Multivariate Methods (Final Report)

Supti Biswas

Abstract

In this analysis, I performed Principal Component Analysis, Factor Analysis, Canonical Correlation and Clustering Method to figure out which variables are mostly responsible to explain the variation in the data. Also, which variables are correlated with each other in order to reduce the dimension of the data. I used R programming language.

Introduction, Problem Formulation & Literature Review

The package 'HDclassif' in R has a large dataset named 'wine' containing 178 rows and 13 variables. The wines are grown in 3 different cultivars and the 13 variables are representing the results of a chemical analysis of the wines. As there are a lot of ingredients which are forming the wines, our question is which variables are mostly related to each other to form a specific type of wine.

As the multivariate methods of statistics show us how the variables are related to each other when the dimension is high and there are lot of variables which we are assuming to be correlated, we considered this dataset to see how the variables are related to each other, which variables are mostly responsible for the variation in the wines.

Data Characterization, Descriptive Analysis and Visualization

'Wine' data has a dimension of 178 X 14. These data are the results of a chemical analysis of wines grown in the same region in Italy but derived from three different cultivars. The analysis determined the quantities of 13 constituents found in each of the three types of wines. The data is available in the UCI Machine Learning Repository https://archive.ics.uci.edu/ml/datasets/Wine. There is no missing value in the dataset. Here are the first 5 rows of the data –

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13
1	14.23	1.71	2.43	15.6	127	2.80	3.06	0.28	2.29	5.640000	1.040	3.92	1065
2	13.20	1.78	2.14	11.2	100	2.65	2.76	0.26	1.28	4.380000	1.050	3.40	1050
3	13.16	2.36	2.67	18.6	101	2.80	3.24	0.30	2.81	5.680000	1.030	3.17	1185
4	14.37	1.95	2.50	16.8	113	3.85	3.49	0.24	2.18	7.800000	0.860	3.45	1480
5	13.24	2.59	2.87	21.0	118	2.80	2.69	0.39	1.82	4.320000	1.040	2.93	735

Here are the details of the variables -

class - The class vector, the three different cultivars of wine are reprensented by the three integers: 1 to 3. V1 - Alcohol

