HUDM6122 Homework_08

Chenguang Pan

2023-04-22

0.1 Github Address

All my latest homework can be found on Github: https://github.com/cgpan/hudm6122_homeworks . Thanks for checking if interested.

0.2 Exercise 1

Do a correspondence analysis for the car-ratings (file cars.txt). Explain how this table can be considered as a contingency table.The data are averaged ratings for 24 car types from a sample of 40 persons. The marks range from 1 (very good) to 6 (very bad).

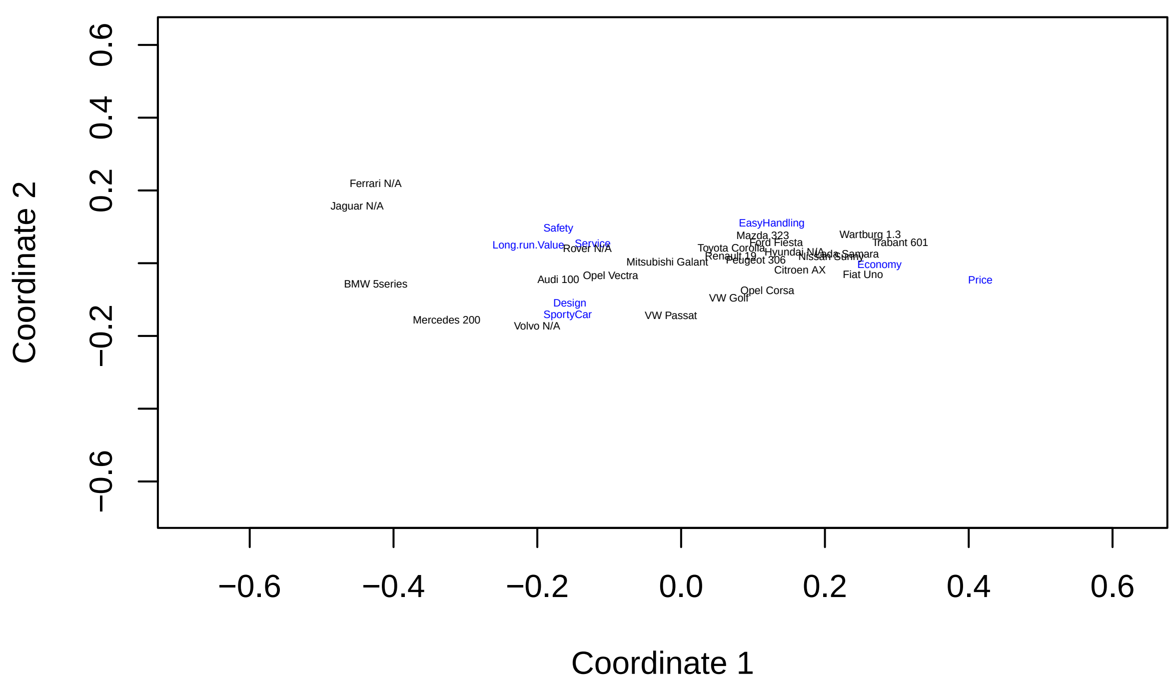
MY SOLUTION:

```
> # import the data
> df <- read.table("cars.txt", header = T)
> head(df)
     Type   Model Economy Service Long.run.Value Price Design SportyCar Safety
1    Audi     100     3.9     2.8            2.2   4.2    3.0       3.1    2.4
2     BMW 5series     4.8     1.6            1.9   5.0    2.0       2.5    1.6
3 Citroen      AX     3.0     3.8            3.8   2.7    4.0       4.4    4.0
4 Ferrari     N/A     5.3     2.9            2.2   5.9    1.7       1.1    3.3
5    Fiat     Uno     2.1     3.9            4.0   2.6    4.5       4.4    4.4
6    Ford  Fiesta     2.3     3.1            3.4   2.6    3.2       3.3    3.6
  EasyHandling
1          2.8
2          2.8
3          2.6
4          4.3
5          2.2
6          2.8
```

