Assignment #7: Factor Analysis
Prabhat Thakur

Introduction:

The objective of this assignment is to explore application of Factor Analysis. This assignment is based on a classic factor analysis example from Stoetzel's 'A Factor Analysis of Liquor Preference' (Journal of Marketing Research). It is also a useful example for those who are likely to see factor analysis in the context of marketing, and in particular marketing segmentation.

Data: The data for this assignment will be the correlation matrix input from the Stoetzel article in the Course reserves.
The liquor preference sample population data comprised of completed survey responses from 1,442 men and women in February 1956. The respondent constituted a representative cross-section of the French population selected by quota sampling methods. The data was collected by conducting in-person interview at home in 161 different localities by the interviewers employed by French Institute of Market Research (ETMAR). The 1442 completed responses constituted 70% completion rate of the total 2,014 surveyed. The data contained ranking (preference) of 9 different liquors from most liked to least liked by the respondents.

Code: Sample code for factor analysis on liquor preference data was provided in StoetzelSkeletonCode.R

Assignment Tasks:

(1) Loading correlation matrix in R:
In this application, the factor analysis is performed using the correlation matrix provided in Stoetzel study. The correlation matrix represents pairwise correlation coefficients between the nine simultaneous choices of liquors requested from each respondent.

I first loaded all correlation coefficients values in a single vector object in R and then converted it into 9x9 matrix object which is nothing but correlation matrix of preferences on 9 liquor options. I also validated that the object created is a matrix object and is symmetric. It is used in below tasks for factor analysis.

(2) Three factor model with VARIMAX rotation:
Stoetzel estimated a three factor model in his paper. I will estimate a three factor model with a VARIMAX rotation using maximum likelihood factor analysis. Below is the output from my factor analysis model in R:

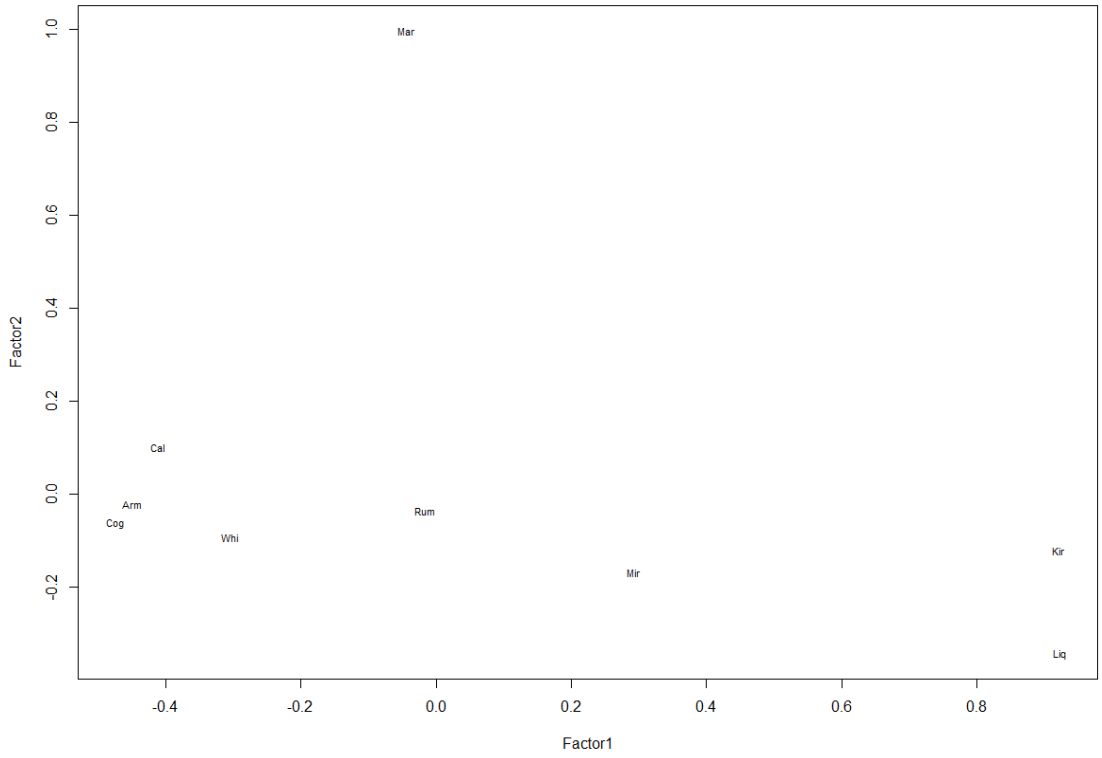
```
Call:
factanal(factors = 3, covmat = cor.matrix, n.obs = 1442, rotation = "varimax")

Uniquenesses:
[1] 0.759 0.792 0.739 0.134 0.005 0.005 0.933 0.890 0.005

Loadings:
      Factor1 Factor2 Factor3
 [1,] -0.450           0.193
 [2,] -0.411   0.100   0.172
 [3,] -0.473           0.183
 [4,]  0.921  -0.121
 [5,]          0.996
 [6,]  0.293  -0.169  -0.938
 [7,]                 0.256
 [8,] -0.305
 [9,]  0.923  -0.344   0.158

                Factor1 Factor2 Factor3
SS loadings      2.477   1.179   1.082
Proportion Var   0.275   0.131   0.120
Cumulative Var   0.275   0.406   0.527

Test of the hypothesis that 3 factors are sufficient.
The chi square statistic is 1820.72 on 12 degrees of freedom.
The p-value is 0
```

