Applying Multivariate Methods on ‘wine’ data using R

BY SUPTI BISWAS

Data Description

- The package 'HDclassif' in R has a large dataset named 'wine' containing 178 rows and 13 variables representing the results of a chemical analysis of wines grown in 3 different countries. The variables taken into consideration are as follows -

class - The class vector, the three different cultivars of wine are represented 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

V12 – OD280/OD315 of diluted wines

V13 – Proline analysis of the wines

Question from the dataset

As we know that the variables in the dataset are related to each other, our question lies in –

- -Which variables are mostly responsible for the variation in the wines?
- -If yes, how they are correlated with each other?
- -Is it possible to make a set of variables which are mostly correlated with another one?

Descriptive Analysis

- 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

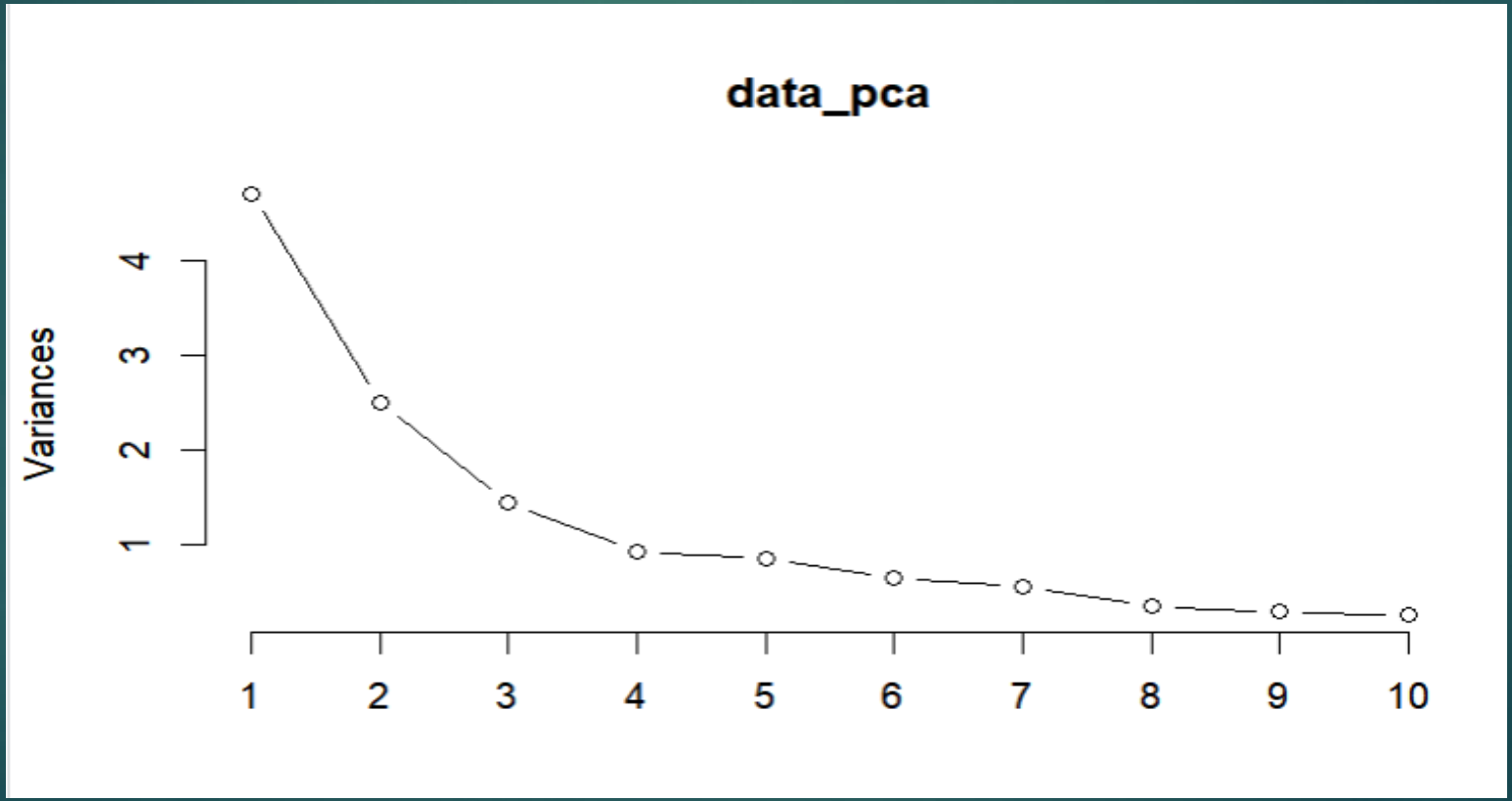
- All the variables except ‘class’ (removed) are continuous variables
- There is no missing value in the dataset
- Variables are correlated with each other