V2 - Malic acid

V3 - Ash

V4 - Alcalinity of ash

V5 - Magnesium

V6 - Total phenols

V7 - Flavanoids

V8 - Nonflavanoid phenols

V9 - Proanthocyanins

V10 - Color intensity

V11 - Hue

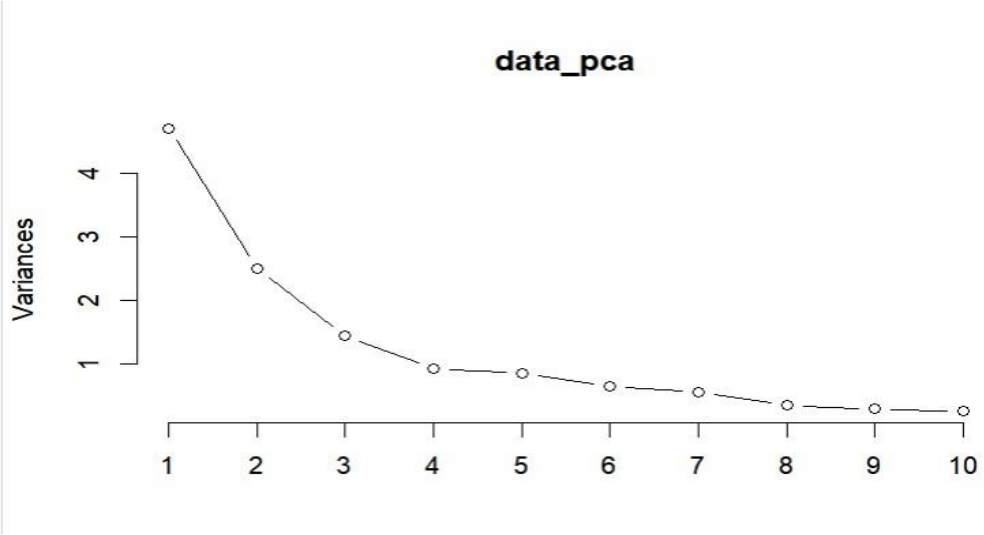
V13 – Proline

Methods

- There are 13 variables in the data. So, Principal Component Analysis would be appropriate to see how many PCs are sufficient to explain most of the variation in the data.
- Factor Analysis would be appropriate to visualize which variables are mostly correlated with each other among these large number of variables.
- Canonical Correlation would be appropriate because it simultaneously identifies the sources of variation that bear strongest statistical associations between both sources of variation
- As the Factor Analysis is performed and 3 factors are found there, we can compare this with the number of clusters which can be found in cluster analysis.
- The Classification methods wouldn't be appropriate here because there is no categorical variables here, all of them are quantitative variables.

Data Analysis and Main Results

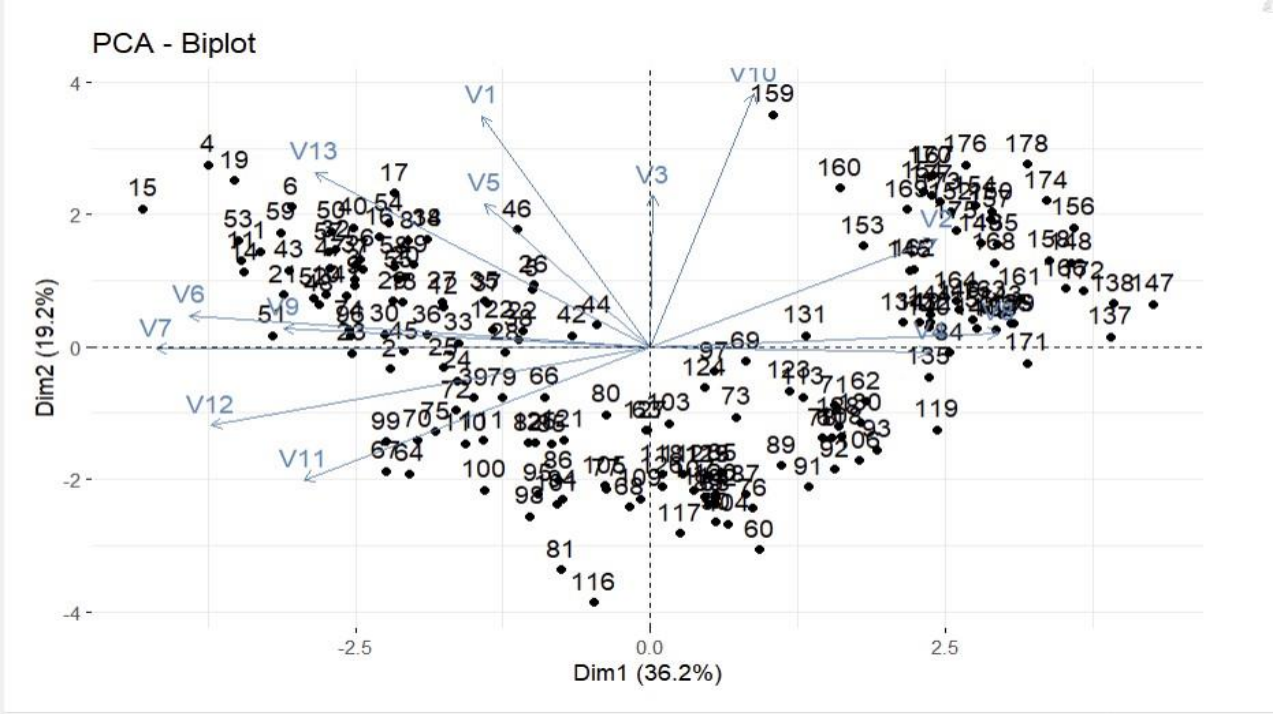
- **Principal Component Analysis –** After performing the PCA, we can take the first 3 PCAs as almost 67% variation of the dataset can be explained by the 3 PCAs. This is also clear from the screeplot below –



From the loading vectors plot and biplot, we see that Ash(V3), Color intensity(V10), Alcohol(V1), Magnesium(V5) and Proline(V13) are highly correlated, and they are forming the first principal component that explains around 37% of variation of the whole dataset.