This dataset is a two-way table that shows the values of different attributes or characteristics (columns) for each type or model of a car (rows). Each row represents a different car model and each column represents a different attribute or characteristic of the car. Therefore, it can be seen as a contingency table.

```
> # import the packages
> library(FactoMineR)
> library(factoextra)
>
> # combine the first two columns into one
> df[,1] <- paste(df$Type,df$Model)
```

```
> df <- df[,-2]
>
> # write a function to calculate the chi-squared distance
> D <- function(x){
+   a <- t(t(x)/colSums(x))
+   ret <- sqrt(colSums((a[,rep(1:ncol(x),ncol(x))]-
+                           a[,rep(1:ncol(x),rep(ncol(x), ncol(x)))])^2*
+                          sum(x)/rowSums(x)))
+   matrix(ret, ncol = ncol(x))
+ }
> # chi-squared distance for columns
> dcols <- D(df[,-1])
> drows <- D(t(df[,-1]))
>
> # run CA
> r1 <- cmdscale(dcols, eig=T)
> c1 <- cmdscale(drows, eig=T)
> plot(r1$points, xlim = range(r1$points[,1], c1$points[,1]) * 1.5,
+      ylim = range(r1$points[,1], c1$points[,1]) * 1.5, type = "n",
+      xlab = "Coordinate 1", ylab = "Coordinate 2", lwd = 2)
> text(r1$points, labels = colnames(df[,-1]), cex = 0.3, col="blue")
> text(c1$points, labels = df[,1], cex = 0.3)
```



0.3 Exercise 2

Write an R function to compute the chi-square statistic of independence. Test the null using for the bachelor data (file bachelors.txt). The data consists of observations of 202,100 bachelors from France and give the frequencies for different sets of modalities classified into regions.

MY SOLUTION:
Write the function first.

```
> # Function to compute chi-square statistic of independence
> chi_square <- function(table) {
+   # Compute row and column totals
+   row_totals <- apply(table, 1, sum)
+   col_totals <- apply(table, 2, sum)
+   n <- sum(table) # Total number of observations
+
+   # Compute expected values
+   expected <- outer(row_totals, col_totals) / n
+
+   # Compute chi-square statistic
+   chi_sq <- sum((table - expected)^2 / expected)
+
+   # Return result
+   return(chi_sq)
+ }
```

Next, import the data and get the test.

```
> # import and clean the data
> df <- read.table("bachelors.txt", header = T)
> rownames(df) <- df$Abbrev.
> df <- df[,-c(1,2,ncol(df))]
>
> # run the function on this cleaned data
> (chi_sq <- chi_square(df))
[1] 4354.548
```

The function looks good. Next, get the pvalue.

```
> # get the degree of freedom
> dg_fd <- (nrow(df) - 1) * (ncol(df) - 1)
> # get the p-value
> (p_value <- 1 - pchisq(chi_sq, dg_fd))
[1] 0
```

Finally, compare the result with R-built-in function.

```
> test <- chisq.test(df)
> test

	Pearson's Chi-squared test

data:  df
X-squared = 4354.5, df = 147, p-value < 2.2e-16
```

The results are identical.That is, we reject the null hypothesis. The variables are not independent with each other!

0.4 Exercise 3

Do correspondence analysis of the U.S. crime data (file UScrime.txt), and determine the absolute contributions for the first three axes. How can you interpret the third axis? Try to identify the states with one of the four regions to which it belongs. Do you think the four regions have a different behavior with respect to crime?