From above we can see that high uniqueness values for Armagnac, Cognac, Calvados, Rum and Whiskey. The loadings are not ordered but appears consistent with Stoetzel's findings. At initial glance, there is a separation between a grouping with negative loadings including Armagnac, Cognac, Calvados, and Whiskey as in the original report. This seems to make sense as these are traditionally 'strong' drinks which Stoetzel also identified.

a. While the correlation matrix is same as Stoetzel, numeric values of factor loading are different from Stoetzel analysis. The possible reason could be due to use of different factor analysis method. It seems, Stoetzel used Thurstone centroid factor analysis method in his work whereas we are using maximum likelihood method, also it's not clear which rotation method was used. I may possibly be able to reproduce a factor analysis and get the same factor loadings if factor analysis and rotation methods are same as original study.

b. Factor1 shows high loading values for Kir and Liq which is similar to Stoetzel analysis. Also loadings of Armagnac, Cognac, Calvados, and Whiskey are negative as in the original report. Out of 9 variables, Factor 1 has six variables as most high or low compared to other factors. Factor 1 is the most important and contributes most to the variance in preference. This is consistent with original report.
   Factor2 shows high loading value for only one variable Marc and all other loadings are negative. In the original report, a likely interpretation of this factor was related to the price, low or high, of the different items. Similar interpretation can be derived from Factor 2. In factor 3, it is maximum in the case of Rum and minimum with Mirabelle and Marc. This gives

similar interpretation as original report. Below is the plot for factor 1 and factor 2. We can see a clear distinction between left and right side grouping.



c. Statistical inference for the maximum likelihood factor analysis and three factors suggestion can be tested using hypothesis testing by the chi-square test statistic.

Null hypothesis for the chi-square test statistic can be stated as below:
H0: 3 factors are sufficient
HA: More factors are needed.

The factor analysis states that for three factors, chi-squared stat is 1820.72 on 12 degrees of freedom and the p-value is zero. P-value is 0 suggests that **we reject the null hypothesis**, which means 3 factors are not sufficient and more factors are needed.

(3) # of Factors:

If we run the factor analysis model using different number of factors and compare total variance explained by each, we can possibility determine the ideal model. I am afraid I do not have good knowledge on various types of liquor so it seems difficult to interpret the factors logically.

Below are the values of total variance explained by each model; we can see that % of total variance explained is increasing with number of factors in the model. For 9 variables, it is not possible to general factor analysis model with 6 factors.

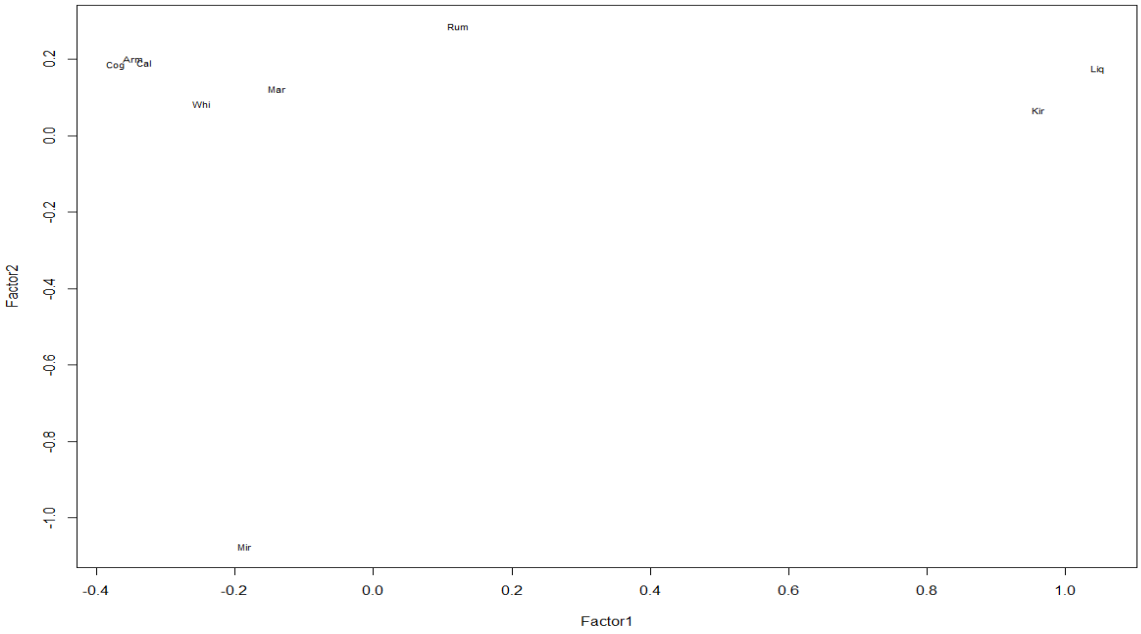
# of Factors	Total Variance
1	27.70%
2	38.90%
3	52.70%
4	59.70%
5	69.80%

