

Assignment2_Description

Neli Noykova

June 11, 2017

MULTIPLE CORRESPONDANCE ANALYSIS - Assignment 2

The data

As in the previous Assignment 1, here we again use **Finnish** sample from ISSP 2012 survey “Family and Changing Gender Roles IV”. Original data involve 1171 observations of 8 variables (4 substantive and 4 demographic). All variables are categorical.

The 4 **substantive** variables, which values are measured in 1-5 scale, are:

A: Married people are generally happier than unmarried people. **B:** People who want children ought to get married. **C:** It is all right for a couple to live together without intending to get married. **D:** Divorce is usually the best solution when a couple can't seem to work out their marriage problems.

The **demographic** variables are: **g:** gender (1=male, 2=female) **a:** age group (1=16-25, 2=26-35, 3=36-45, 4=46-55, 5=56-65, 6= 66+) **e:** education (1=Primary, 2=Comprehensive, primary and lower secondary, 3= Post-comprehensive, vocational school or course, 4=General upper secondary education or certificate, 5= Vocational post-secondary non-tertiary education, 6=Polytechnics , 7= University, lower academic degree, BA, 8=University, higher academic degree, MA **p:** Living in steady partnership (1=Yes, have partner; live in same household, 2=Yes, have partner; don't live in same household, 3=No partner)

Here we do not provide preliminary data treatment. We include all original variables, and also all missing data, denoted with number 9. During the data analyses we simply assign a new category, 9, to the missing data.

Graphical overview of the data and summaries of the variables

The original data, including missing values denoted by number 9, look as:

```
Finland2 <- read.table("FinlandWithMissing.txt")
head(Finland2)
```

```
##   A B C D g a e p
## 1 3 3 1 2 1 2 4 3
## 2 3 2 3 2 1 4 2 3
## 3 3 3 1 3 1 3 8 1
## 4 9 4 1 1 2 2 4 2
## 5 3 2 2 3 2 2 6 1
## 6 9 4 2 1 2 5 6 3
```

```
dim(Finland2)
```

```
## [1] 1171    8
```

```
str(Finland2)
```

```
## 'data.frame':    1171 obs. of  8 variables:
## $ A: int  3 3 3 9 3 9 2 1 3 3 ...
## $ B: int  3 2 3 4 2 4 2 1 3 3 ...
```

```
## $ C: int 1 3 1 1 2 2 2 1 1 1 ...
## $ D: int 2 2 3 1 3 1 3 9 3 2 ...
## $ g: int 1 1 1 2 2 2 2 2 1 2 ...
## $ a: int 2 4 3 2 2 5 4 2 3 1 ...
## $ e: int 4 2 8 4 6 6 5 5 7 4 ...
## $ p: int 3 3 1 2 1 3 1 3 3 3 ...
```

Task 1: Multiple correspondence analysis (MCA) on the four substantive questions. Comparing the results with the case of excluded missing data.

As it was noted during the lectures, since we analyze the data at the nominal level, a missing value is treated as an additional category (in the example here category 9), which is included during the subset analysis.

We show the results for both data sets (without and with missing data) using the default symmetric plot, where both coordinates are principle. We recall the main formulas from MCA (from lecture slides) in order to show the mathematical expression of these coordinates.

If we denote the data table (Finland.txt or FinlandWithMissing.txt) as \mathbf{N} , then the correspondence matrix \mathbf{P} is obtained as:

$$\mathbf{P} = \left(\frac{1}{n} \right) \mathbf{N}$$

If we denote the row and column marginal totals (masses) of \mathbf{P} as \mathbf{r} and \mathbf{c} respectively, and

$$\mathbf{D}_{\mathbf{r}}$$

and

$$\mathbf{D}_{\mathbf{c}}$$

are the diagonal matrices of these masses, then after applying singular value decomposition (SVD) we obtain:

$$\mathbf{S} = \mathbf{D}_{\mathbf{r}}^{(-1/2)} (\mathbf{P} - \mathbf{r}\mathbf{c}^T) \mathbf{D}_{\mathbf{c}}^{(-1/2)}$$

which is equivalent to

$$\mathbf{S} = \mathbf{D}_{\mathbf{r}}^{(1/2)} \left(\mathbf{D}_{\mathbf{r}}^{(-1)} \mathbf{P} \mathbf{D}_{\mathbf{c}}^{(-1)} - \mathbf{1}\mathbf{1}^T \right) \mathbf{D}_{\mathbf{c}}^{(1/2)}$$

According SVD

$$\mathbf{S} = \mathbf{U} \mathbf{D}_{\alpha} \mathbf{V}^T$$

Then the principal coordinates are presented as: - for rows:

$$\mathbf{F} = \mathbf{D}_{\mathbf{r}}^{(-1/2)} \mathbf{U} \mathbf{D}_{\alpha}$$

- for columns:

$$\mathbf{G} = \mathbf{D}_{\mathbf{c}}^{(-1/2)} \mathbf{V} \mathbf{D}_{\alpha}$$

The total variance (inertia) is the sum of squares of the elements of

$$\mathbf{S} = \text{trace}(\mathbf{S}\mathbf{S}^T) = \chi^2/n$$

The Chi-square distance

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(fo_{ij} - fe_{ij})^2}{fe_{ij}}$$

In this expression the observed and expected frequencies of the cell in row i and column j are denoted by

$$fo_{ij}$$

and

$$fe_{ij}$$

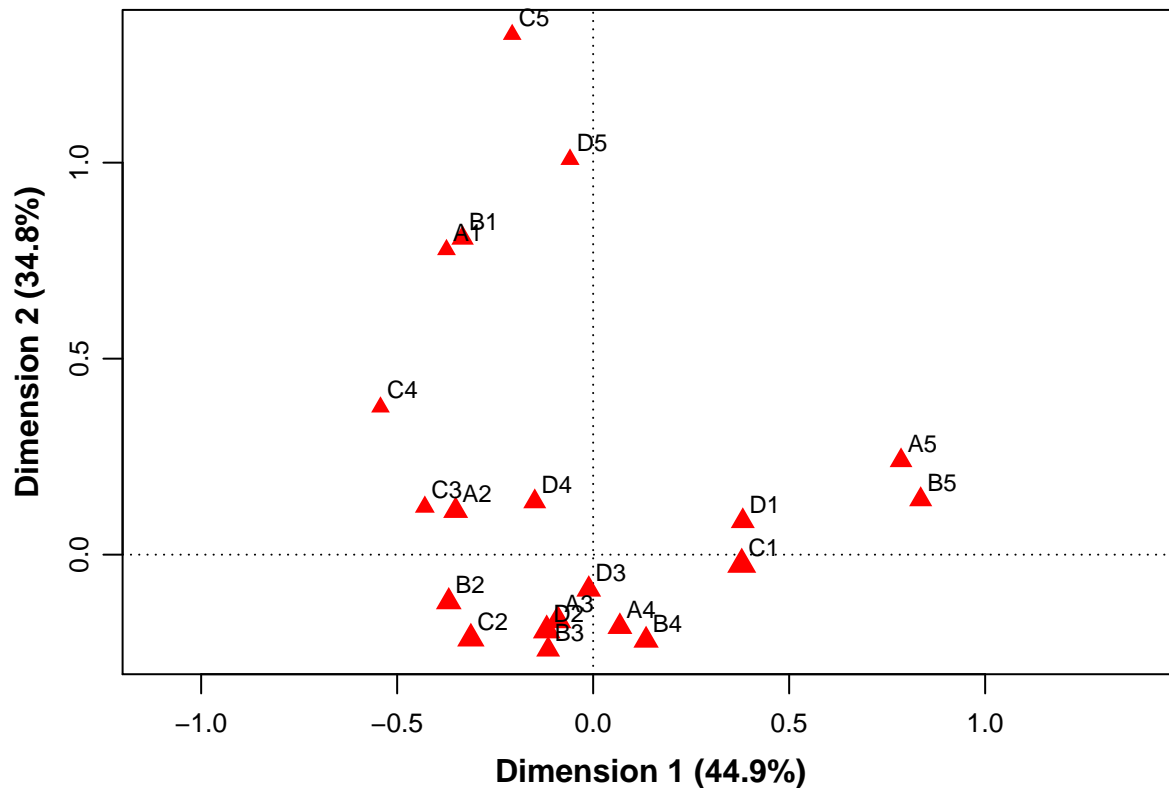
respectively.

We first provide MCA on data Finland.txt, where missed data are excluded and draw the default symmetric plot.

```
require(ca)
```

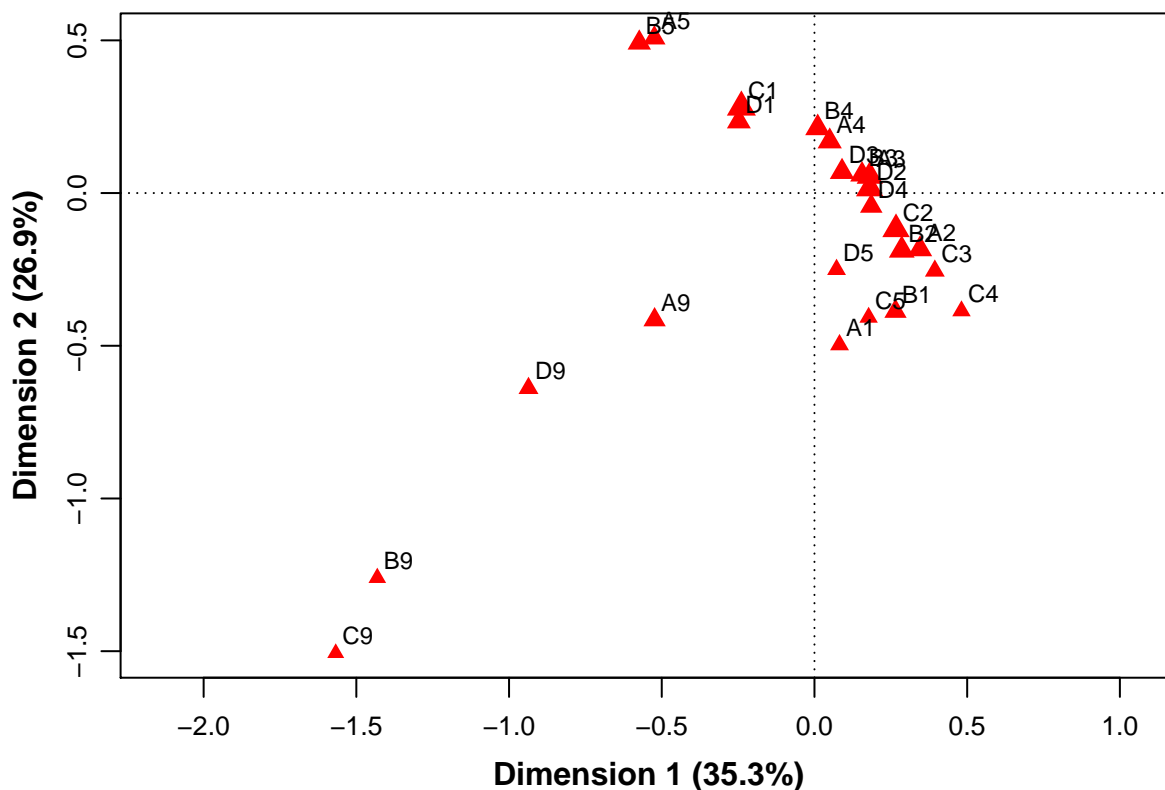
```
## Loading required package: ca
```

```
Finland <- read.table("Finland.txt")
par(mar=c(4.2,4,1,1), mgp=c(2,0.7,0), cex.axis=0.8, font.lab=2, mfrow=c(1,1))
plot(mjca(Finland[,1:4], ps=""), mass=c(F,T))
```



Next perform the same MCA analysis and plot, but using also missed data as separate category = 9:

```
par(mar=c(4.2,4,1,1), mgp=c(2,0.7,0), cex.axis=0.8, font.lab=2, mfrow=c(1,1))
plot(mjca(Finland2[,1:4], ps=""), mass=c(F,T))
```



We observe that the missing category 9 for all 4 substantive questions is situated on the right down part of the biplot, where both principal coordinates take negative values.

The symmetric plot represents the row and column profiles simultaneously in a common space. According Mike Bendixen (Marketing Bulletin, 2003, 14) this plot may lead to misinterpretation if examined in isolation or only visually because principal coordinates are presented for both rows and columns. These coordinates represent the row and column profiles and not the apexes for which the standard coordinates are required. Therefore in such cases asymmetric plots are recommended.

Task 2: Burt matrix. Performing CA using a subset of non-missing rows and columns. Confidence regions of the demographic groups.

