GROUP PROJECT

MULTIVARIATE DATA ANALYSIS

Professor Eunice Carrasquinha

Alberto Fallocco | 59378 | Erasmus
Ana Araújo | 59457 | MCD
Christopher Anaya | 60566 | MCD
Leonor Ferreira | 55708 | MAEG

Faculdade de Ciências da Universidade de Lisboa

2022/2023

Ciências
ULisboa

TABLE OF CONTENTS

1. INTRODUCTION

A dataset was given in order to apply two multivariate methods: **Principal Component Analysis** (PCA) and **Cluster Analysis** (CA). The main objectives of this project are:

- Make a preliminary analysis of the data;
- Conduct a principal component analysis exploring the potentialities of this method;
- Conduct a cluster analysis exploring the hierarchical approach;
- Compare the results obtained in the two analysis methodologies.

1.1. Descriptive Analysis

The data in data_8.csv file contains the evaluation obtained in 6 tests performed by twenty engineers and twenty pilots. The dataset consists in discrete data, where there are 6 variables and 40 observations. The rows correspond to each worker (pilot or engineer) and the columns correspond to the test (variables) performed. The variables are:

- T1 - Test 1
- T2 - Test 2
- T3 - Test 3
- T4 - Test 4
- T5 - Test 5
- T6 - Test 6

The central tendency of the data, where for numeric variables, the minimum, maximum, quartiles, median, and mean values are obtained. Also, dispersion measures were evaluated, where the standard deviation for each variable was calculated. Table 1 presents the central tendency of the dataset and the standard deviation (SD) of the variables.

Variables	**T1**	**T2**	**T3**	**T4**	**T5**	**T6**
Minimum	47.0	17.00	52.00	152.0	209.0	27.00
1st Quartile	116.5	30.75	75.50	186.0	242.0	37.50
Median	129.5	36.50	82.50	223.0	262.5	47.50

Mean	126.9	34.90	81.80	214.7	265.2	48.94
3rd Quartile	142.0	39.00	88.25	240.2	293.2	58.00
Maximum	164.0	55.00	105.00	291.0	324.0	88.00
SD	23.0	8.30	12.05	37.0	32.7	16.97

Table 1. Descriptive Statistics of the Data

As per Table 1, it can be observed that the mean values present a different scale between the variables (for example, a different scale is observed between variable T2 and T5). Also, the SD values obtained are higher for some variables than others. This comparison can be affected by the fact that the variables are in different scales (the variables T4 and T5, that present a higher scale, also present the higher SD values). Thus, the standardization of the data was performed for proper results before performing the PCA and CA.

Standardization is an important technique that is mostly performed as a pre-processing step, to standardize the range of variables of an input data set.

In principal component analysis, features with high variances or wide ranges get more weight than those with low variances, and consequently, they end up dominating the first principal components (components with maximum variance). One of the reasons these variables can have high variances compared to the other ones is just because they can be measured in different scales. Thus, standardization of the data can prevent this, by giving the same weightage to all variables.

Clustering models are distance-based methods analysis. In order to measure similarities between observations and form clusters they use a distance metric. So, variables with high ranges will have a bigger influence on the clustering. Therefore, standardization is required before building a clustering model.

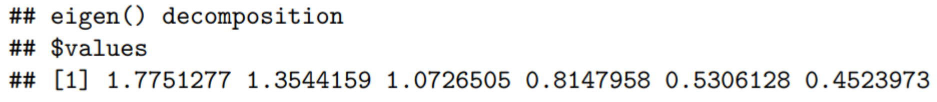
2. PRINCIPAL COMPONENT ANALYSIS

When large multivariate datasets are analyzed, it is often desirable to reduce their dimensionality. One of the techniques used to achieve it is the Principal Component Analysis.

Principal Component Analysis (PCA) is a multivariate technique that analyzes a data table in which observations are described by several inter-correlated quantitative dependent variables. Using mathematical projection, the original variables are replaced by another set of variables, the principal components (p.c), that are uncorrelated and with smaller dimension than the original. It is therefore often the case that an examination of the reduced dimension data set will allow the user to spot trends, patterns, and outliers in the data. It is also possible to interpret the meaning of these new variables and to understand the original data in terms of the new ones.