P-values for first 3 models are 0 and for 4th and 5th model is close to 0 which suggests that more factors are required but due to computational limitations it is not possible to create model with 6 or more factors for 9 variables. From above, the best model appears to be model 5 but it also depends on whether the factor interpretations are logical and makes sense. While interpreting the data visually, it seems that including more than three factors does not provide significant additional value that can be used to form any conclusions about the liquor preferences.

(4) Factor Analysis using PROMAX rotation:

The VARIMAX factor rotation is an example of an orthogonal factor rotation. We also have oblique factor rotations. One example of an oblique factor rotation is the PROMAX rotation. We will fit a three factor model with a PROMAX rotation using maximum likelihood factor analysis.

```
Call:
factanal(factors = 3, covmat = cor.matrix, n.obs = 1442, rotation = "promax")

Uniquenesses:
[1] 0.759 0.792 0.739 0.134 0.005 0.005 0.933 0.890 0.005

Loadings:
      Factor1 Factor2 Factor3
 [1,] -0.347   0.202  -0.102
 [2,] -0.331   0.193
 [3,] -0.371   0.186  -0.146
 [4,]  0.961
 [5,] -0.139   0.123   0.977
 [6,] -0.186  -1.075  -0.101
 [7,]  0.123   0.287
 [8,] -0.248          -0.146
 [9,]  1.047   0.175  -0.186

                Factor1 Factor2 Factor3
SS loadings       2.518   1.407   1.057
Proportion Var    0.280   0.156   0.117
Cumulative Var    0.280   0.436   0.554

Factor Correlations:
        Factor1 Factor2 Factor3
Factor1  1.0000 -0.0591  0.0147
Factor2 -0.0591  1.0000  0.4787
Factor3  0.0147  0.4787  1.0000

Test of the hypothesis that 3 factors are sufficient.
The chi square statistic is 1820.72 on 12 degrees of freedom.
The p-value is 0
```



a. Factor 1 from Promax rotation has similar interpretation as Varimax but highest and lowest loading in factor 2 and factor 3 are switched. For example Mir was in factor 2 as highest loading but now in factor 3.
Looking at the plot of the Factor 1 vs Factor 2 loadings, it seems we get about the same information although inverted from the plot of the varimax analysis. Groupings are similar as we can still see Kir and Liq at the far positive end of the Factor1 spectrum and the cluster of Arm, Cal and Cog at the far left end of the Factor1 scale. It does not appear that the promax provides better interpretability than varimax, but at least similar interpretability. From the factor correlation matrix, we can see that Factor 2 and Factor 3 are more correlated compared to factor 1.

b. Total Variance explained by 3 factors in promax rotation method is 55.4% which is better than varimax rotation. However, looking at the error mean absolute error produced from the promax, we see it underperforms the varimax version with a total MAE of 0.1103 which is more than double the 3 factor MAE from varimax. This could be due to the fact that this rotation is an oblique method, meaning it does not restrict the factors to an uncorrelated set of loadings. Instead, this method allows some correlation between the loadings and thus more error should be assumed. P-value of 0 suggests that more factor could be needed. Factor rotation does effect the statistical inference of factor analysis.

(5) Correlation matrix approximation:

Factor loadings and the specific (or unique) variances can be used to approximate the correlation matrix. Fit of these approximations can be calculated using the Mean Absolute Error of the residual matrix. After running the approximations for both the varimax and the promax

loading results, I obtained a mean absolute error for f1 (varimax) of 0.0486 and for g1 (promax) of 0.11027. Both of these were done using the three-factor analysis. The varimax outperformed the promax model and as was expected, so be used as the preferable model choice.

The original correlation matrix is based on 9 variables. The reduced model from factor analysis has 3 factors. We are using these 3 factors to approximation correlation matrix and by calculating MAE to find how good these three factors explains the correlation between the original 9 variables as compare to original 9 variable correlation matrix.

Reflections & Conclusion:

I found using Factor Analysis quite challenging to understand and hard to interpret model output. It certainly seems to be a valuable and powerful tool to gain a better understanding of the underlying motivations and signals in the data. However, for meaningful interpretation of factors, domain knowledge of investigation subject is very important.