Application of Factor Analysis on Car Perception Data

This notebook shows how to perform Factor Analysis with R and how to use the output to construct a perceptual map. To perform this analysis you need to install these R packages: 1. nFactors 2. ggplot2 3. ggrepel 4. psych 5. scatterplot3d

Reading and outputing data

The dataset includes data from 92 respondents regarding their perception of six sedans (Acura, BMW, Cadillac, Lexus, Lincoln, Mercedes) on six attributes (Trendy/innovative, Styling, Reliability, Sportiness, Performance, Comfort), each measured on a ten-point scale (1=low, 10=high rating). The data excerpt below shows the rating data of respondeent (ID=1). It also shows two additional variables: The respondent education level (MBA vs. Undergrad) and his or her preference rating for each sedan (also measured) on a 10 point scale (1=low and 10=high preference). We use this dataset to answer the following questions: 1. How many factors underlyy consumers' perceptions of the six sedans? What are they? 2. Does consumer perception vary by level of education? 3. How does consumer perception impact consumer preference for cars?

To answer these questions, we first apply factor analysis on the attribute rating data to derive the factors and construct a perceptual map. We then compare the maps of MBA and undergraduate students to see if they are different. Finally, we will run a regression where preference is the dependent variables and the factor scores are the independent variables. The regression output will reveal the relative importance of the factors in preduicting consumer preference for cars.

Hide

```
cardata = read.csv(file = "cardata.csv")
head(cardata)
```

The first analysis decision in factor analysis is to determine the number of factors to retain. The second decision is to understand the factor structure of the data.

Let's determine the number of factors. Note that we will factor analyze only the attribute ratings (variables 4:9).

Determining the Number of Factors

We can use nScree() to answer this question.

Hide

```
library(nFactors)
nScree(cardata[,4:9], cor=TRUE) #This function help you determine the number of factors
```

```
   noc naf nparallel nkaiser
1   1   1         1       1
```

The Kaiser criterion (nkaiser) corresponds to the eigenvalue > 1 criterion when factor analysis is applied on the correlation matrix. The acceleration factor criterion (naf) is based on the elbow criterion in the scree plot. Parallel analysis (nparallel) and optimal coordinates (noc) are two other methods. As you can see all four criteria indicated one factor to retain.

The decision of how many factors also involves judgement on the part of the marketing analyst. It hinges on what the analyst wants to highlight from the data. It is quite possible that one factor may fail to portray important information in the data. In such a case, it is worthwile to run factor analysis, say from 1 to 5 factors (without Varimax rotation), to see how much variance it is explained as we extract varying number of factors. We do this analysis below.

Hide

```
fit <- principal(cardata[,4:9], nfactors=5, rotate="none")
fit$loadings
```

```
Loadings:
            PC1    PC2    PC3    PC4    PC5
Trendy      0.886 -0.246  0.192 -0.140
Styling     0.888 -0.173  0.179 -0.217 -0.192
Sportiness  0.795 -0.448         0.187  0.354
Reliability 0.779  0.255 -0.522 -0.208  0.111
Performance 0.869  0.133 -0.154  0.367 -0.263
Comfort     0.552  0.749  0.330         0.156


                 PC1   PC2   PC3   PC4   PC5
SS loadings    3.873 0.935 0.475 0.280 0.272
Proportion Var 0.645 0.156 0.079 0.047 0.045
Cumulative Var 0.645 0.801 0.880 0.927 0.972
```

The SS loadings in the bottom panel of the output above are the eigenvalues of the correlation matrix. The eigenvalue criterion (eigenvalue less than one) suggests one factor. However, the second eigenvalue (=0.935) is close to one, suggesting that factor 2 may have important information. The cumulative Var criterion suggests that one factor captures 64.5% of the variance of the data; two factors explain 80.1% of thew data; and three factors capture 88% of the data. However, the improvement from three to four factors is very marginal. The fourth factor explains only 4.7% of the variance. This analysis suggests that a two- or three-factor solution is worthwile examining. Let's look at a rotated three-factor solution.

Hide

```
fit <- principal(cardata[,4:9], nfactors=3, rotate="varimax")
fit$loadings
```

```
Loadings:
            RC1   RC3   RC2
Trendy      0.881 0.249 0.209
Styling     0.839 0.281 0.261
Sportiness  0.870 0.271
Reliability 0.294 0.902 0.211
Performance 0.549 0.621 0.330
Comfort     0.126 0.234 0.951


                 RC1   RC3   RC2
SS loadings    2.641 1.468 1.173
Proportion Var 0.440 0.245 0.196
Cumulative Var 0.440 0.685 0.880
```

It looks from this analysis that a three-factor solution has a good interpretation. The first factor loads high on Trendy, Styling, and Sportiness. We call this factor Car Appearance. The second factor loads high on Relaiability and Performance. We call it Car Performance. The third factor loads only on Comfort. Given the clear interpretation of the three factors, we retain this solution for further analysis.