The part of Burt matrix on first 4 substantive questions can be obtained as:

```
require(ca)
Finland2.B <- mjca(Finland2[,1:4])$Burt
```

Since the missing values are involved in this matrix as separate category, we need to take the following part of Burt matrix:

```
Finland2.ABCD = Finland2.B[c(1:24), c(1:24)]
Finland2.ABCD
```

```
##      A:1 A:2 A:3 A:4 A:5 A:9 B:1 B:2 B:3 B:4 B:5 B:9 C:1 C:2 C:3 C:4 C:5
## A:1  59   0   0   0   0   0  32  14   5   4   1   3  21  15   7   7   6
## A:2   0 233   0   0   0   0  53  99  38  27  12   4  69 109  20  17  18
```

```

## A:3  0  0 344  0  0  0 28 100 105 87 22  2 153 134 32 14  7
## A:4  0  0  0 231  0  0 16 47 30 102 33  3 112 93 11 10  3
## A:5  0  0  0  0 152  0  9 12  7 27 96  1 121 17  1  0 12
## A:9  0  0  0  0  0 152 12 41 23 23 28 25 69 49  8  3  4
## B:1 32 53 28 16  9 12 150  0  0  0  0  0 29 41 21 18 38
## B:2 14 99 100 47 12 41  0 313  0  0  0  0 66 175 34 29  4
## B:3  5 38 105 30  7 23  0  0 208  0  0  0 98 88 19  1  1
## B:4  4 27 87 102 27 23  0  0  0 270  0  0 161 99  4  3  1
## B:5  1 12 22 33 96 28  0  0  0  0 192  0 178  6  0  0  6
## B:9  3  4  2  3  1 25  0  0  0  0  0 38 13  8  1  0  0
## C:1 21 69 153 112 121 69 29 66 98 161 178 13 545  0  0  0  0
## C:2 15 109 134 93 17 49 41 175 88 99  6  8  0 417  0  0  0
## C:3  7 20 32 11  1  8 21 34 19  4  0  1  0  0 79  0  0
## C:4  7 17 14 10  0  3 18 29  1  3  0  0  0  0  0 51  0
## C:5  6 18  7  3 12  4 38  4  1  1  6  0  0  0  0  0 50
## C:9  3  0  4  2  1 19  3  5  1  2  2 16  0  0  0  0  0
## D:1 15 31 58 45 41 26 30 39 34 46 58  9 154 43  8  4  4
## D:2 17 97 121 86 44 47 32 146 82 104 45  3 158 208 27 13  3
## D:3  6 49 96 41 27 25 24 61 55 59 41  4 110 83 28 13  7
## D:4  7 37 43 44 12 16 33 39 26 41 19  1 65 55  9 13 16
## D:5  7 13 14  5 13  6 25 11  3  9  9  1 23  4  6  7 17
## D:9  7  6 12 10 15 32  6 17  8 11 20 20 35 24  1  1  3
##      C:9 D:1 D:2 D:3 D:4 D:5 D:9
## A:1  3 15 17  6  7  7  7
## A:2  0 31 97 49 37 13  6
## A:3  4 58 121 96 43 14 12
## A:4  2 45 86 41 44  5 10
## A:5  1 41 44 27 12 13 15
## A:9 19 26 47 25 16  6 32
## B:1  3 30 32 24 33 25  6
## B:2  5 39 146 61 39 11 17
## B:3  1 34 82 55 26  3  8
## B:4  2 46 104 59 41  9 11
## B:5  2 58 45 41 19  9 20
## B:9 16  9  3  4  1  1 20
## C:1  0 154 158 110 65 23 35
## C:2  0 43 208 83 55  4 24
## C:3  0  8 27 28  9  6  1
## C:4  0  4 13 13 13  7  1
## C:5  0  4  3  7 16 17  3
## C:9 29  3  3  3  1  1 18
## D:1  3 216  0  0  0  0  0
## D:2  3  0 412  0  0  0  0
## D:3  3  0  0 244  0  0  0
## D:4  1  0  0  0 159  0  0
## D:5  1  0  0  0  0 58  0
## D:9 18  0  0  0  0  0 82

```

```
summary(Finland2.ABCD)
```

```

##      A:1      A:2      A:3      A:4
## Min.   : 0.000   Min.   : 0.00   Min.   : 0.00   Min.   : 0.00
## 1st Qu.: 2.500   1st Qu.: 3.00   1st Qu.: 3.50   1st Qu.: 2.75
## Median : 6.500   Median : 19.00  Median : 25.00  Median : 13.50
## Mean   : 9.833   Mean    : 38.83   Mean    : 57.33   Mean    : 38.50

```

```
## 3rd Qu.:14.250 3rd Qu.: 50.00 3rd Qu.: 97.00 3rd Qu.: 45.50
## Max. :59.000 Max. :233.00 Max. :344.00 Max. :231.00
## A:5 A:9 B:1 B:2
## Min. : 0.00 Min. : 0.00 Min. : 0.00 Min. : 0.00
## 1st Qu.: 0.75 1st Qu.: 3.75 1st Qu.: 5.25 1st Qu.: 4.75
## Median : 12.00 Median : 21.00 Median : 22.50 Median : 31.50
## Mean : 25.33 Mean : 25.33 Mean : 25.00 Mean : 52.17
## 3rd Qu.: 27.00 3rd Qu.: 29.00 3rd Qu.: 32.00 3rd Qu.: 62.25
## Max. :152.00 Max. :152.00 Max. :150.00 Max. :313.00
## B:3 B:4 B:5 B:9
## Min. : 0.00 Min. : 0.00 Min. : 0.0 Min. : 0.000
## 1st Qu.: 1.00 1st Qu.: 1.75 1st Qu.: 0.0 1st Qu.: 0.000
## Median : 13.50 Median : 17.00 Median : 10.5 Median : 2.500
## Mean : 34.67 Mean : 45.00 Mean : 32.0 Mean : 6.333
## 3rd Qu.: 42.25 3rd Qu.: 66.00 3rd Qu.: 35.0 3rd Qu.: 8.250
## Max. :208.00 Max. :270.00 Max. :192.0 Max. :38.000
## C:1 C:2 C:3 C:4
## Min. : 0.00 Min. : 0.0 Min. : 0.00 Min. : 0.0
## 1st Qu.: 19.00 1st Qu.: 5.5 1st Qu.: 0.75 1st Qu.: 0.0
## Median : 67.50 Median : 42.0 Median : 7.50 Median : 3.5
## Mean : 90.83 Mean : 69.5 Mean :13.17 Mean : 8.5
## 3rd Qu.:129.00 3rd Qu.: 94.5 3rd Qu.:20.25 3rd Qu.:13.0
## Max. :545.00 Max. :417.0 Max. :79.00 Max. :51.0
## C:5 C:9 D:1 D:2
## Min. : 0.000 Min. : 0.000 Min. : 0.00 Min. : 0.00
## 1st Qu.: 0.750 1st Qu.: 0.750 1st Qu.: 3.75 1st Qu.: 3.00
## Median : 4.000 Median : 2.000 Median : 28.00 Median : 38.00
## Mean : 8.333 Mean : 4.833 Mean : 36.00 Mean : 68.67
## 3rd Qu.: 8.250 3rd Qu.: 3.250 3rd Qu.: 43.50 3rd Qu.: 98.75
## Max. :50.000 Max. :29.000 Max. :216.00 Max. :412.00
## D:3 D:4 D:5 D:9
## Min. : 0.00 Min. : 0.0 Min. : 0.000 Min. : 0.00
## 1st Qu.: 3.75 1st Qu.: 1.0 1st Qu.: 1.000 1st Qu.: 1.00
## Median : 26.00 Median : 16.0 Median : 6.500 Median : 9.00
## Mean : 40.67 Mean : 26.5 Mean : 9.667 Mean :13.67
## 3rd Qu.: 56.00 3rd Qu.: 39.5 3rd Qu.:13.000 3rd Qu.:18.50
## Max. :244.00 Max. :159.0 Max. :58.000 Max. :82.00
```

For our work we need also to calculate the indicator matrix, obtained also via the MCA function `mjca`:

```
Finland2.Z <- mjca(Finland2[,1:4], ps="", reti=T)$indmat
head(Finland2.Z)
```

```
## A1 A2 A3 A4 A5 A9 B1 B2 B3 B4 B5 B9 C1 C2 C3 C4 C5 C9 D1 D2 D3 D4 D5 D9
## 1 0 0 1 0 0 0 0 0 1 0 0 0 1 0 0 0 0 0 1 0 0 0 0
## 2 0 0 1 0 0 0 0 1 0 0 0 0 0 1 0 0 0 0 1 0 0 0 0
## 3 0 0 1 0 0 0 0 0 0 1 0 0 0 1 0 0 0 0 0 1 0 0 0
## 4 0 0 0 0 0 1 0 0 0 1 0 0 1 0 0 0 0 0 1 0 0 0 0
## 5 0 0 1 0 0 0 0 1 0 0 0 0 0 1 0 0 0 0 0 1 0 0 0
## 6 0 0 0 0 0 1 0 0 0 1 0 0 0 1 0 0 0 0 1 0 0 0 0
```

```
dim(Finland2.Z)
```

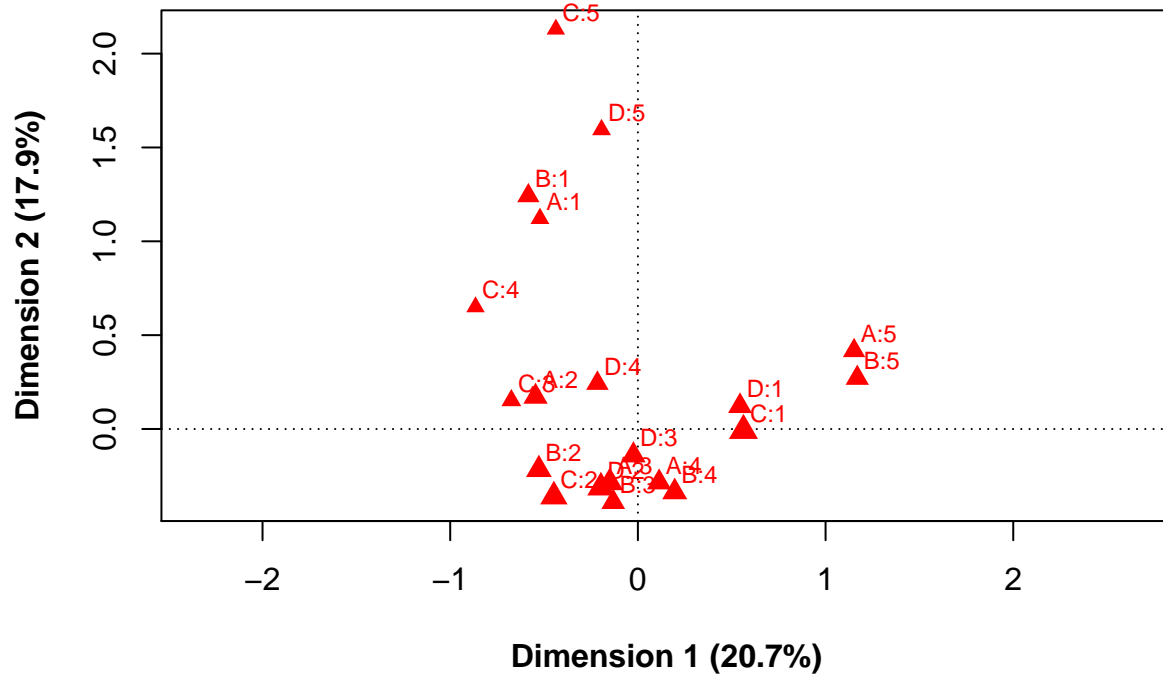
```
## [1] 1171 24
```

The relationship between both matrices is:

