Practical 3

Hsuan Lee

1. Take home exercise

The data to be analyzed in this exercise can be found in the following file.

- `GBMobility.txt`

The data in this file constitute a contingency table of counts, the classic 1949 Great Britain five-by-five son's by father's occupational mobility table. Import the data into R. The warning message that might show up in using the function read.table() can be ignored.

```
X <- read.table("data/GBMobility.txt")
```

```
## Warning in read.table("data/GBMobility.txt"): incomplete final line found by
## readTableHeader on 'data/GBMobility.txt'
```

The rows of the data table correspond to five different categories of father's occupation and the columns to the same five different categories of son's occupation. The cells in the main diagonal of the table refer to fathers and sons with the same occupational category, and this group is important because it measures the total amount of mobility exhibited by the sons. The categories for both nominal variables are:

1. upper nonmanual (UN; self-employed professionals, salaried professionals, managers, nonretail salespersons)
2. lower nonmanual (LN; proprietors, clerical workers, retail salespersons)
3. upper manual (UM; manufacturing craftsmen, other craftsmen, construction crafts- men)
4. lower manual (LM; service workers, other operatives, manufacturing operatives, ma- nufacturing laborers, other laborers)
5. farm (F; farmers and farm managers, farm laborers)

If the table is called X, then the row and column labels can be assigned by executing

```
rownames(X) <- c('UN F','LN F','UM F','LM F','F F')
colnames(X) <- c('UN S','LN S','UM S','LM S','F S')
```

Obtain the correspondence table using the function prop.table(). Use the function sum() to check whether the sum of all elements of the correspondence table equals one. The matrix of row profiles can be obtained by using the argument margin = 1 in the function prop.table() and the matrix of column profiles by using the argument margin = 2. Use the functions rowSums() and colSums() to check whether the sums of the profiles are all equal to one. Install and load the R package ggpubr and execute ggballoonplot(X, fill ='value').

```r
# make the data as matrix
X <- as.matrix(X)

# check if the sum of all elements of the correspondence table is 1
sum(prop.table(X))
```

```
## [1] 1
```

```r
# check the row sum = 1 or not
rowSums(prop.table(X, margin = 1))
```

```
## UN F LN F UM F LM F  F F
##    1    1    1    1    1
```

```r
# check the column sum = 1 or not
colSums(prop.table(X, margin = 2))
```

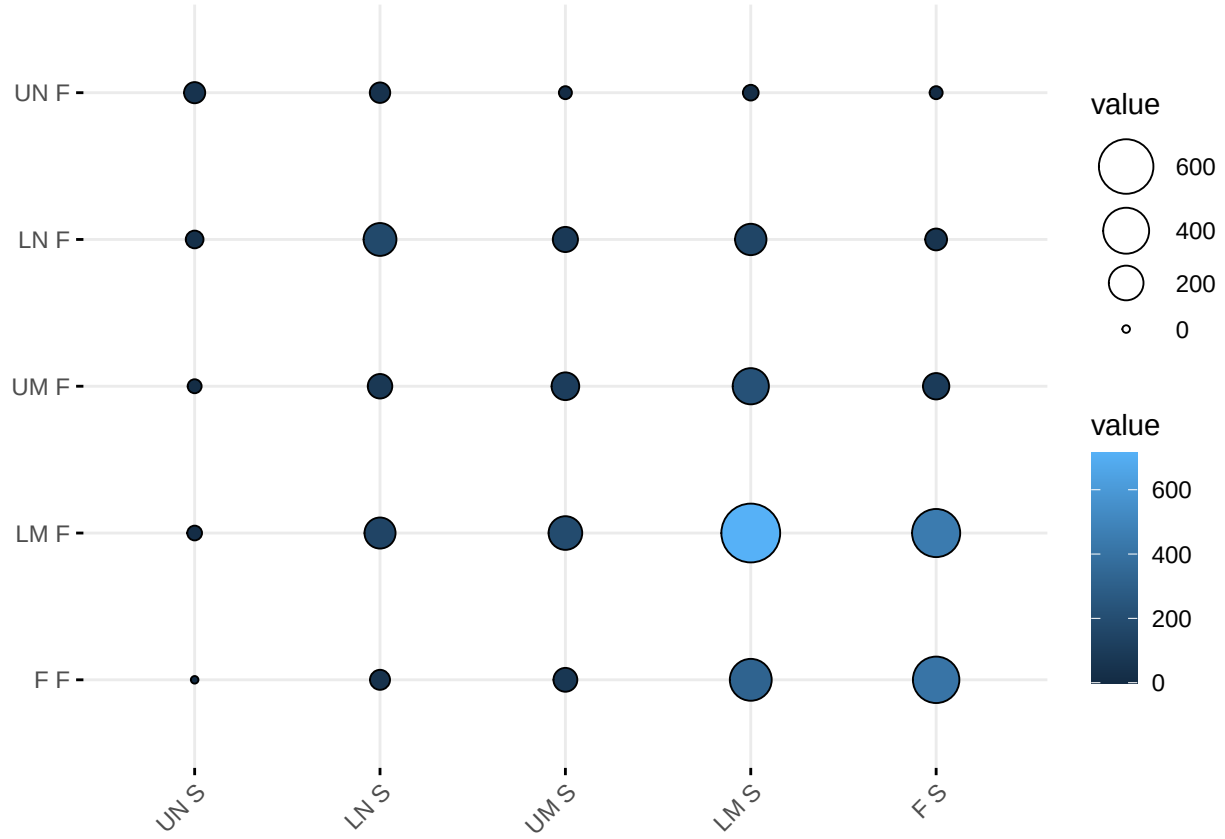
```
## UN S LN S UM S LM S  F S
##    1    1    1    1    1
```

