# Exploratory Factor Analysis and Confirmatory Factor Analysis Report

ST405 – Assignment 01

S/18/805

## 1. Introduction

Factor analysis is a statistical method that is used to reduce large number of variables into fewer number of underlying (latent) factors. It is a method of dimension reduction. This reduction is done by uncovering the inter-correlation between variables. It is possible that variation of few observed variables mainly reflect the variation in one unobserved variable (factor). By doing factor analysis we will be able to interpret the dataset more simply and find the hidden pattern.

We can conduct a factor analysis on a dataset in two ways.

I. Exploratory Factor Analysis (EFA)
II. Confirmatory Factor Analysis (CFA)

Exploratory Factor Analysis is used when we have no knowledge or a little knowledge about factors and relationship between variables. In EFA we explore the underlying structure of the observed variables and determine the number of factors needed to explain the observed variance sufficiently. And also the relationship between each variable and factor.

Confirmatory Factor Analysis is used after identifying the factors and creating the model. CFA specify how well the data are fitted to the measured model. It is hypothesis testing method.

### 1.1 Major Question

In here we asses a dataset and find how many factors are needed, how well those factors explain the variability of the data. Then we fit a model and check adequacy of the fitted model.

### 1.2 Purpose of the study

Purpose of this study is to perform Exploratory Factor Analysis and Confirmatory Factor Analysis on **Wine** dataset.
By doing a factor analysis we aim to simplify the data in order to interpret the relationships among variables. Also we need to find the hidden pattern of the data so that we can understand the data accurately.

## 2. Methodology

Dataset used in this study is a result of a chemical analysis of wine grown in the same region in Italy those are derived from three different cultivars for the purpose of determine the origin of the wine. After analyzing quantities of 13 constituents found in each of the three types of wine were recorded.
There are 14 attributes (variables) and 178 instances. First attribute is a class identifier. It is labeled 1-3. All other attributes are continuous. There are no missing values in the dataset.

Attribute list:

1. Class
2. Alcohol
3. Malic acid
4. Ash
5. Alkalinity of ash
6. Magnesium
7. Total phenols
8. Flavanoids
9. Non-flavanoid phenols
10. Proanthocyanins
11. Color intensity
12. Hue
13. OD280/OD315 of diluted wines
14. Proline

Before doing the factor analysis we remove first attribute which is "class" because it is a class identifier and treated as a categorical variable. Since there are 13 feature attributes analyze and interpret is complex. Thus, we perform EFA and CFA. In EFA, both principal component method (PC) and maximum likelihood method (ML) are used with “varimax” rotation method.

Since variables did not measured under same units, for all the calculations correlation matrix of the variables are used (Because variance-covariance of standardized data are equal to the correlation matrix of data which are not standardized).

Then we focus on confirmatory factor model with some latent variables and adequacy of the model.

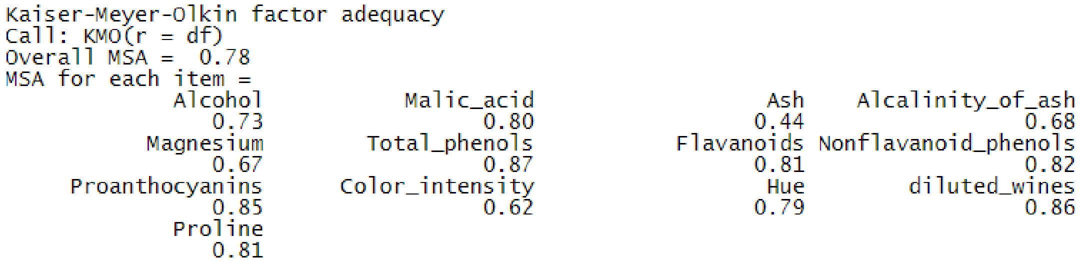
## 3. Results and Discussion

### 3.1 Exploratory Factor Analysis(EFA)

#### 3.1.1. Suitability of the dataset

First, we need to determine whether the selected dataset is suitable for apply the factor analysis methods. In order to do that, we can use Kaiser-Meyer-Olkin (KMO) test and Bartlett’s test.