$$\mathbf{B} = \mathbf{Z}^T \mathbf{Z}$$

The non-missing categories are obtained by excluding the missing ones. Further analysis is provided using subset CA:

```
Finland2.nonmissing <- -c(6,12,18,24)
Finland2.mca1 <- ca(Finland2.B, subsetrow=c(Finland2.nonmissing), subsetcol=c(Finland2.nonmissing))
plot(Finland2.mca1, what=c("none", "all"), mass=c(F,T), font.lab=2)
```



```
summary(Finland2.mca1)
```

```
##
## Principal inertias (eigenvalues):
##
## dim    value      %    cum%    scree plot
## 1      0.243700  20.7  20.7    *****
## 2      0.210498  17.9  38.6    ****
## 3      0.103194   8.8  47.4    **
## 4      0.095390   8.1  55.5    **
## 5      0.075342   6.4  61.9    **
## 6      0.070659   6.0  67.9    **
## 7      0.062970   5.4  73.3    *
## 8      0.057778   4.9  78.2    *
## 9      0.049824   4.2  82.4    *
## 10     0.047380   4.0  86.5    *
## 11     0.043140   3.7  90.1    *
## 12     0.036612   3.1  93.2    *
## 13     0.029151   2.5  95.7    *
## 14     0.022491   1.9  97.6
```

```

## 15      0.017302   1.5 99.1
## 16      0.008924   0.8 99.9
## 17      0.001268   0.1 100.0
## 18      0.000285   0.0 100.0
## 19      5.3e-050   0.0 100.0
## 20      1.3e-050   0.0 100.0
## -----
## Total: 1.175975 100.0
##
##
## Rows:
##      name  mass  qlt  inr   k=1 cor ctr   k=2 cor ctr
## 1 |  A1 |  13 293  56 | -522 52 14 | 1121 241 75 |
## 2 |  A2 |  50 299  46 | -545 271 61 | 172 27 7 |
## 3 |  A3 |  73 166  39 | -149 36 7 | -285 130 28 |
## 4 |  A4 |  49 88  45 | 113 12 3 | -284 76 19 |
## 5 |  A5 |  32 680  61 | 1152 602 177 | 417 79 27 |
## 6 |  B1 |  32 805  64 | -584 145 45 | 1243 660 235 |
## 7 |  B2 |  67 397  47 | -528 340 76 | -216 57 15 |
## 8 |  B3 |  44 136  47 | -132 14 3 | -390 122 32 |
## 9 |  B4 |  58 163  45 | 196 42 9 | -336 122 31 |
## 10 | B5 |  41 782  64 | 1169 742 230 | 271 40 14 |
## 11 | C1 | 116 797  39 | 563 796 151 | -8 0 0 |
## 12 | C2 | 89 588  42 | -448 358 73 | -360 231 55 |
## 13 | C3 | 17 130  53 | -674 123 31 | 152 6 2 |
## 14 | C4 | 11 197  55 | -865 126 33 | 650 71 22 |
## 15 | C5 | 11 658  65 | -437 27 8 | 2130 631 230 |
## 16 | D1 | 46 262  47 | 544 250 56 | 122 13 3 |
## 17 | D2 | 88 277  37 | -197 79 14 | -312 198 41 |
## 18 | D3 | 52 21  43 | -24 1 0 | -140 20 5 |
## 19 | D4 | 34 65  47 | -216 28 6 | 243 36 10 |
## 20 | D5 | 12 467  58 | -194 7 2 | 1595 461 150 |
##
## Columns:
##      name  mass  qlt  inr   k=1 cor ctr   k=2 cor ctr
## 1 |  A1 |  13 293  56 | -522 52 14 | 1121 241 75 |
## 2 |  A2 |  50 299  46 | -545 271 61 | 172 27 7 |
## 3 |  A3 |  73 166  39 | -149 36 7 | -285 130 28 |
## 4 |  A4 |  49 88  45 | 113 12 3 | -284 76 19 |
## 5 |  A5 |  32 680  61 | 1152 602 177 | 417 79 27 |
## 6 |  B1 |  32 805  64 | -584 145 45 | 1243 660 235 |
## 7 |  B2 |  67 397  47 | -528 340 76 | -216 57 15 |
## 8 |  B3 |  44 136  47 | -132 14 3 | -390 122 32 |
## 9 |  B4 |  58 163  45 | 196 42 9 | -336 122 31 |
## 10 | B5 |  41 782  64 | 1169 742 230 | 271 40 14 |
## 11 | C1 | 116 797  39 | 563 796 151 | -8 0 0 |
## 12 | C2 | 89 588  42 | -448 358 73 | -360 231 55 |
## 13 | C3 | 17 130  53 | -674 123 31 | 152 6 2 |
## 14 | C4 | 11 197  55 | -865 126 33 | 650 71 22 |
## 15 | C5 | 11 658  65 | -437 27 8 | 2130 631 230 |
## 16 | D1 | 46 262  47 | 544 250 56 | 122 13 3 |
## 17 | D2 | 88 277  37 | -197 79 14 | -312 198 41 |
## 18 | D3 | 52 21  43 | -24 1 0 | -140 20 5 |
## 19 | D4 | 34 65  47 | -216 28 6 | 243 36 10 |

```

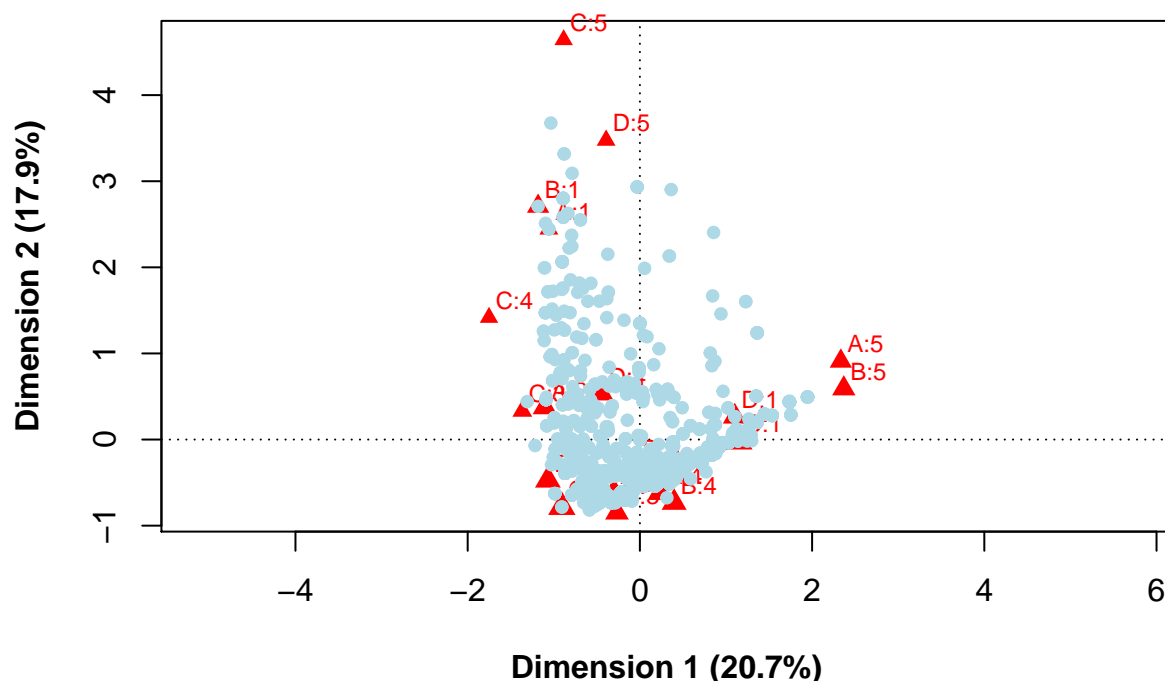


```
## 20 | D5 | 12 467 58 | -194 7 2 | 1595 461 150 |
```

This plot is quite similar to the plot, obtained when missing data are removed from the data set.

Next we add responding points to the plot. For this purpose we first transform the corresponding frequencies in the table and draw a sequence of points at the specified coordinates:

```
Finland2.sum <- apply(Finland2.Z[,c(Finland2.nonmissing)], 1, sum)
Finland2.sum[Finland2.sum==0] <- 1
Finland2.rpc <- Finland2.Z[,c(Finland2.nonmissing)] %*% Finland2.mca1$colcoord / Finland2.sum
plot(Finland2.mca1, what=c("none", "all"), mass=c(F,T), font.lab=2, map="rowprincipal")
points(Finland2.rpc, pch=19, col="lightblue", cex=0.8)
```



Next we compute confidence regions of demographic groups. For this purpose we have to add group ellipses and use Prof. Greenacre's program "confidenceplots", which was given separately.

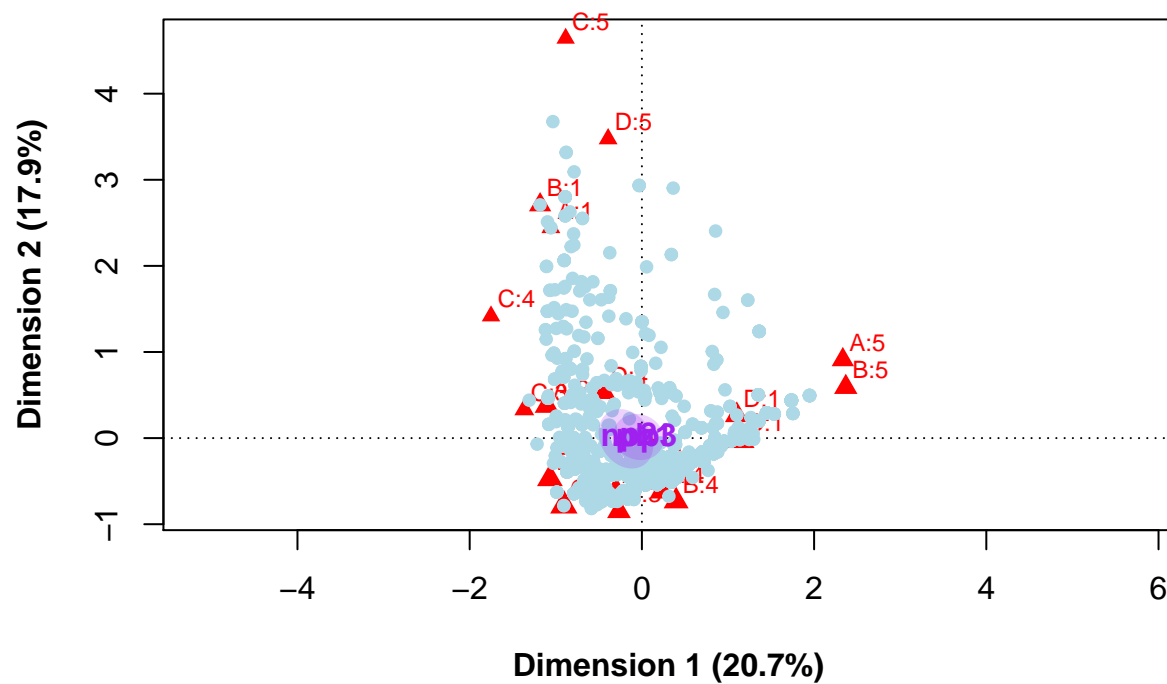
```
require(ellipse)
```

```
## Loading required package: ellipse
```

```
source("confidenceplots.R")
```

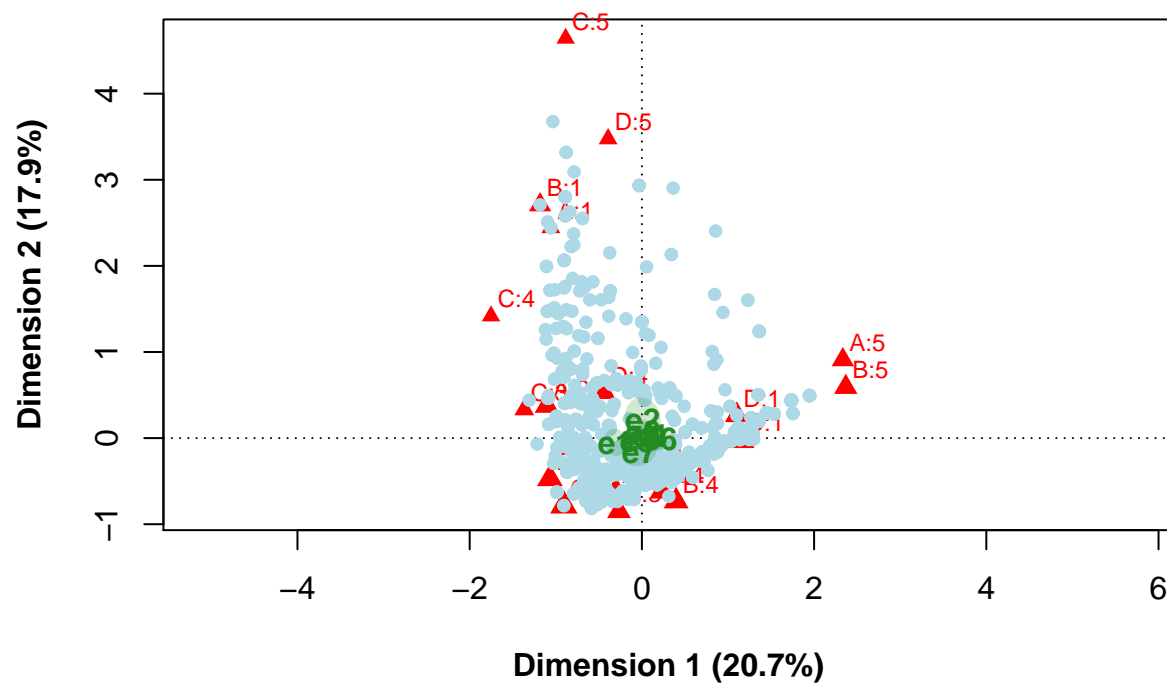
First we draw **confidence plots** for the demographic group **partnership** by generation asymmetric plot, adding points and finally draw confidence plots.

```
plot(Finland2.mca1, what=c("none", "all"), mass=c(F,T), font.lab=2, map="rowprincipal")
points(Finland2.rpc, pch=19, col="lightblue", cex=0.8)
confidenceplots(Finland2.rpc[Finland2$p<4,1], Finland2.rpc[Finland2$p<4,2], group=Finland2$p[Finland2$e
groupnames=c("ph1", "p2", "nop3"), shownames=T, add=T)
```



