High-Dimensional Data Analysis

Assignment 1

Due 7th September, 2018

Data

In this assignment you are required to analyse a data set on causes of death for males in the United States. Each **observation** is a cause of death, while each **variable** is an age range. The measurement in all cases is the probability of each cause of death. For example, assuming that a male aged from 46-55 dies, then the probability that he dies of influenza is 0.01108975. For every age, the probabilities of all causes of death sum to one.

The causes of death are: Accident, Cancer, Diabetes, Heart related, Influenza, Mental, Nephritis, Stroke, Suicide and Other. Mental here refers to conditions such as Alzheimer's disease and dementia and not to mental illnesses such as depression or schizophrenia.

Part 1: To be completed by all students

Provide a preliminary analysis of this data based on a principal components analysis. Your analysis MUST include a biplot. You are also encouraged to present other plots that assist in understanding the features of the data. Excluding plots, your analysis should be roughly a half to one page. R Code used to conduct the analysis can be included in an appendix.

Part 2a: To be completed ETF3500 Students Only!

The following questions can all be answered with short 1-2 sentence answers.

1. Did you standardise the data in Part 1. Why or why not?
2. How much of the total variation in the data is explained by the third principal component on its own?
3. How much of the total variation in the data is explained by the first two principal components together?
4. Use a biplot to find two causes of death that are similar to one another in terms of their age profile. How do you come to this conclusion?

5. Use a biplot to find two causes of death that are different from one another in terms of their age profile. How do you come to this conclusion?

Part 2b: To be completed ETF5500 Students Only!

The following questions can all be answered with short 1-2 sentence answers.

1. Did you standardise the data in Part 1. Why or why not?
2. Use a biplot to determine two causes of death that are similar to one another in terms of their age profile. How do you come to this conclusion? Justify your choice of biplot (if you have not already included this biplot in Part 1 then include it here).
3. Use a biplot to determine two age ranges that are uncorrelated in terms cause of death probabilities. How do you come to this conclusion? Justify your choice of biplot (if you have not already included this biplot in Part 1 then include it here).
4. How much of the total variation in the data is explained by the principal components shown on either biplot?
5. In what ways is this analysis similar/ different to Multidimensional Scaling?