In order to make all variables more comparable, the variables are standardized, meaning that the centered value of each variable is divided by the corresponding standard deviation. As said before, there are some differences in the mean and standard deviation values, which could indicate that they are in different scales. With this pre-processing of the data, making each variable have the same ‘size’, PCA can be performed.

First, to begin the PCA, we have to obtain the eigenvalues and eigenvectors associated with the correlation matrix of the data. Usually, instead of the correlation matrix, the covariance matrix is used. In our case, as said before, the variables in the study can be in different scales so, it’s better to use the correlation matrix.



```
## eigen() decomposition
## $values
## [1] 1.7751277 1.3544159 1.0726505 0.8147958 0.5306128 0.4523973
```

Table 2. Eigenvalues

After the PCA is performed, the question of how many PCs should be retained arises.

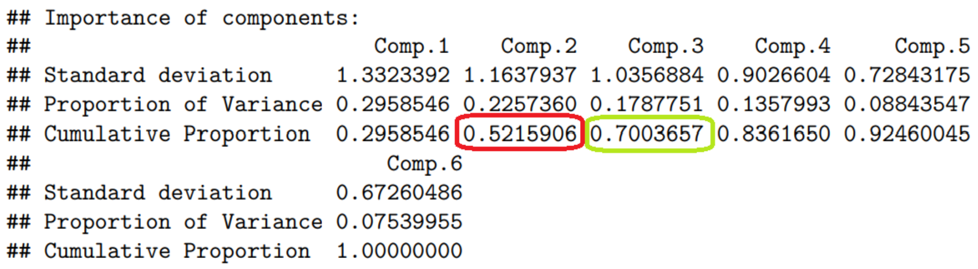
- **Kaiser’s Criterion**

If the data is standardized, each variable has a variance of one. If all variables are orthogonal to each other, then every component in a PCA model would have an eigenvalue of one. It is then fair to say, that if a component has an eigenvalue larger than one, it explains variation of more than one variable.

Based on this criterion, and looking at Table 2, we should retain the first 3 principal components, since those are the ones whose eigenvalues are greater than one. (1.775, 1. 354 and 1.073)

- **High Fraction of Variation Explained**

This criterion suggests that we retain as many pcs as necessary so that the percentage of variance explained by them is greater than a given value α fixed a priori. Considering only 2 principal components, the total of explained variance is 52,2%, which is not high! This value reaches 70% when considering 3 PCs, which is not very high either.