Hide

```
colnames(fit$weights) = c("Appearance", "Performance", "Comfort") # Naming the factors
fit$weights
```

```
             Appearance Performance      Comfort
Trendy       0.451769686 -0.27497882  0.07367905
Styling      0.400831739 -0.23133003  0.12214951
Sportiness   0.443894424 -0.08208276 -0.27113948
Reliability -0.350452175  1.06183977 -0.26457070
Performance -0.009305856  0.41731508  0.02843823
Comfort     -0.099899402 -0.26153179  1.03312622
```

The factor weights show how the factor scores are obtained. For example, the scores on factor 1 are obtained as follows:
Appearance=0.45Trendy + 0.40Styling + … -0.09Comfort

Note that all the attribute ratings are standardized to mean zero and unit variance when computing the factor scores. In such a case, each factor will also be standardized.

Hide

```
colnames(fit$scores) = c("Appearance", "Performance", "Comfort")
reduced_data = cbind(cardata[,1:3],fit$scores)
head(reduced_data)
```

The table above reports the factor scores of respondent ID=1. Note that his or her data are reduced from six attributes to only three factors.

To produce a perceptual map, we need to compute the average factor scores for each car brand. The output is given below.

Hide

```
attach(reduced_data)
brand.mean=aggregate(reduced_data[,c(4:6)], by=list(Brand), FUN=mean, na.rm=TRUE)
detach(reduced_data)
head(brand.mean)
```

We are now ready to produce a three-dimensional perceptual map of the cars.

Three dimensional perceptual map

Hide

```
attach(brand.mean)
```

```
The following objects are masked from brand.mean (pos = 3):

    Appearance, Comfort, Group.1, Performance

The following objects are masked from brand.mean (pos = 4):

    Appearance, Comfort, Group.1, Performance

The following objects are masked from brand_by_edu (pos = 5):

    Appearance, Comfort, Performance

The following objects are masked from brand_by_edu (pos = 6):

    Appearance, Comfort, Performance

The following objects are masked from brand_by_edu (pos = 7):

    Appearance, Comfort, Performance

The following objects are masked from brand_by_edu (pos = 8):

    Appearance, Comfort, Performance

The following objects are masked from brand.mean (pos = 9):

    Appearance, Comfort, Group.1, Performance
```
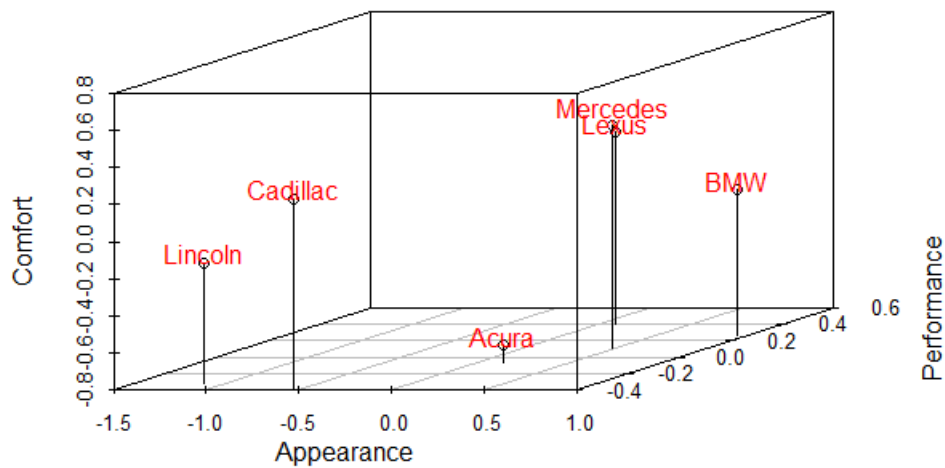
Hide

```
# 3D Scatterplot
library(scatterplot3d)
s3d <- scatterplot3d(Appearance, Performance, Comfort,
                     scale.y = 1, type='h', asp = 1,
                     main="3D Perceptual Map")
text(s3d$xyz.convert(Appearance, Performance, Comfort + c(rep(0.05,5),0.1)),
     labels=(brand.mean[,1]),
     col = 'red')
```

3D Perceptual Map

```
detach(brand.mean)
```

To examine whether MBA and undergraduate students perceive cars differently, we calculate average brand perception for each of these two groups of students.

```
attach(reduced_data)
```

```
The following objects are masked from brand.mean:

    Appearance, Comfort, Performance
```

```
brand_by_edu=aggregate(reduced_data[,c(4:6)], by=list(Brand, Education),
                        FUN=mean, na.rm=TRUE)
colnames(brand_by_edu) = c("Brand", "Edu", "Appearance", "Performance", "Comfort")
detach(reduced_data)
brand_by_edu # Print the average factor scores by brand and education
```

