PCA & K-Means for Customer Segmentation

Kunpei Peng & Kun Qian

January 2021

Intro

A pilot survey was conducted of 30 individuals to understand their different attitudes towards discount stores (e.g., K-Mart), which usually come with low levels of service versus departmental stores (e.g., Macys) which come with higher levels of service. Respondents were asked their opinion on a 0-9 agree-disagree scale (where 0 indicates strongly disagree and 9 indicates strongly agree) on the following questions:
1. I never go back to a store that had treated me with disrespect.
2. I find salespeople who fawn over me irritating, I just expect them to courteous.
3. I have a lot of questions when I shop, I greatly appreciate a salesperson who is willing to spend time answering my questions.
4. I care little for the fancy displays in departmental stores.
5. Discount stores are much more of a bargain than departmental stores.

---

1. Import, scale and summarize the data

```
# import data
df <- read.csv("assg1.csv")

# take a look at the first few rows
head(df)

##   Resp X1 X2 X3 X4 X5
## 1    1  6  0  8  4  4
## 2    2  4  9  2  8  9
## 3    3  2  8  2  6  9
## 4    4  5  7  3  9  6
## 5    5  3  8  3  9  6
## 6    6  0  6  0  2  1

# use the 'psych' library to summarize the stats of the data
library(psych)
describe(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
```

From the statistical summary table we can tell that among the responses from the 30 survey takers, X4 has the highest average score, which indicates in general customers tend to agree that they don't care much about fancy displays in department stores. X5 has the lowest average score, which indicates a relatively weak agreement by that the discount stores are much more of a bargain than department stores.

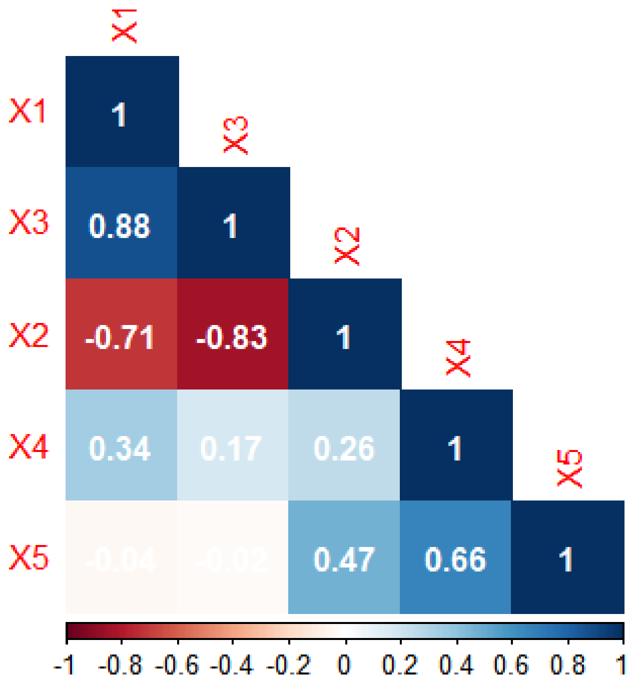
Since each of the 5 questions have different means and standard deviation, we'll then scale the data first before feeding into the PCA model.

```
# scale the data
df.sc <- df
df.sc[,2:6] <- scale(df.sc[,2:6])

# examine the result
describe(df.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
```

After the scaling, the means of all the scoring variables are now zero and the standard deviations are all 1 now.

2. Correlation Matrix

```
library(corrplot)

## corrplot 0.84 loaded

corrplot(cor(df.sc[,2:6]), method = "color", type = "lower", order =
"hclust", addCoef.col = "white")
```



From the correlation plot we can observe that X1 and X3 is highly positively correlated. To put it back in context, it indicates that if a customer would appreciate a salesperson who's willing to answer his/her question, then it's likely that s/he would not go to a store that treat him/her with disrespect. It seems like people who give high score to X1 and X3 care a lot about customer service and would choose between store based on how they're treated.

X2 is highly negatively correlated with both X1 and X3. This means if a customer is likely to find the salespeople overly fawning, then s/he is less likely to appreciate the salesperson's effort to answer question nor would him/her avoid going to a store merely due to disrespect. It seems like people who give a high rating to X2 prefer to be left alone when shopping and are also less sensitive to good customer service.