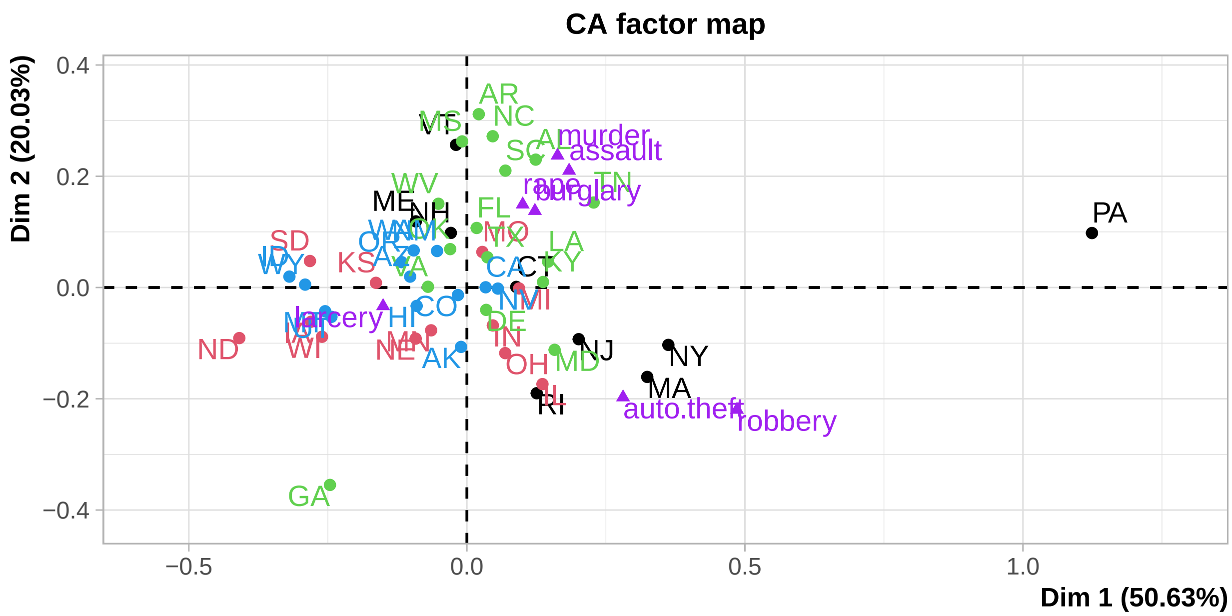
MY SOLUTION:

```
> # import the data
> df <- read.table("UScrime-1.txt", header = T)
> rownames(df) <- df$state
> # RUN CA
> library(FactoMineR)
> ca_ <- CA(df[4:10], graph = F)
> summary(ca_)

Call:
CA(X = df[4:10], graph = F)

The chi square of independence between the two variables is equal to 11287.55 (p-value =  0 ).

Eigenvalues
                       Dim.1   Dim.2   Dim.3   Dim.4   Dim.5   Dim.6
Variance               0.033   0.013   0.012   0.007   0.000   0.000
% of var.             50.632  20.035  17.911  10.689   0.456   0.278
Cumulative % of var.  50.632  70.667  88.577  99.266  99.722 100.000

Rows (the 10 first)
             Iner*1000    Dim.1    ctr   cos2    Dim.2    ctr   cos2    Dim.3
ME         |     0.335 | -0.092  0.272  0.268 |  0.119  1.153  0.450 | -0.091
NH         |     0.415 | -0.029  0.025  0.020 |  0.098  0.734  0.231 | -0.174
VT         |     1.094 | -0.020  0.013  0.004 |  0.256  5.522  0.660 | -0.167
MA         |     4.467 |  0.324  6.895  0.510 | -0.161  4.280  0.125 | -0.252
RI         |     2.489 |  0.126  1.177  0.156 | -0.190  6.817  0.358 | -0.169
CT         |     0.649 |  0.089  0.481  0.245 |  0.001  0.000  0.000 | -0.154
NY         |     5.411 |  0.362 11.019  0.673 | -0.103  2.258  0.055 |  0.146
NJ         |     1.175 |  0.201  2.521  0.709 | -0.093  1.360  0.151 | -0.067
PA         |     8.652 |  1.124 25.596  0.977 |  0.098  0.491  0.007 | -0.129
OH         |     0.389 |  0.069  0.297  0.253 | -0.118  2.198  0.739 |  0.003
              ctr   cos2
ME          0.762  0.266 |
NH          2.573  0.724 |
VT          2.613  0.279 |
MA         11.770  0.308 |
RI          6.050  0.284 |
CT          4.114  0.741 |
NY          5.066  0.109 |
NJ          0.801  0.080 |
PA          0.960  0.013 |
OH          0.002  0.001 |
```

```
Columns
             Iner*1000    Dim.1    ctr   cos2    Dim.2    ctr   cos2    Dim.3
murder     |     0.940 |  0.163  0.160  0.056 |  0.237  0.853  0.119 |  0.393
rape       |     0.503 |  0.100  0.138  0.090 |  0.149  0.770  0.200 |  0.109
robbery    |    14.193 |  0.486 20.136  0.469 | -0.220 10.403  0.096 |  0.368
assault    |     9.570 |  0.184  4.003  0.138 |  0.210 13.213  0.180 |  0.326
burglary   |    11.226 |  0.122 11.952  0.352 |  0.138 38.339  0.446 | -0.079
larcery    |    13.410 | -0.151 38.231  0.942 | -0.033  4.737  0.046 |  0.017
auto.theft |    15.402 |  0.281 25.381  0.544 | -0.197 31.683  0.269 | -0.121
              ctr   cos2
murder      2.624  0.326 |
rape        0.461  0.107 |
robbery    32.689  0.269 |
assault    35.508  0.434 |
burglary   14.138  0.147 |
larcery     1.335  0.012 |
auto.theft 13.244  0.100 |
> # table of eigenvalues
> row_coord <- ca_$row$coord[,3]
> row_coord[order(row_coord, decreasing = TRUE)[1:7]]
       GA        MS        NC        MD        IL        NY        LA
0.3473941 0.2309334 0.2182393 0.2081974 0.1719769 0.1461266 0.1306424
> ca_$col$coord[,3]
     murder        rape     robbery     assault    burglary     larcery
 0.39332624  0.10927656  0.36809412  0.32560788 -0.07913001  0.01674929
 auto.theft
-0.12065495
```

