R instructions for the 8th seminar

In a data set *Juice.RData* there are results of market research whose purpose was to examine an effect of particular criteria leading to purchase of an particular juice brand. The respondents were surveyed which brand they usually buy and which of offered criteria mostly affect their brand choice.
a) Interpret associations among categories of the variable "Brand"; the same for the variable "Criteria"'.
b) Interpret associations among categories of the prime variable "Brand" and categories of the variables "Criteria" and "Education".

R Instructions for the problem 1:

•Get familiar with data (create useful contingency tables)

```
load("Juice.RData")
table(Juice[,1:2])
addmargins(table(Juice[,1:2]))
prop.table(table(Juice[,1:2]),1)
```
row profiles (points)
```
prop.table(table(Juice[,1:2]),2)
```
column profile (points)
```
addmargins(prop.table(table(Juice[,1:2])))
```

R Instructions for the problem 2:

•Perform Pearson χ^2 test.

```
ChiResults<-chisq.test(table(Juice[,1:2]))
ChiResults$observed
ChiResults$expected
ChiResults$residuals
```
standardized residuals
```
ChiResults
```
p-value=9.918e-05 proved at $\alpha = 0,05$ that "Brand" and "Criteria" are dependent.

R Instructions for the problem 3:

•Create an object in R bearing all essential results of Correspondence analysis.

```
!library(ca)
CaResults<-ca(table(Juice[,1:2]),nd=4)
```

R Instructions for the problem 4:

•Find out the total inertia ($I = V/n$); •What is the maximum dimension solution? •What percentage of total inertia is attributed to particular axes of new dimension system? (=eigenvalues/total innertia); •How many axes is enough to explain more then 90% of total inertia? - denote it as m. •Asses the scree plot.

```
CaResults
sum(summary(CaResults)$scree[,2])
summary(CaResults,scree=TRUE)
```
($I = 0.006207 + 0.040311 + 0.008041 + 0.008843 + 0.020140 = 0.083542$)

R Instructions for the problem 5:

•Describe results for raw profiles when $m = 4$.

```
.....................................................................
 Rows:
          Cappy     Hello    Rauch     Relax     Toma
Mass     0.190909  0.140000 0.174545  0.247273 0.247273
ChiDist  0.180317  0.536593 0.214639  0.189106 0.285393
Inertia  0.006207  0.040311 0.008041  0.008843 0.020140
Dim. 1   0.033971 -2.018189  0.813566 -0.478250 1.020392
```

```
Dim. 2  -1.077383  1.359917  0.081323 -1.021544 1.025990
Dim. 3   1.745140  0.267081 -0.727249 -1.122589 0.137376
Dim. 4  -0.175135 -0.386087 -1.879301  0.715287 0.965086
.....................................................................
```

*In a line `Mass` there are relative frequences of particular brands (the same was calculated in the first problem)

*In a line `Inertia` there are contributions of particular brands to total inertia. (Brand Hello is mostly responsible for the fact, that "Brand" and "Criterion" are depending. To get the relative contribution to particular brands divide the line Inertia by I=0,08354)

*In a line `ChiDist` there are row chi-square distances to the centroid (points which are far from centroid are responsible for dependence between variables).

*In lines `Dim.  1 - 4` there are the coordinates (standardized) of raw profiles in a new system of $m = 4$ axes.

R Instructions for the problem 6:

•Calculate the quality of representation of particular raw points when using only $m = 2$ "new axes" (principal coordinates). !For voluntiers!

This value "quality" is equal to sum of correlations between i-th row point with the first and the second "new axes" (principal coordinate). The correlation between i-th row point with first axes can be interpreted as proportion of i-th row point inertion explained by the first "new" axes (principal coordinate).

```
summary(CaResults)$rows
```

```
.....................................................................
rows
  name mass  qlt  inr  k=1 cor ctr  k=2 cor ctr  k=3 cor ctr  k=4 cor ctr
1 Cppy  191 1000   74    9   2   0 -125 477 222  130 520 581   -2   0   6
2 Hell  140 1000  483 -513 913 570  157  86 259   20   1  10   -4   0  21
3 Rach  175 1000   96  207 927 116    9   2   1  -54  64  92  -18   7 616
4 Relx  247 1000  106 -121 413  57 -118 390 258  -84 196 312    7   1 127
5 Toma  247 1000  241  259 825 257  119 173 260   10   1   5    9   1 230
.....................................................................
```

```
A<-summary(CaResults)$rows
A[,c(3,6,9)]
```

`(A[,6]+A[,9])/1000` ((cor when k=1 + cor when k=2)/qlt)

The results: Cappy 0.479; Hell 0.999; Rauch 0.929; Relax 0.803; Toma 0.998

It means that when using only two new axes, Cappy is represented only with 48%, but other brands are represented very well.

R Instructions for the problem 7:

•Represent graphicaly both raw and column points in a system of first two principal axes (which preserves more then 90% of total inertia)

`plot(CaResults, dim=c(1,2),map="symmetric")` a symetric [1] map for the 1st and the 2nd principal axes.

`plot(CaResults, dim=c(1,3),map="symmetric")` a map for the 1st and the 3rd principal axes.

[1] In a symmetric map both the row and column points are scaled to have inertias equal to the total inertia along the principal axes, that is both rows and columns are in pricipal coordinates. Other options are as follows: "rowprincipal" or "colprincipal" - these are the so-called asymmetric maps, with either rows in principal coordinates and columns in standard coordinates, or vice versa.