the data for your chosen value of k).
- Estimate the loading matrix $\boldsymbol{\Lambda}$ and the specific variances $\{\psi_{jj}\}_{j=1}^{p}$.
- Interpret the factors in the estimated model.
- Estimate the factors for all the n observations in the dataset.
- Letting $k = 2$, compare the estimated factors with the two dimensional reduction of the data obtained with principal component analysis.
- Discuss the relevance of the factor model for your dataset. In particular, for your dataset would you recommand to use factor analysis or principal component analysis?

To estimate the model you need to work with the correlation matrix $\boldsymbol{R}$ and to use the iterated principal factor analysis algorithm introduced in Chapter 2. You can either implement the algorithm yourself or use an `R` package (e.g. the command `fapa` of the package `fungible`). In this latter case you need to clearly explain which matrix $\hat{\boldsymbol{\Psi}}^{(0)}$ of initial specific variances is used.

Ideally, you need to choose a dataset for which the factor model is a reasonable assumption. But using any dataset that you used in Portfolio 1 is OK.

Task 2

Perform independent component analysis (ICA) in `R` using a dataset of your choice. To perform this task you need to use the FastICA approach covered in Chapter 3, which is for instance implemented in the `fastICA` package. (Following a comment made in Chapter 3, in this package the matrix $\tilde{\boldsymbol{B}}$ is not estimated as described in the lecture notes.)

More precisely, you need to

- Plot the original variables,
- Estimate the signals using the FastICA approach, specifying which function φ you use.
- Plot the estimated signals.

For this task you can for instance use the `icamusical` dataset[1], which contains $p = 3$ audio files. If you listen these files you will remark that each of them is a combination of three different recordings. The goal is then to use ICA to recover the original recordings from these the audio files.

Assuming that the three original audio files are saved as `audio1.wav`, `audio2.wav` and `audio3.wav`, they can be converted into a numerical data matrix $\boldsymbol{X}^0$ as follows:

```
library(tuneR)

f1<-readWave('audio1.wav')
X1<-f1@left

f2<-readWave('audio2.wav')
X2<-f2@left

f3<-readWave('audio3.wav')
X3<-f3@left

X0<-cbind(X1,X2,X3)
```

Assuming that your estimated signals are in the $n \times p$ matrix $\boldsymbol{S}$, you can save them as audio files as follows:

```
library(seewave)

savewav(S[,1], f=f1@samp.rate, channel = 1, filename = "signal1.wav")

savewav(S[,2], f=f1@samp.rate, channel = 1, filename = "signal2.wav")

savewav(S[,3], f=f1@samp.rate, channel = 1, filename = "signal3.wav")
```

You can now listen the three recordings extracted from the data using ICA and appreciate the result! (Two recordings should be a nice piece of music while the last one should be a static note.)

[1] available at `https://www.kaggle.com/chittalpatel/icamusical/activity`