Intelligent Data Analysis - Homework 1.3

Panagiotis Michalopoulos, Javier de la Rua, Michail Gongolidis,
Ignacio Rodriguez, Daniel Minguez

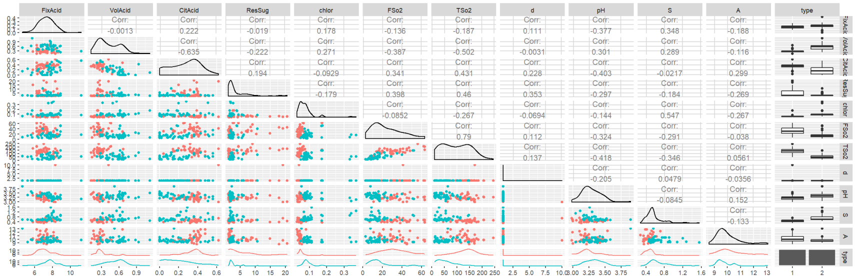
November 4th, 2018

1 Dimensionality reduction: PCA Analysis

1.1 Perform a Principal Component Analysis on the wine data set

1.1.1 Decide whether to use matrix S or R to extract the principal components.

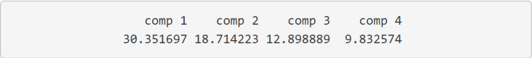
After we have explored the data, we have decided to use the R matrix (correlation) because the variables have different scales. The covariance matrix (S) is used when the variable scales are similar and the correlation matrix (R) when variables are on different scales in order to have an objective measure to compare. We have used the following plot to explore the data:



1.1.2 Describe the 4 first components, the proportion of variance explained and which variables constribute the most to each one.

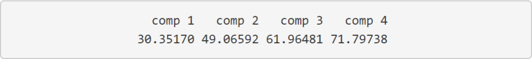
We can see that the first 4 components reflect approx. 71.7% of the variation of data. Components 1 and 2 are the components that mostly reflect the variation of data with a value of 48%, we can use those values to represent the data.

First 4 components



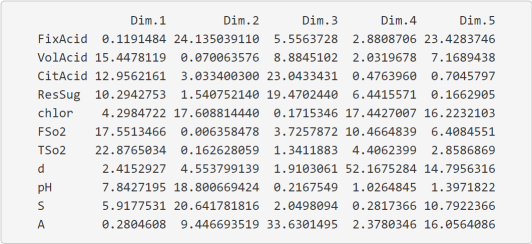
```
    comp 1    comp 2    comp 3    comp 4
30.351697 18.714223 12.898889  9.832574
```

First 4 cumulative components



```
  comp 1   comp 2   comp 3   comp 4
30.35170 49.06592 61.96481 71.79738
```

By analyzing the results we can say which variables contribute the most to each one, for example *FixAcid, chlor, pH* and *S* contribute the most to Dim2:



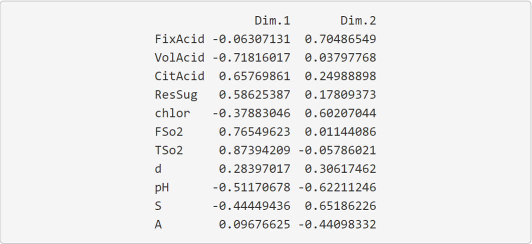
```
             Dim.1        Dim.2      Dim.3      Dim.4      Dim.5
FixAcid  0.1191484 24.135039110  5.5563728  2.8808706 23.4283746
VolAcid 15.4478119  0.070063576  8.8845102  2.0319678  7.1689438
CitAcid 12.9562161  3.033400300 23.0433431  0.4763960  0.7045797
ResSug  10.2942753  1.540752140 19.4702440  6.4415571  0.1662905
chlor    4.2984722 17.608814440  0.1715346 17.4427007 16.2232103
FSo2    17.5513466  0.006358478  3.7257872 10.4664839  6.4084551
TSo2    22.8765034  0.162628059  1.3411883  4.4062399  2.8586869
d        2.4152927  4.553799139  1.9103061 52.1675284 14.7956316
pH       7.8427195 18.800669424  0.2167549  1.0264845  1.3971822
S        5.9177531 20.641781816  2.0498094  0.2817366 10.7922366
A        0.2804608  9.446693519 33.6301495  2.3780346 16.0564086
```