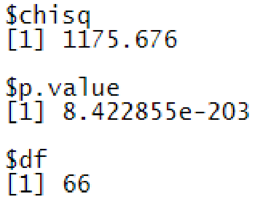
- KMO Test Result



```
Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = df)
Overall MSA =  0.78
MSA for each item = 
             Alcohol           Malic_acid                  Ash    Alcalinity_of_ash 
                0.73                 0.80                 0.44                 0.68 
           Magnesium        Total_phenols           Flavanoids Nonflavanoid_phenols 
                0.67                 0.87                 0.81                 0.82 
     Proanthocyanins      Color_intensity                  Hue        diluted_wines 
                0.85                 0.62                 0.79                 0.86 
             Proline 
                0.81 
```

Overall MSA value is 0.78 which is good and it indicates that the variables in dataset are moderately to highly correlate with each other. It suggest there are underlying factors that explain the relationship among variables. Thus, factor analysis is an appropriate technique for this dataset.
But since there are variable with low individual MSA value(less than 0.5), remove those low contribution variables.
After removing those, overall MSA is 0.81 which is much better than previous MSA with all the remaining variables have individual MSA greater than or equal to 0.6.

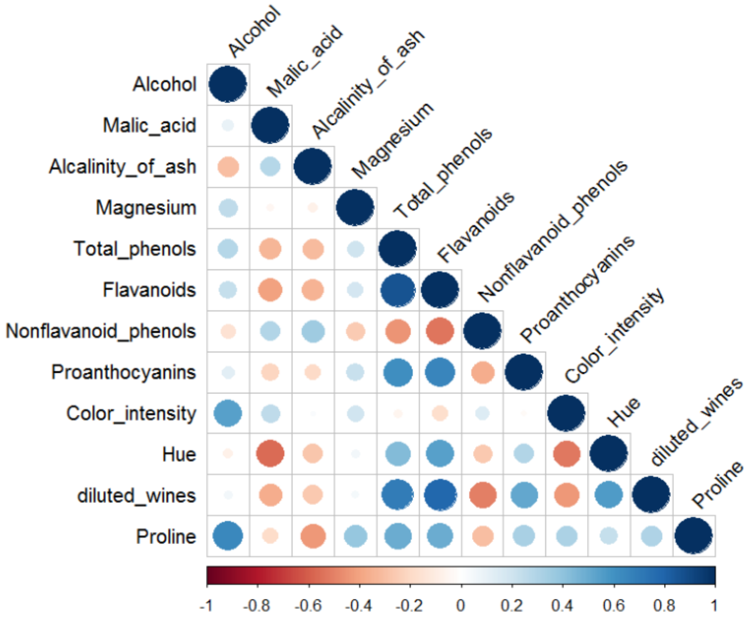
- Bartlett Test Result



```
$chisq
[1] 1175.676

$p.value
[1] 8.422855e-203

$df
[1] 66
```

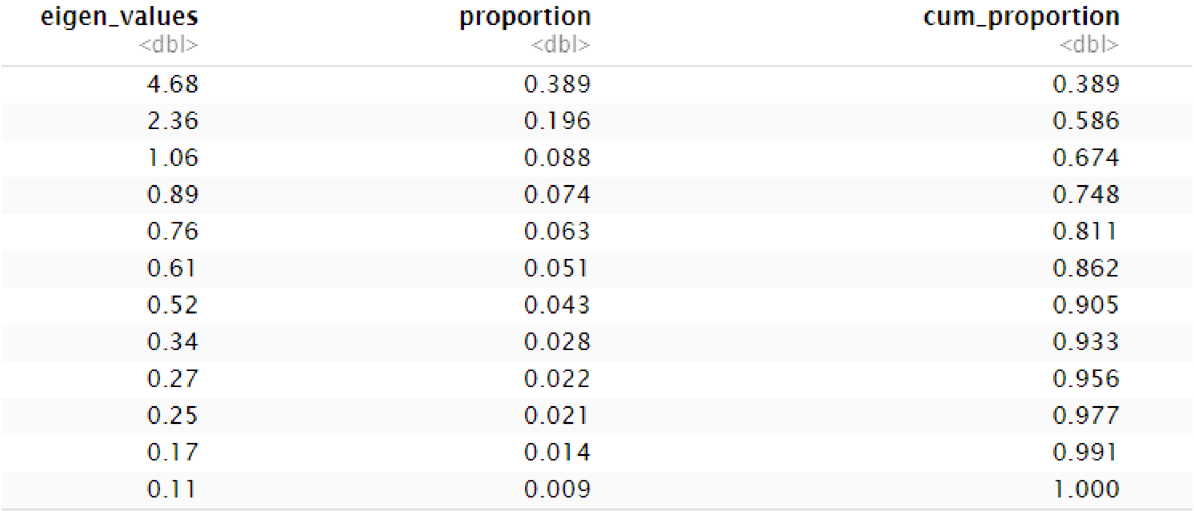
According to the Bartlett test result, significance level is less than 0.05. Variable are correlated. Hence, we conclude that the chosen “Wine” dataset is suitable for apply factor analysis technique.

