Introduction to Data Science: Group Assignment

Group number: 10
Ernani Hazbolatow 2023708/304318
Julia van Bon 2014511/872721
Andrey Peshev 2023683/138297
Sarah Via 2035640/135598

April 2020

Answers Short Questions: PCA

Are all assumptions met to run a PCA?

The correlation matrix confirms that a large amount of variables are linearly correlated (Figure 1). More ellipsoid shape means that there is stronger linearity between the variables. The p-value of the Bartlett test is smaller than 0.05 and KMO index is greater than 0.6, which means we can proceed with the PCA.

How many components do you choose?

Based on the scree plot (Figure 2), we chose 5 components. The 6th and 7th components have eigenvalues larger than 1, but they don't seem to add that much to the explained variance and would make the interpretation more difficult.

Which form of rotation do you choose?

The simple structure is is not attained better by OBLIMIN and and there is no correlation between the components that is larger than $|0.30|$, therefore we choose VARIMAX rotation.

How much % of the variance can be explained by the PCA solution you choose?

The 5 components explained 58% of the variance.

How many variables violate the rules of thumb of a simple structure?

None of the variables violate the simple structure in the VARIMAX model because all of them have loadings higher than $|0.30|$ on only one component. Only, the variable inquisitive is close to this cut-off value with 0.175 for component 4 and 0.34 for component 5.

Think of a label for every component

Inspired by the big five model of normal personality traits (Goldberg, 1993), we used the same construct names to summarize the correlated variables.

- **Component 1:** Extraversion
- **Component 2:** Conscientiousness
- **Component 3:** Neuroticism
- **Component 4:** Agreeableness
- **Component 5:** Openness to experience

Report PCA

Introduction

The first part of the analysis focuses on a data set of 295 observations on 30 personality trait variables from second-year bachelor students. A principal component analysis (PCA) was performed, in order to understand which variables are correlated and can be grouped together. The goal of this analysis is to reduce the dimensionality of data and make it easier to interpret.

Method

The analysis of the data was performed with the programming language R in the IDE RStudio. Before conducting a PCA, the linearity of the data was evaluated with a correlation matrix plot (Figure 1), Bartlett sphericity test and Kaiser-Meyer-Olkin (KMO) test. According to the correlation matrix plot, the relationship between all the variables is not perfectly linear but it is sufficient for proceeding with the analysis. The Bartlett sphericity test had a p-value smaller than the alpha level of 0.05 and the KMO index was 0.79, which is indeed larger than the desirable value of 0.6. The three assumptions were successfully met and three different PCA's were conducted. First, a PCA with no rotation was used in order to find a sufficient number of components for grouping the variables in the dataset. After that in order to transform the results into a simpler structure that is easier to interpret, VARIMAX and OBLIMIN rotation were applied.

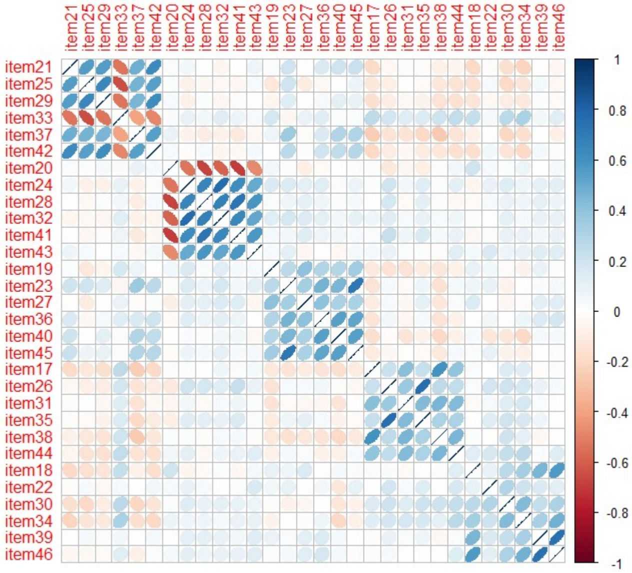


Figure 1: Correlation matrix

Results

According to the scree plot (Figure 2) of the first PCA, 5 components were chosen. The 6th and 7th component have eigenvalues a little higher than 1, meaning that they contain more information than one observed variable. However, all components after the 6th one add little information compared to the previous. This is seen in the graph with the clear "elbow" structure of the line.