On the other hand, Alcalinity of Ash(V4), Nonflavanoid Phenols(V8), Total Phenols(V6), Flavanoids(V7), Proanthocyanins(V9), Hue(V11) and OD280/OD315 of Diluted Wines(V12) variables are highly correlated with each other and forming the second principal component that explains around 20% of the total variation of the dataset.

The biplot is shown below -

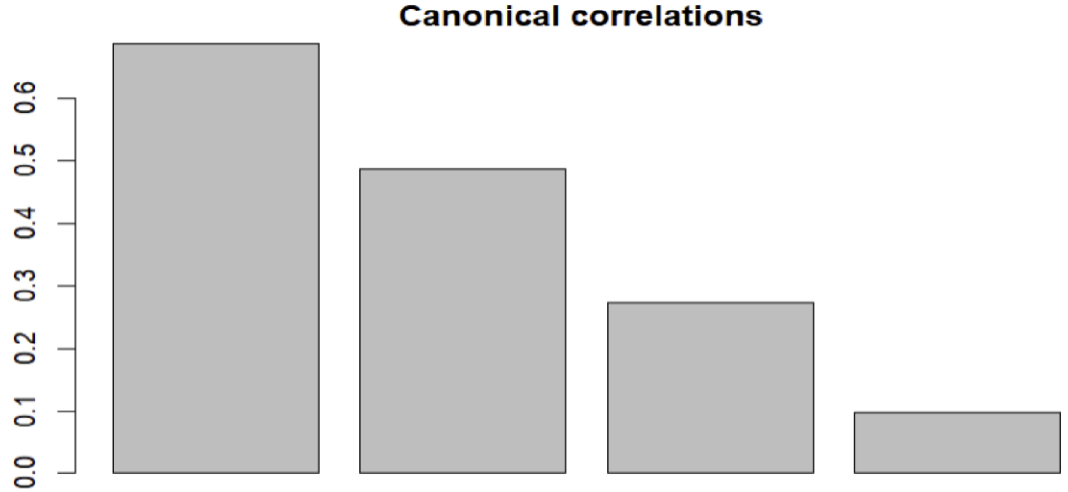


- **Canonical Correlation** - To conduct canonical correlation, we need two sets of variables. From the Principal Component Analysis, I have taken 4 variables from each principal components which are mostly correlated to each other and trying to see the canonical correlation among them**.**

  Set 1 – Ash, Color Intensity, Alcohol and Magnesium

  Set 2 – Alcalinity of Ash, Flavanoids, Nonflavanoid Phenols, Proanthocyanins

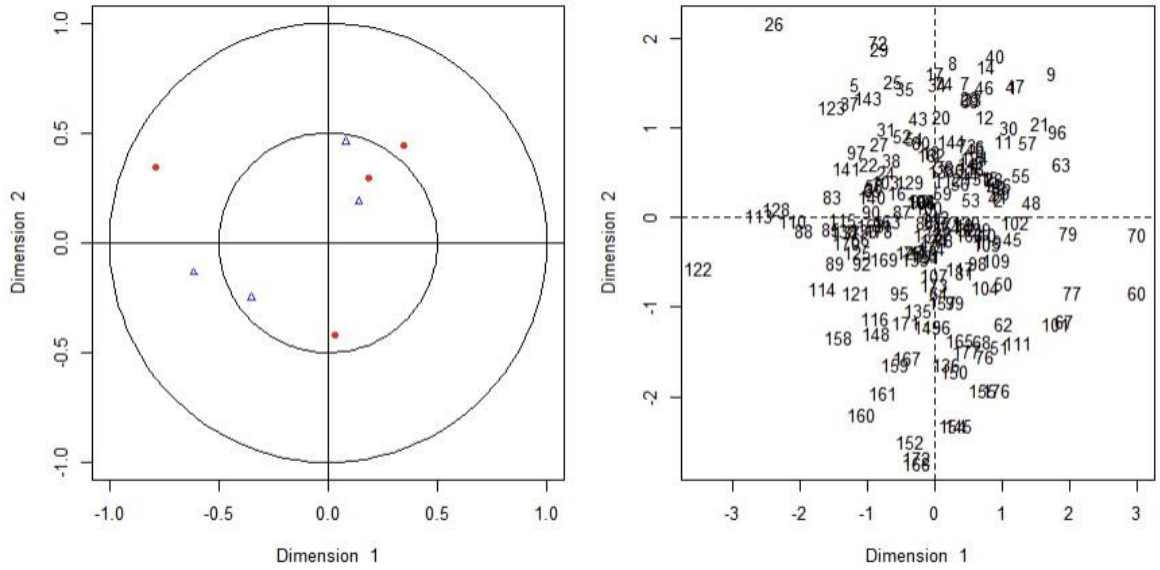
The following barplot shows the canonical correlation among these 2 groups of variables