Correlation among variables in processed dataset are shown in following plot.
There are 12 variables remain in the dataset.



Now, we need to determine the number of factors that are suitable to explain the variation of data. There are three ways to do that.

- Eigen values



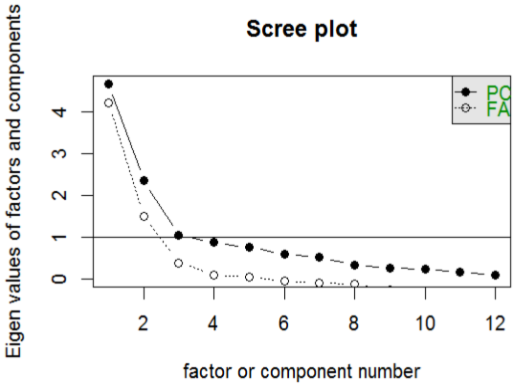
| eigen_values<br><dbl> | proportion<br><dbl> | cum_proportion<br><dbl> |
|---|---|---|
| 4.68 | 0.389 | 0.389 |
| 2.36 | 0.196 | 0.586 |
| 1.06 | 0.088 | 0.674 |
| 0.89 | 0.074 | 0.748 |
| 0.76 | 0.063 | 0.811 |
| 0.61 | 0.051 | 0.862 |
| 0.52 | 0.043 | 0.905 |
| 0.34 | 0.028 | 0.933 |
| 0.27 | 0.022 | 0.956 |
| 0.25 | 0.021 | 0.977 |
| 0.17 | 0.014 | 0.991 |
| 0.11 | 0.009 | 1.000 |

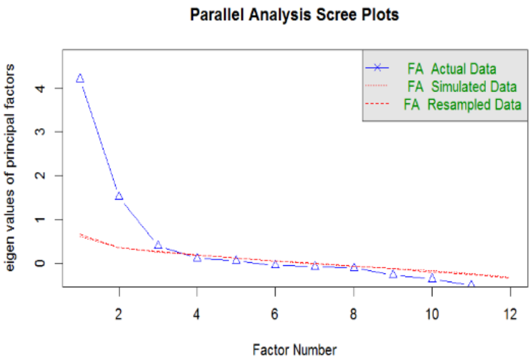
I) Cumulative proportion of at least 0.8 ( or 80% explained variance)
Looking at above table we can see that we should retain 5 factor in order to obtain at least 0.8 cumulative proportion. *(Joliffe's method)*

II) Eigen values greater than 1.
There are three Eigen values greater than 1. Altogether those 3 explains 67.4% of total variation. *(Kiser criteria)*

- Scree plot

III) Based on the "elbow" of the plot.
Under this we are looking at the Scree plot and Parallel scree plot to determine the number of factors.





Scree plot suggest that three factors are sufficient to explain the data.

Since both Kiser criteria and scree plot suggest number of factors to be three we consider three factors and further analyze the dataset.

### 3.1.2. Estimate the parameters of the factor model

We can estimate the parameters (factor loadings) using Principal component method and Maximum likelihood estimation.

- Principal Component (PC) Method

| Variable | Original factor loadings | | | Rotated factor loadings | | |
|---|---|---|---|---|---|---|
| | Factor 1 | Factor 2 | Factor 3 | Factor 1 | Factor 2 | Factor 3 |
| Alcohol | 0.29 | 0.74 | -0.11 | 0.10 | 0.79 | -0.09 |
| Malic_acid | -0.48 | 0.26 | 0.28 | -0.23 | 0.00 | -0.57 |
| Alcalinity_of_ash | -0.45 | -0.15 | 0.31 | -0.16 | -0.37 | -0.40 |
| Magnesium | 0.25 | 0.31 | -0.01 | 0.17 | 0.36 | 0.00 |
| Total_phenols | 0.85 | 0.08 | 0.24 | 0.81 | 0.28 | 0.23 |
| Flavanoids | 0.94 | -0.03 | 0.21 | 0.87 | 0.21 | 0.35 |
| Nonflavanoid_phenols | -0.57 | -0.01 | 0.00 | -0.45 | -0.19 | -0.29 |
| Proanthocyanins | 0.63 | 0.04 | 0.29 | 0.67 | 0.15 | 0.09 |
| Color_intensity | -0.20 | 0.82 | 0.10 | -0.16 | 0.64 | -0.53 |
| Hue | 0.63 | -0.42 | -0.33 | 0.33 | -0.08 | 0.76 |
| diluted_wines | 0.80 | -0.28 | 0.13 | 0.74 | -0.04 | 0.44 |
| Proline | 0.61 | 0.59 | -0.26 | 0.27 | 0.80 | 0.26 |

As you can see in the table we calculated original factor loadings and observed that there are not a single variables associated with Factor 3. All the variables are associated with either Factor 1 or Factor 2. Thus, original loadings do not readily interpretable.