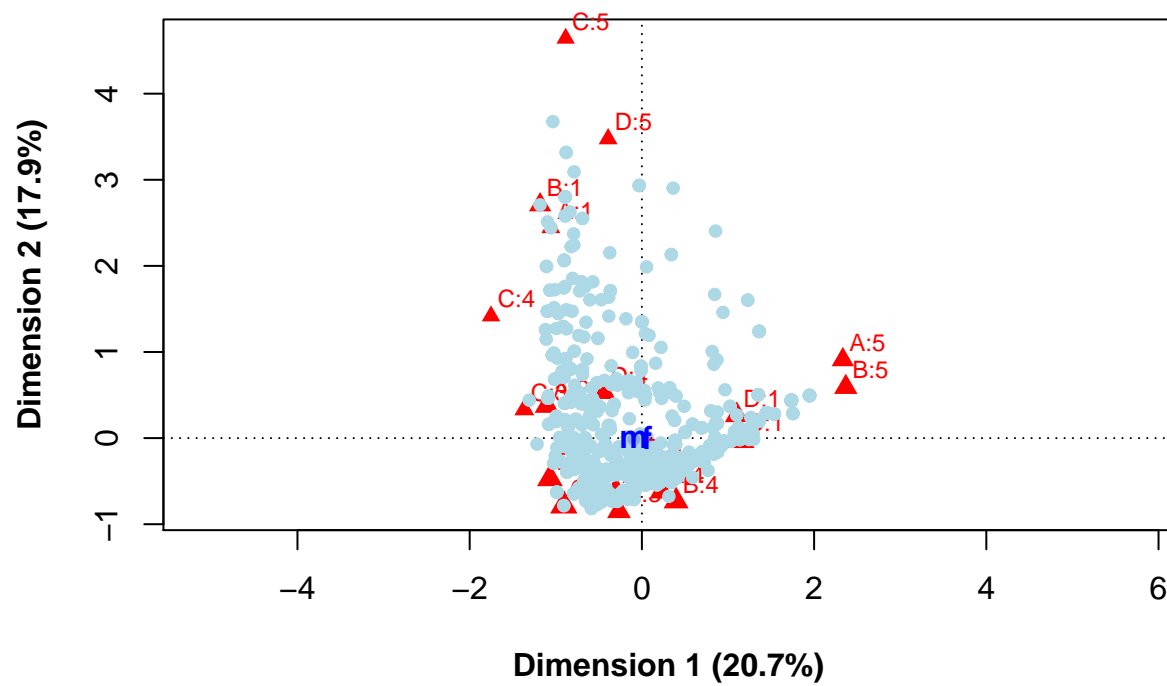
Next we draw confidence plot for the demographic group **education**.

```
plot(Finland2.mca1, what=c("none", "all"), mass=c(F,T), font.lab=2, map="rowprincipal")
points(Finland2.rpc, pch=19, col="lightblue", cex=0.8)
confidenceplots(Finland2.rpc[Finland2$e<9,1], Finland2.rpc[Finland2$e<9,2], group=Finland2$e[Finland2$e
groupnames=c("e1","e2","e3","e4","e5","e6","e7","e8"), shownames=T, add=T)
```



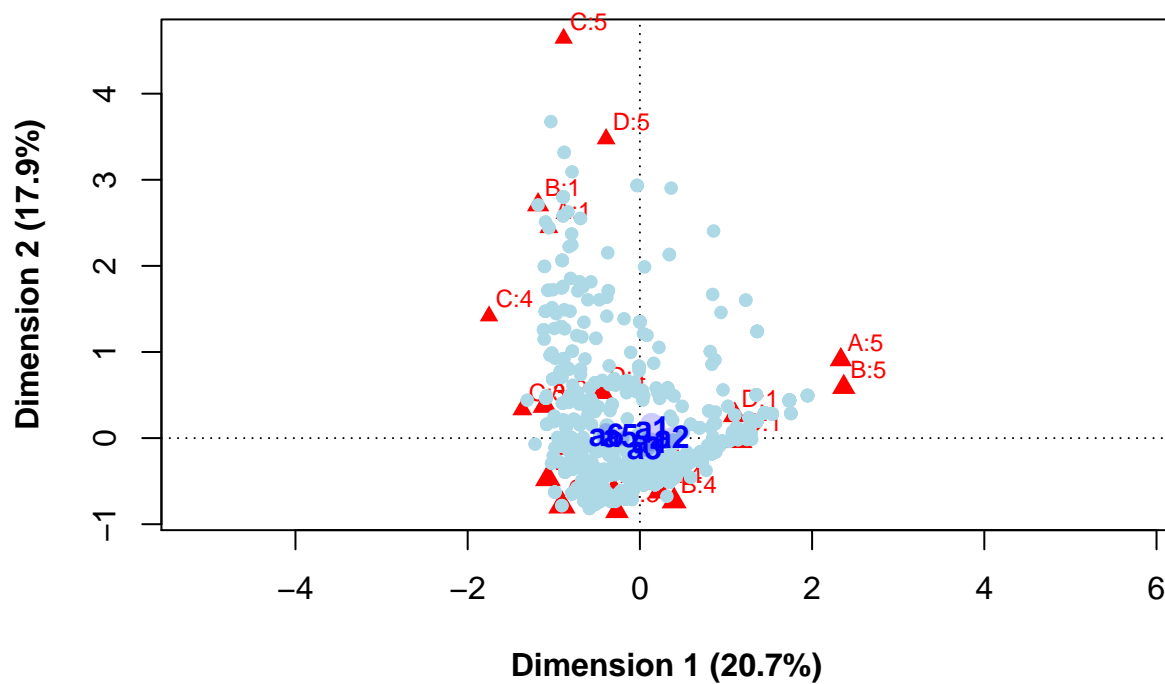
The confidence plot for the demographic group **gender** is:

```
plot(Finland2.mca1, what=c("none", "all"), mass=c(F,T), font.lab=2, map="rowprincipal")
points(Finland2.rpc, pch=19, col="lightblue", cex=0.8)
confidenceplots(Finland2.rpc[Finland2$g<3,1], Finland2.rpc[Finland2$g<3,2], group=Finland2$g[Finland2$g<3],
groupnames=c("m","f"), shownames=T, add=T)
```



The confidence plot for the demographic group **age** is:

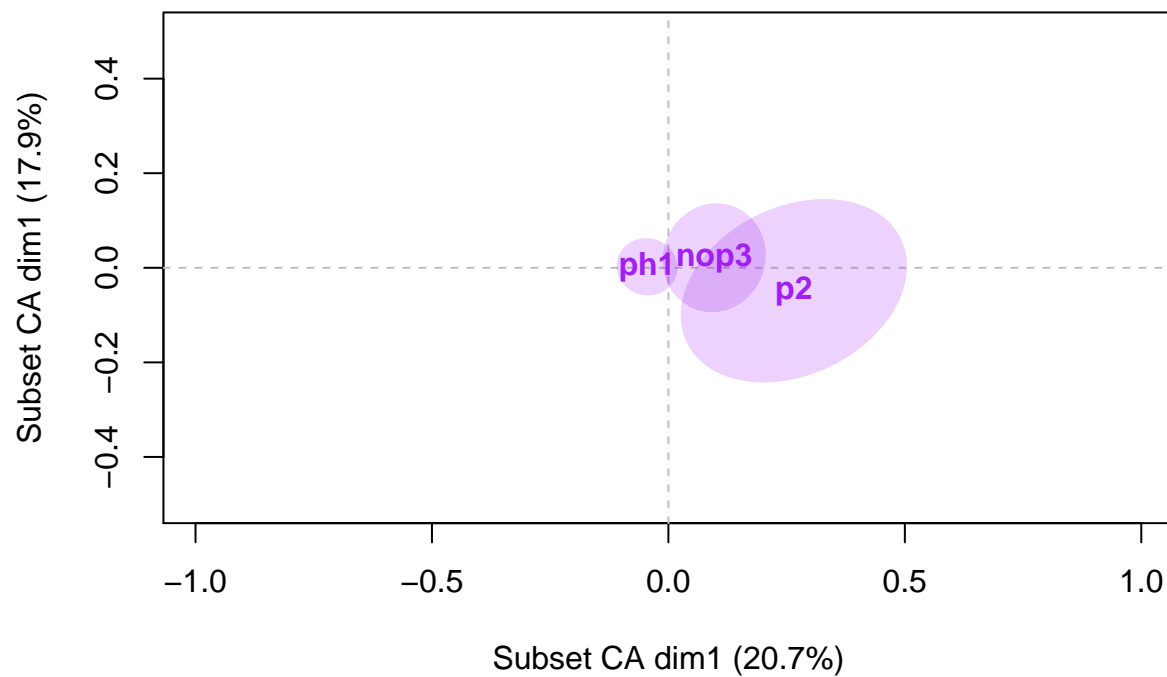
```
plot(Finland2.mca1, what=c("none", "all"), mass=c(F,T), font.lab=2, map="rowprincipal")
points(Finland2.rpc, pch=19, col="lightblue", cex=0.8)
confidenceplots(Finland2.rpc[Finland2$a<7,1], Finland2.rpc[Finland2$a<7,2], group=Finland2$a[Finland2$a
groupnames=c("a1","a2","a3","a4","a5","a6"), shownames=T, add=T)
```



If we plot **only the confidence plots (just the ellipses)** for every demographic group, the plots will look as follow:

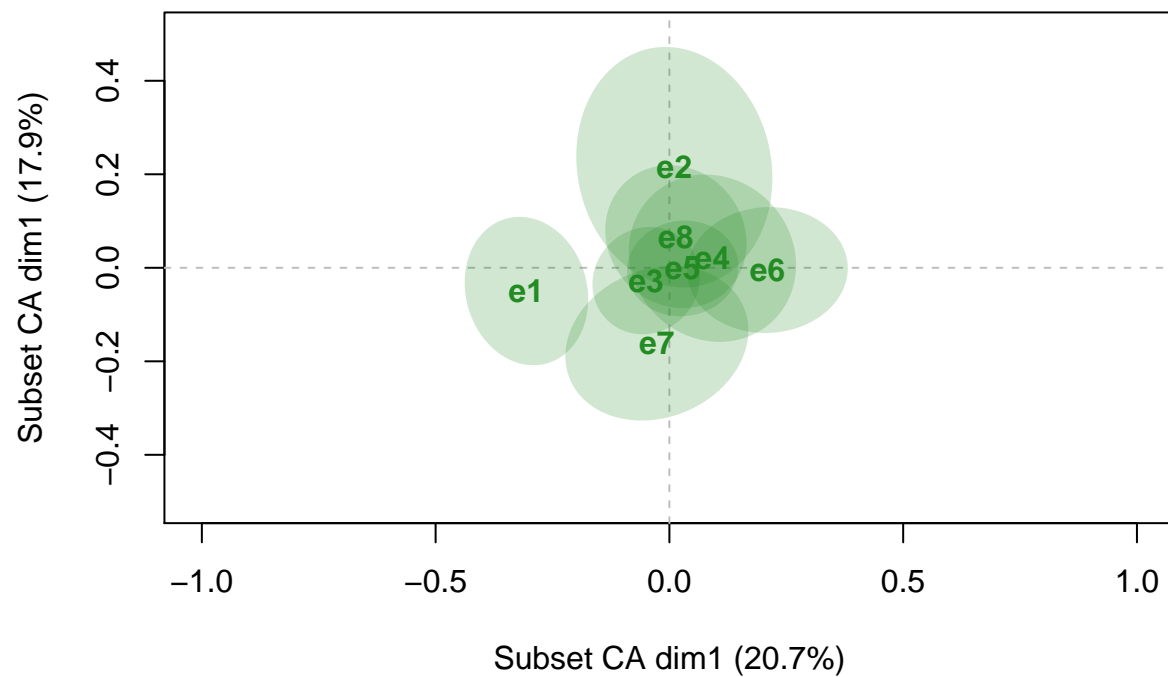
- for the demographic group **partnership**:

```
plot(Finland2.rpc, type="n", asp=1, xlab="Subset CA dim1 (20.7%)", ylab="Subset CA dim1 (17.9%)", xlim=-4, ylim=-1)
abline(v=0, h=0, lty=2, col="grey")
confidenceplots(Finland2.rpc[Finland2$p<4,1], Finland2.rpc[Finland2$p<4,2], group=Finland2$p[Finland2$p<4,1],
               groupnames=c("ph1", "p2", "nop3"), shownames=T, add=T)
```



