Assignment #7

Andrew Knight

Introduction:

The goal of this assignment is to explore Factor Analysis using Stoetzel's 'Factor Analysis of Liquor Preference' in R.

Calculations File: StoetzelSkeletonCode_Knight.R

Assignment Tasks

1: Load & Review Correlation Matrix in R

Factor analysis was used due to its ability to explain correlations between measures by looking at the underlying factors. The goal of the study was to study the correlations between individual drink choices and see if a forecast could be constructed for future choices based on his or her preference. The preferences were determined by examining the correlations between these underlying factors in the aggregate for nine different drinks.

The liquor preference data is comprised of completed survey responses from 1,442 adults who supposedly represented a good cross-section of French population in February 1956. The data were collected by conducting in-home interviews in which adults were asked to rank the nine liquor choices in order of their preference. An interesting note also stated that the 1442 completed responses constituted 70% completion rate of the total 2,014 surveyed, which seems low for a simple task. My assumption is that the others were determined to either not know enough about the liquor options to be able to rank them all or that they did not rank them because they had no opinion on one or more of the options listed. The study does not explicitly state what the rejection criteria was but I think a reasonable assumption for data included in a Factor Analysis study should have no selection included if any data points were missing.

The resulting data are essentially ranking values for each option. The correlation matrix from the Stoetzel study was provided and used for the analysis that follows.

2: Estimate a Three-Factor Model with Varimax Rotation

My factor analysis output in R is displayed here.

```
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
> |
```

First, looking at the variance of each variable, I see high uniqueness values for Armagnac, Cognac, Calvados, Rum and Whiskey. The loadings, although ordered differently, appear 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. Marc and Mirabella show high loading values for Factor 2 and Factor 3 respectively as compared to the other variables. I don't kow much about these drinks but my plots show a dichotomy between them that is similar to the original report. These two are almost always at far ends of the spectrum from one another regardless of the Factor comparison as shown in the scatter matrix below. This appears to confirm the contrast between the drinks on opposite ends of the proce spectrum in the original report as well.

a. While the correlation matrix is the same as Stoetzel in part 1, I do not get the same numerical results for the factor analysis. I would expect to get a similar factor interpretation however the load values may not be the same. I would attribute this to using a different rotation method perhaps.

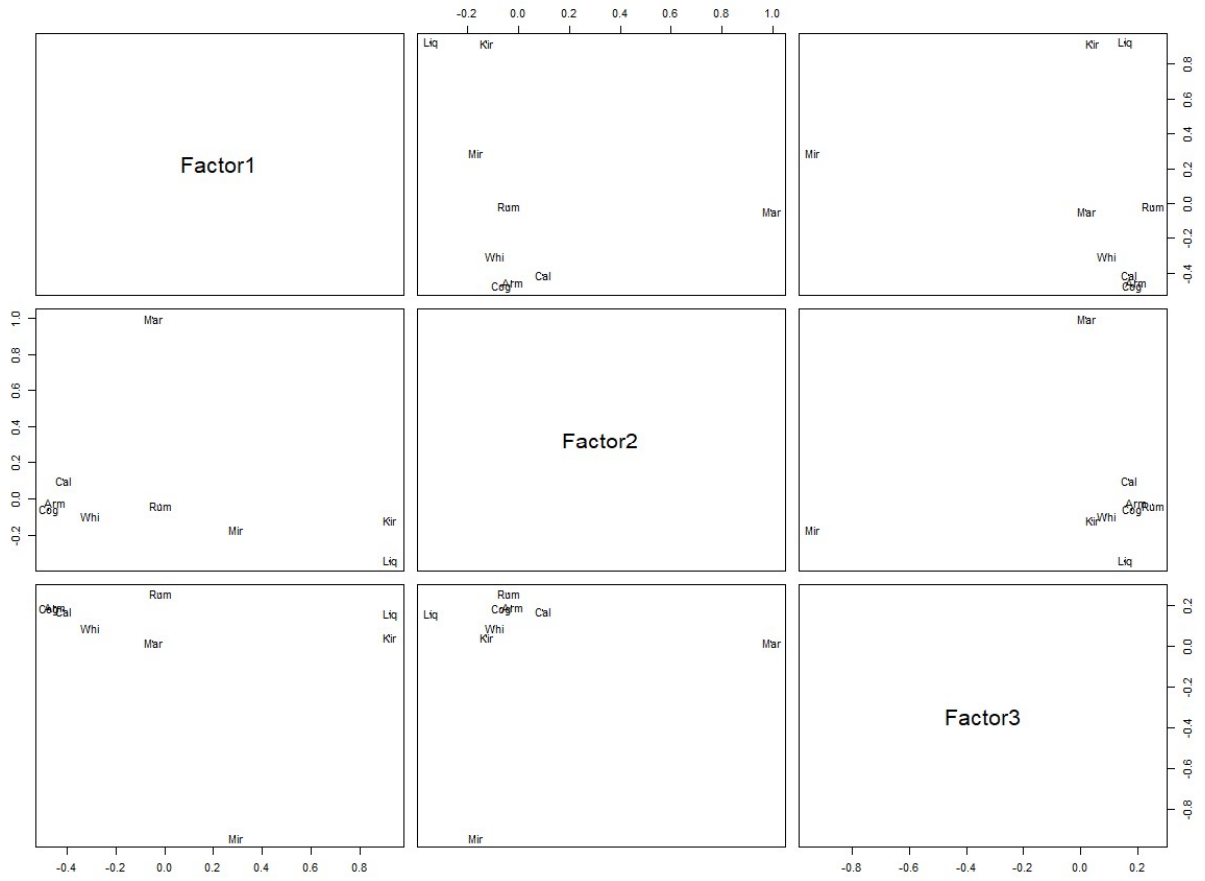
b. From the factanal summary results, we see that the common variances are low but some of the individual variances are quite high. When we look at the loadings we see high loadings (high correlations) between the first two factors and at least one of the variables. Factor1 shows high loading