To visualize the correspondence table using a balloon plot. One of the R packages for correspondence analysis is ca. Install and load this package.

```r
library(ca)
library(ggpubr)
```

```
## Loading required package: ggplot2
```

```r
ggballoonplot(X, fill ='value')
```



1. Apply a correspondence analysis to the GB mobility table. The function to be used is ca().

```
cor_ana <- ca(X)
```

2: Explore the arguments and values of the function ca() using ?ca. Obtain the row and column standard coordinates.

```
#?ca

# row standard coordinates
cor_ana$rowcoord
```

```
##             Dim1        Dim2       Dim3        Dim4
## UN F -4.21774838  2.76259596 -0.8247861  0.08316923
## LN F -1.07057462 -1.51349359  1.5820254 -0.35362317
## UM F -0.09618024 -0.90600267 -1.1383626  1.90394443
## LM F  0.32798260 -0.02322569 -0.6331580 -0.89827224
## F F   0.74389540  1.06175694  1.0284440  0.63250001
```

```
# column standard coordinates
cor_ana$colcoord
```

```
##             Dim1       Dim2       Dim3       Dim4
## UN S -4.57435069  3.1286811 -1.4321499  0.4325194
## LN S -1.23068956 -1.3000285  1.6060874 -0.6059222
```

```
## UM S -0.08655863 -1.0888712 -0.7366755  2.2097314
## LM S  0.29172981 -0.1641592 -0.8337655 -0.7999665
## F S   0.68418273  1.0303197  0.8768138  0.3742693
```

3. Use the function summary() to determine the proportion of total inertia explained by the first two extracted dimensions.