Methods of Data Analysis

- Principal Component Analysis
- Canonical Correlation
- Factor Analysis
  - Comparison between FA and PCA
- Clustering Method

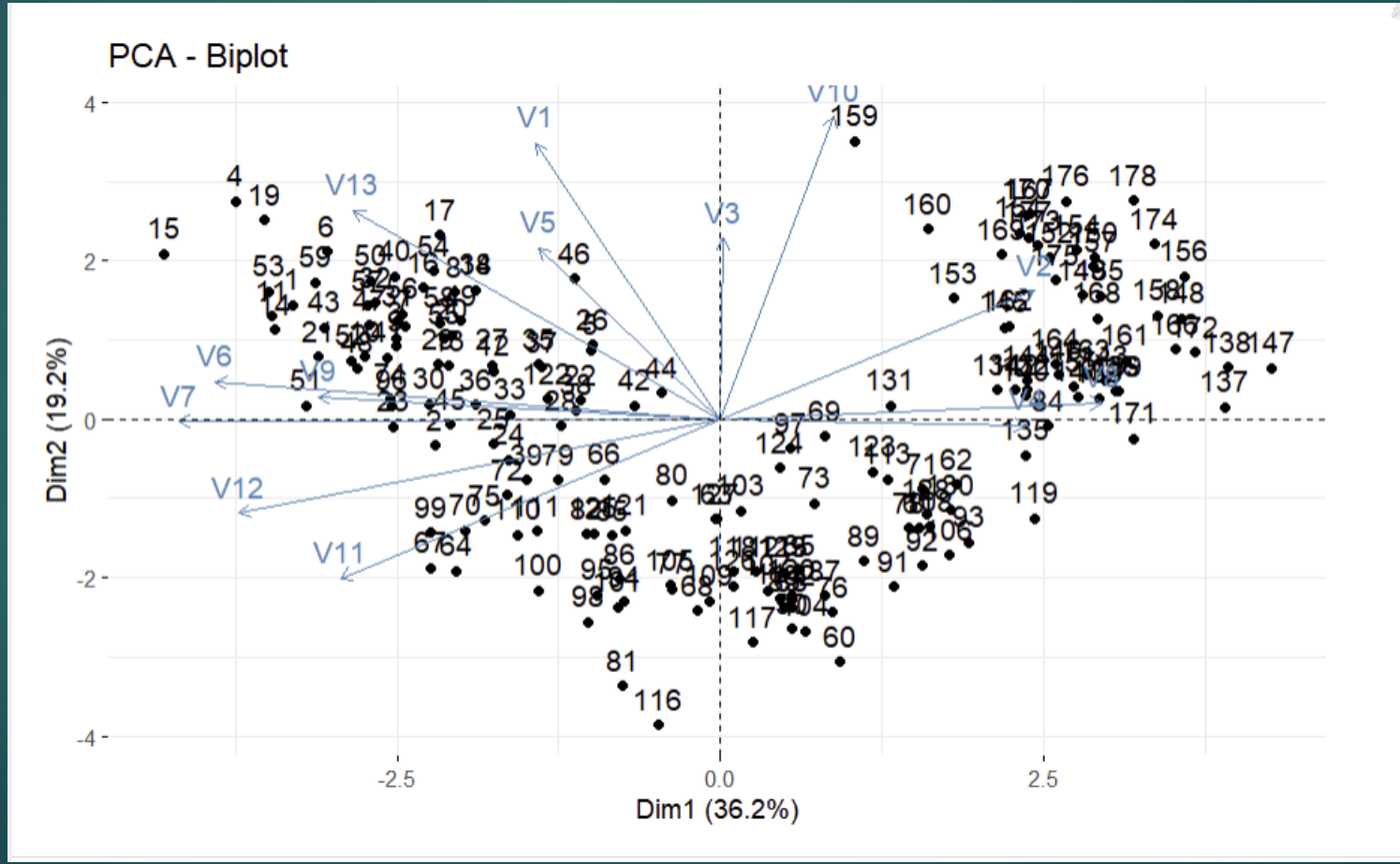
Principal Component Analysis

- Screeplot