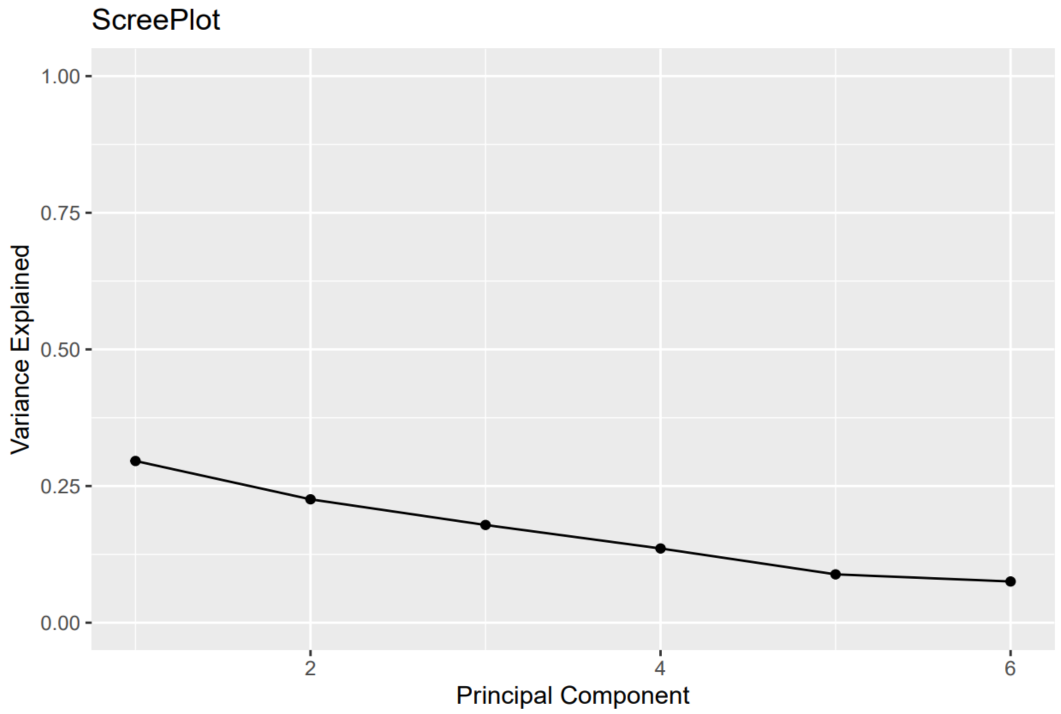


```
## Importance of components:
##                           Comp.1    Comp.2    Comp.3    Comp.4     Comp.5
## Standard deviation     1.3323392 1.1637937 1.0356884 0.9026604 0.72843175
## Proportion of Variance 0.2958546 0.2257360 0.1787751 0.1357993 0.08843547
## Cumulative Proportion  0.2958546 0.5215906 0.7003657 0.8361650 0.92460045
##                            Comp.6
## Standard deviation     0.67260486
## Proportion of Variance 0.07539955
## Cumulative Proportion  1.00000000
```

Table 3. PCA Summary

- **Screeplot**

The r PCs that contribute the most should be retained, standing out sharply from the others. In our case, the slope does not change that much between PC 2 and PC 3, as seen in the graphic below.



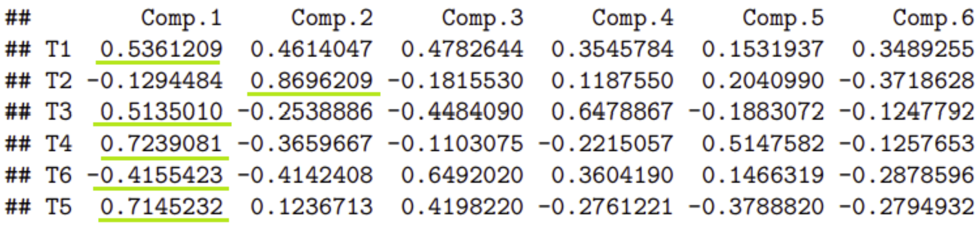
Graphic 1. ScreePlot

According to some authors, the screeplot criterion should be used, preferably, when the number of variables is less than 30. Considering that we only have 6 variables (T1, T2, T3, T4, T5, T6), we're discarding the other two criteria and focus on this one.

With that being said, considering this analysis is the one that prevails, we will only consider 2 PCs.

2.1. Interpretation of the PCs

T1, T3, T4, T5 and T6 are the variables that contribute the most for the explanation of 1st principal component retained. In addition, T2 is the most important variable for the explanation of the 2nd PC.



```
##          Comp.1     Comp.2     Comp.3     Comp.4     Comp.5     Comp.6
## T1  0.5361209  0.4614047  0.4782644  0.3545784  0.1531937  0.3489255
## T2 -0.1294484  0.8696209 -0.1815530  0.1187550  0.2040990 -0.3718628
## T3  0.5135010 -0.2538886 -0.4484090  0.6478867 -0.1883072 -0.1247792
## T4  0.7239081 -0.3659667 -0.1103075 -0.2215057  0.5147582 -0.1257653
## T6 -0.4155423 -0.4142408  0.6492020  0.3604190  0.1466319 -0.2878596
## T5  0.7145232  0.1236713  0.4198220 -0.2761221 -0.3788820 -0.2794932
```