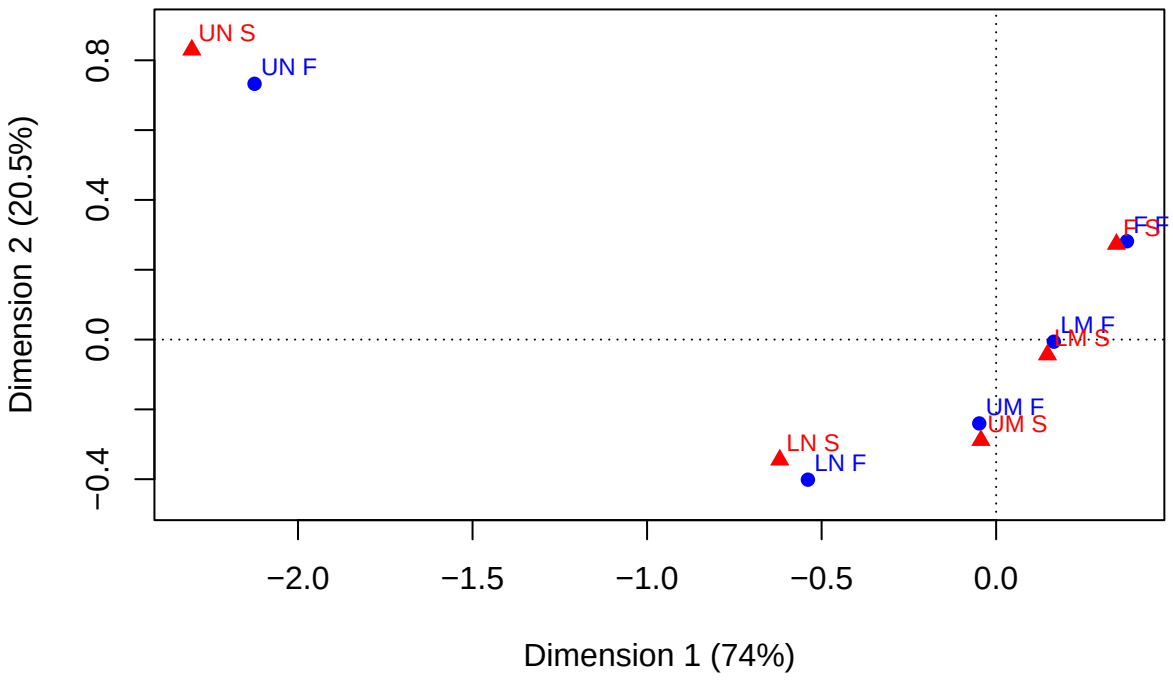
```
summary(cor_ana)
```

```
##
## Principal inertias (eigenvalues):
##
##  dim    value      %   cum%   scree plot
##  1      0.253729  74.0  74.0  ******************
##  2      0.070329  20.5  94.5  *****
##  3      0.015439   4.5  99.0  *
##  4      0.003471   1.0 100.0
##         -------- -----
##  Total: 0.342969 100.0
##
##
## Rows:
##     name   mass  qlt  inr     k=1 cor ctr    k=2 cor ctr
## 1 |  UNF |   37  998  544 | -2125 892 656 |  733 106 282 |
## 2 |  LNF |  142  920  203 |  -539 592 162 | -401 328 324 |
## 3 |  UMF |  148  648   40 |   -48  25   1 | -240 623 122 |
## 4 |  LMF |  432  752   46 |   165 751  46 |   -6   1   0 |
## 5 |   FF |  242  925  167 |   375 591 134 |  282 334 272 |
##
## Columns:
##     name   mass  qlt  inr     k=1 cor ctr    k=2 cor ctr
## 1 |  UNS |   29  995  518 | -2304 880 616 |  830 114 288 |
## 2 |  LNS |  140  924  222 |  -620 706 212 | -345 218 236 |
## 3 |  UMS |  131  771   42 |   -44  17   1 | -289 754 156 |
## 4 |  LMS |  409  645   43 |   147 593  35 |  -44  52  11 |
## 5 |   FS |  291  940  174 |   345 577 136 |  273 363 309 |
```

94.5% of total inertia explained by the first two extracted dimensions.