1.1.3 Try to give an interpretation of the first two PC's.

If we analyze the first two PC, we can check which variables are strongly correlated with the components, so values with high values either in positive or negative directions will be considered highly correlated, ($> |0.5|$).The first component is highly correlated with *'total sulfur dioxide', 'free sulfur dioxide', 'citic acid' and 'residual sugar'* in a positive direction, and *'Volatile acidity'* in a negative direction. We can also appreciate that in the previous plot.

The second component can be viewed as a measure of *fixed acidity, chlorides and sulphates* , while having a low ph and alcohol , so this component can be viewed as a term of how un-alcoholic this wine could be.

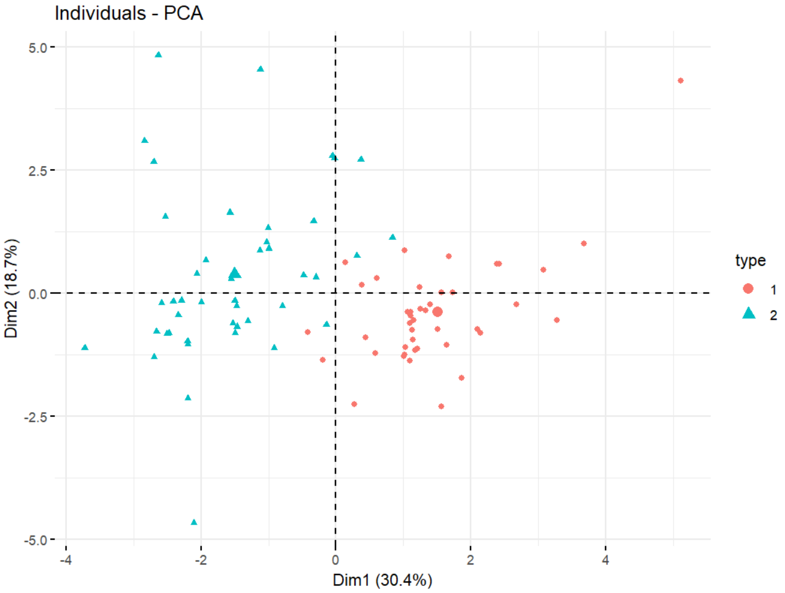
Most important components



```
              Dim.1       Dim.2
FixAcid -0.06307131  0.70486549
VolAcid -0.71816017  0.03797768
CitAcid  0.65769861  0.24988898
ResSug   0.58625387  0.17809373
chlor   -0.37883046  0.60207044
FSo2     0.76549623  0.01144086
TSo2     0.87394209 -0.05786021
d        0.28397017  0.30617462
pH      -0.51170678 -0.62211246
S       -0.44449436  0.65186226
A        0.09676625 -0.44098332
```

1.1.4 Plot the observations on the space spanned by the two first components and colour them according to "type".

We have 2 wine types and as we can see in the below plot we can cluster the wine variables based on their types. We can asssume that Dim1 separates wines according to type more that Dim2.

