PCA Analysis

Aakanksha Baid

January 2020

**Import, scale, and summarize the data**

#import  
assgn1\_DF = read.csv("Data/assg1.csv")  
  
#summary  
library(psych)  
summary(assgn1\_DF)

## Resp X1 X2 X3 X4   
## Min. : 1.00 Min. :0.0 Min. :0.000 Min. :0.000 Min. :0.0   
## 1st Qu.: 8.25 1st Qu.:2.0 1st Qu.:2.500 1st Qu.:2.000 1st Qu.:2.5   
## Median :15.50 Median :5.0 Median :6.000 Median :3.000 Median :6.0   
## Mean :15.50 Mean :4.4 Mean :5.133 Mean :3.667 Mean :5.4   
## 3rd Qu.:22.75 3rd Qu.:7.0 3rd Qu.:7.750 3rd Qu.:5.750 3rd Qu.:8.0   
## Max. :30.00 Max. :9.0 Max. :9.000 Max. :9.000 Max. :9.0   
## X5   
## Min. :0.000   
## 1st Qu.:2.000   
## Median :4.000   
## Mean :4.533   
## 3rd Qu.:6.000   
## Max. :9.000

describe(assgn1\_DF)

## vars n mean sd median trimmed mad min max range skew kurtosis se  
## Resp 1 30 15.50 8.80 15.5 15.50 11.12 1 30 29 0.00 -1.32 1.61  
## X1 2 30 4.40 2.94 5.0 4.46 2.97 0 9 9 -0.18 -1.53 0.54  
## X2 3 30 5.13 3.06 6.0 5.29 2.97 0 9 9 -0.46 -1.18 0.56  
## X3 4 30 3.67 2.75 3.0 3.50 2.97 0 9 9 0.49 -0.94 0.50  
## X4 5 30 5.40 2.91 6.0 5.58 2.97 0 9 9 -0.48 -1.17 0.53  
## X5 6 30 4.53 2.85 4.0 4.50 2.97 0 9 9 0.19 -1.19 0.52

#scaling  
assgn1.sc <- assgn1\_DF  
assgn1.sc[, 2:6] <- scale(assgn1.sc[,2:6])  
describe(assgn1.sc)

## vars n mean sd median trimmed mad min max range skew kurtosis  
## Resp 1 30 15.5 8.8 15.50 15.50 11.12 1.00 30.00 29.00 0.00 -1.32  
## X1 2 30 0.0 1.0 0.20 0.02 1.01 -1.50 1.56 3.06 -0.18 -1.53  
## X2 3 30 0.0 1.0 0.28 0.05 0.97 -1.68 1.26 2.94 -0.46 -1.18  
## X3 4 30 0.0 1.0 -0.24 -0.06 1.08 -1.34 1.94 3.28 0.49 -0.94  
## X4 5 30 0.0 1.0 0.21 0.06 1.02 -1.86 1.24 3.10 -0.48 -1.17  
## X5 6 30 0.0 1.0 -0.19 -0.01 1.04 -1.59 1.57 3.16 0.19 -1.19  
## se  
## Resp 1.61  
## X1 0.18  
## X2 0.18  
## X3 0.18  
## X4 0.18  
## X5 0.18

The data has ratings of 30 respondents on 5 questions related to their attitudes towards discount and departmental stores.The ratings are all on a scale of 0-9. The average ratings of X1, X2, X3, X4 and X5 are 4.4, 5.13, 3.67, 5.4 and 4.53. As mean ratings are not same as median ratings, the distributon of ratings is skewed.The median ratings are generally higher than the mean ratings except for questions 3 and 5. The first quartile for X1, X3 and X5 is 2 while for X2 and X4 is 2.5, meaning 25 percent of the ratings are at most 2 for X1, X3 and X5. The third quartile shows variability across questions, for X1 its 7, or 75 percent of ratings for X1 were up to 7.

After scaling the data to ensure that the data is comparable, means of the ratings in questions X1 - X5 are now zero and the standard deviations are all 1.

**matrix of correlations**

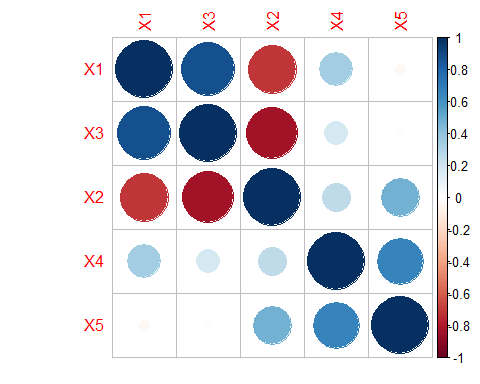
#correlation  
cor(assgn1.sc[, 2:6])

## X1 X2 X3 X4 X5  
## X1 1.00000000 -0.7107386 0.87895393 0.3432966 -0.03865034  
## X2 -0.71073859 1.0000000 -0.83182655 0.2573651 0.47015073  
## X3 0.87895393 -0.8318265 1.00000000 0.1727462 -0.02056604  
## X4 0.34329663 0.2573651 0.17274619 1.0000000 0.66421386  
## X5 -0.03865034 0.4701507 -0.02056604 0.6642139 1.00000000

#correlation plots  
library(corrplot)

## corrplot 0.84 loaded

corrplot(cor(assgn1.sc[, 2:6]), order="hclust")