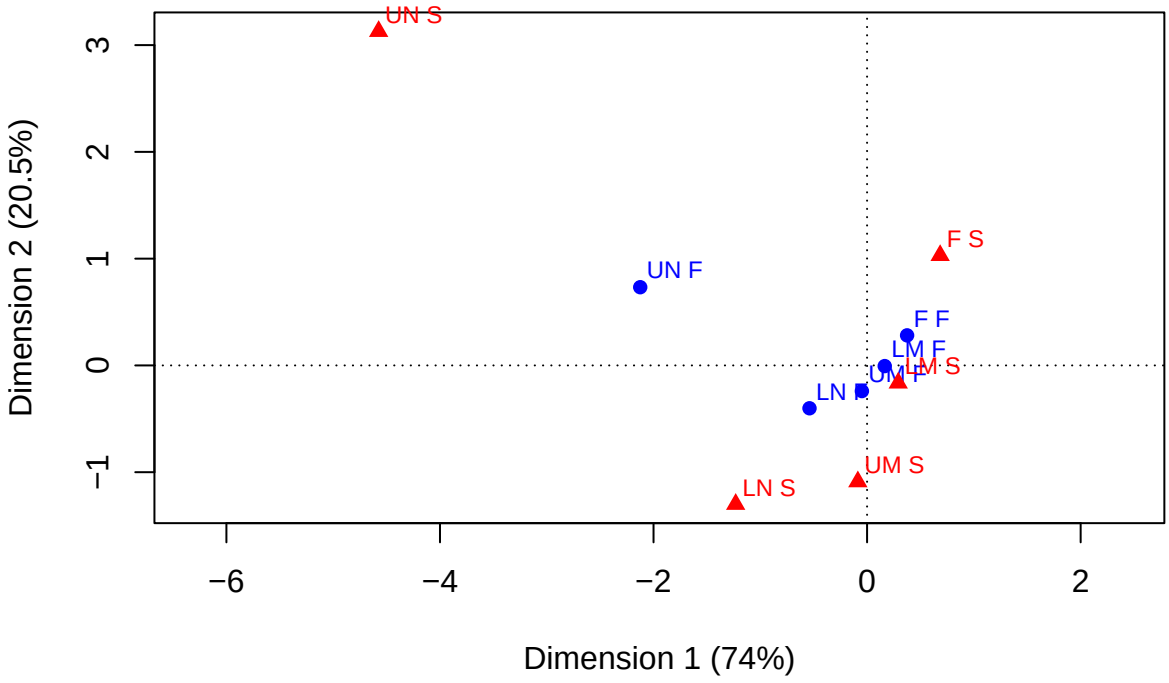
4. Use the function plot() to obtain a symmetric map.

```
plot(cor_ana, map = "symmetric")
```



5. Use the argument map='rowprincipal' to obtain an asymmetric map with principal coordinates for rows and standard coordinates for columns.

```
plot(cor_ana, map = "rowprincipal")
```



Part 2: Lab exericse

For the lab exercises, you will use the file

- `EcoActivity.txt`

This data contains a two-way contingency table that can be used to analyze economic activity of the Polish population in relation to gender and level of education in the second quarter of 2011. The rows of the table refer to different levels of education, that is:

1. tertiary (E1),
2. post-secondary (E2),
3. secondary (E3),
4. general secondary (E4),
5. basic vocational (E5),
6. lower secondary, primary and incomplete primary (E6).

The columns refer to the levels:

1. full-time employed females (A1F),

2. part-time employed females (A2F),
3. unemployed females (A3F),
4. economically inactive females (A4F),
5. full-time employed males (A1M),
6. part-time employed males (A2M),
7. unemployed males (A3M),
8. economically inactive males (A4M).