X5 and X4 also has a moderate positive correlation: Customers who don't care much about the departmental stores' fancy display also agree that discount stores are much of a bargain than departmental stores. It seems like this group of customer has a preference of discount store over departmental stores.

3. PCA

```
# Perform PCA on survey scorings
df.pc <- prcomp(df.sc[,2:6])
summary(df.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
```

PCA maps the original six features on a new 6 dimension space which each dimension is orthogonal to each other. The first principal factor explains 53.4% of the total variation while the second explains 37.0% The first two principal components combined explain 90.5% of the total variation.

```
# Show the relationship between the principal factors and original features
df.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
```

4. Equations of the Original Features and the Principal Factors

$$PC1 = -0.555 \cdot X1 + 0.571 \cdot X2 - 0.580 \cdot X3 - 0.017 \cdot X4 + 0.174 \cdot X5$$

$$PC2 = 0.232 \cdot X1 + 0.212 \cdot X2 + 0.157 \cdot X3 + 0.686 \cdot X4 + 0.637 \cdot X5$$

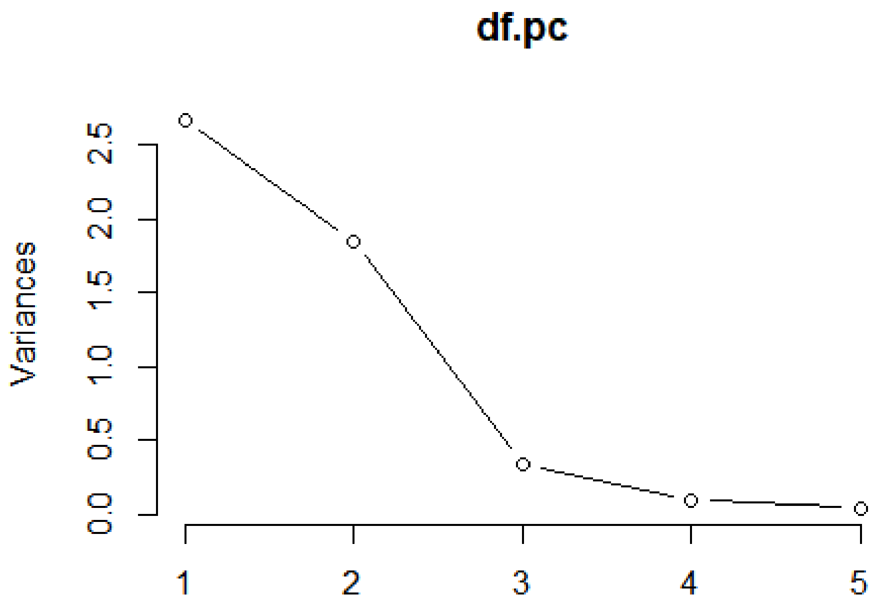
$$PC3 = -0.280 \cdot X1 - 0.183 \cdot X2 + 0.311 \cdot X3 - 0.561 \cdot X4 + 0.691 \cdot X5$$

$$PC4 = -0.723 \cdot X1 - 0.462 \cdot X2 + 0.187 \cdot X3 + 0.461 \cdot X4 - 0.125 \cdot X5$$

$$PC5 = 0.191 \cdot X1 - 0.618 \cdot X2 - 0.713 \cdot X3 + 0.041 \cdot X4 + 0.267 \cdot X5$$

5. How many Factors Should we Retain?

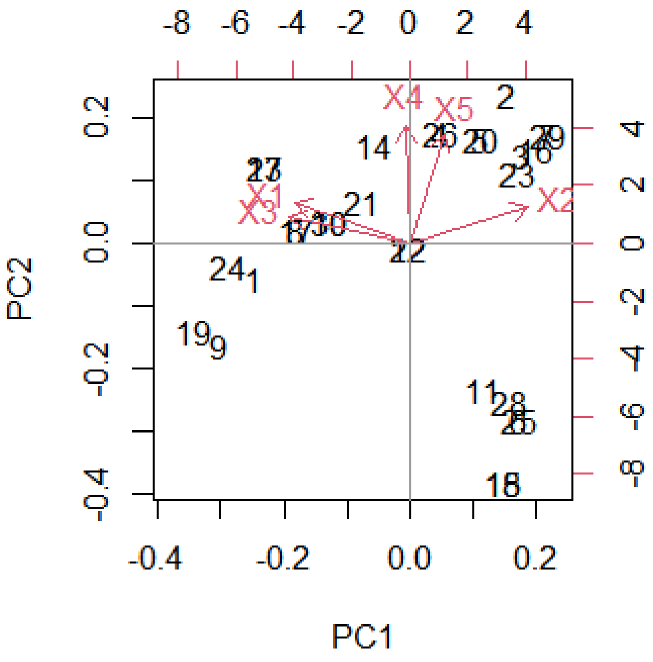
```
# make the 'elbow plot'
plot(df.pc, type = "l")
```



