Assignment 4

NAME: Your Name
DUE DATE: April 4th, 6pm

Problem 1 (100 pts)

In the folder Assignment 4, you will find the data set called data-final.csv. This data set is from the Five Personality Data Set, and it collects on-line personality test (take a look to the codebook.txt in the folder Assignment 4).

(a) (40 points) Consider the first 50 variables of this data set (this should correspond to the codebook.txt variables). Perform the Principal Component (PC) analysis after having scaled the data. How many components will you retain based on the total variance explained by each component? Plot a bar plot (in ggplot) showing the proportion of variance explained by each PC (consider just the first 10 PC). Then, plot the PC that you have chosen in an heatmap, choose your own colour in three different tonality (where one should be white). How can you interpret this plot and the PC? Is there any link with the name of the data set "the five big personalities"?

(b) (40 points) Perform a factor analysis model with 5 factors with no rotation. How is the total variance explained from the model? Now perform the factor analysis model with 5 factors and with the varimax rotation (remember to not scale the data). Will you keep the model with 5 common factors or will you add another one? Explain why. Plot in an heatmap the matrix of factor loadings matrix (similar to Figure 1) Again choose your own colour by considering three different tonality. Interpretation: Now interpret the factors. Explain what each factor represents and give a name to each factor based on its high loadings.

(c) (20 points) Perform a bootstrap of 50 samples. For each of the bootstrapped sample save the proportion of variance explained by each factor (consider just the first five factors). Plot the proportion of variance explained by each of the five factors with a boxplot in the ggplot and then perform the histogram for each proportion. What can you say about these five distributions obtained? If we bootstrap the loadings we will obtain something no sense in a statistical framework. Explain why.

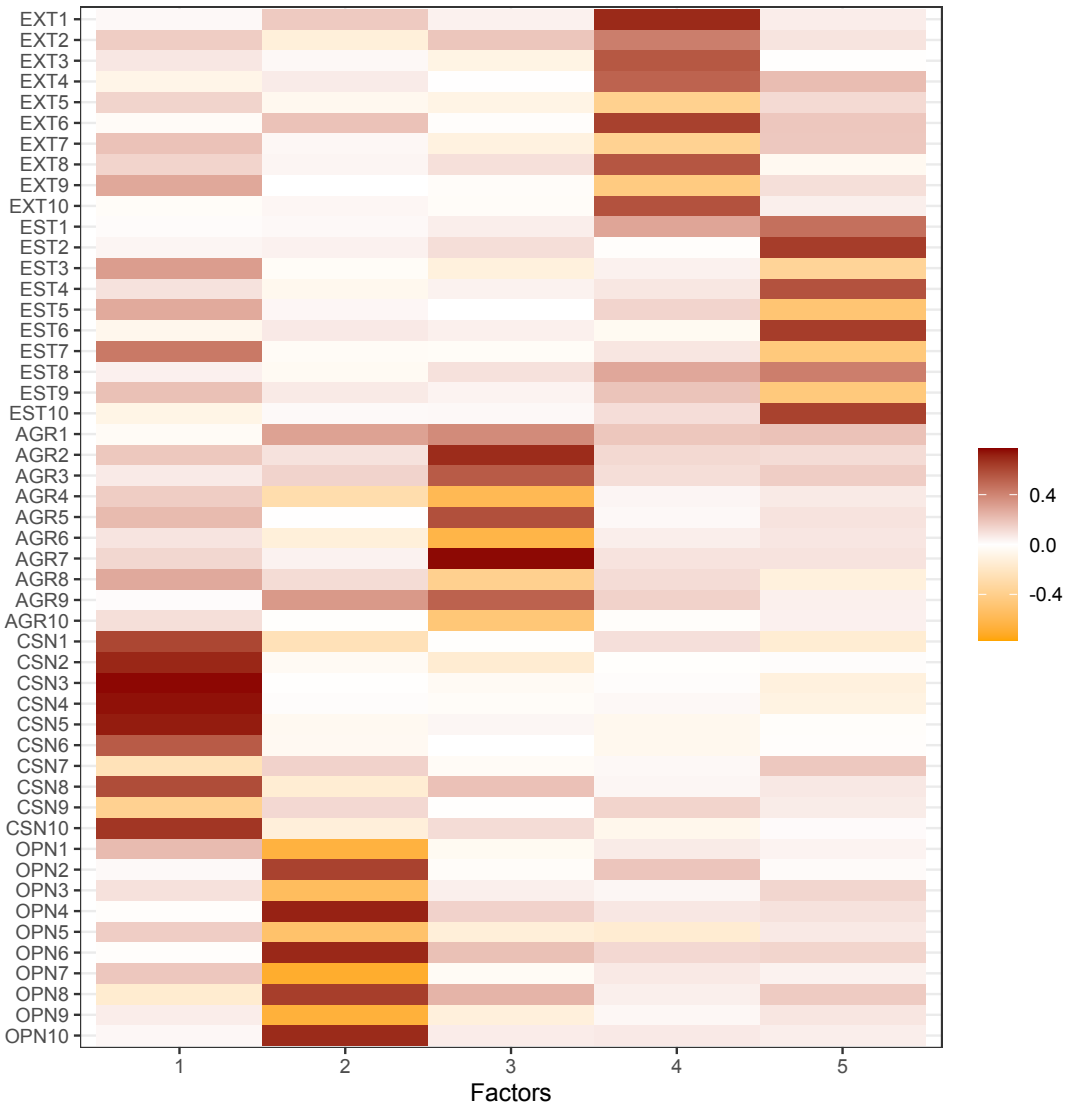

Figure 1: Estimate