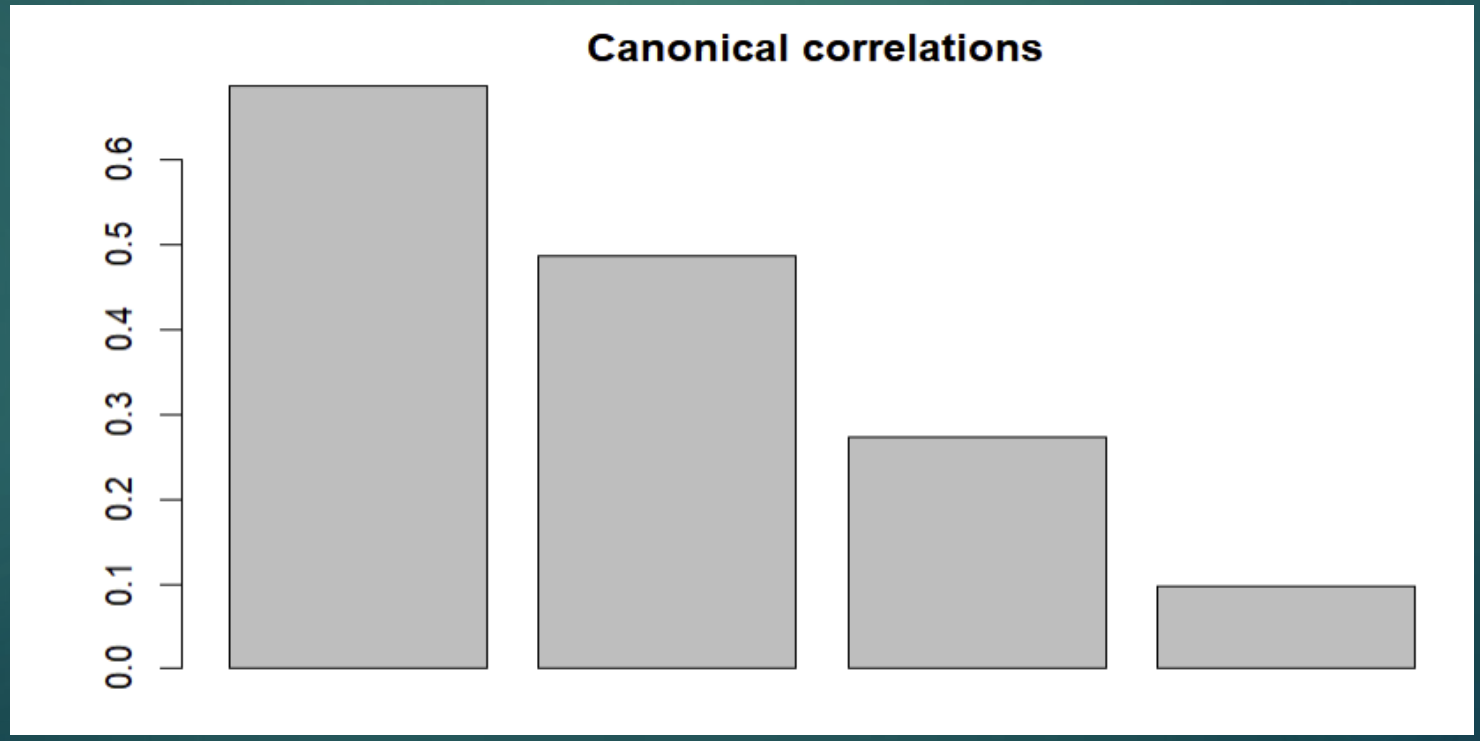
Principal Component Analysis

- Biplot –



Canonical Correlation

- Set 1 – Ash, Color Intensity, Alcohol and Magnesium
- Set 2 – Alcalinity of Ash, Flavanoids, Nonflavanoid Phenols, Proanthocyanins
- Barplot –