It seems like the “kink” happens to be at the 3rd factor. Additionally, only the 1st and 2nd factors have variances over 1. Therefore we proceed with the 1st and 2nd principal factors.

6. Naming the factors

```
# Take a look at the bi-plot
biplot(df.pc)
abline(h = 0, v = 0, col = "gray60")
```



Recall the relationship between the origianl features and the chosen factors:

$$PC1 = -0.555 \cdot X1 + 0.571 \cdot X2 - 0.580 \cdot X3 - 0.017 \cdot X4 + 0.174 \cdot X5$$

The first factor is positively correlated with X2, and negatively correlated with X1 and X3. X2 represents aversion of over fawning salespeople; X1 and X3 represent high sensitivity to good customer service. Therefore, we intuitively name this principal factor as “Insensitivity to Service”.

$$PC2 = 0.232 \cdot X1 + 0.212 \cdot X2 + 0.157 \cdot X3 + 0.686 \cdot X4 + 0.637 \cdot X5$$

The second factor is strongly positively correlated with both X4 and X5, which represent a less sensitivity of fancy display in the departmental stores and a preference of the discount stores over departmental stores regards of price. We conclude this factor as “Driven by Deals”

```
pc1_name = "Insensitivity to Service"
pc2_name = "Driven by Deals"
```

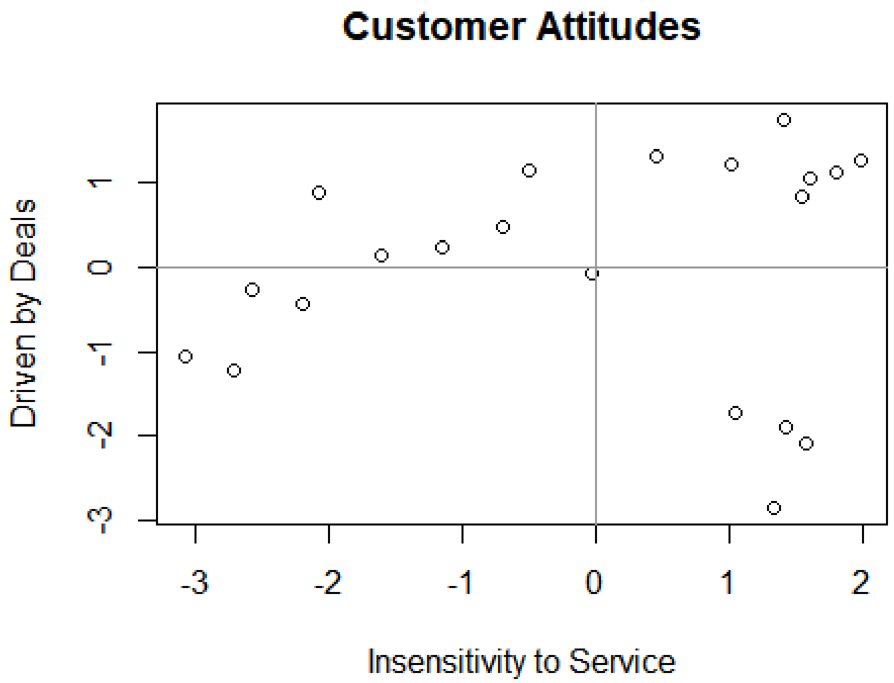
7. Plot Consumers’ Attitude based on the 2 Chosen Factors

```
# Generate a new data set which each survey taker's scores are now
transformed into the new factors
df.pred <- predict(df.pc, df.sc)
df.pca <- as.data.frame(df[,1])
df.pca[,2:6] <- df.pred
colnames(df.pca) <- c("Resp", "PC1","PC2","PC3","PC4","PC5")
```

```
# Take a look at the first few rows
head(df.pca)
```

```
##   Resp        PC1        PC2        PC3         PC4         PC5
## 1    1 -2.1981356 -0.4310847  0.7865926  0.47806275 -0.05326375
## 2    2  1.4055057  1.7530293  0.1986795 -0.38238728  0.08050699
## 3    3  1.6077723  1.0540027  0.8344928 -0.05690111  0.12459818
## 4    4  0.4441288  1.3157762 -0.5834355  0.03189228  0.02291299
## 5    5  1.0077284  1.2270052 -0.4533662  0.37243649 -0.30904699
## 6    6  1.5696986 -2.0890595 -0.2489927  0.31691943  0.11194402
```

```
# Plot customer attitudes
plot(df.pca$PC1, df.pca$PC2, xlab = pc1_name, ylab = pc2_name, main =
"Customer Attitudes")
abline(h = 0, v = 0, col = "gray60")
```