Below is a joint map of MBA and undergraduate students perception of cars.

```
attach(brand_by_edu)
```

```
The following objects are masked from brand_by_edu (pos = 3):

    Appearance, Brand, Comfort, Edu, Performance

The following objects are masked from brand_by_edu (pos = 4):

    Appearance, Brand, Comfort, Edu, Performance

The following objects are masked from brand_by_edu (pos = 5):

    Appearance, Brand, Comfort, Edu, Performance

The following objects are masked from brand.mean:

    Appearance, Comfort, Performance
```

```r
# 3D Scatterplot
library(scatterplot3d)
s3d1 <- scatterplot3d(brand_by_edu[,3], brand_by_edu[,4],
        brand_by_edu[,5], xlab = "Appearance",
        ylab = "Performance", zlab = "Comfort",
        scale.y = 1, type='h', asp = 1,
        main="3D Perceptual Map")
tmp <- brand_by_edu[which(brand_by_edu$Edu == 'MBA'),]
text(s3d1$xyz.convert(tmp$Appearance, tmp$Performance,
        tmp$Comfort + c(rep(0.05,5),0.1)),
        labels=(tmp$Brands), col = 'darkgreen')
```

```r
tmp <- brand_by_edu[which(brand_by_edu$Edu=='Undergrad'),]
text(s3d1$xyz.convert(tmp$Appearance, tmp$Performance,
        tmp$Comfort + c(rep(0.05,5),0.1)),
        labels=(tmp$Brands), col = 'red')
legend(-3, 8,
        legend=c("MBA", "Undergrad"),
        col=c("red", "darkgreen"), lty=1, cex=0.8)
```
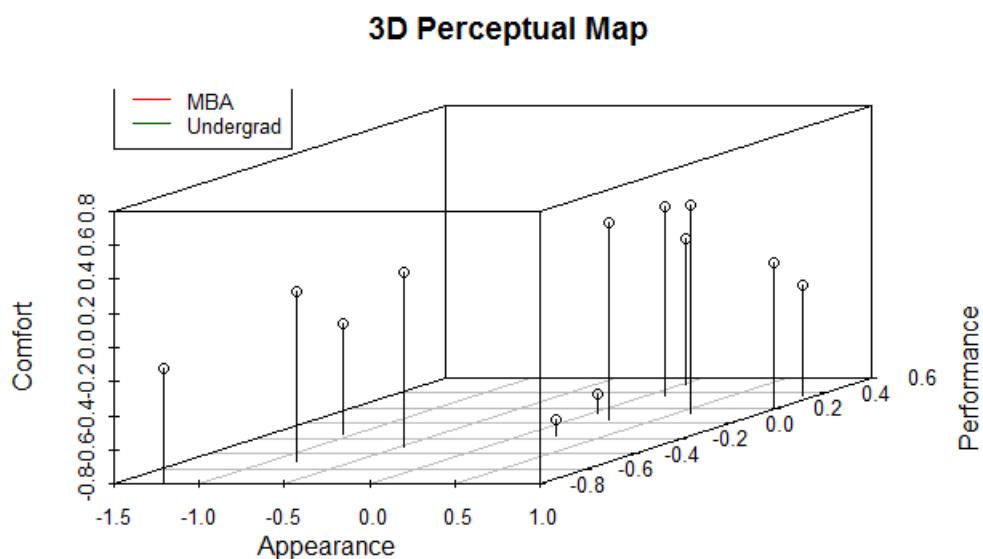

Except for Cadillac and Lincoln, there is a little difference in perceptions between these two groups of students about the remaining five car brands. We now examine the drivers of car preference using regression analysis.

```r
# Append preference data
red_data = cbind(cardata[,c(2,3,10)], fit$scores)
colnames(red_data) = c("Brand", "Edu", "Preference", "Appearance", "Performance", "Comfort")
head(red_data)
```

```r
# Multiple Linear Regression
regfit <- lm(Preference ~ Appearance + Performance + Comfort, data=red_data)
summary(regfit) # show results
```

```
Call:
lm(formula = Preference ~ Appearance + Performance + Comfort,
    data = red_data)

Residuals:
   Min     1Q Median     3Q    Max
-6.616 -0.982 -0.381  0.344 44.218

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   7.5181     0.1752  42.918  < 2e-16 ***
Appearance    1.5609     0.1753   8.903  < 2e-16 ***
Performance   1.1064     0.1753   6.310 5.75e-10 ***
Comfort       0.5364     0.1753   3.059  0.00233 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.116 on 548 degrees of freedom
Multiple R-squared:  0.1899,    Adjusted R-squared:  0.1854
F-statistic: 42.81 on 3 and 548 DF,  p-value: < 2.2e-16
```

The regression output suggests that car Appearance is the most important driver of student preference for cars ($\beta=1.56; \ p,0.01$), followed by Performance ($\beta=1.10; \ p,0.01$), and Comfort ($\beta=0.53; \ p,0.01$). Cadillac, for example, should improve its appearance (e.g., styling) and perormance if it wants to have a better appeal in this market segment. The R-Squared for this regression is only 18.99%. This suggests that there are other drivers of preference beyond consumer perception (e.g., price).