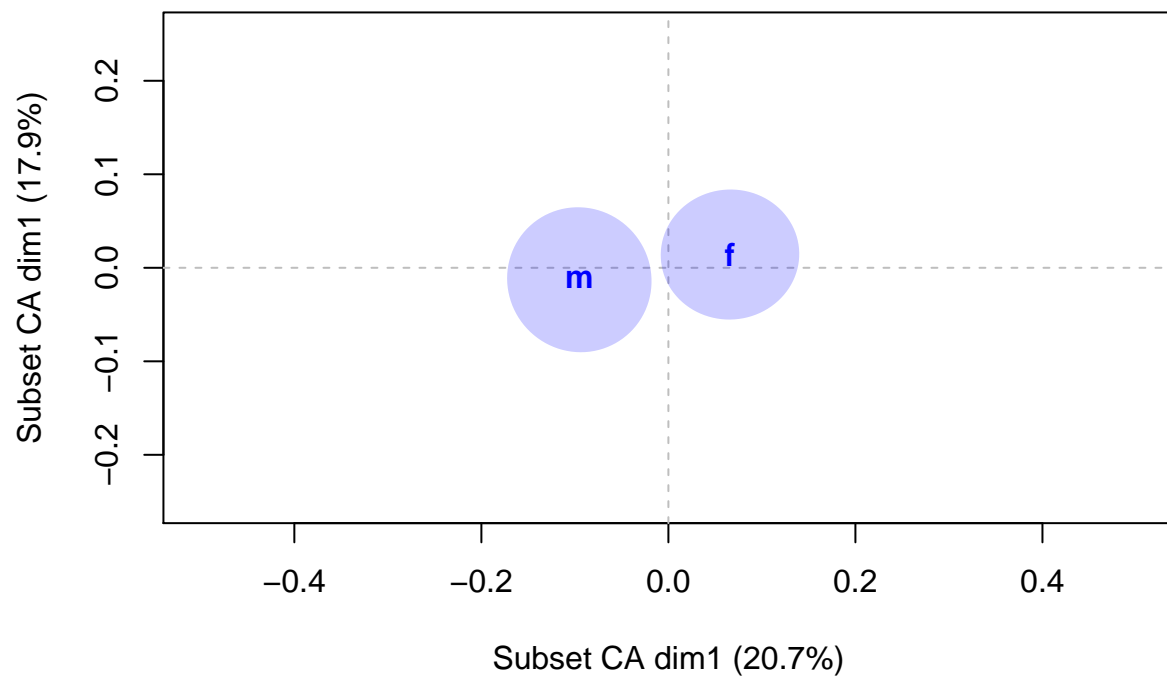
- for the demographic group **education**:

```
plot(Finland2.rpc, type="n", asp=1, xlab="Subset CA dim1 (20.7%)", ylab="Subset CA dim1 (17.9%)", xlim=
abline(v=0, h=0, lty=2, col="grey")
confidenceplots(Finland2.rpc[Finland2$e<9,1], Finland2.rpc[Finland2$e<9,2], group=Finland2$e[Finland2$e
groupnames=c("e1", "e2", "e3", "e4", "e5", "e6", "e7", "e8"), shownames=T, add=T)
```



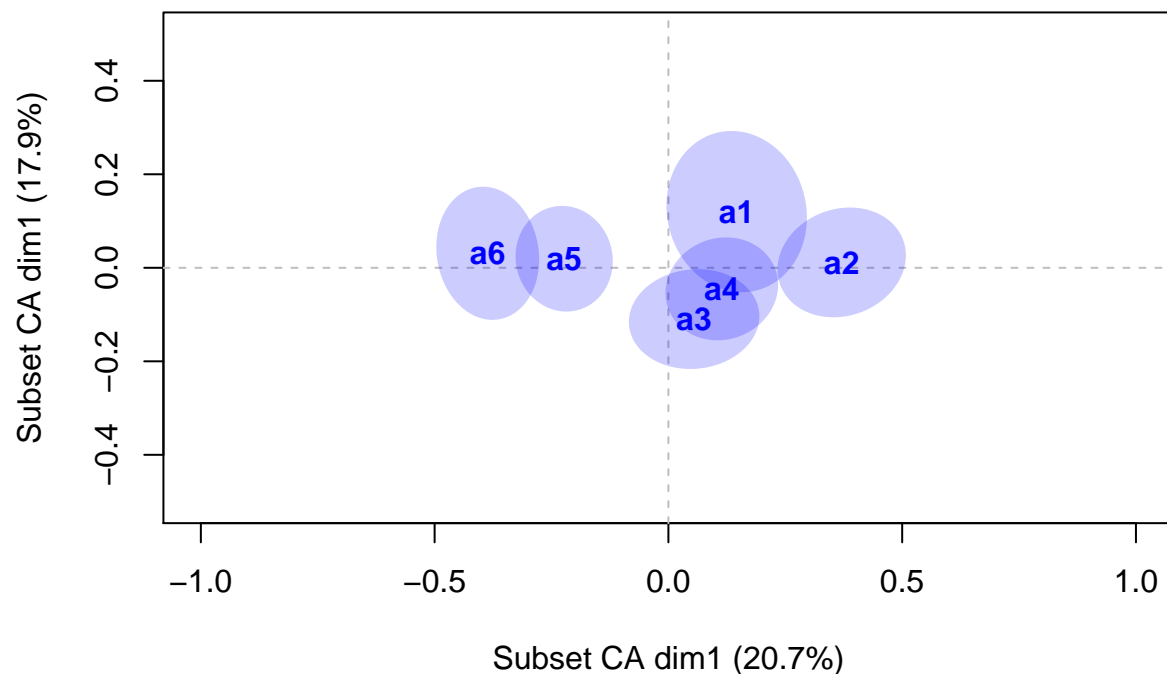
- for the demographic group **gender**:

```
plot(Finland2.rpc, type="n", asp=1, xlab="Subset CA dim1 (20.7%)", ylab="Subset CA dim1 (17.9%)", xlim=-1, ylim=-0.4,
abline(v=0, h=0, lty=2, col="grey")
confidenceplots(Finland2.rpc[Finland2$g<3,1], Finland2.rpc[Finland2$g<3,2], group=Finland2$g[Finland2$g<3],
groupnames=c("m", "f"), shownames=T, add=T)
```



- for the demographic group **age**:

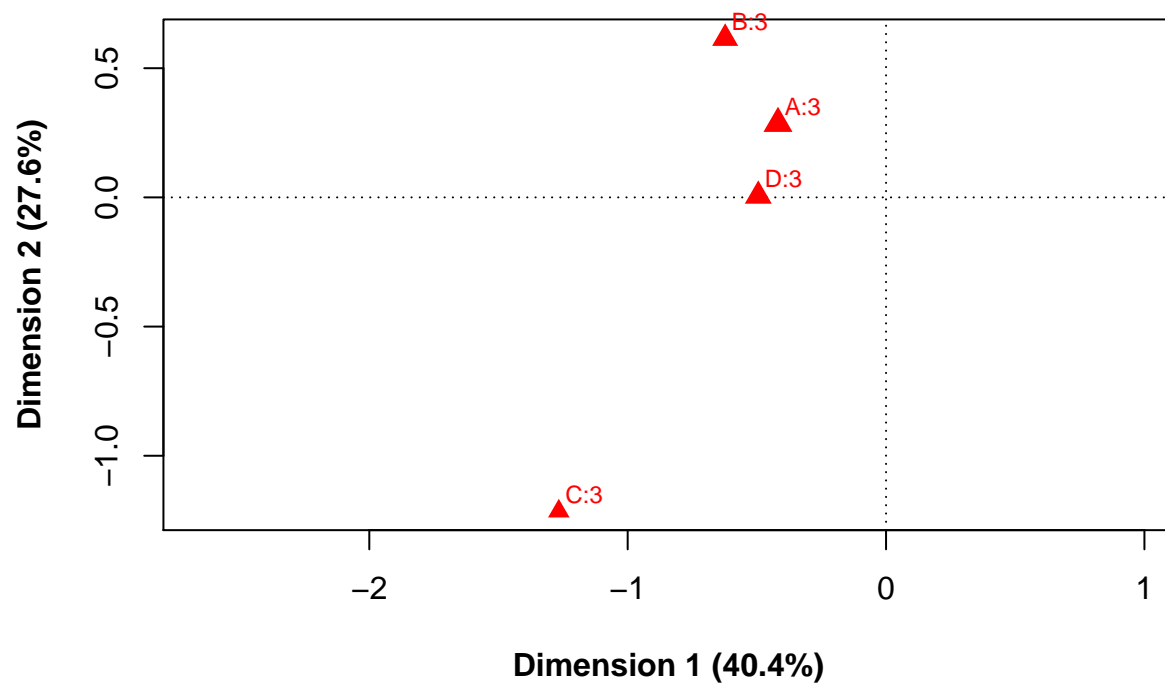
```
plot(Finland2.rpc, type="n", asp=1, xlab="Subset CA dim1 (20.7%)", ylab="Subset CA dim1 (17.9%)", xlim=-0.5, ylim=-0.3,
abline(v=0, h=0, lty=2, col="grey")
confidenceplots(Finland2.rpc[Finland2$a<7,1], Finland2.rpc[Finland2$a<7,2], group=Finland2$a[Finland2$a<7],
groupnames=c("a1","a2","a3","a4","a5","a6"), shownames=T, add=T)
```

Task 3: Analysis of the “middle” and “missing” values of the substantive variables. Confidence intervals of the demographic groups.

Analysis of “middle” category could be provided via CA of Burt matrix.

```
# investigation of "middle" category, via the Burt matrix
#seq() - generates a sequence of numbers
#seq(from=3, to=46, by=6): 3,9,15,21,27,33,39,45
Finland2.middle <- seq(3,22,6)
Finland2.mca2 <- ca(Finland2.B, subsetrow=Finland2.middle, subsetcol=seq(3,22,6))
#what - Vector of two character strings specifying the contents of the plot.
#First entry sets the rows and the second entry the columns. Allowed values # are "all" (all available)
#"active" (only active points are displayed)
#"passive" (only supplementary points are displayed)
#"none" (no points are displayed)
#The status (active or supplementary) of columns is set in mjca using the option
#supcol.
# mass - area, first - rows, second - columns
#font.lab - labels for font?
plot(Finland2.mca2, what=c("none", "all"), mass=c(F,T), font.lab=2)
```



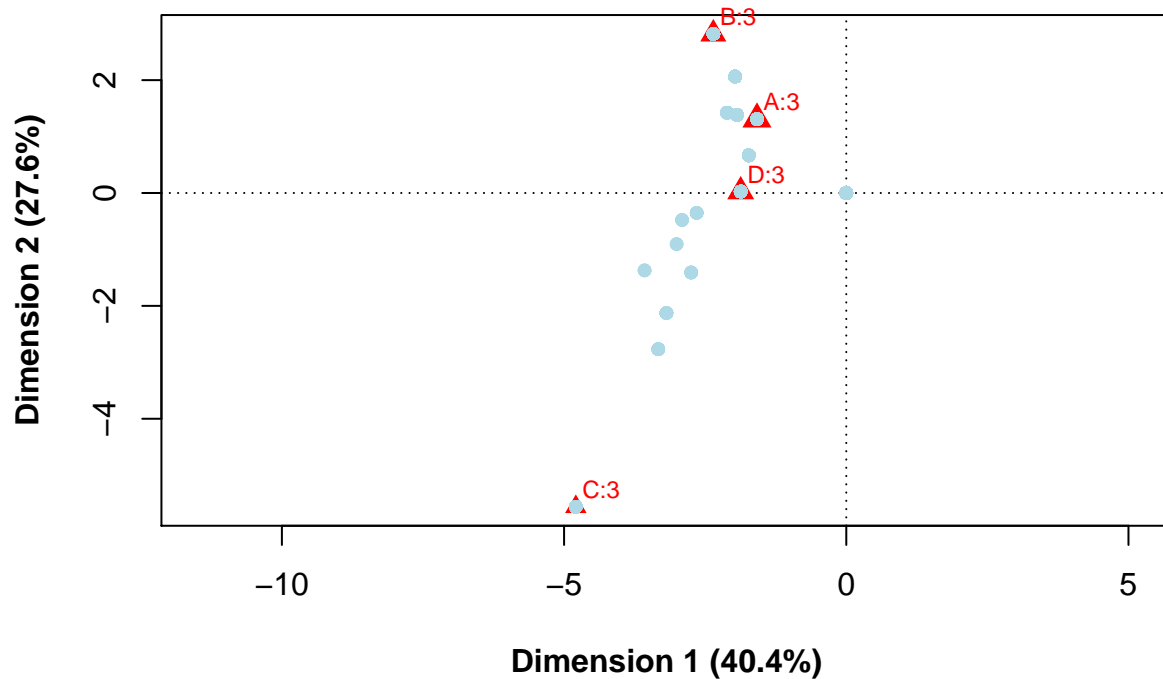
```
summary(Finland2.mca2)
```

```
##
## Principal inertias (eigenvalues):
##
## dim    value      %   cum%   scree plot
## 1      0.069858  40.4  40.4  *****
## 2      0.047719  27.6  68.0  *****
## 3      0.034199  19.8  87.8  *****
## 4      0.021173  12.2 100.0  ***
## -----
## Total: 0.172949 100.0
##
## Rows:
##   name  mass  qlt  inr    k=1 cor ctr    k=2 cor ctr
## 1 |  A3 |   73  563  194 |  -418 384 184 |   286 179 126 |
## 2 |  B3 |   44  770  256 |  -623 389 246 |   616 381 353 |
## 3 |  C3 |   17  943  318 | -1267 491 387 | -1215 452 522 |
## 4 |  D3 |   52  317  232 |  -495 317 182 |     6   0   0 |
##
## Columns:
##   name  mass  qlt  inr    k=1 cor ctr    k=2 cor ctr
## 1 |  A3 |   73  563  194 |  -418 384 184 |   286 179 126 |
## 2 |  B3 |   44  770  256 |  -623 389 246 |   616 381 353 |
## 3 |  C3 |   17  943  318 | -1267 491 387 | -1215 452 522 |
```

```
## 4 | D3 | 52 317 232 | -495 317 182 | 6 0 0 |
```

