MS – E 2112 – Multivariate statistical analysis – Home exercise 6

Srikanth Gadicherla, 546195

February 26, 2016

Problem 1

Data: A 2-dimensional frequency table SMOKING.txt data was provided same as the last exercise, where the details of smoking of employees in a company was tabulated. The employees were categorized into Senior Managers (SM), Junior managers (JM), Senior Employee (SE), Junior Employee (JE), Secretaries (SC), which were further categorized into None, Light, Medium and Heavy smokers.

The below balloon plot shows the data. The size of the bubble shows the magnitude of the value in each category for employees.

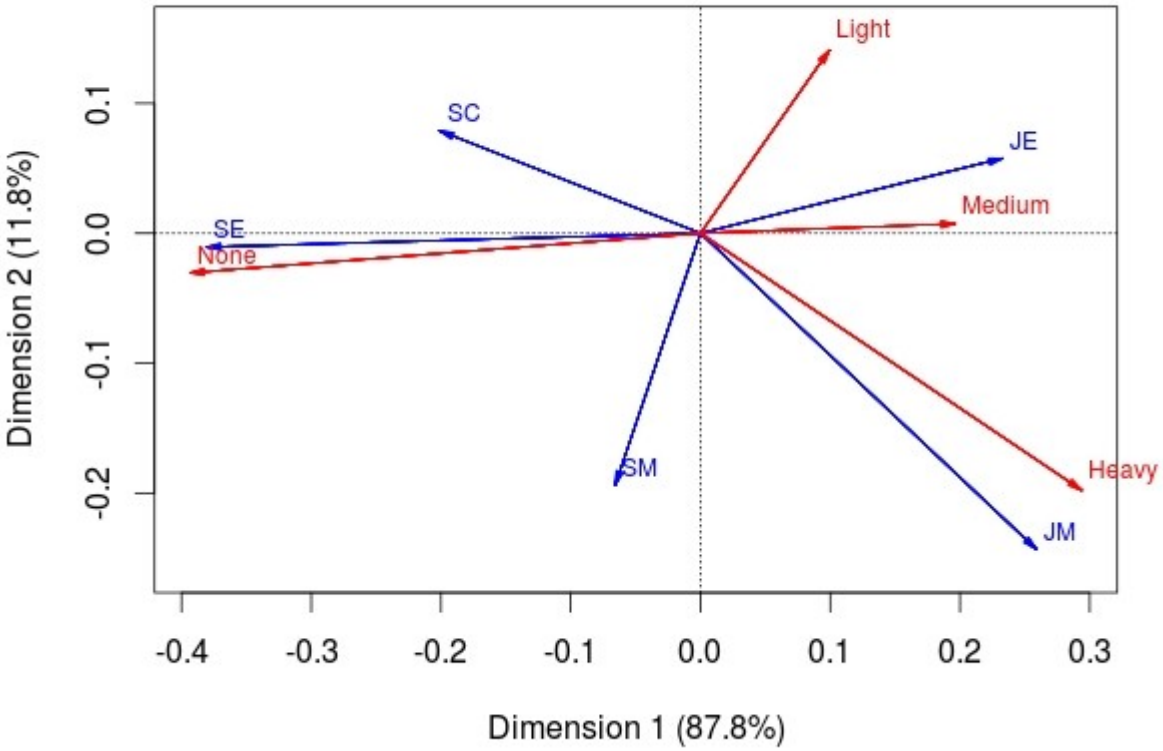
a) The row profile is row wise normalization where sum of each row is 1. For our data, the row profile is given in the below figure.

	Row Profiles			
	None	Light	Medium	Heavy
SM	0.3636	0.1818	0.2727	0.1818
JM	0.2222	0.1667	0.3889	0.2222
SE	0.4902	0.1961	0.2353	0.0784
JE	0.2045	0.2727	0.3750	0.1477
SC	0.4000	0.2400	0.2800	0.0800

Similarly, The column profile is column wise normalization where sum of each column is 1. For our data, the column profile is given in the below figure.

	Column Profiles			
	None	Light	Medium	Heavy
SM	0.06557377	0.04444444	0.0483871	0.08
JM	0.06557377	0.06666667	0.1129032	0.16
SE	0.40983607	0.22222222	0.1935484	0.16
JE	0.29508197	0.53333333	0.5322581	0.52
SC	0.16393443	0.13333333	0.1129032	0.08

b) The correspondence analysis was performed on the data with the use of package “ca” in R. The bi-plot for our data is given below.



The main observations from the bi-plot are:

1. The Junior Managers(JM) are the most frequent smokers in Heavy category as it shown by the JM and Heavy arrows with less angle between them.