8. Segmentation

We decided to use the unsupervised clustering algorithm, K-means, to assign each of the respondent to a customer segment.

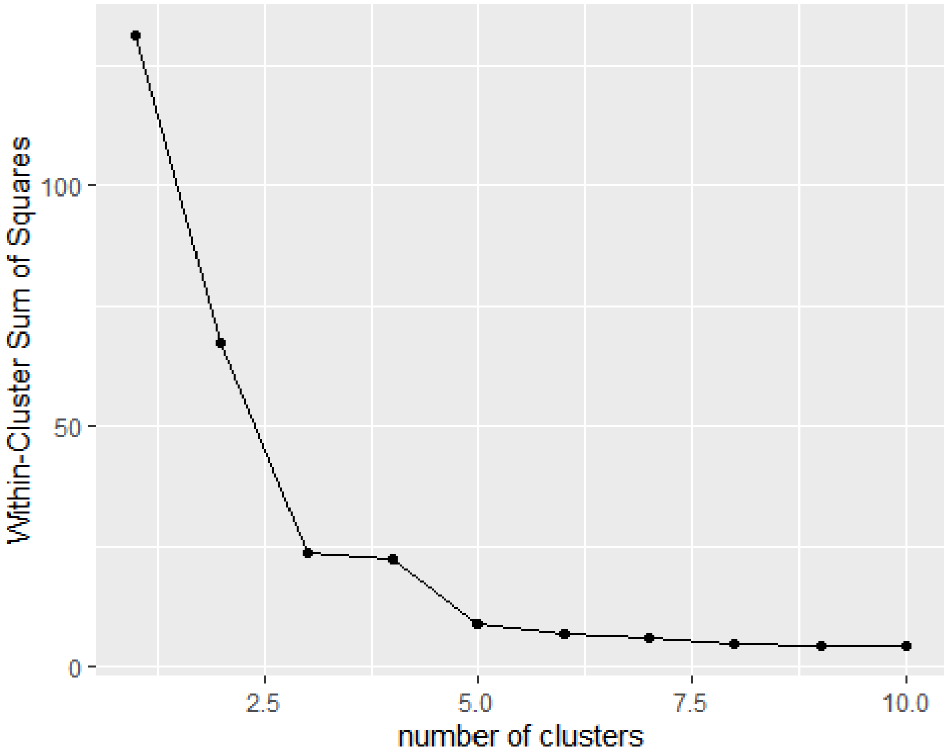
```
# initiate the randomizer
set.seed(1)

# determine the optimal number of clusters by examining the within-cluster
sum of squares (wss) at different k
wss <- numeric()
```

```
for (i in 1:10) {
  wss[i] <- kmeans(df.pca[,2:3], centers=i)$tot.withinss
}
# plot wss
library(ggplot2)
```

```
##
## Attaching package: 'ggplot2'
```