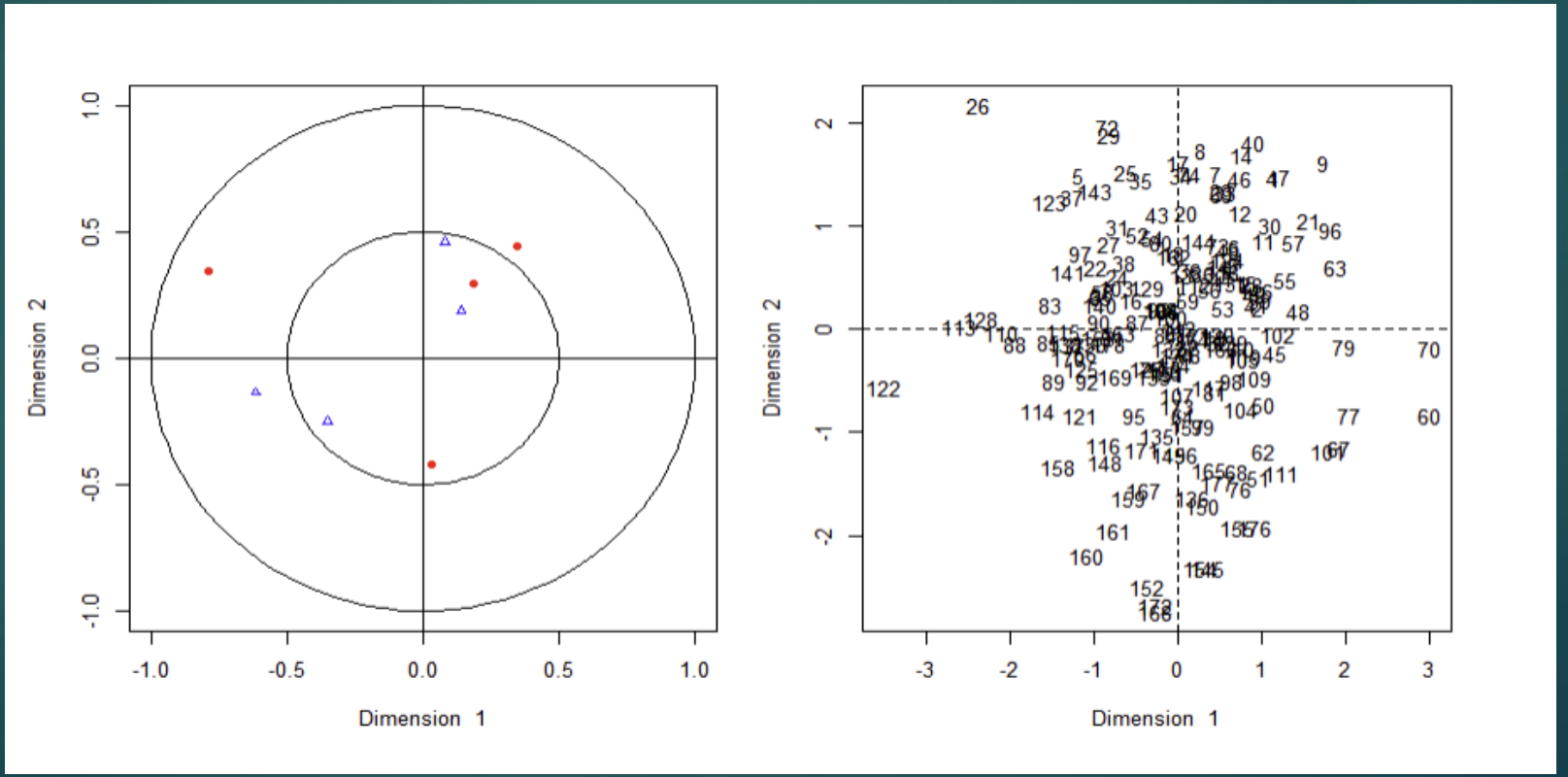


Canonical Correlation

- Correlation values are