[Refer to Xue Yu's solution]
The absolute contributions for the first three axes are 50.63%, 20.03%, and 17.91% respectively. The third axis can better represent crimes related to personal injury (murder, robbery, and assault) in states such as GA, MS, W, MD, etc.

```
> plot(ca_, col.row = df$region, col.col="purple")
```



Northeast region (black points) is more related to auto.theft and robbery. Some states in mid-west (red points), such as ND, IA, WI, and SD are more related to larcery, while there are still some states in mid-west, such as IN, OH,and IL are more related to auto.theft and robbery. Most states from south (green points) are more related to roe, burglary, murder and assault, Most states from west (blue points) are more related to larcery.

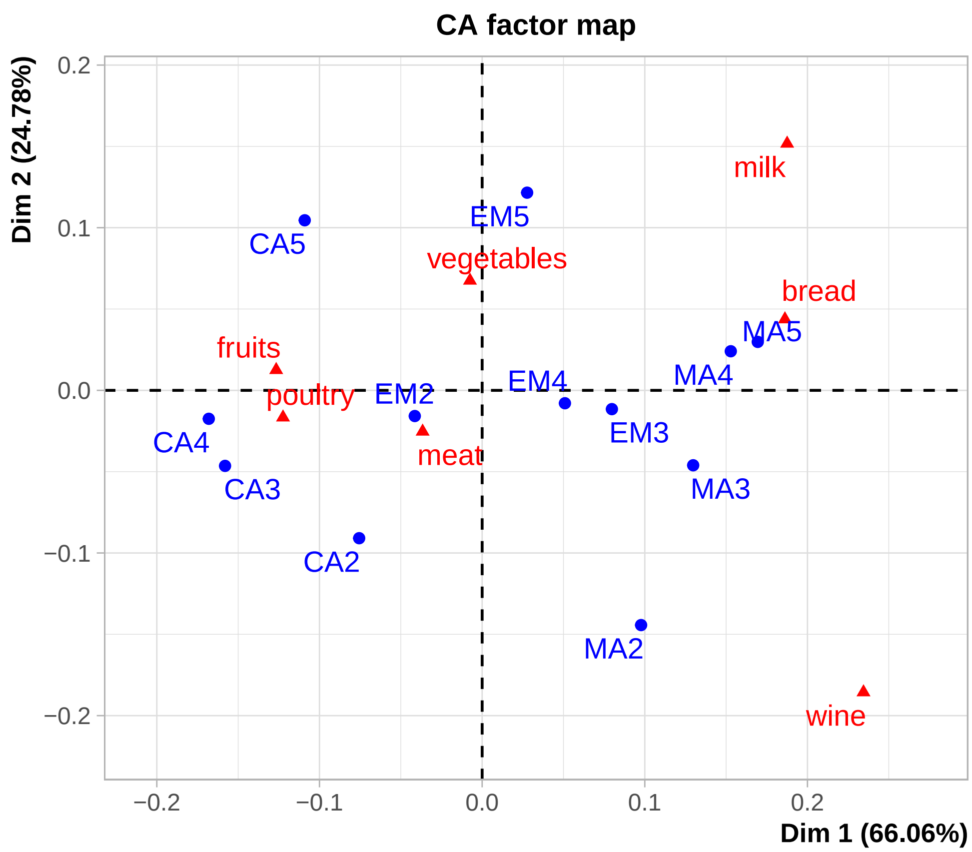
0.5 Exercise 4

Consider the food data (file food.txt). Given that all of the variables are measured in the same units (dollars), explain how this table can be considered as a contingency table. Perform a correspondence analysis and compare the results to those obtained with the PCA analysis of the correlation matrix. The data set consists of the average expenditures on food for several different types of families (manual workers = MA, employees = EM, managers = CA) with different numbers of children (2,3,4 or 5 children).