To solve that issue we calculated loadings after rotate the factors using “varimax” method.

Interpretation:

- Factor 1 is highly correlated with Total_phenols, Flavanoids and moderately correlated with diluted_wines, Proanthocyanins, Nonflavanoid_phenols.
- Factor 2 is strongly correlated with Alcohol, Proline and moderately correlated with Color_intensity while Magnesium is weakly correlated. All variables which are correlated with Factor 2 are positive.
- Factor 3 is strongly correlated with hue and moderately correlated with Malic_acid, Alcalinity_of_ash.

- Maximum Likelihood (ML) Method (Using “varimax” rotation)

| Variable | Factor 1 | Factor 2 | Factor 3 |
|---|---|---|---|
| Alcohol | 0.12 | 0.76 | -0.18 |
| Malic_acid | -0.26 | -0.07 | -0.47 |
| Alcalinity_of_ash | -0.16 | -0.41 | -0.33 |
| Magnesium | 0.12 | 0.38 | -0.01 |
| Total_phenols | 0.83 | 0.29 | 0.21 |
| Flavanoids | 0.87 | 0.24 | 0.34 |
| Nonflavanoid_phenols | -0.45 | -0.19 | -0.28 |
| Proanthocyanins | 0.67 | 0.15 | 0.09 |
| Color_intensity | -0.09 | 0.56 | -0.71 |

| | | | |
|---|---|---|---|
| Hue | 0.33 | 0.02 | 0.72 |
| diluted_wines | 0.70 | 0.03 | 0.51 |
| Proline | 0.26 | 0.83 | 0.18 |

Interpretation:

- Factor 1 is highly correlated with Total_phenols, Flavanoids and moderately correlated with diluted_wines, Proanthocyanins, Nonflavanoid_phenols.
- Factor 2 is strongly correlated with Alcohol, Proline and moderately correlated with, Alcalinity_of_ash Magnesium is weakly correlated.
- Factor 3 is strongly correlated with hue, Color_intensity and moderately correlated with Malic_acid.

### 3.1.3. Compare the Results of PC method and ML method

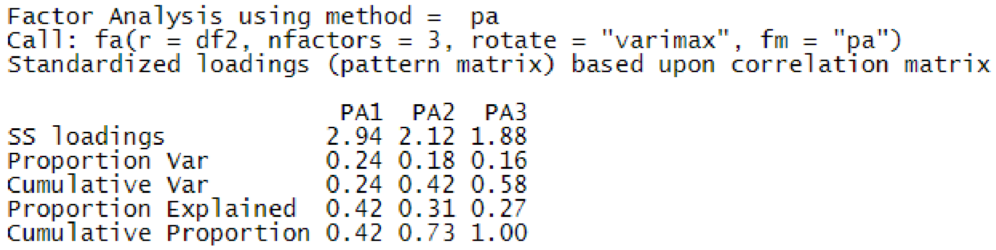
- Communalities

| Variable | PC Method | ML Method |
|---|---|---|
| Alcohol | 0.640 | 0.629 |
| Malic_acid | 0.375 | 0.291 |
| Alcalinity_of_ash | 0.321 | 0.302 |
| Magnesium | 0.154 | 0.160 |
| Total_phenols | 0.789 | 0.811 |
| Flavanoids | 0.933 | 0.938 |
| Nonflavanoid_phenols | 0.328 | 0.314 |
| Proanthocyanins | 0.480 | 0.476 |
| Color_intensity | 0.719 | 0.823 |
| Hue | 0.682 | 0.631 |
| diluted_wines | 0.743 | 0.749 |
| Proline | 0.779 | 0.797 |

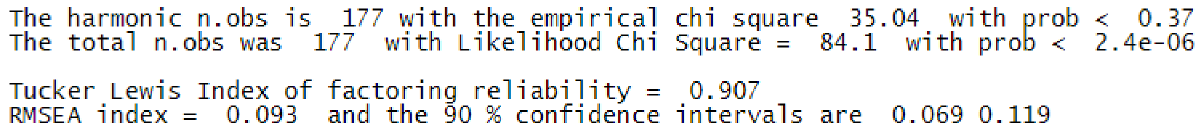
Communalities generated from both methods exhibit closely comparable values with minor deviations. Thus, in both methods other than Malic_acid, Alcalinity_of_ash, Magnesium, Nonflavanoid_phenols, Proanthocyanins high variance of all other variables are explained.
Total communalities are 6.943 and 6.921 for PC method and ML method respectively.