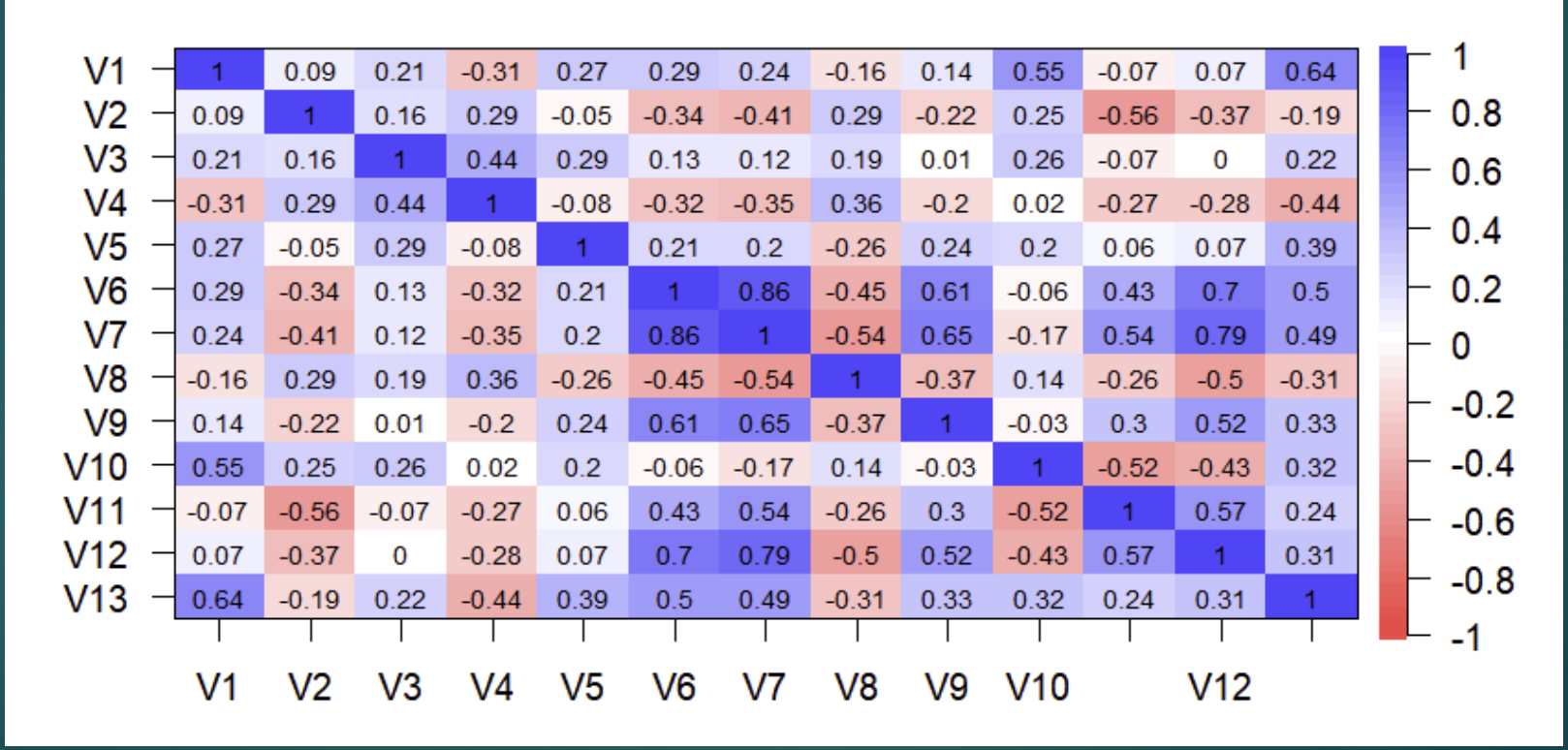
  0.68744483 0.48619581 0.27343023 0.09646641