Table 4. Component Matrix

Loadings' Analysis:

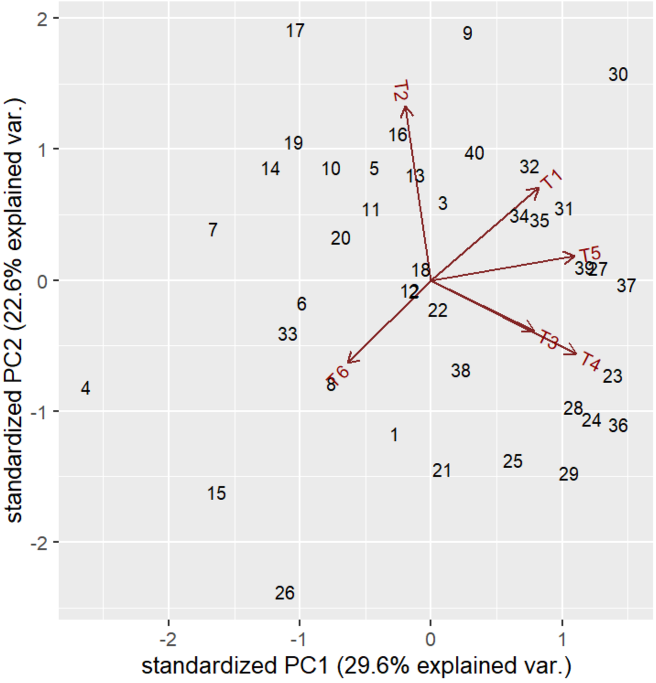
- In the horizontal axis is represented the correlation of the variables with the 1st principal component. (T1, T3, T4, T5, T6)
- In the vertical axis is represented the correlation of the variables with the 2nd principal component. (T2)
- T1, T3, T4 and T5 are on the right side of the horizontal axis indicating high and positive correlations. T6, one the other hand, is on the left side indicating a negative correlation.
- T2 is in the top left of the plot, indicating a high correlation.

Note that the size of the arrows the size of the arrows are proportional to the variability associated to each principal component. In this case, for example, T4 and T3 have the same direction although T4 is a bit longer than T3, indicating that T4 is better explained by the 1st PC than T3.

The angles between two variables (represented by the arrows) indicate how much are the variables correlated: if the angle is small, it means that the variables are strongly correlated. On the other hand, a bigger angle implies that the variables are not correlated.

Scores' Analysis:

- Individuals 4 and 26 present lower T1 and T5 values when compared to the others.
- Individuals 9, 17 and 30 present lower T2 when compared to the other individuals.



Graphic 2. Biplot

3. CLUSTER ANALYSIS

Cluster Analysis (CA) is another technique used to obtain the reduction of the dimensionality of the data. It's a method of exploratory analysis that works by organizing items into groups, or clusters, based on how closely associated they are.

Unlike many other statistical methods, Cluster Analysis is typically used when there is no assumption made about the likely relationships within the data. It provides information about where associations and patterns in data exist, but not what those might be or what they mean.

3.1. Hierarchical Cluster Analysis Methods

In our case, being required of applying a hierarchical approach, we chose to implement three of the most widely used methods, such as Single Linkage, Average Linkage and Ward.