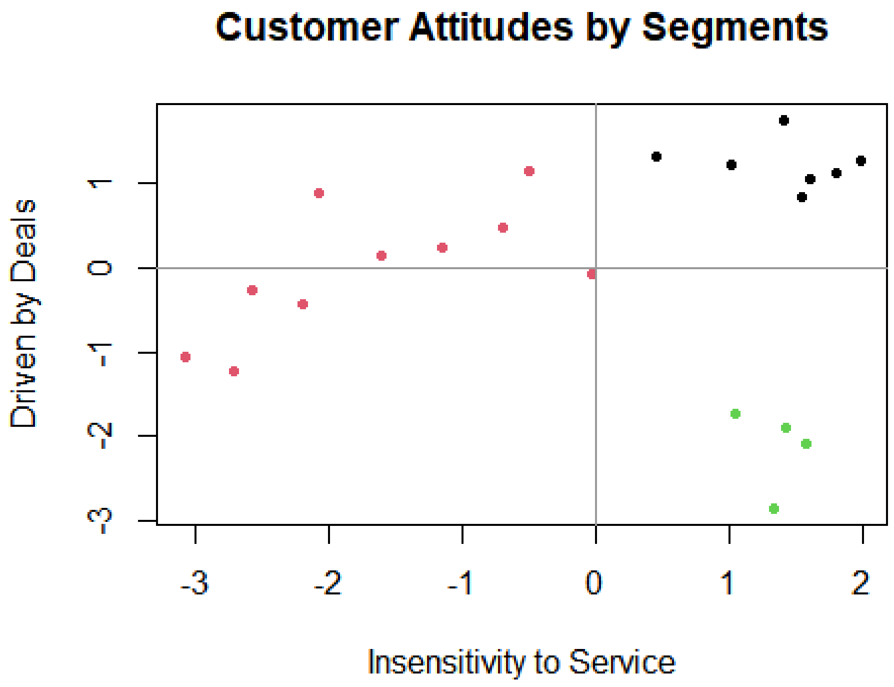
```
## The following objects are masked from 'package:psych':
##
##     %+%, alpha
```

```
qplot(1:10, wss, geom=c('point','line'), xlab="number of clusters",
ylab="Within-Cluster Sum of Squares")
```



This plot of ‘Within-Cluster Sum of Squares’ shows a clear ‘kink’ at 3 clusters. We’ll then use 3 as the number of customer segments. Now let’s apply K-means and plot the result.

```
# perform k-means with 3 clusters
clus <- kmeans(df.pc$x[,1:2],3)

# plot the segments
plot(df.pc$x[,1:2], pch=20, col=clus$cluster, xlab=pc1_name, ylab=pc2_name,
main="Customer Attitudes by Segments")
abline(h = 0, v = 0, col = "gray60")
```



```
# dataEllipse(df.pc$x[,1],df.pc$x[,2], center = as.factor(clus$cluster),
lwd=1,
#     group.labels = NULL, plot.points=FALSE, add=TRUE,
#     fill=TRUE, fill.alpha=0.02)
```

As we can see from the graph, each of the respondent is now assigned to one of the three segments indicating by the color of either red, black, or green of the dot.

9. Describe the characteristics of each segment. Discuss the relationship between the segment characteristics and the original attributes

Bargain Hunters: The black group at the top right corner is insensitive to excellent customer services while having a strong perception that discount stores usually have better deals. We suspect this group of customers care more about deals and bargain rather than the shopping experience. They're likely to be loyal customers of the discount stores. Connecting back with the original attributes, they score high in X2, X4, and X5, and low in X1 and X3.

Mr/Mrs I Don't Care: The green group at the bottom right corner is both insensitive to the service and insensitive to the deals. They seem not to have a strong opinion on how stores should work and maybe shopping in stores is less of an important thing for them to take time to consider. Looking at the original attributes, these are the people who gave low scores in X1, X3, X4, and X5, and high score in X2 only.

Better Treat Me Well: The red group at the right is at the negative side of 'Insensitivity to Service'. This implies a great concern of customer service. Recall the constitution of the first factor, negativity of the first factor is link with positivity in both X1 and X3. These two attributes express aversion of disrespect and appreciation of help. Therefore, this group of customers is very experience driven, and they will choosing store based on the customer service. However, within this group people can be either slightly driven by deals or slightly indifferent of deals. We believe this group would enjoy shopping at departmental stores more.

10. Which segment is the most profitable? What is the size of this segment?

We expect the most profitable segment to be Segment: **"Better Treat me well"**. They value service quality over price of the products they consume. Under the premise that it's relatively cheap for stores to re-train staffs, it makes sense for stores to attract these shoppers in making high margin purchases by offering above average service quality and attentions to the customers. Moreover, the "Better Treat Me Well" segment is also the biggest segment with relatively lower sensitivity to "bargain pricing", as the range of variation retains between -1 and 1 along the y-axis. This indicates that stores can potentially raise the profit margin by increasing the prices of their products while improving their service quality because their customers in this segment are not as likely to be influenced by the slight rise in prices.

The size of this segment is

$$\frac{10}{30} = 33.3\%$$