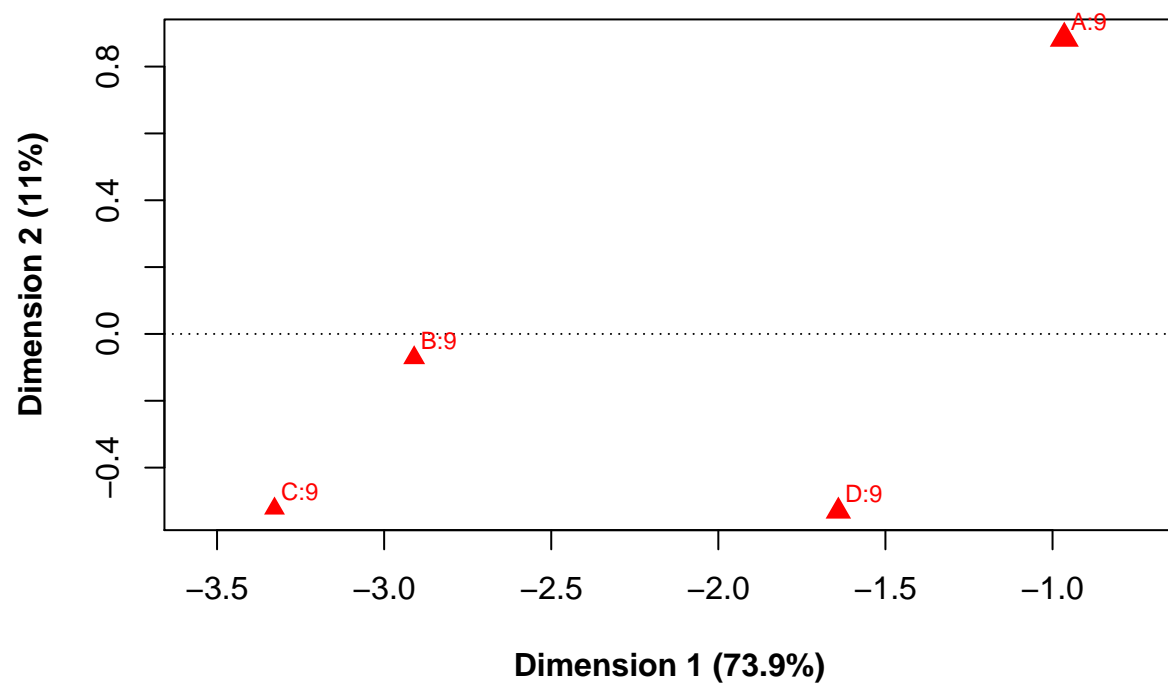
Adding respondent points only for **middle** categories:

```
#Finland2.Z - indicator matrix
# apply(variable, margin, function). - Returns a vector or array or list of values obtained by applying
# a function to margins of an array or matrix
# margin specifies if you want to apply by row (margin = 1),
# by column (margin = 2), or for each element (margin = 1:2).
Finland2.sum4 <- apply(Finland2.Z[,c(Finland2.middle)], 1, sum)
Finland2.sum4[Finland2.sum4==0] <- 1
Finland2.rpc5 <- Finland2.Z[,c(Finland2.middle)] %*% Finland2.mca2$colcoord / Finland2.sum4
plot(Finland2.mca2, what=c("none", "all"), mass=c(F,T), font.lab=2, map="rowprincipal")
points(Finland2.rpc5, pch=19, col="lightblue", cex=0.8)
```



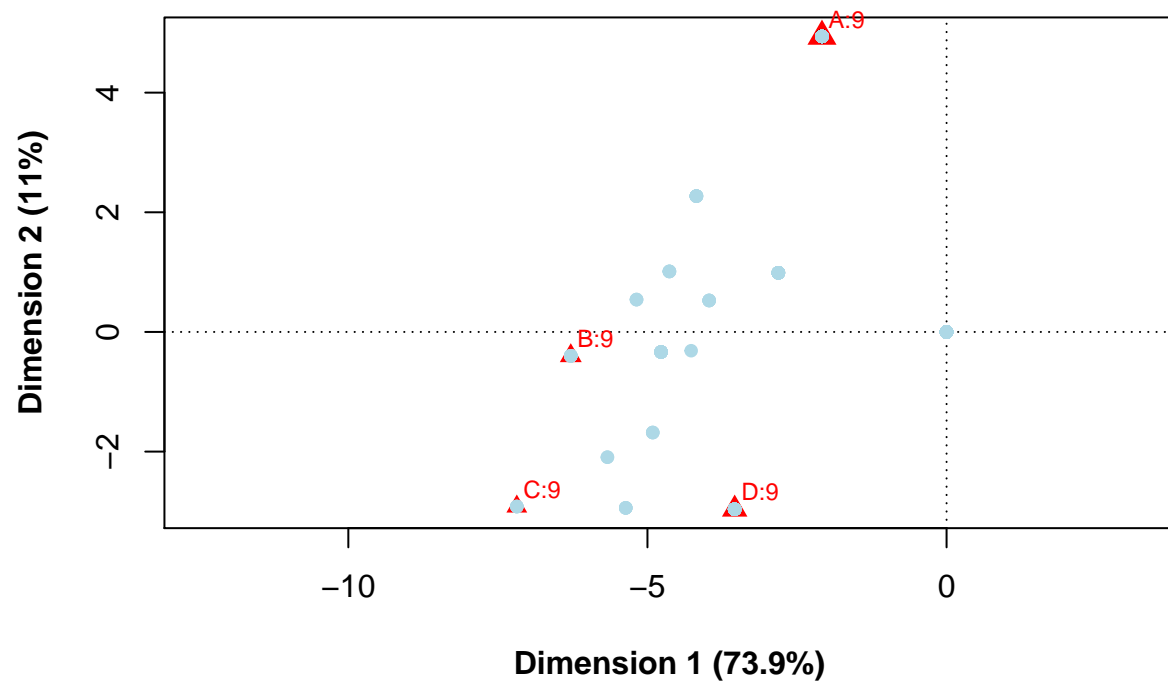
Only **missing** categories are obtained as:

```
Finland2.missing <- c(6,12,18,24)
Finland2.mca4 <- ca(Finland2.B, subsetrow=c(Finland2.missing), subsetcol=c(Finland2.missing))
plot(Finland2.mca4, what=c("none", "all"), mass=c(F,T), font.lab=2)
```



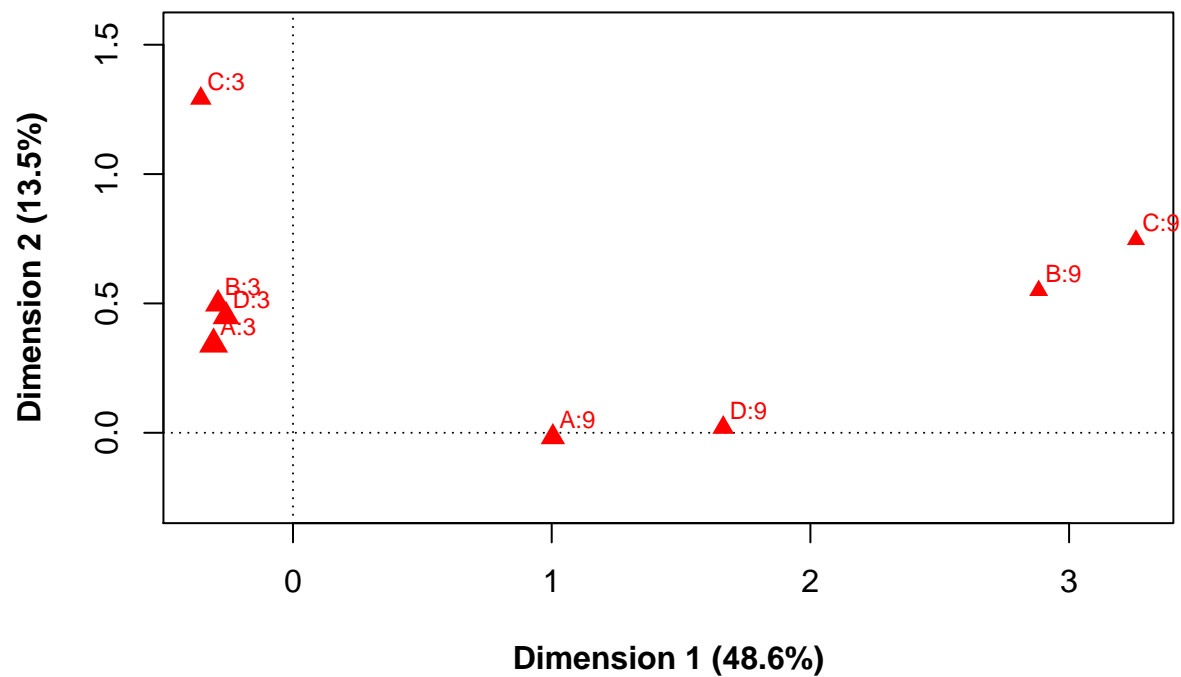
Adding respondent points only for **missing** categories:

```
Finland2.sum2 <- apply(Finland2.Z[,c(Finland2.missing)], 1, sum)
Finland2.sum2[Finland2.sum2==0] <- 1
Finland2.rpc3 <- Finland2.Z[,c(Finland2.missing)] %*% Finland2.mca4$colcoord / Finland2.sum2
plot(Finland2.mca4, what=c("none", "all"), mass=c(F,T), font.lab=2, map="rowprincipal")
points(Finland2.rpc3 , pch=19, col="lightblue", cex=0.8)
```



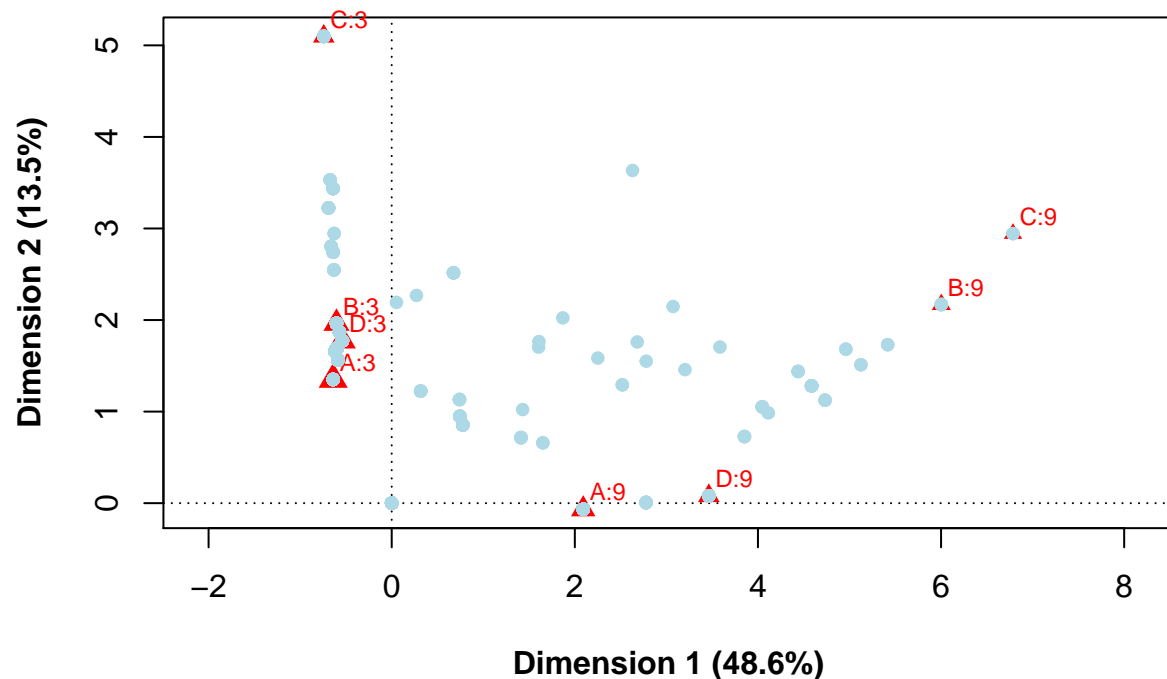
Both “middle” and missing categories are investigated again using CA of Burt matrix.

```
Finland2.missing <- c(6,12,18,24)
Finland2.mca3 <- ca(Finland2.B, subsetrow=c(Finland2.middle,Finland2.missing), subsetcol=c(Finland2.mid
plot(Finland2.mca3, what=c("none", "all"), mass=c(F,T), font.lab=2)
```