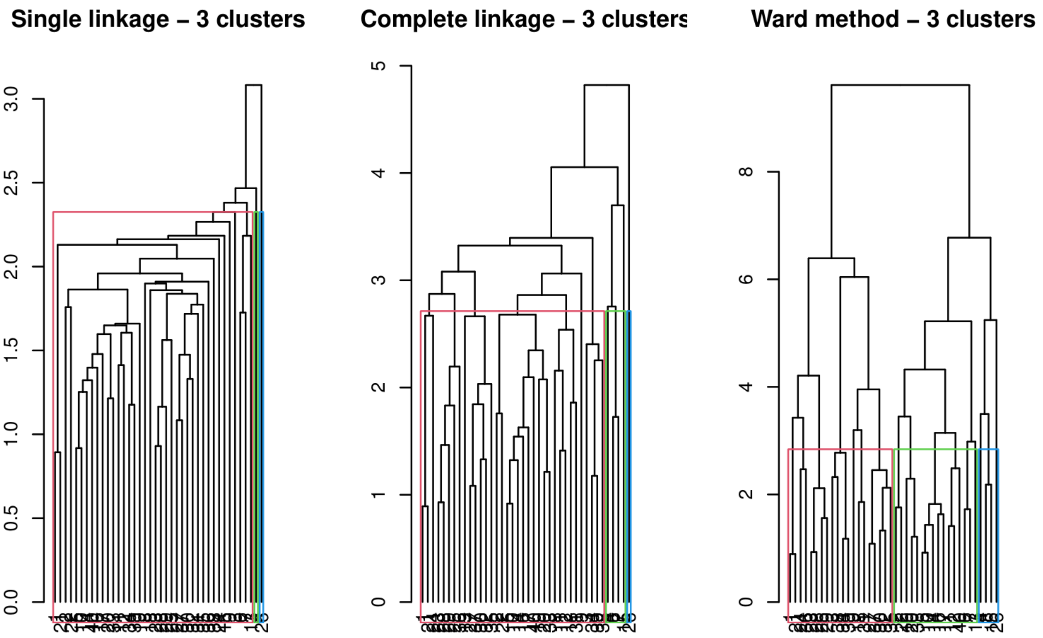
The main factor that discriminates hierarchical approaches is the way the similarity/dissimilarity between the observations is computed: let C_r and C_s be two classes, in the Single Linkage method the smallest dissimilarity between an element of C_r and an element of C_s is used, while the Average and Ward methods use an average of the dissimilarities between the elements of all pairs that can be formed with an element of C_r and an element of C_s , just differently computed.

We implemented the analysis, then checked which clusters were merged at each step of the algorithm. After that, we also checked the distances at which the clusters were merged, and the indices of the points in the original data that are ordered according to the clustering result. Lastly, results from the analysis were plotted in a dendrogram for better visualization. The dendrograms for the three methods were then plotted together for comparing and contrasting, for both a 3-clusters and a 4-clusters division.

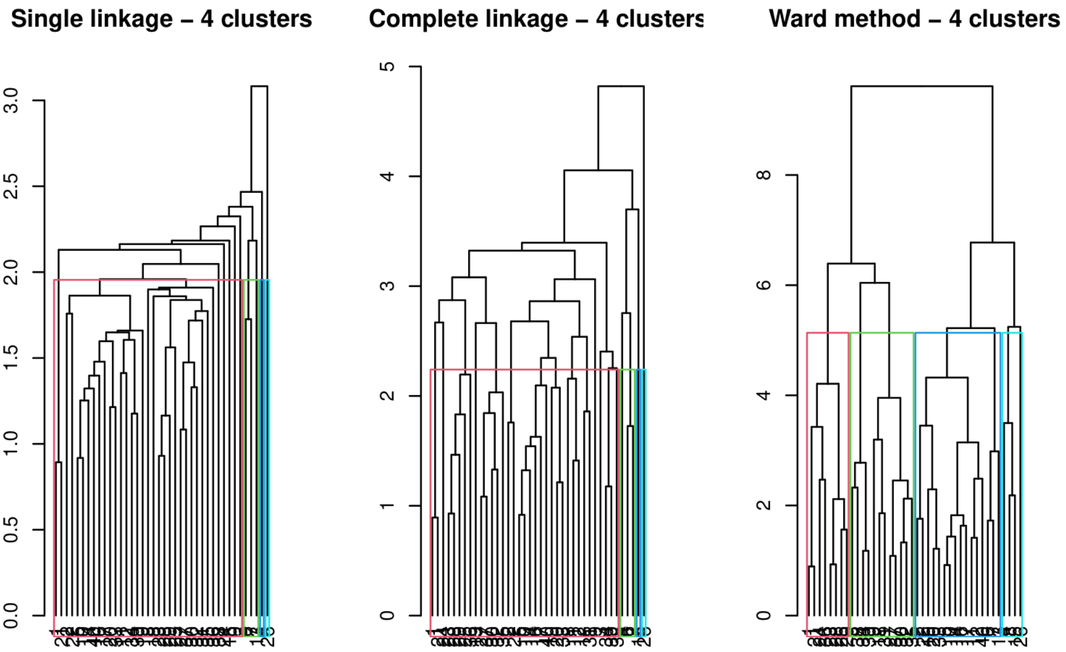
A closer evaluation of the average cohesion is due: it represents the average distance between the points within a cluster, and it is obtained as the sum of all the distances between the points within a cluster divided by the number of pairs of points within the cluster.