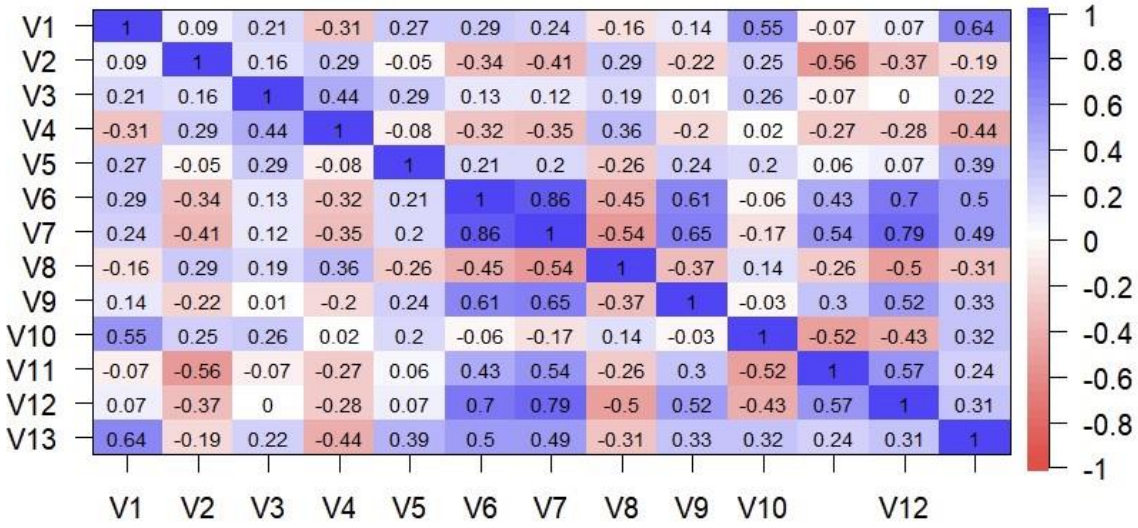
- The correlation values are as follows –

  0.68744483 0.48619581 0.27343023 0.09646641

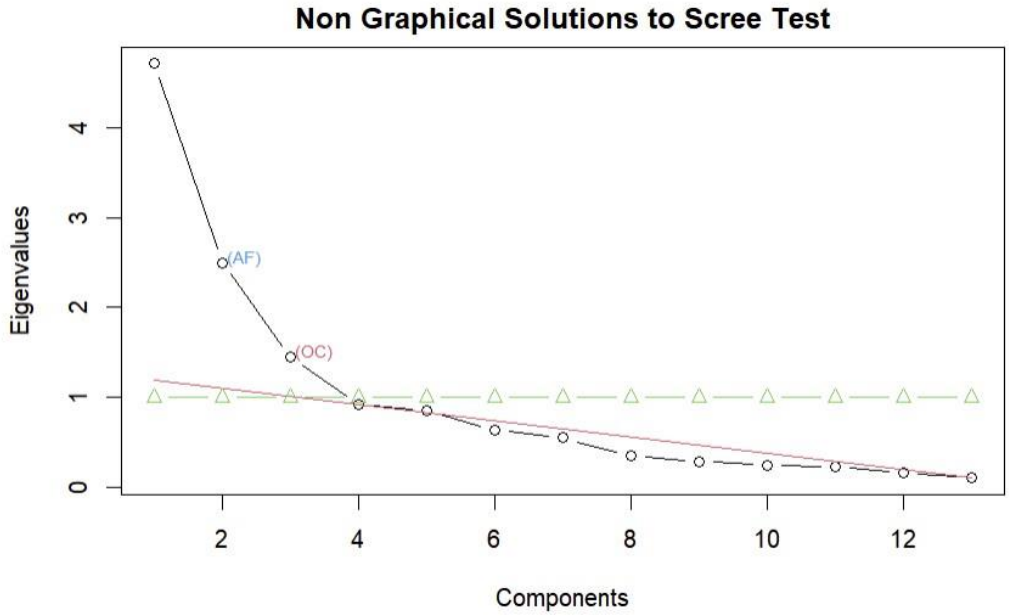
Another graphical representation of the canonical correlation among the variables –