- Graphical Representation using Dimensions –



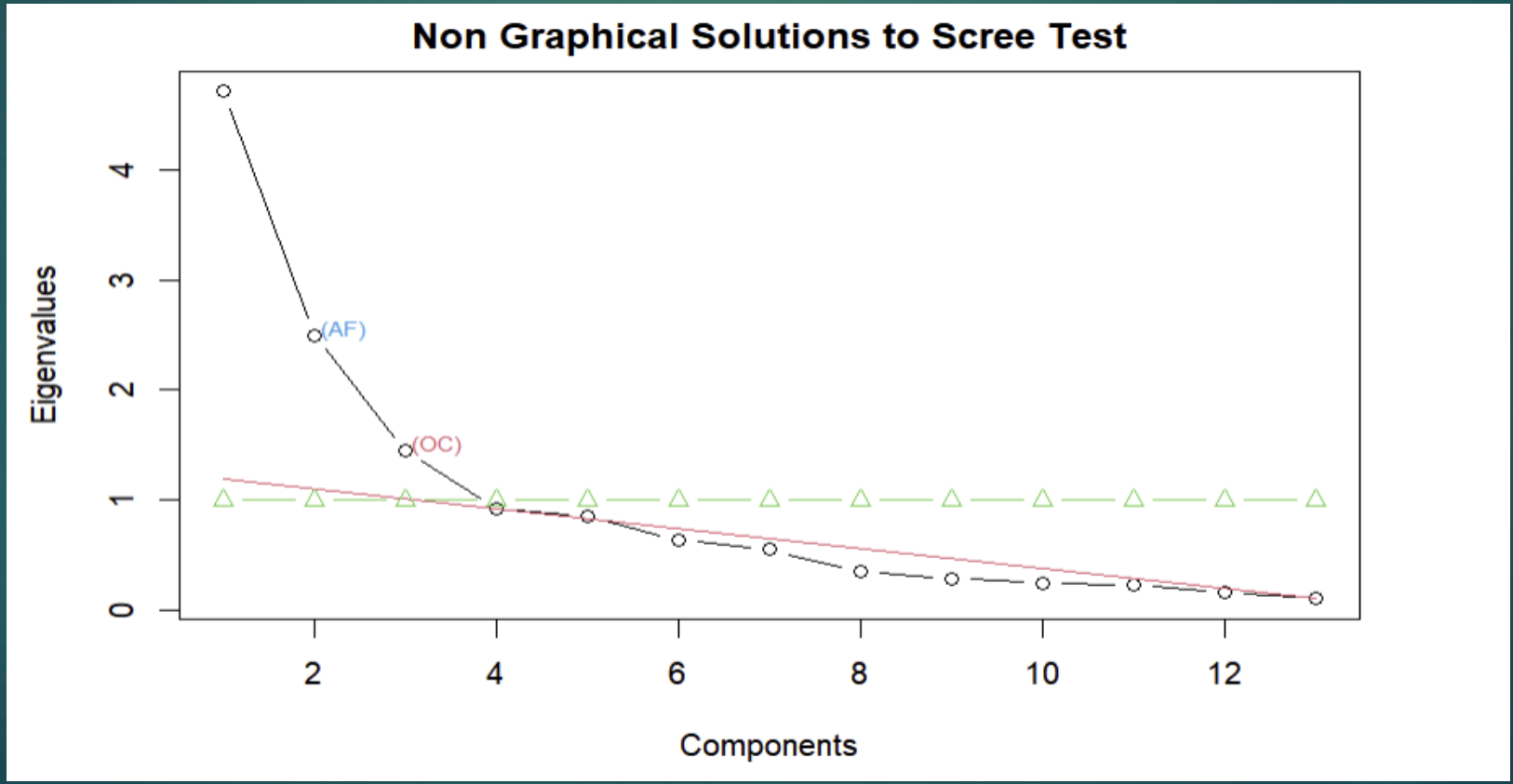
Canonical Correlation

- Correlation Matrix Plot



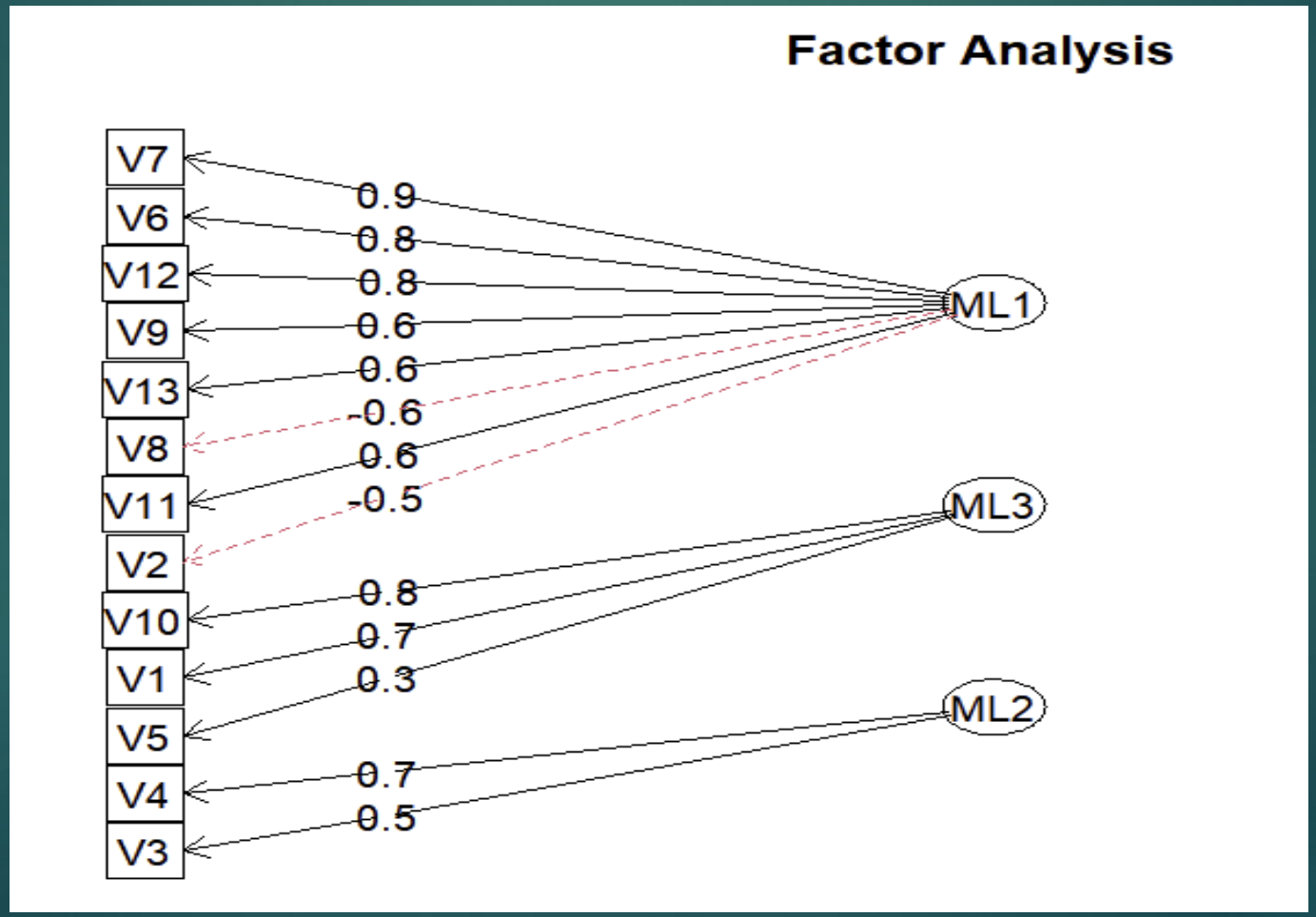
Factor Analysis

- Screeplot



Factor Analysis

- FA Diagram



Factor Analysis

- 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. Nonflavanoid 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.