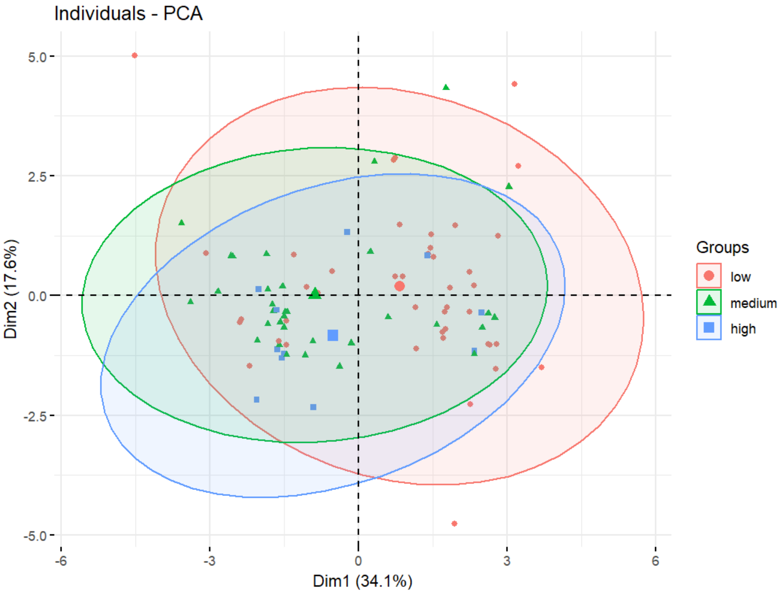


1.1.5 Plot the observations on the space spanned by the two first components and colour them according to "quality".

We can appreciate how the second principal component seems to divide the zones with high quality versus the zone with lower quality, being the mayority in the bottom left quadrant.

Also the first principal component seems to divide the lower quality wines, being the mayority in the positive part, although there are low quality elements among all the quadrants.

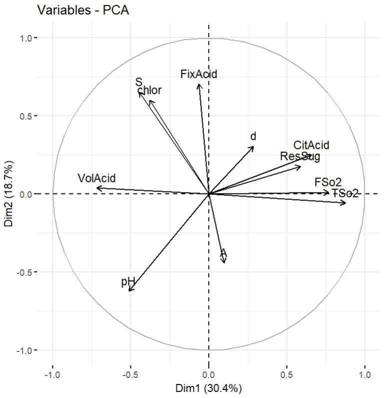
Finally the medium level group are present in all quadrants except the top right one. We can say that top-medium quality tend to be in the bottom left quadrant.



1.1.6 Plot the circle of correlations and draw up some conclusion(s).

In the circle correlation plot we can see visually what we described previously in the interpretation of the first 2 PCA components, that is, *'total sulfur dioxide'*, *'free sulfur dioxide'*, *'citic acid'* and *'residual sugar'* are highly correlated with Dim1 and negative correlated with *VolAcid*.

On the other hand Dim2 is positive correlated with *fixed acidity, chlorides* and *sulphates*, while having a negative correlation with *ph* and *alcohol.*

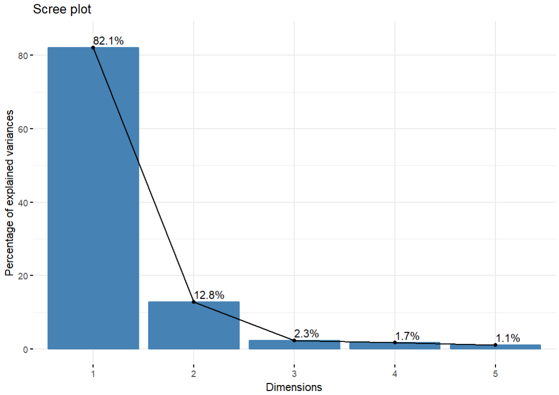


1.2 Perform a Principal Component Analysis on the cars data set (on the set of 5 quantitative variables).

1.2.1 Interpret your results.

After performing the principal component analysis, we can see that the eigenvalues measure the amount of variation retained by each principal component. Eigenvalues are large for the first PCs and small for the subsequent PCs.