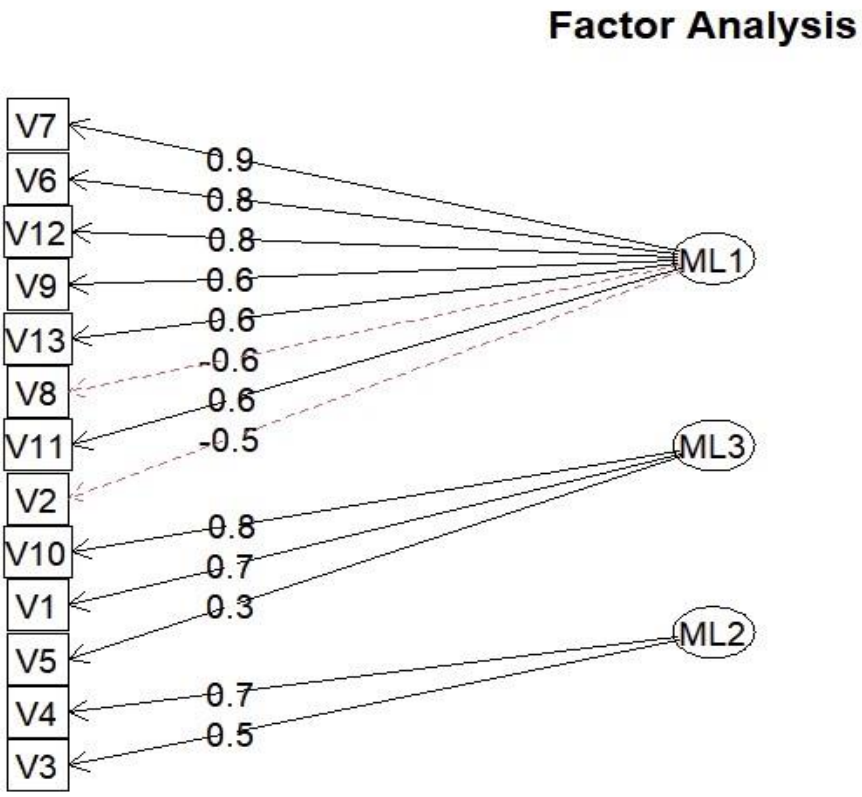
The correlation matrix plot among all of the variables are shown below –



- **Factor Analysis –** To group the variables, I will take 3 factors based on the criteria that the eigen values are greater than 1. The green triangles in the graph below is showing the line where eigen value is 1.



The diagram below shows the factors that are mostly correlated with each other –



Factor 1 - Flavanoids (V7), Total Phenols (V6), OD280/OD315 of diluted wines (V12), Proanthocyanins (V9), Proline (V13) and Hue (V11) are making the first factor, i.e., they are mostly correlated with each other to explain the variation among the 13 variables. Nonfalavanoid Phenols (V8) and Malic Acid (V2) are negatively correlated with the others in this factor.

Factor 2 – Alcalinity of Ash (V4) and Ash (V3) are mostly correlated by making the second factor.

Factor 3 – Color Intensity (V10), Alcohol (V1) and Magnesium (V5) are mostly correlated by making the third factor.

