**RESEARCH AND SURVEY STATISTICS – STA3022F**

**COMPUTER PRACTICAL 2**

**PRINCIPAL COMPONENT ANALYSIS AND FACTOR ANALYSIS**

1. Open Rstudio
2. Save the following files from VULA

prac2.R

PCAbiplot.R

Eurojobs.txt OR Eurojobs.csv (this depends on your R version)

donor.csv (this is your data for the practical assignment)

Prac 2 PCA.doc

1. Open the **script** file **prac2.R**.

From the top menu, select File > Open File

Select the file **prac2.R** which you saved from Vula

Click Open

The script file appears in the Window in the top left.

We will execute the commands step by step by selecting (highlighting) the relevant lines in the script file and clicking on Run at the top right of this window.

1. Example PCA

Load the data set Eurojobs.txt into R, ensuring Heading is ‘Yes’ and Separator is ‘Tab’. The data are the percentage employed in different industries in Europe countries during 1979. It is downloaded from <http://lib.stat.cmu.edu/DASL/Datafiles/EuropeanJobs.html> and the original source is Euromonitor (1979), European Marketing Data and Statistics, London: Euromonitor Publications, 76-77. The data is also found in: Manly, B.F.J. (1986) Multivariate Statistical Methods: A Primer, New York: Chapman & Hall, 11 and : Hand, D.J., et al. (1994) A Handbook of Small Data Sets, London: Chapman & Hall, 303.

The first column contains the names of the countries. The remaining nine columns contains the percentages of people employed in the different industries.

Country: Name of country

Agr: Percentage employed in agriculture

Min: Percentage employed in mining

Man: Percentage employed in manufacturing

PS: Percentage employed in power supply industries

Con: Percentage employed in construction

SI: Percentage employed in service industries

Fin: Percentage employed in finance

SPS: Percentage employed in social and personal services

TC: Percentage employed in transport and communications

Construct data matrix of the 9 variables on the 26 countries.

4.1. Select lines 5 and 6 and click Run to execute the commands.

Perform principal component analysis and create a scree plot. Even though all variables are in percentage form, some economic sectors are bigger than others, so to place the variables on equal footing, we will standardise the data.

4.2. Select lines 10 to 12 and click Run to execute the commands.

For this data set, it appears that 3 principal components are sufficient.

* 1. Select line 16 and Run to set the number of pc’s desired.

Perform PCA and inspect the loadings and principal component scores.

* 1. Select lines 18 to 21 and Run.

From the loadings we see that the first principal component is a contrast between Agriculture and Mining and all the other sectors. The second principal component contrasts the service sectors (Service, Finance and Social and personal services) with the other sectors. The third principal component is more difficult to interpret with large positive loadings on Construction, Social and personal services and Transport and communication. Large negative loadings are observed for Power supply, Financial services as well as Mining, Agriculture and the Service industry. The first two principal components account for only 62% of the total variation, while the first three account for almost 75% of the variation. The third principal component is therefore making a notable contribution even though it is not easy to interpret.

Based on the interpretation of the loadings, large positive principal component scores for the first component indicate countries with very little mining and agriculture while those with large negative scores, especially Turkey, had many people employed in Mining or Agriculture. *Note: You can have large positive scores for countries like Turkey, if you had positive loadings for Agr and Min, it all depends on the output from the specific algorithm. See section 2.3 of the notes.*

* 1. Access the PCA biplot code in the file PCAbiplot.R and construct a PCA biplot of the data. Select File, Open File . . , Select the file PCAbiplot.R (which you saved from Vula) and click Open. Now we have two script files open, see the tabs at the top of the script files. Select all the code in PCAbiplot.R (CTRL + A) and click Run.

Return to your prac2.R script file by clicking on the appropriate Tab.

* 1. Now that the functions in PCAbiplot.R have been defined, select line 24 and Run.

We see in the biplot that the horizontal ordering gives an indication of the degree to which Agriculture and Mining plays a role in the economies. Vertically we have the degree to which the services plays a role, with countries like Denmark, Netherlands and Sweden towards the bottom (in my case, can be towards the top) have many people employed in service sectors.

**PRACTICAL QUIZ DATA and INSTRUCTIONS:**

We will use a data set freely available at <http://kdd.ics.uci.edu/databases/kddcup98/> for analyses with the different methods in the exploratory analysis part of the course.

The data set is donor.csv

It was originally used for a competition of the International Conference on Knowledge Discovery and Data Mining held in 1998. The data set consists of on donations made by members of the Paralyzed Veterans of America Society (PVA). The aim of the competition was to find a model which most accurately predicted which persons in the data base will donate money after receiving an invitation to donate via mail. We will not focus on this goal here, but use different parts of the data to practice the analysis of the methods discussed in the first part of the course.

The file donor.csv contains a selection of the variables in the full data set. Read the data into R.

1. Donor data analysis for Practical quiz:
2. Consider the following demographic and donor information and perform a PCA to explore relationships between donor behaviour and demographics. I.e. try to highlight to ‘your manager’ which demographic groups tend to donate more money, more often, etc.

AGE

INCOME

WEALTH1 A rating of wealth

NUMCHLD Number of children in household

NUMPROM Lifetime number of promotions received to date

RAMNTALL Dollar amount of lifetime donations to date

AVGGIFT Average dollar amount of donations to date

RFA\_2F Number of donations made recently

The code and comments in the script file, lines 27 to 45 will assist you. Remember to carefully consider whether to standardise your data.

1. Complete the Quiz: Prac2 - Quiz on Vula.