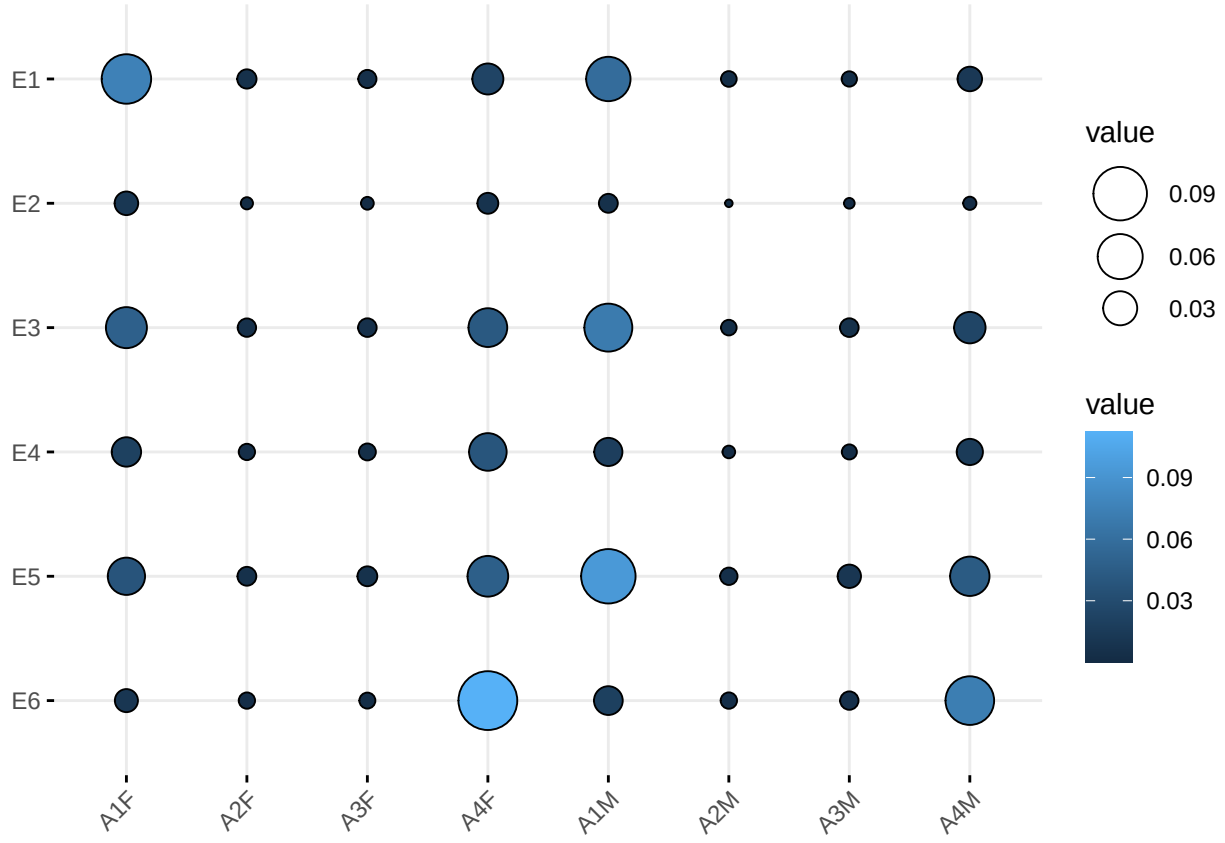
Import the data into R and respond to the following items.

```
EcoActivity <- read.table("data/EcoActivity.txt")
```

6. Give the rows 1 to 6 the labels E1 to E6, respectively. Give the columns 1 to 4 the labels A1F to A4F, and the columns 5 to 8 the labels A1M to A4M, respectively. Give a visualization of the correspondence matrix.

```
rownames(EcoActivity) <- c('E1', 'E2', 'E3', 'E4', 'E5', 'E6')
colnames(EcoActivity) <- c('A1F', 'A2F','A3F', 'A4F', 'A1M', 'A2M', 'A3M', 'A4M')

cor_mat <- prop.table(EcoActivity)
ggballoonplot(cor_mat, fill = 'value')
```



7. Give the proportion of full-time employed females with secondary level of education.

```r
cor_mat[3, 1]
```

```
## [1] 0.04787468
```

```r
cor_mat
```

```
##           A1F         A2F         A3F         A4F         A1M          A2M
## E1 0.07496704 0.006027500 0.004991524 0.023607710 0.057826333 0.0030137502
## E2 0.01117599 0.001130156 0.001381302 0.007785521 0.005682175 0.0002197526
## E3 0.04787468 0.005211276 0.005368243 0.041658818 0.069567401 0.0029509638
## E4 0.02046839 0.003484649 0.003955547 0.038613675 0.018239468 0.0013185157
## E5 0.03782884 0.005619388 0.006623972 0.046932881 0.093677403 0.0043322660
## E6 0.01045395 0.003484649 0.003233503 0.112042444 0.019338231 0.0035160419
##             A3M         A4M
## E1 0.0027626044 0.012463113
## E2 0.0006278646 0.001569662
## E3 0.0054624223 0.024549507
## E4 0.0026056382 0.015005965
## E5 0.0108306649 0.043291266
## E6 0.0052112764 0.072016073
```