The loading vector outputs are as follows –

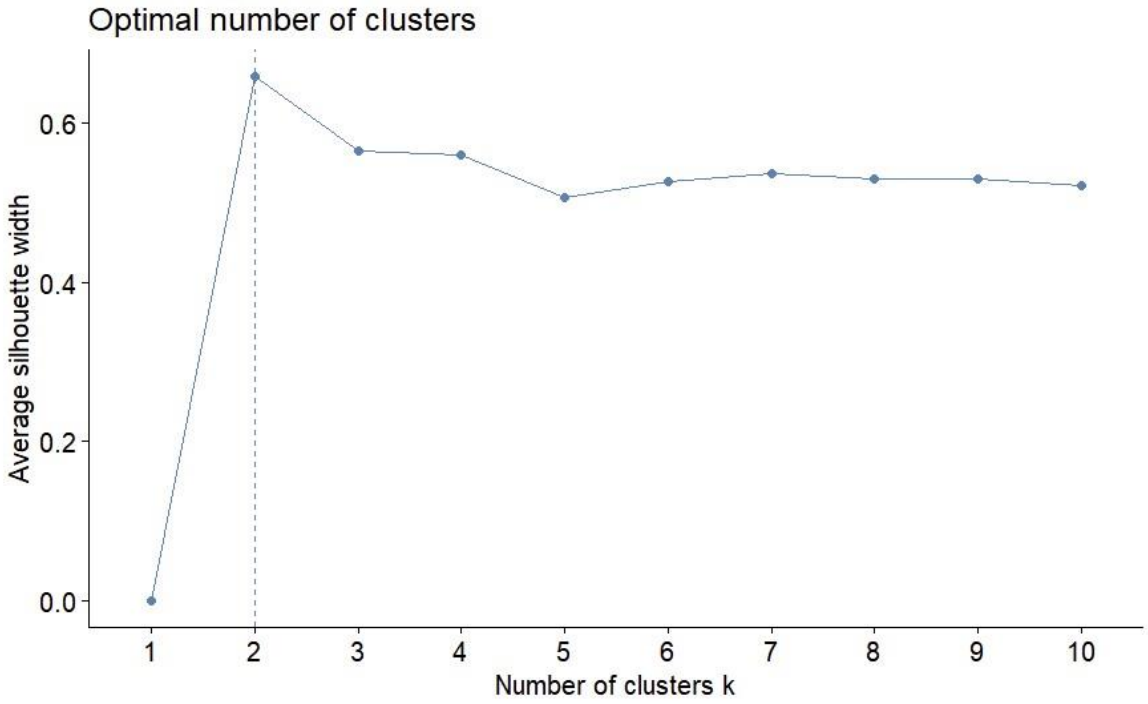
```
Loadings:
    ML1    ML3    ML2
V1   0.318  0.675 -0.235
V2  -0.454  0.260
V3          0.480  0.493
V4  -0.628         0.725
V5   0.212  0.342
V6   0.840  0.150  0.270
V7   0.915         0.305
V8  -0.582
V9   0.614         0.242
V10 -0.165  0.825 -0.215
V11  0.568 -0.389  0.154
V12  0.773 -0.231  0.311
V13  0.586  0.498 -0.156

                 ML1   ML3   ML2
SS loadings    4.300 2.048 1.241
Proportion Var 0.331 0.158 0.095
Cumulative Var 0.331 0.488 0.584
```

- **Comparison between PCA and FA -** From the results of these two methods, we see that two groups of variables are mostly correlated by both methods –

  Group 1 – Color Intensity (V10), Alcohol (V1) and Magnesium (V5)

  Group 2 – Total Phenols (V6), Flavanoids (V7), Proanthocyanins (V9), Hue (V11) and OD280/OD315 of diluted wines (V12)

- **Clustering Method –** By using clustering method, we can decide the optimal number of clusters. The result below is the elbow method -



So, there should be 2 clusters for this dataset though this part is subjective in clustering method.

Comparing Factor Analysis and Clustering Method - We found that these two results do not match. 3 factors are found in the FA to explain the variation but 2 clusters will be optimal in the clustering method.

Conclusions and Research Directions

There could be some further research in this dataset. We can add another variable named 'Price' that will show the price of these wines. According to the prices, it can be divided into 3 or 4 categories based on their price (Low, Average, High). Then classification methods would be appropriate to analyze the data and we could run a logistic regression model as well taking the categorical variable as the response.