- Standardized loadings

PC Model

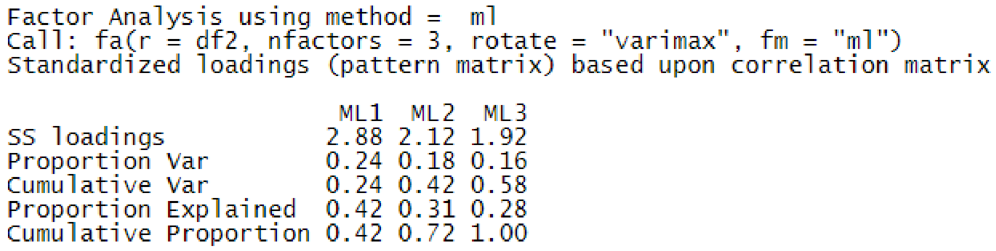


```
Factor Analysis using method =  pa
Call: fa(r = df2, nfactors = 3, rotate = "varimax", fm = "pa")
Standardized loadings (pattern matrix) based upon correlation matrix

                        PA1  PA2  PA3
SS loadings            2.94 2.12 1.88
Proportion Var         0.24 0.18 0.16
Cumulative Var         0.24 0.42 0.58
Proportion Explained   0.42 0.31 0.27
Cumulative Proportion  0.42 0.73 1.00
```

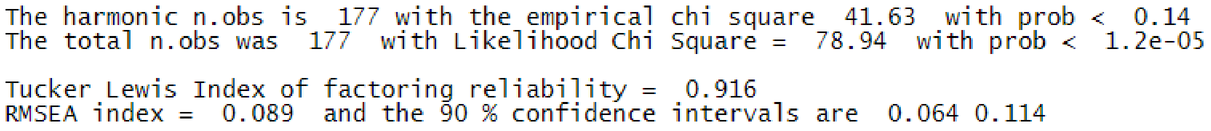


```
The harmonic n.obs is  177 with the empirical chi square  35.04  with prob <  0.37
The total n.obs was  177  with Likelihood Chi Square =  84.1  with prob <  2.4e-06

Tucker Lewis Index of factoring reliability =  0.907
RMSEA index =  0.093  and the 90 % confidence intervals are  0.069 0.119
```

ML Model



```
Factor Analysis using method =  ml
Call: fa(r = df2, nfactors = 3, rotate = "varimax", fm = "ml")
Standardized loadings (pattern matrix) based upon correlation matrix

                       ML1  ML2  ML3
SS loadings           2.88 2.12 1.92
Proportion Var        0.24 0.18 0.16
Cumulative Var        0.24 0.42 0.58
Proportion Explained  0.42 0.31 0.28
Cumulative Proportion 0.42 0.72 1.00
```

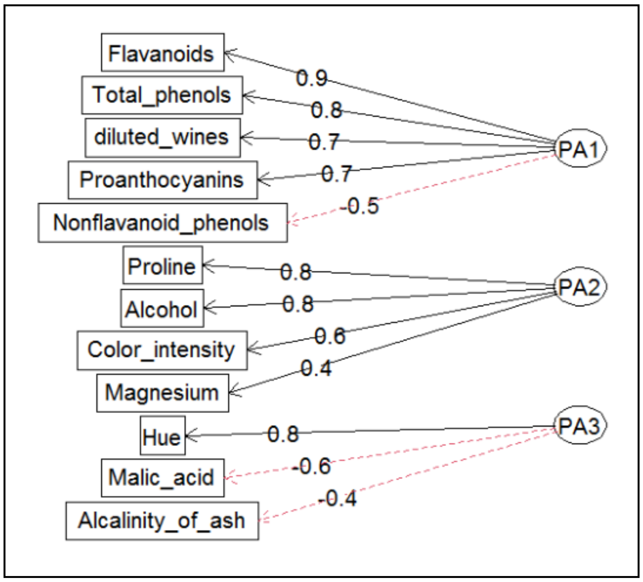


```
The harmonic n.obs is  177 with the empirical chi square  41.63  with prob <  0.14
The total n.obs was  177  with Likelihood Chi Square =  78.94  with prob <  1.2e-05

Tucker Lewis Index of factoring reliability =  0.916
RMSEA index =  0.089  and the 90 % confidence intervals are  0.064 0.114
```