values with two variables, 4 & 9. Factor2 shows high loading value for only one, variable 5 in this case. Factor3 shows one high negative loading value for variable 6.

When comparing the qualitative outcomes from Stoetzel's report I see some correlation between Kirsh and Mirabelle as he pointed out but they do not seem as strongly correlated as Kirsh and Liquor.

c. The factor analysis states that for three factors I have chi-squared stat of 1820.72 on 12 degrees of freedom and the p-value is zero. When I perform factor analysis on four factors, I have a chi-squared value of 968.7 on 6 degrees of freedom and a p-value of zero. The cumulative variance covered by the first four factors is nearly about 60%, which as opposed to the total variance for three factors of just over 50%.

Looking at the loading plots it seems that the first factor is the measure of the sweet vs strong preference among the drinks. With liquors, kirsh and (to a slightly lesser extent) Mirabelle on the sweeter side showing positive loading values for Factor1 and the strong drinks listed above showing negative values for Factor 1, this indicates the first preferences scale is closely tied to this attribute. Stoetzel's report also pointed out that this first factor also carries the most significance in explaining the correlations between the variables. Also, the second factor seems to exhibit a measure of cost variation between the drinks. This too contributes to the drink preference but based on the variation, less so than Factor1. One difference I found difficult to discern was the contribution of the third factor. The original report calls out this factor for sociological or regional influence in the preferences but I could not find strong evidence of this in my plots.



3: Determine the Number of Factors

If we run the factor analysis model using different number of factors and compare the MAE for each, we can determine the ideal model. In my case I tested running the between 1 and 6 factors with the results in the table below.

# of Factors	MAE	Total Variance
1	0.0862	27.7%
2	0.0705	38.9%
3	0.0486	52.7%
4	0.0325	59.7%
5	0.0200	69.8%
6	?	

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. Due to this I have decided to restrict the analysis to three factors. Also, I receive an error when attempting to perform factanal on nine variables for six factors.

4: Try Promax Rotation

Using the oblique rotation, Promax, we get a slightly different interpretation. We also start by using three factors just as Stoetzel showed.

a. Looking at the plot of the Factor 1 vs Factor 2 loadings, it seems we get about the same information albeit 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. We also still see Mir alone on its own on the negative far end of the Factor2 loading whereas everything else is centered close to zero similar to the varimax plot. It does not appear that the promax provides better interpretability than varimax, but at least similar interpretability.

b. However, looking at the error mean absolute error produced from the promax plot, we see it underperforms the varimax version with a total MAE of 0.1103 which is more than double the three-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.

So, we know that the model is not as accurate but then we should review the loading matrix to determine if we have better interpretability than the varimax method. Here is my correlation matrix for the three-factor promax model.

```
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
>
```

I see that we have a high unique correlation on Factor 4 with variable 5, which is Mirabelle. But I also get a high value when using only three factors for Mirabelle. This gives me the impression that factor loading is shifting the variance out as more factors are added. This seems inconsistent with Stoetzel's report.

5: Correlation Matrix Approximation

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. We are computing the mean error between the approximated correlation matrix given by the three factors versus the original correlation matrix provided in the report.

Reflections & Conclusion

In summary, I found using Factor Analysis a bit more challenging to implement to quickly perform exploratory analysis on the data. It certainly seems to be a valuable and powerful tool to gain a better understanding of the underlying motivations and signals in the data. Rotations provide a better way to interpret the data in a useful way, at least the first couple factors. Using factor analysis does require discipline in checking assumptions about the data and in checking variances. Furthermore, it informs the analyst of useful trends but I think favors those who have at least some domain knowledge of the content under investigation. I see now why this is used primarily as an exploratory step used to gain additional insight before moving on building the model.