Factor Analysis

- Loading Vector Output (using R) –

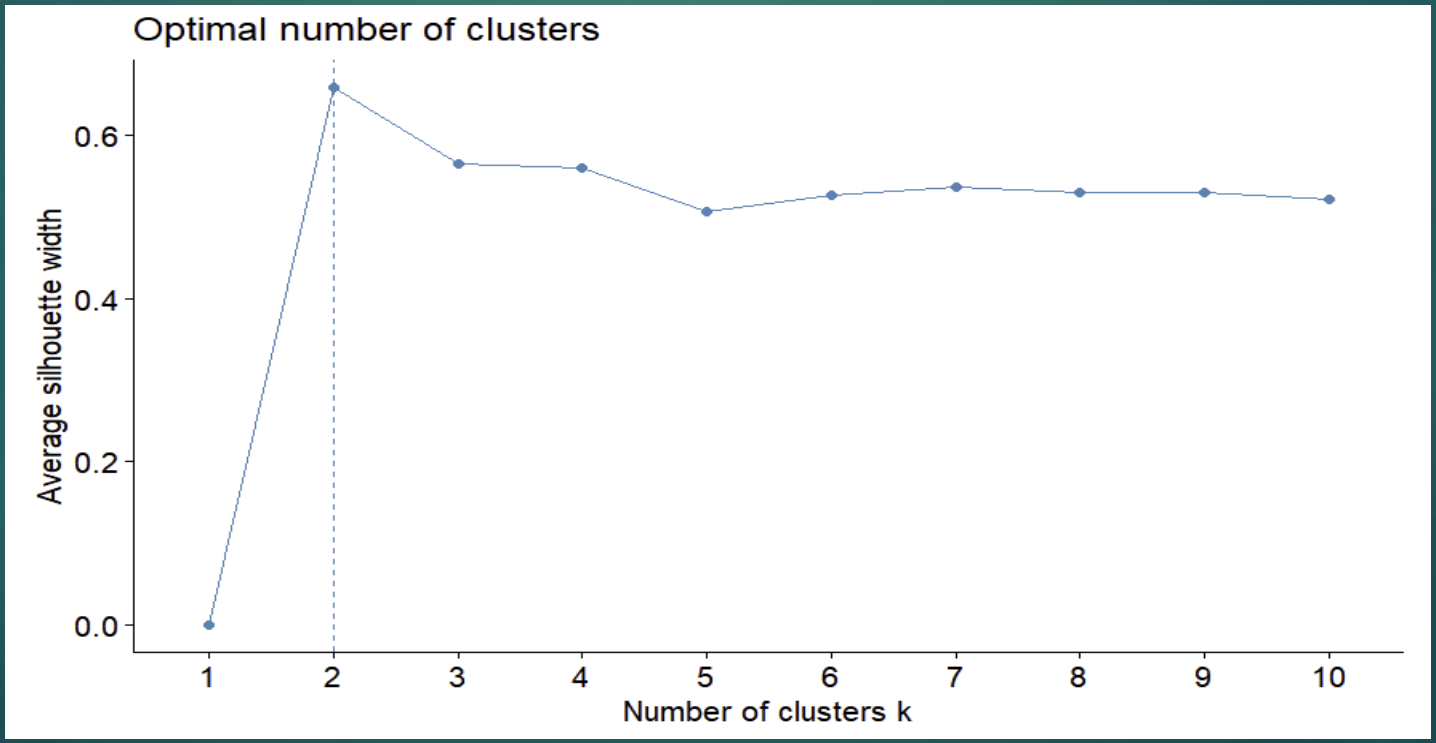
```
Loadings:
     ML1    ML3    ML2
V1    0.318  0.675 -0.235
V2   -0.454  0.260
V3           0.480  0.493
V4   -0.628         0.725
V5    0.212  0.342
V6    0.840  0.150  0.270
V7    0.915         0.305
V8   -0.582
V9    0.614         0.242
V10 -0.165  0.825 -0.215
V11  0.568 -0.389  0.154
V12  0.773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

- Graph using elbow method
- FA has 3 clusters



Conclusion

- Adding Another Variable named ‘Price’
- Using Classification
- Logistic Regression Model