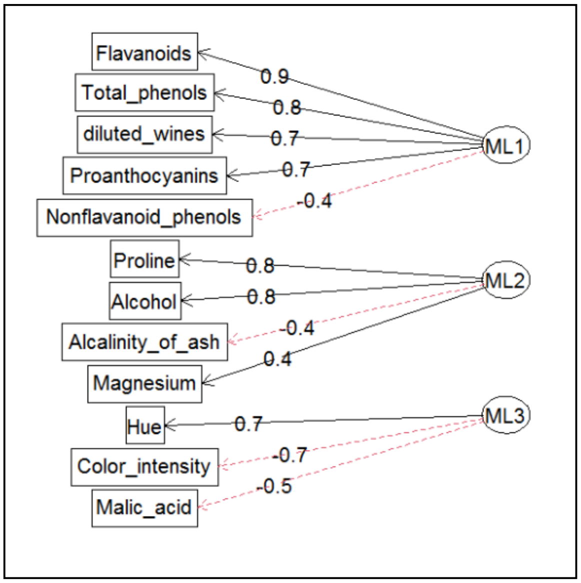
Based on empirical chi square test results we can say that both models are statistically significant.

In both models cumulative variance explained is 58% of total variance. Tucker Lewis Index for both methods are close to 1 (greater than 0.9) which indicates good fit between hypothesized model and observed data. According to RMSEA indices with values 0.093 and 0.089 both PC and ML model indicate adequate fit.

PC Factors



ML Factors



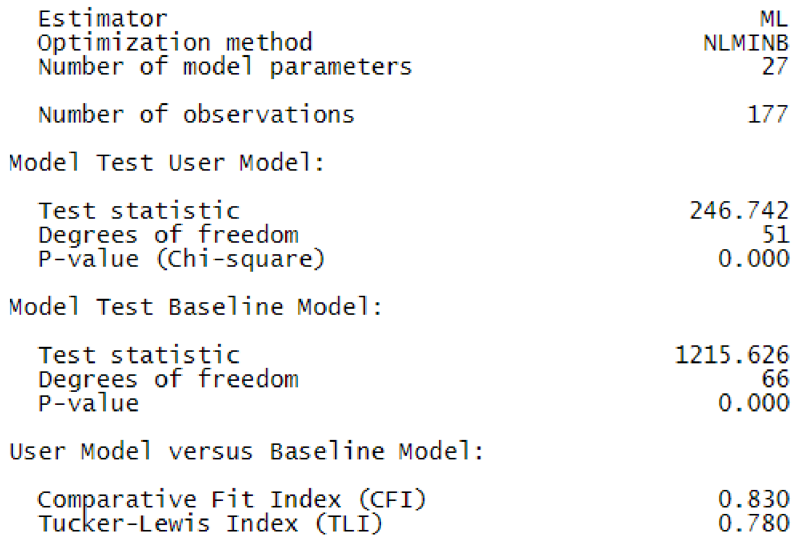
Since Tucker Lewis Index and RMSEA index for ML model is slightly better, we choose factors according to ML method.

### 3.2 Confirmatory Factor Analysis (CFA)

Factor1 =~ Flavanoids+Total_phenols+diluted_wines+Proanthocyanins+Nonflavanoid_phenols

Factor2 =~ Proline+Alcohol+Alcalinity_of_ash+Magnesium

Factor3 =~ Hue+Malic_acid+Color_intensity



```
  Estimator                                         ML
  Optimization method                           NLMINB
  Number of model parameters                        27

  Number of observations                           177

Model Test User Model:

  Test statistic                               246.742
  Degrees of freedom                                51
  P-value (Chi-square)                           0.000

Model Test Baseline Model:

  Test statistic                              1215.626
  Degrees of freedom                                66
  P-value                                        0.000

User Model versus Baseline Model:

  Comparative Fit Index (CFI)                    0.830
  Tucker-Lewis Index (TLI)                       0.780
```

Comparative Fit Index (CFI) and Tucker-Lewis Index (TLI) assess whether the user model provides a better fit to the data compared to baseline model (which is a reference for evaluate the performance of the user model).

In here, CFI = 0.83 and TLI = 0.78

High values for CFI and TLI close to 1 indicates the adequate fit of the model to the data.