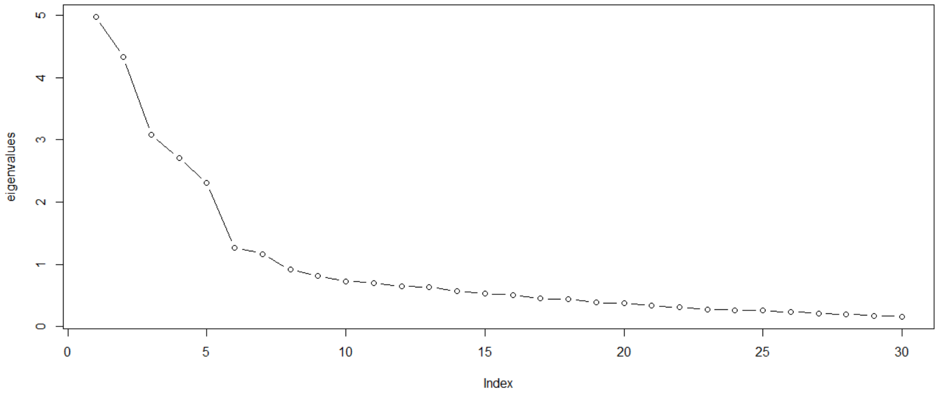
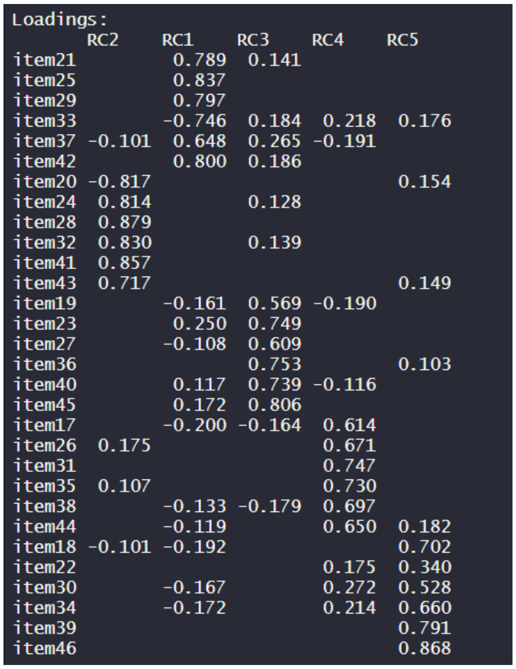


Figure 2: Eigenvalue screeplot

In order to make the components easier to interpret, PCA's with VARIMAX and OBLIMIN rotation were performed. Out of them both VARIMAX was preferred in the final interpretation of the components, because the simple structure was not attained better by OBLIMIN and none of the components had a correlation larger than |.30|. The loadings for the 5 components of the VARIMAX rotation can be seen in Figure 3. The 5 components strongly represent the Big Five personality traits (Goldberg, 1993). Accordingly, the following labels were chosen for each component:

- **Component 1:** Extraversion
- **Component 2:** Conscientiousness
- **Component 3:** Neuroticism
- **Component 4:** Agreeableness
- **Component 5:** Openness to experience



```
Loadings:
       RC2    RC1    RC3    RC4    RC5   
item21         0.789  0.141              
item25         0.837                     
item29         0.797                     
item33        -0.746  0.184  0.218  0.176
item37 -0.101  0.648  0.265 -0.191       
item42         0.800  0.186              
item20 -0.817                       0.154
item24  0.814         0.128              
item28  0.879                            
item32  0.830         0.139              
item41  0.857                            
item43  0.717                       0.149
item19        -0.161  0.569 -0.190       
item23         0.250  0.749              
item27        -0.108  0.609              
item36                0.753         0.103
item40         0.117  0.739 -0.116       
item45         0.172  0.806              
item17        -0.200 -0.164  0.614       
item26  0.175                0.671       
item31                       0.747       
item35  0.107                0.730       
item38        -0.133 -0.179  0.697       
item44        -0.119         0.650  0.182
item18 -0.101 -0.192                0.702
item22                       0.175  0.340
item30        -0.167         0.272  0.528
item34        -0.172         0.214  0.660
item39                              0.791
item46                              0.868
```

Figure 3: Loadings of 5 principal components for the 30 variables

Discussion

It is safe to say that the analysis succeeded in grouping the variables of the personality test into components that are easily interpretable and make future studies of this dataset easier to work with. The five components (i.e. Extraversion, Conscientiousness, Neuroticism, Agreeableness, Openness) help to visualize the significant personality traits of the 295 bachelor students.