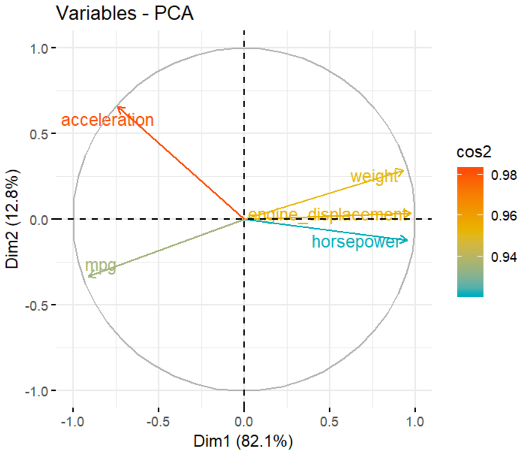
We examine the eigenvalues to determine the number of principal components to be considered. In our case about 95% of the variation is explained by the first two PCs. We will explore deeper in the following sections.



1.2.2 Plot the circle of correlations and draw some conclusion(s).

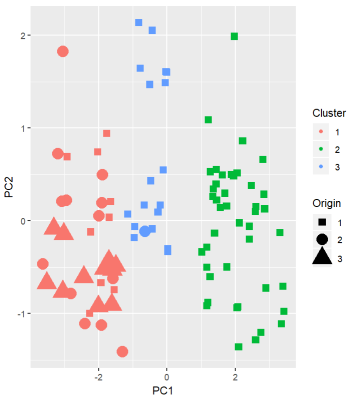
This kind of plot shows the relationships between all variables. It can be interpreted as follow:

- It shows the correlation of each variable versus the PC.
- Positively correlated variables are grouped together.
- Negatively correlated variables are positioned on opposite sides of the plot origin (opposed quadrants).
- *Acceleration* and *Engine_ diplacement* are the variables that contribute the most to the representation.
- A high cos2 indicates a good representation of the variable on the principal component. In this case the variable is positioned close to the circumference of the correlation circle. A low cos2 indicates that the variable is not perfectly represented by the PCs. In this case the variable is close to the center of the circle. In our case every variable is very well represented by only two PCs. The sum of cos2 for the acceleration variable equals 1 which indicates perfect representation by the two PCs (Check the scale).

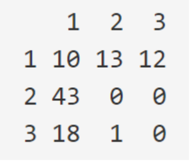


1.2.3 Use the k-means function in R on the five principal component scores with $k = 3$, $nstart = 25$and iter.max=100. Before running it, set the random seed to 12345 to obtain reproducible results ($set.seed(12345)$). Explore visually how well the clustering recovers the actual origin of the cars. Conclusions.

After applying the k-means function we obtain the following:



We can observe that the second and third clusters have almos all elements from origin 1 and the cluster 1 have elements from all 3 origins.



```
     1  2  3
1 10 13 12
2 43  0  0
3 18  1  0
```

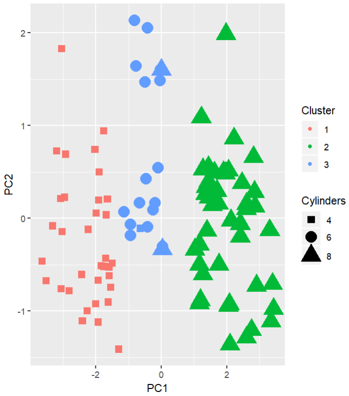
This means that the cluster does not recover well the actual origin of the cars, since the origins are mixed among the clusters.

1.2.4 k-mean clustering - Lumping and splitting errors.

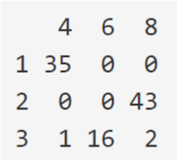
After creating the function for calculating the lumpong and splitting errors we have obtained a lumping error of 424 and a splitting error of 1397.

1.2.5 Repeat the same analysis using cylinders as the categorical variable to be matched by the clustering algorithm. As there is only one car with 3 cylinders, don't use it in the analysis and consider the variable cylinder taking values 4, 6 and 8.

In this case we obtain a better classification:



With a lumping error of 50 and a splitting error of 121. Here is the classification:



```
     4  6  8
1 35  0  0
2  0  0 43
3  1 16  2
```

This could be due to a better explanation of the variance of the dataset from the variable cyl, so when we apply PCA to the observations with different sources of variance are separated and the k-mean cluster is able to differenciate them better.