

Intelligent Data Analysis - Homework 1.3

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

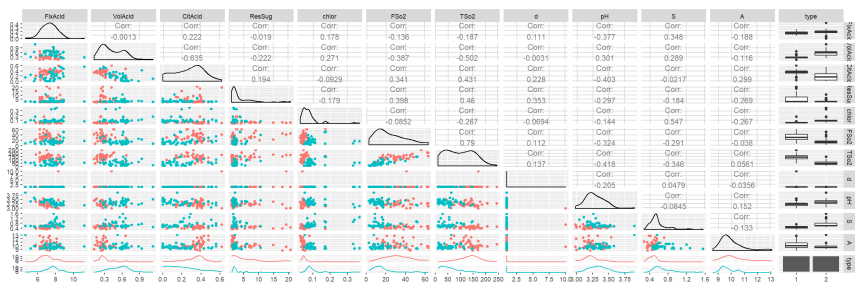
November 4th, 2018

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 contribute 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. By analyzing the results we can say which variables contribute the most to each one:

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	

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

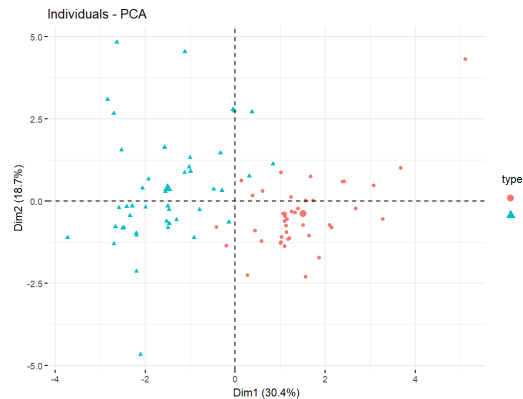
If we analyze the first two PC, we can check which variables are strongly correlated with the components, so values with high values either in positive or negative directions will be considered highly correlated, ($> |0.5|$). The first component is highly correlated with 'total sulfur dioxide', 'free sulfur dioxide', 'citric 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

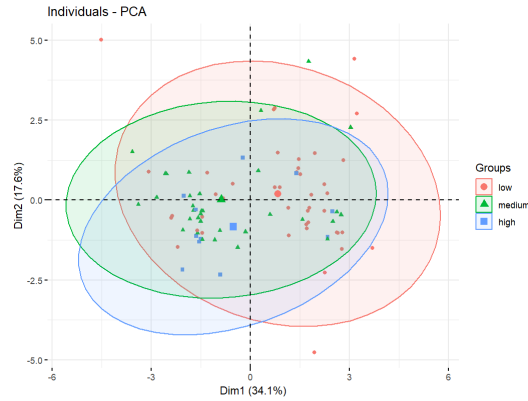
1.1.4 Does any of the two first components separate wines by their "type"? Plot the observations on the space spanned by the two first components and colour them according to "type". What do you observe?

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 assume that Dim1 separates wines according to type more than Dim2.



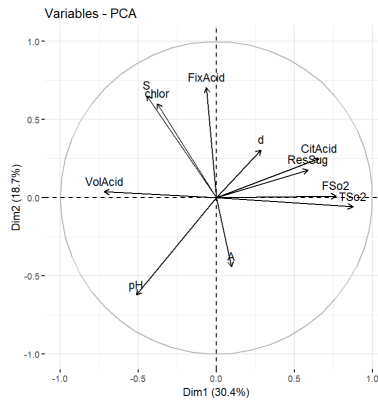
1.1.5 Plot the observations on the space spanned by the two first components and colour them according to "quality". Is there anything that deserves a comment?

We can appreciate how the second principal component seems to divide the zones with high quality versus the zone with lower quality, being the majority in the bottom right quadrant. Also the top left quadrant seems to albergate mostly low quality, although low quality elements are more or less equally distributed among all the quadrants. Finally the medium level group are present equally distributed in all quadrants except the top left one. We can say that top-medium quality tend to be in the bottom right quadrant.



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

In the circle correlation plot we can see the reflection of what we described previously in the interpretation of the first 2 PCA components, the closer to the PC1 components the wines are the more the sulfur dioxide and acidic acid they will have, aslo the lower they close to DIM2, the less alcoholic those wines would be.

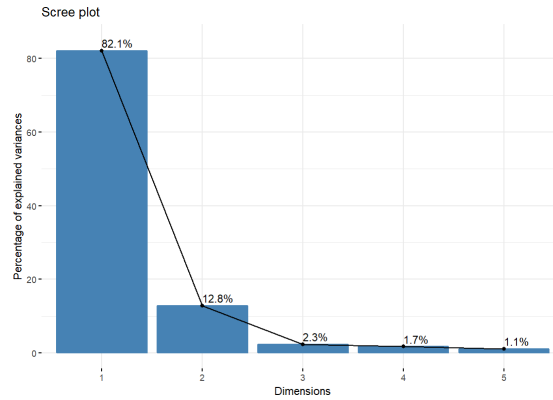


1.2 Perform a Principal Component Analysis on the cars data set (on the set of 5 quantitative variables).

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

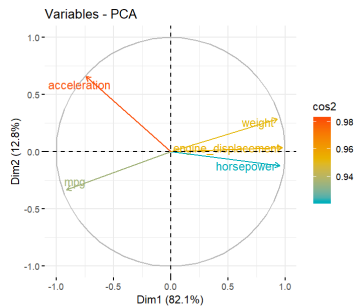
be considered. In our case about 95% of the variation is explained by the first 2 Pcs. We will explore deeper in the following sections.



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

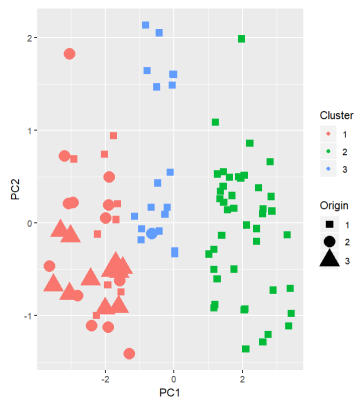
This kind of plot shows the relationships between all variables. It can be interpreted as follow:

- Positively correlated variables are grouped together.
- Negatively correlated variables are positioned on opposite sides of the plot origin (opposed quadrants).
- The distance between variables and the origin measures the quality of the variables on the factor map. Variables that are away from the origin are well represented on the factor map.
- Acceleration and Engine_displacement are the 2 variables that contribute the most to the representation.
- A high \cos^2 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 \cos^2 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 2 PCs. The sum of \cos^2 for the acceleration variable equals 1 which indicates perfect representation by the 2 PCs.



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 2nd and 3rd clusters have almos all elements from origin 1 and the cluster 1 have elements from all 3 origins.

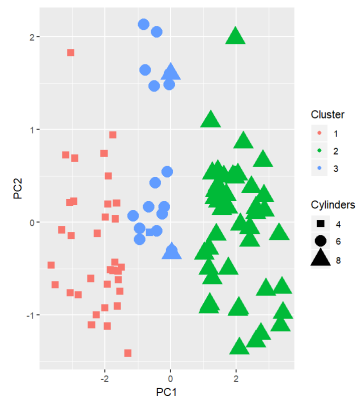
	1	2	3
1	10	13	12
2	43	0	0
3	18	1	0

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