8. Give the matrices of row profiles and column profiles.

```r
EcoActivity <- as.matrix(EcoActivity)

(row_pro <- prop.table(EcoActivity, margin = 1))
```

```
##           A1F        A2F        A3F       A4F        A1M         A2M        A3M
## E1 0.40378762 0.03246534 0.02688536 0.1271559 0.31146432 0.016232668 0.01487995
## E2 0.37791932 0.03821656 0.04670913 0.2632696 0.19214437 0.007430998 0.02123142
## E3 0.23625097 0.02571650 0.02649109 0.2055771 0.34329977 0.014562355 0.02695585
## E4 0.19739631 0.03360581 0.03814714 0.3723887 0.17590070 0.012715713 0.02512867
## E5 0.15183972 0.02255544 0.02658770 0.1883821 0.37600806 0.017389113 0.04347278
## E6 0.04559146 0.01519715 0.01410186 0.4886364 0.08433735 0.015334064 0.02272727
##           A4M
## E1 0.06712885
## E2 0.05307856
## E3 0.12114640
## E4 0.14471692
## E5 0.17376512
## E6 0.31407448
```

```r
rowSums(row_pro)
```

```
## E1 E2 E3 E4 E5 E6
##  1  1  1  1  1  1
```

```r
(col_pro <- prop.table(EcoActivity, margin = 2))
```

```
##           A1F        A2F        A3F        A4F        A1M        A2M        A3M
## E1 0.36971667 0.24150943 0.19533170 0.08722886 0.21876485 0.19631902 0.10045662
## E2 0.05511689 0.04528302 0.05405405 0.02876696 0.02149644 0.01431493 0.02283105
## E3 0.23610466 0.20880503 0.21007371 0.15392646 0.26318290 0.19222904 0.19863014
## E4 0.10094442 0.13962264 0.15479115 0.14267486 0.06900238 0.08588957 0.09474886
## E5 0.18656139 0.22515723 0.25921376 0.17341376 0.35439430 0.28220859 0.39383562
## E6 0.05155597 0.13962264 0.12653563 0.41398910 0.07315914 0.22903885 0.18949772
##           A4M
## E1 0.07379182
## E2 0.00929368
## E3 0.14535316
## E4 0.08884758
## E5 0.25631970
## E6 0.42639405
```

```
colSums(col_pro)
```

```
## A1F A2F A3F A4F A1M A2M A3M A4M
##   1   1   1   1   1   1   1   1
```

9. What is the conditional proportion of full-time employed females given tertiary level of education and what is the conditional proportion of full-time employed males given tertiary level of education?

```
row_pro[1, 1]
```

```
## [1] 0.4037876
```

```
row_pro[1, 5]
```

```
## [1] 0.3114643
```

10. What is the conditional proportion of females with the lowest level of education given economically inactive? What is the conditonal proportion of males with the lowest level of education given economically inactive?

```
col_pro[6, 4]
```

```
## [1] 0.4139891
```

```
col_pro[6, 8]
```

```
## [1] 0.4263941
```

11. Apply a correspondence analysis to the data. How large is the total inertia?

```
cor_ana.2 <- ca(EcoActivity)

# total inertia
sum(cor_ana.2$rowinertia)
```

```
## [1] 0.2449547
```

The total inertia is 0.24.