Adding respondent points only for both “middle” and missing categories:

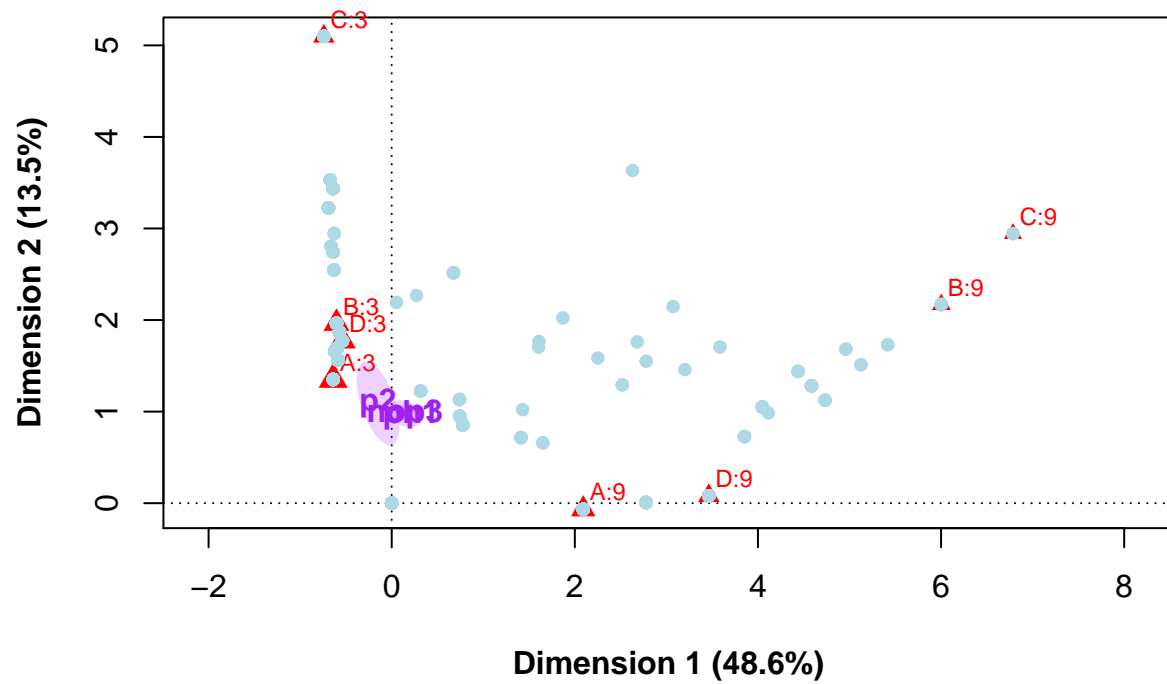
```
Finland2.sum1 <- apply(Finland2.Z[,c(Finland2.middle,Finland2.missing)], 1, sum)
Finland2.sum1[Finland2.sum1==0] <- 1
Finland2.rpc2 <- Finland2.Z[,c(Finland2.middle,Finland2.missing)] %*% Finland2.mca3$colcoord / Finland2
plot(Finland2.mca3, what=c("none", "all"), mass=c(F,T), font.lab=2, map="rowprincipal")
points(Finland2.rpc2 , pch=19, col="lightblue", cex=0.8)
```



Next we compute confidence regions of demographic groups for **both missing and “middle”** groups.

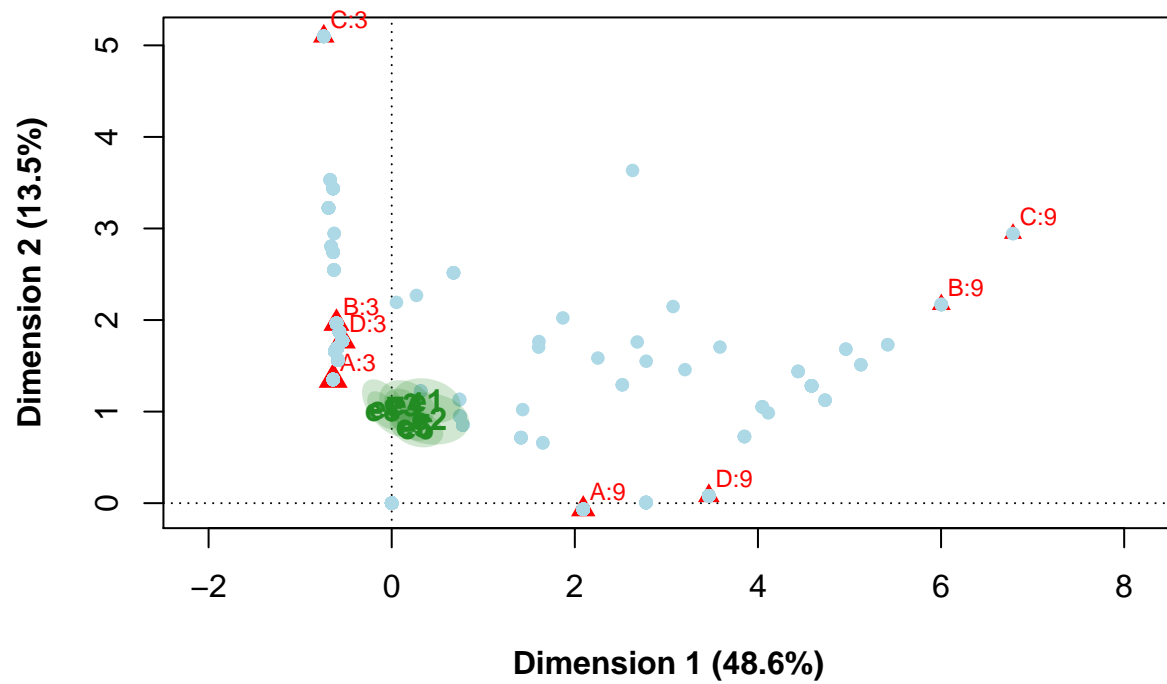
- confidence plot for the demographic group **partnership**

```
#generated asymmetric plot
plot(Finland2.mca3, what=c("none", "all"), mass=c(F,T), font.lab=2, map="rowprincipal")
#add points
points(Finland2.rpc2, pch=19, col="lightblue", cex=0.8)
#draw confidence plots
confidenceplots(Finland2.rpc2[Finland2$p<4,1], Finland2.rpc2[Finland2$p<4,2], group=Finland2$p[Finland2$
groupnames=c("ph1", "p2", "nop3"), shownames=T, add=T)
```



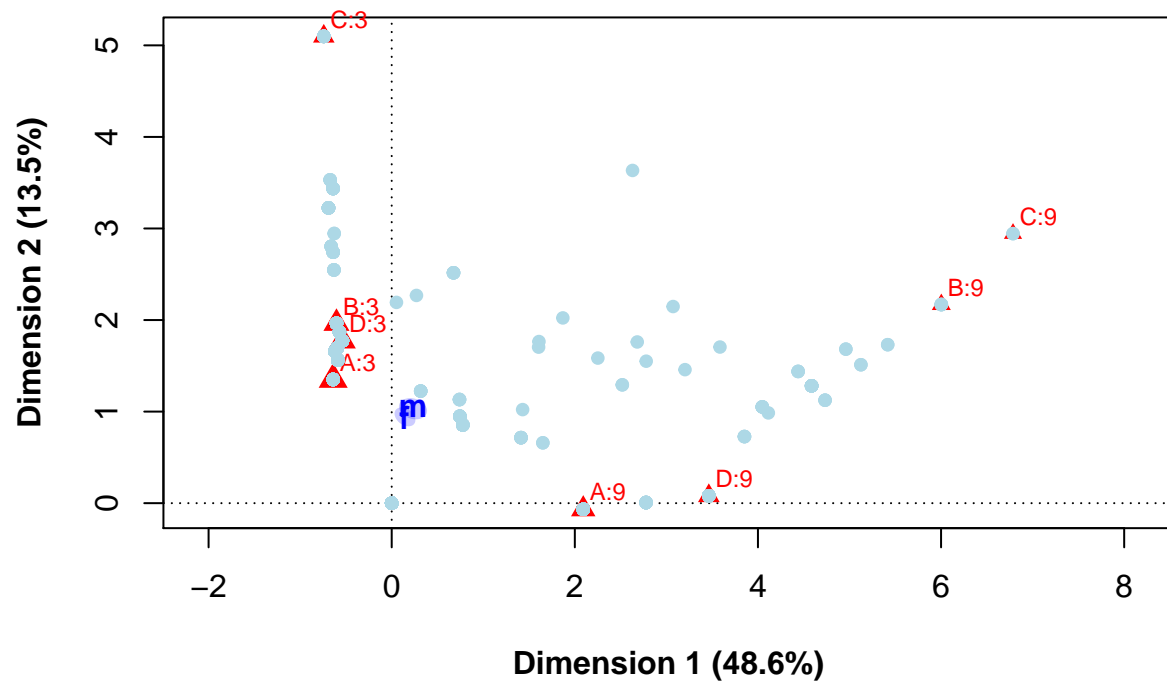
- confidence plot for the demographic group **education**

```
plot(Finland2.mca3, what=c("none", "all"), mass=c(F,T), font.lab=2, map="rowprincipal")
points(Finland2.rpc2, pch=19, col="lightblue", cex=0.8)
confidenceplots(Finland2.rpc2[Finland2$e<9,1], Finland2.rpc2[Finland2$e<9,2], group=Finland2$e[Finland2$e<9],
               groupnames=c("e1","e2","e3","e4","e5","e6","e7","e8"), shownames=T, add=T)
```

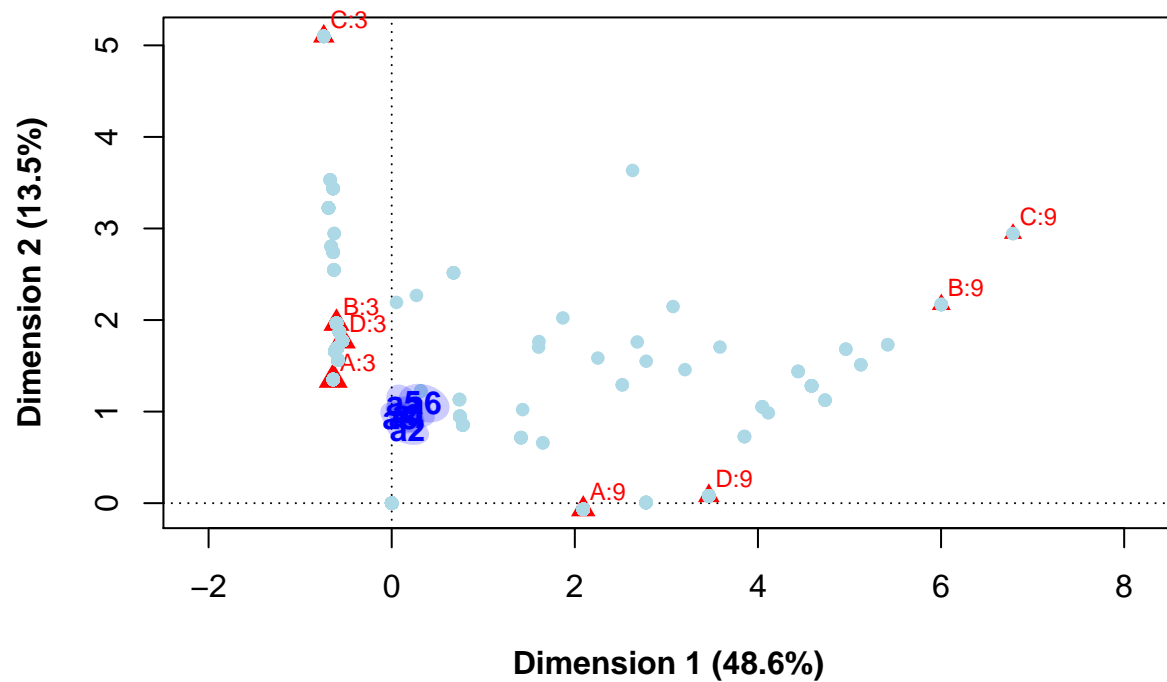
- confidence plot for the demographic group **gender**

```
plot(Finland2.mca3, what=c("none", "all"), mass=c(F,T), font.lab=2, map="rowprincipal")
points(Finland2.rpc2, pch=19, col="lightblue", cex=0.8)
confidenceplots(Finland2.rpc2[Finland2$g<3,1], Finland2.rpc2[Finland2$g<3,2], group=Finland2$g[Finland2$g<3],
               groupnames=c("m","f"), shownames=T, add=T)
```



