G0O00a: Statistical Data Analysis
Academic year 2019-2020: Project 1

Clément Cerovecki, Sebastiaan Höppner

The project consists of analyzing the *Steel production* data set. This data set contains the geometric measurement data of different types of errors that occur with steel plates during the production process. The first 25 variables of the dataset contain these measurement data ($X\_Minimum$, ..., $SigmoidOfAreas$). The last variable Flt indicates what type of fault each observation is. A detailed description of the variables is not available because the data has been made available anonymously.

Each group of 1, 2 or 3 students chooses the data set as determined according to the group division on Toledo. For example, group 1 will use the data set "Group1.csv".

You will answer the questions by performing an appropriate analysis with `R`. You will process the discussion of the results and the necessary figures in a written report that consists of a maximum of 6 pages (12pt font size). Only report results and interpretations, do not repeat theory from the course!

One single folder containing your report and `R` script should be uploaded on Toledo before **23 April 2020**. This project is graded on 5 points.

Exercise 1: Explorative analysis and transformation to normality

1. Perform an exploratory analysis on the dataset. State your main findings.

2. Now consider only the continuous variables. If there are variables whose distribution is very different from a normal distribution, you try to transform it so that they are closer to a normal distribution. Briefly report which variables you transform, why and how you perform the transformation.

For the following exercises you will only continue working with these, whether or not transformed, continuous variables.

Exercise 2: PCA

1. Perform a PCA analysis on your data. The data matrix should only consist of the continuous measurements. Please argue why you base the analysis on the correlation or covariance matrix of the data. Explain how you choose the number of components.

2. Make scatter plots of the first three scores (or only the first two if you have kept 2 components). Can you recognize the different types of errors (i.e. variable Flt)? Use a different symbol (and possible color) for each type of error in the scatter plots.
   Next, discuss whether outliers are found through the diagnostic plot. Again, use a different symbol as before for each type of error in the diagnostic plot.

Exercise 3: Clustering

1. Perform a cluster analysis on the scores obtained in the PCA analysis according to one partitioning method and one hierarchical cluster algorithm of your choice. Explain your choice. Take the number of clusters equal to the number of types of errors (variable Flt) in your data set.

2. Discuss the resulting silhouette plots, cluster plots, and the dendrogram.

3. Determine and discuss to what extent the obtained clusters correspond to the different types of errors (i.e. variable Flt) in your data set.

Exercise 4: Testing multivariate normality

1. Consider the scores (obtained in the PCA analysis) of the largest cluster according to your chosen partitioning method. Do these scores include outliers? (to this end you might use the MCD estimator). If yes remove these outliers.

2. Can you assume that the underlying distribution of the resulting scores is a multivariate normal distribution?

Good luck!