## 4. Conclusion and Recommendations

- In original data set there are 14 variables with one class variable. In order to perform factor analysis accurately class variable were removed. Only the continuous feature variables are remain in the dataset.
- According to KMO test and Bartlett test dataset is adequate to perform factor analysis technique.
- There are 12 variables remain in the dataset for analyze.
- Based on Eigen values greater than 1 and scree plot, 3 factors are sufficient to explain the dataset.
- After evaluating the loadings of PC method and ML method, ML results are slightly better to fit the model. Thus, variables associated with factors are decided from ML method.
- Proportion of total variance explained by those factors are 58%.
- By doing CFA, it is confirmed that fitted model is adequate to interpret the dataset.

- ❖ Looking at all the details we can conclude that although the proportion of total variance explained by the model is not good, it will give a better understanding about the underlying factors in the dataset.

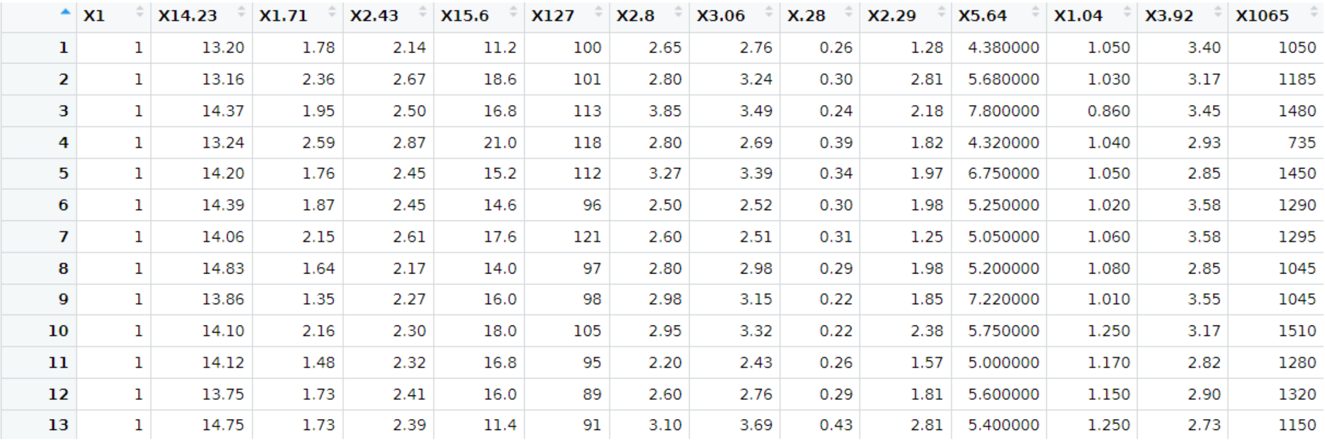
## 5. References

- https://statisticsbyjim.com/basics/factor-analysis/
- https://scholarworks.calstate.edu/downloads/qr46r2957
- https://bookdown.org/sz_psyc490/r4psychometics/factor-analysis.html#exploratory-factor-analysis-efa
- https://www.promptcloud.com/blog/exploratory-factor-analysis-in-r/#:~:text=analysis%20and%20interpretation.-,What%20is%20exploratory%20factor%20analysis%20in%20R%3F,a%20smaller%20number%20of%20variables.
- https://radiant-rstats.github.io/docs/multivariate/pre_factor.html#:~:text=The%20KMO%20and%20Bartlett%20test,is%20correlated%20with%20other%20variables.

Dataset:
https://archive.ics.uci.edu/dataset/109/wine

## 6. Appendices

Dataset:



| | X1 | X14.23 | X1.71 | X2.43 | X15.6 | X127 | X2.8 | X3.06 | X.28 | X2.29 | X5.64 | X1.04 | X3.92 | X1065 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 13.20 | 1.78 | 2.14 | 11.2 | 100 | 2.65 | 2.76 | 0.26 | 1.28 | 4.380000 | 1.050 | 3.40 | 1050 |
| 2 | 1 | 13.16 | 2.36 | 2.67 | 18.6 | 101 | 2.80 | 3.24 | 0.30 | 2.81 | 5.680000 | 1.030 | 3.17 | 1185 |
| 3 | 1 | 14.37 | 1.95 | 2.50 | 16.8 | 113 | 3.85 | 3.49 | 0.24 | 2.18 | 7.800000 | 0.860 | 3.45 | 1480 |
| 4 | 1 | 13.24 | 2.59 | 2.87 | 21.0 | 118 | 2.80 | 2.69 | 0.39 | 1.82 | 4.320000 | 1.040 | 2.93 | 735 |
| 5 | 1 | 14.20 | 1.76 | 2.45 | 15.2 | 112 | 3.27 | 3.39 | 0.34 | 1.97 | 6.750000 | 1.050 | 2.85 | 1450 |
| 6 | 1 | 14.39 | 1.87 | 2.45 | 14.6 | 96 | 2.50 | 2.52 | 0.30 | 1.98 | 5.250000 | 1.020 | 3.58 | 1290 |
| 7 | 1 | 14.06 | 2.15 | 2.61 | 17.6 | 121 | 2.60 | 2.51 | 0.31 | 1.25 | 5.050000 | 1.060 | 3.58 | 1295 |
| 8 | 1 | 14.83 | 1.64 | 2.17 | 14.0 | 97 | 2.80 | 2.98 | 0.29 | 1.98 | 5.200000 | 1.080 | 2.85 | 1045 |
| 9 | 1 | 13.86 | 1.35 | 2.27 | 16.0 | 98 | 2.98 | 3.15 | 0.22 | 1.85 | 7.220000 | 1.010 | 3.55 | 1045 |
| 10 | 1 | 14.10 | 2.16 | 2.30 | 18.0 | 105 | 2.95 | 3.32 | 0.22 | 2.38 | 5.750000 | 1.250 | 3.17 | 1510 |
| 11 | 1 | 14.12 | 1.48 | 2.32 | 16.8 | 95 | 2.20 | 2.43 | 0.26 | 1.57 | 5.000000 | 1.170 | 2.82 | 1280 |
| 12 | 1 | 13.75 | 1.73 | 2.41 | 16.0 | 89 | 2.60 | 2.76 | 0.29 | 1.81 | 5.600000 | 1.150 | 2.90 | 1320 |
| 13 | 1 | 14.75 | 1.73 | 2.39 | 11.4 | 91 | 3.10 | 3.69 | 0.43 | 2.81 | 5.400000 | 1.250 | 2.73 | 1150 |

R codes:

```
library(tidyverse) library(skimr) library(corrplot) library(readxl) library(psych) library(lavaan)
#read dataset
df <- read.csv("../Data/wine.data",sep=",") #%>% view()
#remove class variable
df <- df[,-1]
#correlation matrix
round(cor(df),3)
#KMO test
df_kmo <- KMO(r=df)
df_kmo
```

```
#remove variables
df2<- df[,df_kmo$MSAi>0.5]
df2_kmo <- KMO(r=df2)
df2_kmo
cor_2 <- round(cor(df2),5)
cor_2
#Bartlett test
cortest.bartlett(cor_2, n= nrow(df2))
corrplot(cor_2,type = "lower", tl.col = "black",tl.srt = 45)
#eigen values
eigen <- eigen(cor_2)
eigen_values <- round(eigen$values,2)
eigen_values
#summary details of eigen values
proportion <- round(eigen_values/sum(eigen_values),3)
cum_proportion <- round(cumsum(eigen_values)/sum(eigen_values),3)
table <- data.frame(eigen_values,proportion,cum_proportion)
table
#scree plots
scree(df2)
fa.parallel(df2,fm="pa",fa="fa")
#PC Method
#original factor loadings
df2_PC_unrotate<- fa(df2 ,nfactors = 3,rotate = "none",fm = "pa")
print(df2_PC_unrotate)
#rotated factor loadings
df2_PC<- fa(r = df2 ,nfactors = 3,rotate = "varimax",fm = "pa")
print(df2_PC)
#diagram
fa.diagram(df2_PC)
#communalities
rotated_pc_com <- setNames(as.data.frame(unclass(df2_PC$communality)),"Communality")
round(rotated_pc_com,3)
#ML Method
df2_ML<- fa(df2 ,nfactors = 3,rotate = "varimax",fm = "ml")
print(df2_ML)
fa.diagram(df2_ML)
rotated_ml_communality <-as.data.frame(unclass(df2_ML$communality))
round(rotated_ml_communality,3)
#Normalize data
df2 <- scale(df2)
#Model
model <- '
  Factor1 =~ Flavanoids+Total_phenols+diluted_wines+Proanthocyanins+Nonflavanoid_phenols
            Factor2 =~ Proline+Alcohol+Alcalinity_of_ash+Magnesium
            Factor3 =~ Hue+Malic_acid+Color_intensity
  '
fit <- cfa(model, data = df2)
summary(fit,fit.measures= TRUE, standardized= TRUE)
```