Answers Short Question: Clustering

What were the differences and/or similarities in the results of the three different clustering approaches?

Across all approaches using 2 clusters, every approach managed to have the same dog breeds in the same cluster.

Which clustering approach and how many clusters do you choose and why?

The internal and stability measures usually preferred hierarchical with n=2 clusters. This clustering scored the highest on all internal measures and lowest on APN and ADM. K-means with n=6 outperformed hierarchical with n=6 on the AD and FOM measures. Hence, we prefer the hierarchical clustering with n=2. This is also the reason why we pick 2 clusters. Additionally, the scree plot with within sum of squares shows the elbow at n=2, which is another reason to opt for n=2 clusters.

Would you say that the clustering algorithm you choose in the end (as final solution) did a good job in clustering the dogs? Why or why not?

The hierarchical clustering method did a good job in clustering the dogs because it seems to put them into two distinguishable categories. However, fur quality is a strange indicator because we cannot really say that the dogs in the second Cluster have bad fur. They all have shorter fur (except for the yorkshire terrier), but that does not mean it is of a lower quality.

Think of a label for every group of the N (=total number of clusters that you choose as a final solution) dog groups (Group 1 of dogs = dogs for labor (guard dogs), Group 2 of dogs = dogs used as pets).

The dogs in Cluster 1 are often used for labor or difficult tasks. For example, Huskies or Alaskan malamutes are sleigh dogs. Thus, they should have thick fur, should be intelligent and trainable and have high energy levels. The dogs in cluster 2 look more like dogs that are often used as pets. They should be agreeable and tolerant with people.

Report Clustering

Introduction

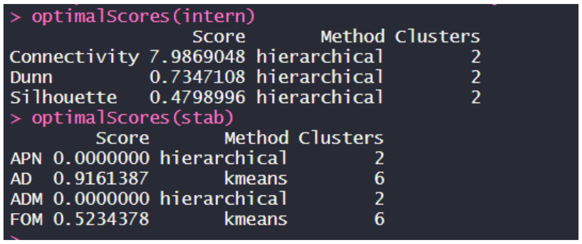
The second part of the analysis focuses on a data set consisting of 12 observations from different dog breeds on 6 different characteristics. The tested characteristics were: intelligence, fur quality, train-ability, energy level, agreeableness and tolerance. In order to cluster the dog breeds based on the different characteristics, three different clustering approaches were performed. Therefore, we expect to find the best clustering algorithm and the total number of clusters, in other words, the findings will help to summarize which dog breed belongs to which group.

Method

The analysis of the data was performed with the programming language R in the IDE RStudio. Before conducting the different clustering analysis, the variables were standardized without the row names. Then, the row names were added back. Three different clustering approaches were performed. First, a k-means analysis was performed with 2 clusters. The K-means result for 2, 3, 4, 5, 6 clusters were compared by the within-cluster variation of different clusters. Secondly, a Partitioning Around Medoids (PAM) analysis was performed with 2, 3 and 4 clusters. This was done by visual inspecting and plotting the medoids. Thirdly, a hierarchical clustering analysis was performed. The hierarchical clustering was run three times for a complete linkage, average linkage, and single linkage method. The hierarchical results were plotted. Finally, for all clustering methods, stability and internal measures were computed. Then, the optimal scores from those measures were used to determine the clustering method. Here, hierarchical clustering was picked. Lastly, the number of clusters was defined based on hierarchical plot.