2. The Senior employees(SE) are least frequent smokers. This is supported by the fact that the correponding arrows in the figure are together and also close to the X-axis.

3. The Junior Employees (JE) are frequent smokers in all the categories, that's the reason the JE arrow is between arrows of categories Light, Medium, Heavy. This can also be substantiated by the fact that JE arrow is almost diametricallty opposite to None arrow.

4. The Secretaries(SC) are not Light smokers (or are not related), this is substantiated by the fact that SC arrow is perpendicular to Light arrow and are least frequent smokers in Heavy and Medium

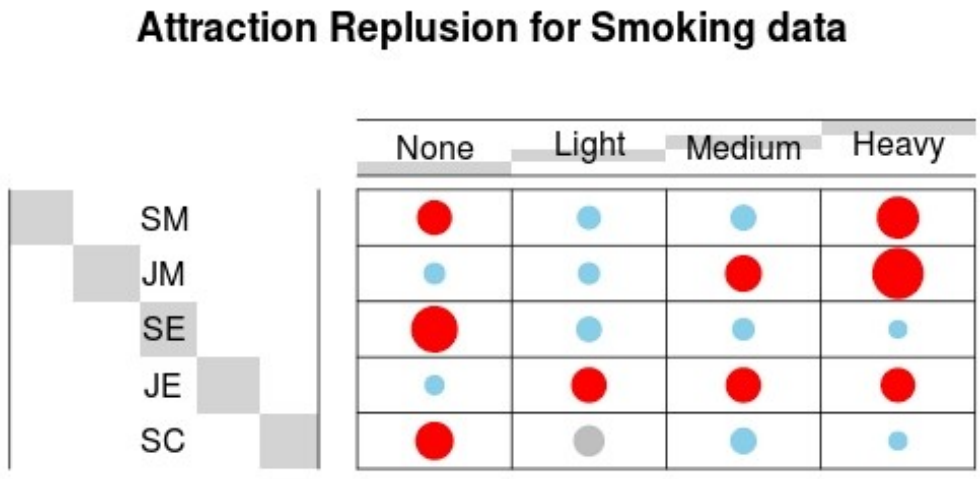
category substantiated by the fact that the SC arrow is (almost) diagonally opposite direction to Medium and Heavy.

5. The Senior Managers (SM) either smoke heavily or dont smoke, that's the reason the SM arrow is between Heavy and None arrows.

c) Are the results in Harmony with Attraction Repulsion matrix?

Below is the balloon plot of the Attraction Repulsion matrix, proving the claims of the correspondence analysis.

Here in the plot, the blue is less frequent(less than 1), red is more frequent (more than 1) and grey is not related (equal to 1).



The R code for the exercise is given in the appendix.

Appendix

The code for the problem solved above.

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
setwd("~/Documents/OneDrive/Aalto/Sem2/MSA/Ex 6")
library(ca)
library("gplots")

data <- read.table("SMOKING.txt",header = T, row.names = 1)

data <- data[,-5]
data <- data[-6,]

D <- as.matrix(data)

# Graph
# 1. convert the data as a table
dt <- as.table(as.matrix(D))
balloonplot(t(dt), main ="Smoking of Employees in a Company", xlab ="", ylab="",
            label = FALSE, show.margins = FALSE)

# standardizing the row and columns so that they add to 1
row_prof <- prop.table(D,1) #row profile
col_prof <- prop.table(D,2) #col profile

#Correspondence Analysis
smoking_ca <- ca(D,nd=NA)

#Summary and Plot
summary(smoking_ca)
plot(smoking_ca,arrows=c(T,T))

## AR Matrix and bubble plot
```

```
v1 <- margin.table(D,1) # Gives you the sum of all the rows

v2 <- margin.table(D,2) # Gives you the sum of all columns


n1 = length(v1)

n2 = length(v2)


V1 = matrix(v1,ncol = 1)

V2 = matrix(v2,ncol=n2)


E = V1 %*% V2 / sum(D)


AR.matrix <- D/E # D = original data (number of observations)

# E = expected number of observations under independence


# Graph

# 1. convert the data as a table

dt <- as.table(as.matrix(AR.matrix))

balloonplot(t(dt), main ="Attraction Replusion for Smoking data", xlab ="",
ylab="",

            label = FALSE, show.margins = FALSE, scale.method=c("diameter"),

            dotcol = c("red","skyblue","skyblue","red","skyblue","skyblue",

                       "red","red","red","skyblue","skyblue","skyblue","skyblue",

                       "red","red","red","red","grey","skyblue","skyblue"))


%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
```