The correlation analysis suggests that – 1) there is a strong correlation between the variables (questions X1 - X3), 2) Strong positive correlation between X1 and X3 suggests that individuals like store salespersons who respect them and answer their questions, 3) Strong negative correlation between X1 and X2 indicates that individuals like to be respected and treated courteously but not to the extent of fawning or flattery, 4) Also, strong negative correlation between X2 and X3 suggests that individuals like store salespersons who answers all their questions but not those who fawn over customers, 5)MOderate positive correlation between X4 and X5 suggests individuals prefer low price over fancy displays

### PCA

#PCA  
assgn1.pc = prcomp(assgn1.sc[,2:6])  
summary(assgn1.pc)

## Importance of components:  
## PC1 PC2 PC3 PC4 PC5  
## Standard deviation 1.6345 1.3610 0.58478 0.30601 0.20148  
## Proportion of Variance 0.5343 0.3704 0.06839 0.01873 0.00812  
## Cumulative Proportion 0.5343 0.9048 0.97315 0.99188 1.00000

#matrix of relationship between factors and original variables  
assgn1.pc$rotation

## PC1 PC2 PC3 PC4 PC5  
## X1 -0.55491712 0.2324895 -0.2795639 -0.7234233 0.19109786  
## X2 0.57063147 0.2117721 -0.1832902 -0.4621638 -0.61833777  
## X3 -0.57964069 0.1567520 0.3106045 0.1866707 -0.71282845  
## X4 -0.01698458 0.6859778 -0.5610983 0.4611287 0.04092556  
## X5 0.17370927 0.6371534 0.6906101 -0.1248104 0.26709683

PCA maps our original 5 variables on another five-dimensional space, where each dimension is orthogonal. The Proportion of variance row of the table above explains how much of the variance in the data is explained by each component, in decreasing order. Thus, the first component explains 53.43% of the variance in the data, the second component explains 37.04%, and so on. The cumulative variance row for the first and second components together explain 90.48% of the variance in the data. In other words, we can reduce the dimensionality of our original set of five variables to two, and still retain 90.48% of the explanatory power.

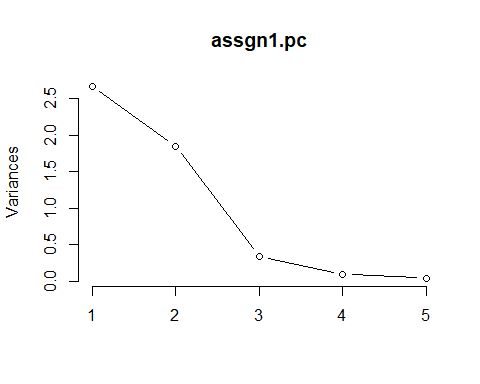
The matrix tell the exact relationship between each factor and the original variables.

### Equations

Equations defining relationship between each factor from PCA and the original variables are -

### Scree Plot

#Using a `scree plot’ to help decide how many factors to retain  
plot(assgn1.pc, type='l')



From the scree plot above, 2 factors lie above the “kink”. Also by the 2nd criteria, 2 is more than 1 and are the optimal factors to be kept in the data.

Thus, we can reduce the dimensionality of our data from five to two and 2 factors should be retained.

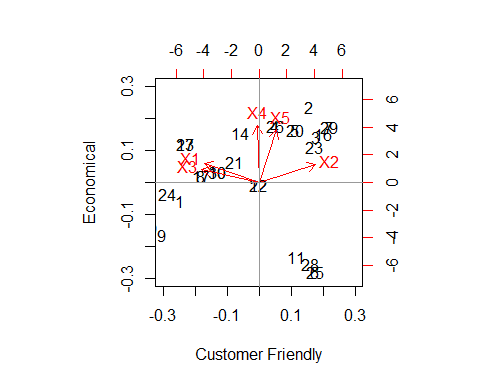
**Factor names**

Intuitive names for the first and second factors can be PC1 = Customer friendly or Good Customer service and PC2 = Economical or thrifty. These factors represent store attributes that customers value. I arrived at them from the equations defining relationship between the factors and the original variables.

The first factor is heavily correlated with the rating on store’s courteous staff who treats customers with repect but do not necessarily spend time answering customer questions or trying to flatter the customer. The PC1 equation is strongly correlated with the above characteristics.The second factor seems most positively correlated with department stores that do not have fancy display and discount stores that offer more bargain and cheap prices showing the importance of low price to customers. This is evidenced from PC2 equation which has strong correlation with above characteristics.

**Plot of all the consumers on a two-dimensional map based on the first two principal components**

#plot the ratings on the two main principal components using `biplot'  
# h and v are horizontal and vertical  
biplot(assgn1.pc, xlim=c(-.3,.3), ylim=c(-0.3,.3),xlab = 'Customer Friendly',ylab = 'Economical')  
abline(h = 0, v = 0, col = "gray60")



### Segments

The above plot suggests presence of 4 customer segments represented by th black dots clusters.

**Characteristics of each segment**

Characteristics of 4 customer segments - The first customer segment seeks good customer service or friendly service, want their questions answered and respected. The second customer segment are the thrifty/economical customers who are price sensitive and seek low cost discount stores than the big fancy displays of departmental stores. The third segment like courteous store salespersons who do not fawn over customers. The fourth segment lies in the bottom right quadrant are customers who don’t have any specific attitudes. The segment characteristics are mostly correlated with original attributes.

### Profitable segment

The fourth customer segment is expected to be more profitable than the rest as they don’t have any specific needs and can be easily pleased. The second segment who seek low prices are expected to be least profitable. The other 2 segments require investment on hiring and training of good salespersons and may not be the most profitable. Assuming the data to be representative, the size of the profitable or fourth segment is 4 or in percent terms is 4/30 = 13.3% of all respondents.