MY SOLUTION:
Since the rows represent different workertypes and the columns represent different food categories, and the values in the table can be seen as the count or frequency of observations in each category for each workertype, therefore this dataset can be considered as a contingency table.

```
> # import the data
> df <- read.table("food.txt", header = T)
> rownames(df) <- df$Workertype
> df <- df[,-c(1,2)]
> colnames(df)
[1] "bread"      "vegetables" "fruits"     "meat"       "poultry"
[6] "milk"       "wine"
```

```
> dim(df)
[1] 12  7
> # RUN CA
> library(FactoMineR)
> ca <- CA(df,graph=T)
```



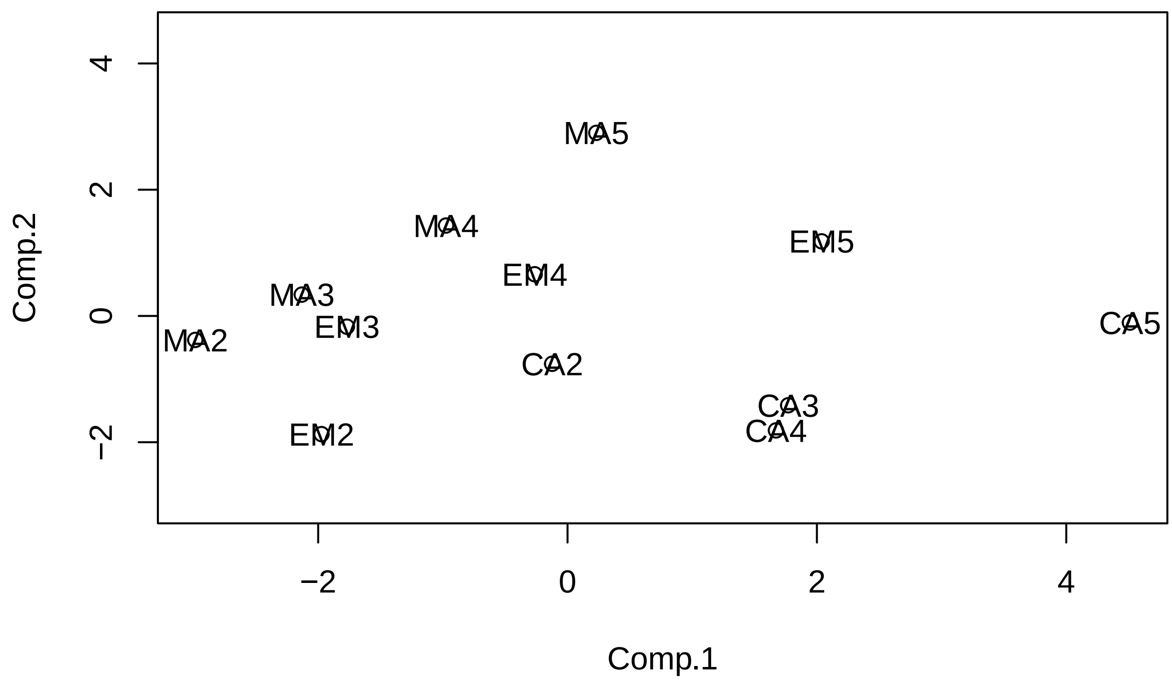

```
> ca$eig
      eigenvalue percentage of variance cumulative percentage of variance
dim 1 0.0139275913           66.0607153                          66.06072
dim 2 0.0052247270           24.7816865                          90.84240
dim 3 0.0009972617            4.7301662                          95.57257
dim 4 0.0005210095            2.4712284                          98.04380
dim 5 0.0002978663            1.4128260                          99.45662
dim 6 0.0001145604            0.5433776                         100.00000
```

```
> # run PCA
> df_scale <- scale(df)
> pca <- princomp(df, cor=T)
> summary(pca)
Importance of components:
                          Comp.1    Comp.2     Comp.3    Comp.4      Comp.5
Standard deviation     2.0816429 1.3528822 0.79425212 0.3582283 0.239908711
Proportion of Variance 0.6190339 0.2614700 0.09011949 0.0183325 0.008222313
Cumulative Proportion  0.6190339 0.8805039 0.97062342 0.9889559 0.997178229
                            Comp.6       Comp.7
```

```
Standard deviation     0.137290211 0.0300632132
Proportion of Variance 0.002692657 0.0001291138
Cumulative Proportion  0.999870886 1.0000000000
```

I will choose the first two components to represent the data.

```
> xlim <- range(pca$scores[,1])
> plot(pca$scores, xlim=xlim, ylim=xlim)
> text(pca$scores, rownames(df))
```



The two graphs show very similar results, since they all captures the similarity between the workertypes, although the data points are shown at different location.