The average cohesion calculated for the clusters resulting from the Single Linkage approach is 0.492, and it stands the lowest of the three components that were calculated. On the other hand, the value calculated for the clusters as they were aggregated by the Ward method is the highest (0.812). Lastly, the Average Linkage approach resulted in a score of 0.63.

These values indicate the average distance between the points within the clusters obtained from the three clustering algorithms. Hence, our analysis shows that the clusters found by the Single Linkage method - at least for our dataframe - are the most similar.



Graphic 3. *3 Dendrogram (3 Clusters)*



Graphic 4. *Dendrogram (4 Clusters)*

4. CONCLUSION

We conducted two complementary analyses of our data: a Principal Components Analysis and a Cluster Analysis.

The aim of the Principal Components Analysis was to generate a reduced number of dimensions of the data, the principal components, that could describe the majority of the variation of the data. We concluded that the first two principal components should be retained, but that cumulatively they only described 52% of the present variance. In short, we could reduce the dimensions and come up with a simplified data representation, but we lose almost half of the variability in the data when doing so.

With the Cluster Analysis, we implemented a cohesion score as an evaluation metric between candidate hierarchical models. We concluded that the best performing hierarchical model was a Single Linkage with three clusters. Examining the resulting dendrogram, we see that the majority of the data groups into a single cluster, with the remaining clusters containing a small number of data points each which we could classify as outliers. Taking a step back, we can conclude that, although there are variations in the proportion of variance explained between variables, all of the variables help to explain the variation in the data. In fact, the variable with the least explanatory power still explains about 7.5% of the variance, just under half of what it would be if all the variables uniformly explained the data to the same degree. We can also conclude that, with regards to the observed subjects, the vast majority cluster into a group of similar features, with a small number of outliers that differ significantly in their measurements.

5. REFERENCES

Abdi, H. and Williams, L.J. (2010) “Principal component analysis,” Wiley Interdisciplinary Reviews: Computational Statistics, 2(4), pp. 433–459. Available at: https://doi.org/10.1002/wics.101.

Bro, R. and Smilde, A.K. (2014) “Principal component analysis,” Anal. Methods, 6(9), pp. 2812–2831. Available at: https://doi.org/10.1039/c3ay41907j.

Chan, J.Y.K. and Bauer, C.F. (2014) ‘Identifying At-Risk Students in General Chemistry via Cluster Analysis of Affective Characteristics’, *Journal of Chemical Education*, 91(9), pp. 1417–1425. Available at: https://doi.org/10.1021/ed500170x.

Clark, N.R. and Ma’ayan, A. (2011) “Introduction to statistical methods to analyze large data sets: Principal Components Analysis,” Science Signaling, 4(190). Available at: https://doi.org/10.1126/scisignal.2001967.

Gewers, F.L. *et al.* (2021) ‘Principal Component Analysis: A Natural Approach to Data Exploration’, *ACM Comput. Surv.*, 54(4). Available at: https://doi.org/10.1145/3447755.

Jolliffe, I. (2005) “Principal component analysis,” Encyclopedia of Statistics in Behavioral Science [Preprint]. Available at: https://doi.org/10.1002/0470013192.bsa501.

Richardson, M. (2009) Principal Component Analysis. Available at: https://people.duke.edu/~hpgavin/SystemID/References/Richardson-PCA-2009.pdf.