Results

All the three clustering approaches suggested that the solution with 2 clusters is the most fitted. Every approach managed to have the same dogs in the same cluster. The hierarchical method with 2 clusters was chosen based on the optimal scores for internal and stability measures. More specific, based on the internal measures a hierarchical method with 2 clusters would be most optimal. Based on the stability measures on APN & ADM a hierarchical method with 2 clusters would be most optimal (Figure 4). Based on the stability measures on AD & FOM the 6 PAM method with 6 clusters would be most optimal.



```
> optimalScores(intern)
                 Score       Method Clusters
Connectivity 7.9869048 hierarchical        2
Dunn         0.7347108 hierarchical        2
Silhouette   0.4798996 hierarchical        2
> optimalScores(stab)
        Score       Method Clusters
APN 0.0000000 hierarchical        2
AD  0.9161387       kmeans        6
ADM 0.0000000 hierarchical        2
FOM 0.5234378       kmeans        6
```

Figure 4: Optimal Internal and Stability output

The output of the hierarchical method is shown in an dendrogram (Figure 5).

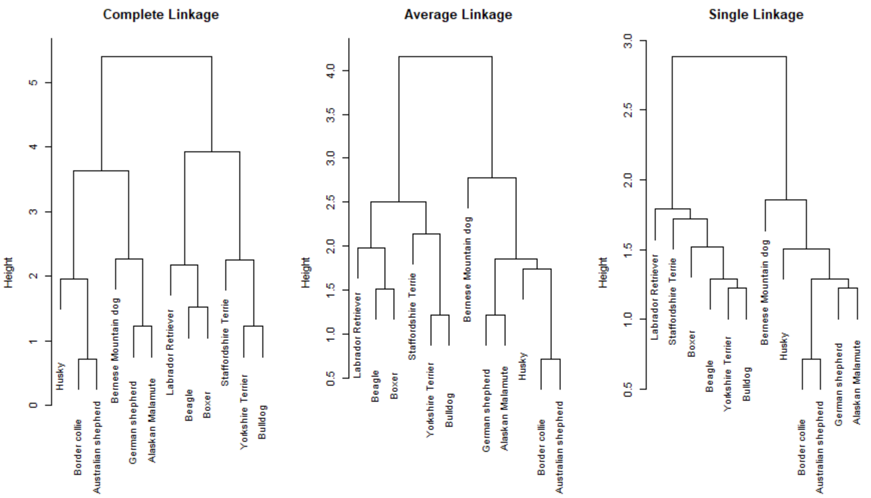


Figure 5: Hierarchical clustering results

Additionally, the scree plot with within-cluster sum of squares (to be minimized) (Figure 6) shows that it is not much improving by adding more clusters after 2 clusters ("elbow" at 2 clusters). This is another reason to opt for 2 clusters.

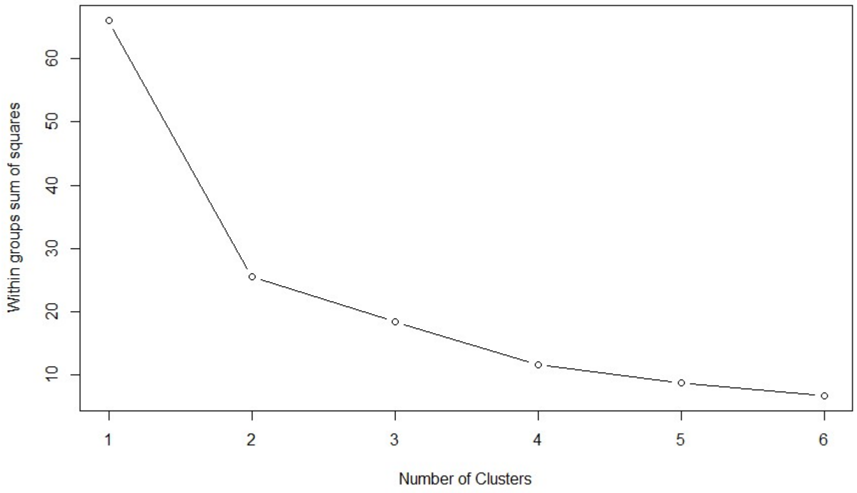


Figure 6: Within sum of squares Scree plot

Discussion

The hierarchical clustering algorithm did a good job clustering the dogs. This is because the 6 different characteristics help to cluster the 12 dog breeds into two distinguishable categories. Namely, dogs that can be used as pets or dogs that are often used for labor or difficult tasks. Even though all three approaches came to the same number of clusters and groups of dogs. Hierarchical clustering is preferred because unlike k-means and PAM it does not require to specify the number of clusters beforehand.

References

Goldberg, L. R. (1993). The structure of phenotypic personality traits. *American psychologist, 48(1)*, 26.