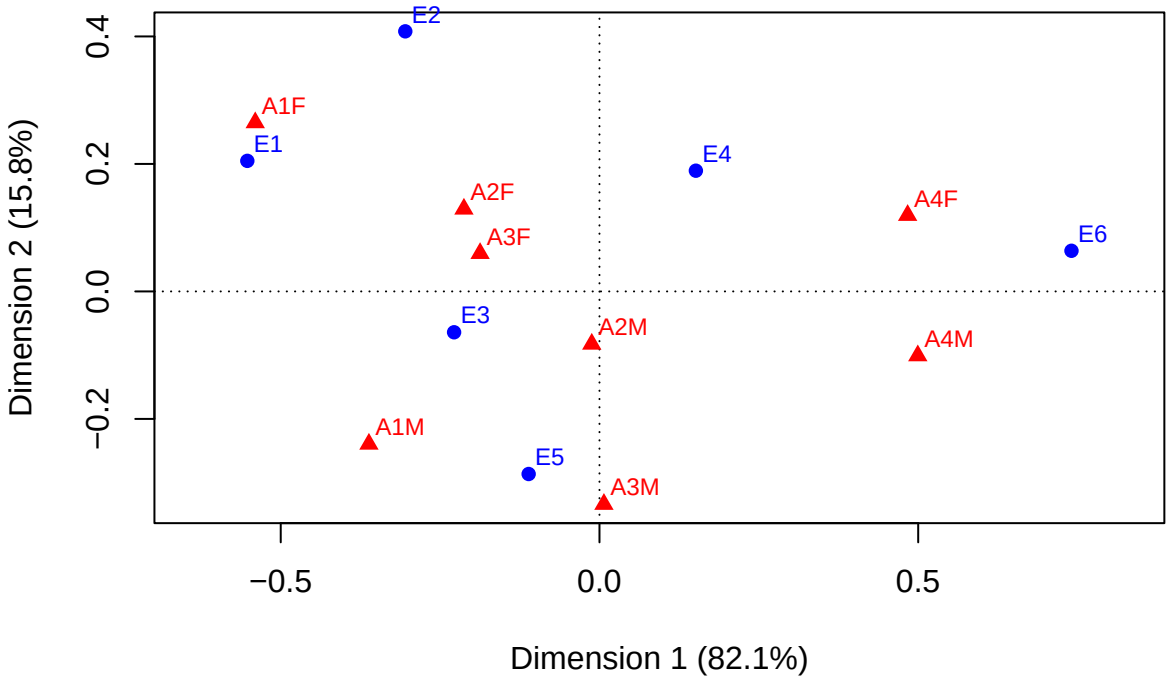
12. Set the desired minimum proportion of explained inertia to .85. How many underlying dimensions are sufficient? What is the proportion of inertia explained by this number of dimensions?

```
summary(cor_ana.2)
```

```
##
## Principal inertias (eigenvalues):
##
##  dim    value      %   cum%   scree plot
##  1      0.201099  82.1  82.1  **********
##  2      0.038632  15.8  97.9  ****
##  3      0.004543   1.9  99.7
##  4      0.000603   0.2 100.0
##  5      7.8e-050   0.0 100.0
##         -------- -----
##  Total: 0.244955 100.0
##
##
## Rows:
##     name   mass  qlt  inr    k=1 cor ctr    k=2 cor ctr
## 1 |   E1 |  186  983  268 | -552 864 282 |  205 119 202 |
## 2 |   E2 |   30  924   34 | -305 331  14 |  408 593 127 |
## 3 |   E3 |  203  961   48 | -228 891  52 |  -64  70  22 |
## 4 |   E4 |  104  722   34 |  151 281  12 |  189 441  96 |
## 5 |   E5 |  249  994   97 | -111 130  15 | -286 864 529 |
## 6 |   E6 |  229  996  519 |  740 989 625 |   64   7  24 |
##
## Columns:
##     name   mass  qlt  inr    k=1 cor ctr    k=2 cor ctr
## 1 |  A1F |  203  995  301 | -540 802 294 |  265 193 368 |
## 2 |  A2F |   25  824    8 | -213 601   6 |  129 222  11 |
## 3 |  A3F |   26  417   10 | -187 378   4 |   60  38   2 |
## 4 |  A4F |  271  991  277 |  483 934 315 |  119  57 100 |
## 5 |  A1M |  264  998  204 | -362 694 172 | -239 304 392 |
## 6 |  A2M |   15  423    1 |  -12   9   0 |  -83 414   3 |
## 7 |  A3M |   28  842   15 |    7   0   0 | -334 842  79 |
## 8 |  A4M |  169  964  186 |  499 926 209 | -101  38  44 |
```

13. Give the symmetric map for the final solution.

```
plot(cor_ana.2)
```


E2
A1F
E1
E4
A2F
A4F
A3F
E6
E3
A2M
A4M
A1M
E5
A3M
0.4
0.2
0.0
−0.2
Dimension 2 (15.8%)
−0.5
0.0
0.5
Dimension 1 (82.1%)