- confidence plot for the demographic group **age**

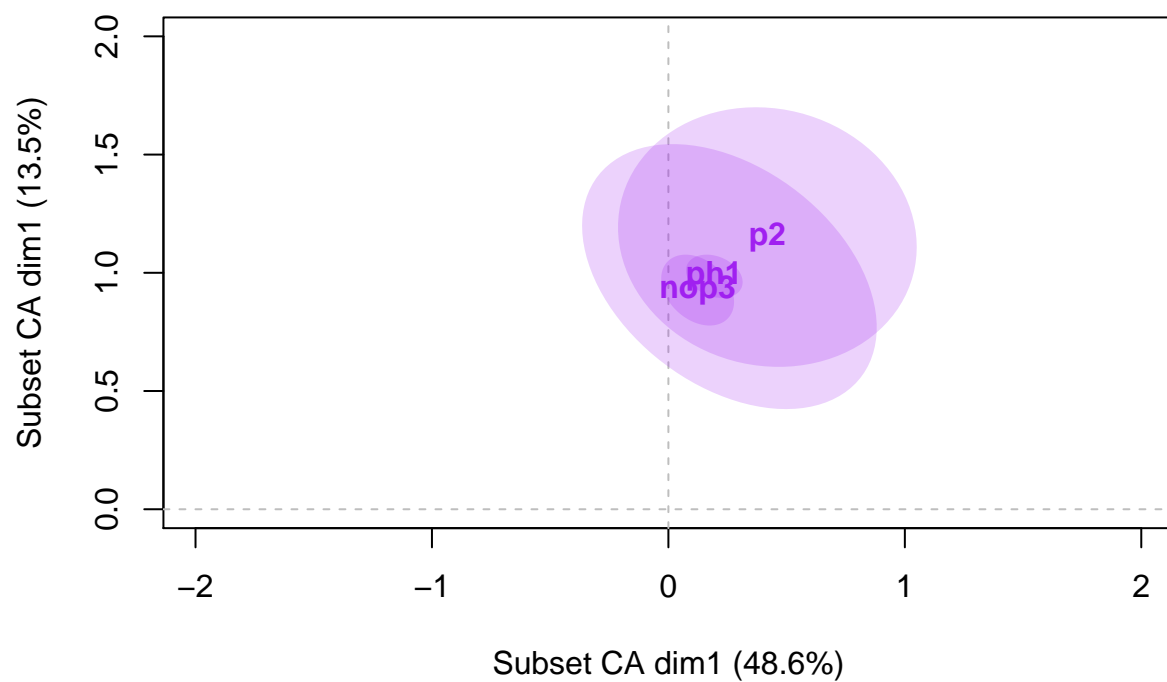
```
plot(Finland2.mca3, what=c("none", "all"), mass=c(F,T), font.lab=2, map="rowprincipal")
points(Finland2.rpc2, pch=19, col="lightblue", cex=0.8)
confidenceplots(Finland2.rpc2[Finland2$a<7,1], Finland2.rpc2[Finland2$a<7,2], group=Finland2$a[Finland2$a<7],
               groupnames=c("a1","a2","a3","a4","a5","a6"), shownames=T, add=T)
```



Since from these plots the confidence regions are not so clear, similarly as in Task 2, we next plot “just the ellipses”.

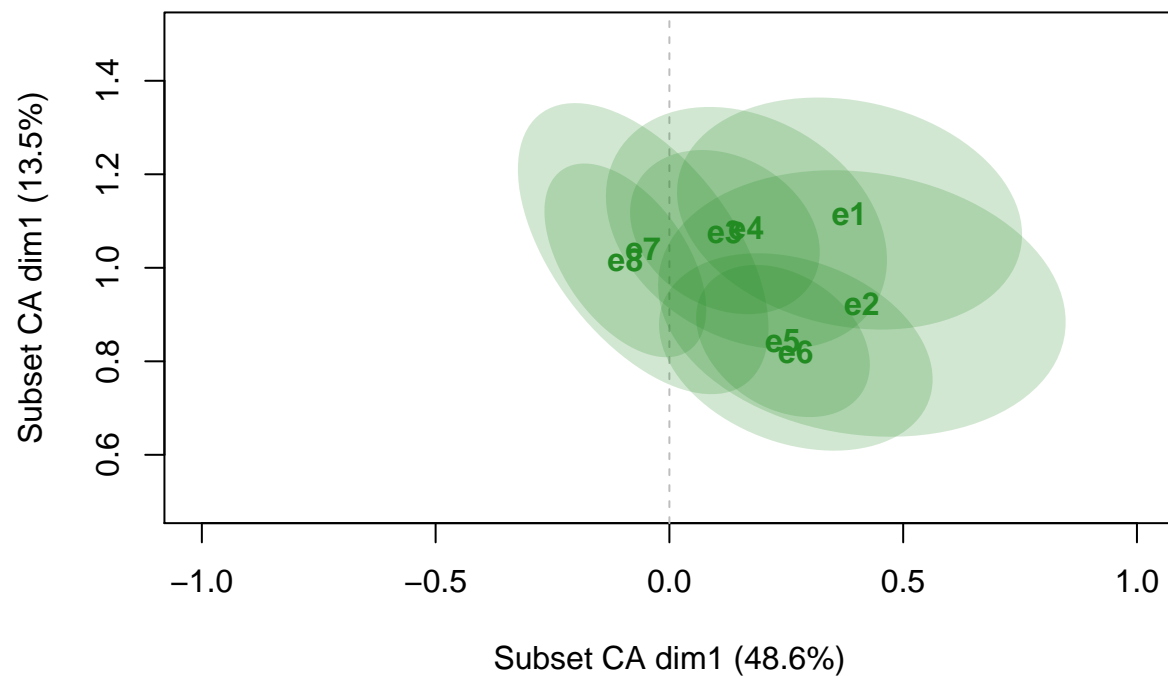
- for the demographic group **partnership**:

```
plot(Finland2.rpc2, type="n", asp=1, xlab="Subset CA dim1 (48.6%)", ylab="Subset CA dim1 (13.5%)", xlim=c(-2, 8), ylim=c(0, 5))
abline(v=0, h=0, lty=2, col="grey")
#draw confidence plots
confidenceplots(Finland2.rpc2[Finland2$p<4,1], Finland2.rpc2[Finland2$p<4,2], group=Finland2$p[Finland2$p<4],
               groupnames=c("ph1", "p2", "nop3"), shownames=T, add=T)
```



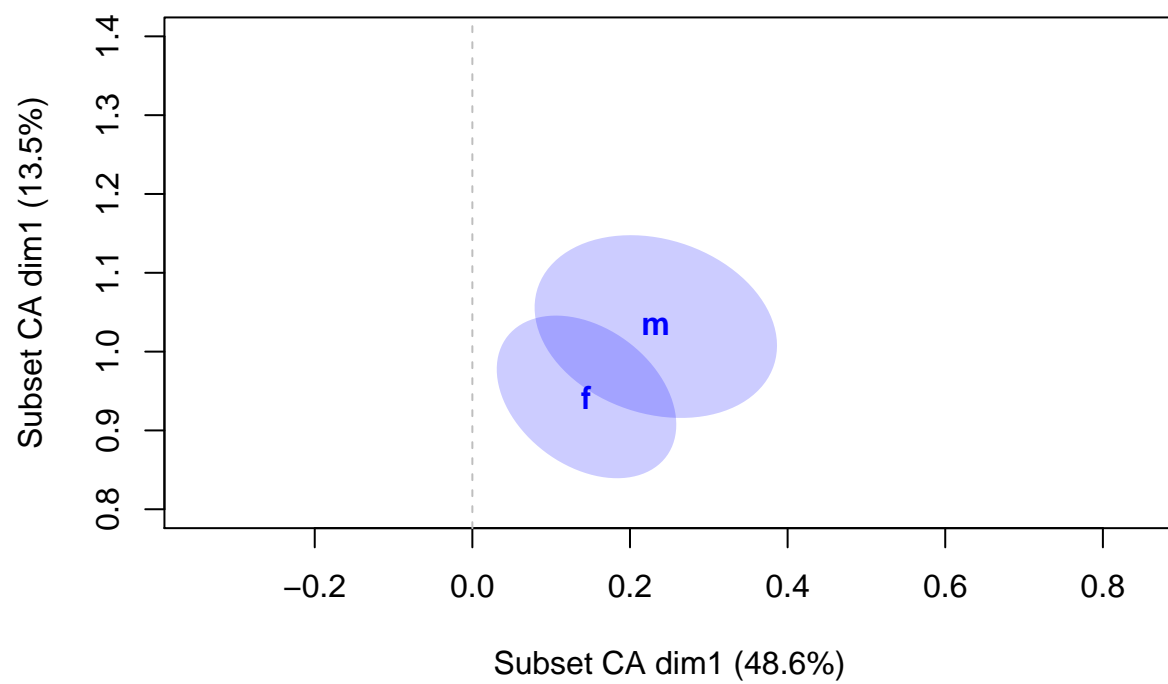
- for the demographic group **education**:

```
plot(Finland2.rpc2, type="n", asp=1, xlab="Subset CA dim1 (48.6%)", ylab="Subset CA dim1 (13.5%)", xlim=
abline(v=0, h=0, lty=2, col="grey")
confidenceplots(Finland2.rpc2[Finland2$e<9,1], Finland2.rpc2[Finland2$e<9,2], group=Finland2$e[Finland2
groupnames=c("e1", "e2", "e3", "e4", "e5", "e6", "e7", "e8"), shownames=T, add=T)
```



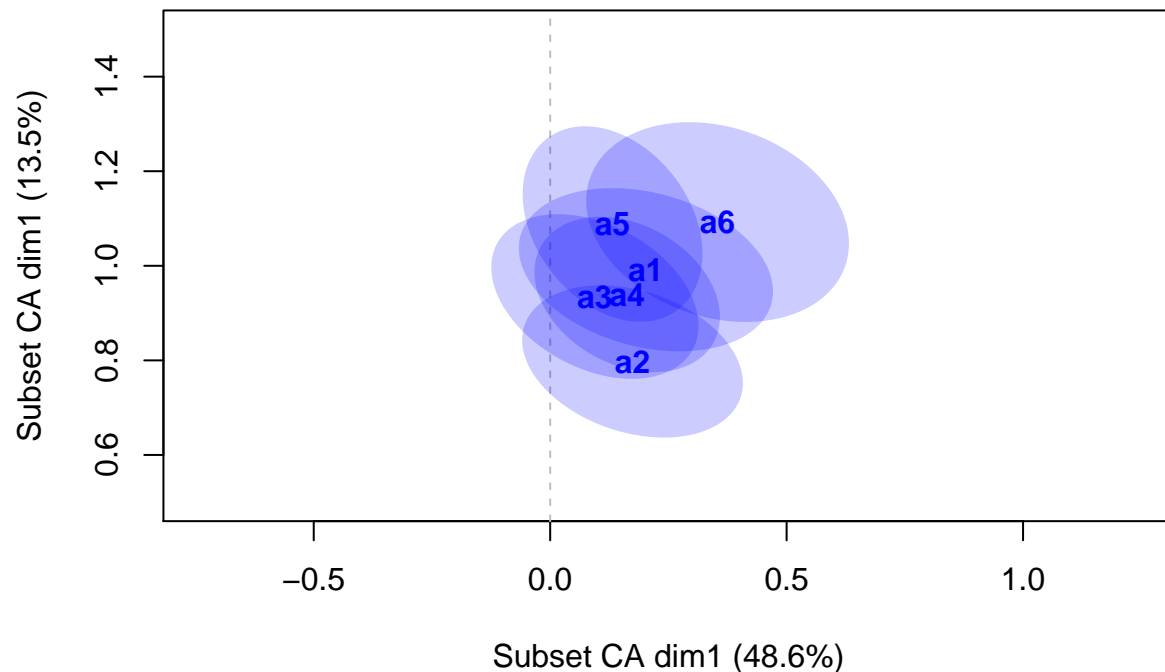
- for the demographic group **gender**:

```
plot(Finland2.rpc2, type="n", asp=1, xlab="Subset CA dim1 (48.6%)", ylab="Subset CA dim1 (13.5%)", xlim=
abline(v=0, h=0, lty=2, col="grey")
confidenceplots(Finland2.rpc2[Finland2$g<3,1], Finland2.rpc2[Finland2$g<3,2], group=Finland2$g[Finland2
groupnames=c("m", "f"), shownames=T, add=T)
```



- for the demographic group **age**:

```
plot(Finland2.rpc2, type="n", asp=1, xlab="Subset CA dim1 (48.6%)", ylab="Subset CA dim1 (13.5%)", xlim=
abline(v=0, h=0, lty=2, col="grey")
confidenceplots(Finland2.rpc2[Finland2$a<7,1], Finland2.rpc2[Finland2$a<7,2], group=Finland2$a[Finland2
groupnames=c("a1", "a2", "a3", "a4", "a5", "a6"), shownames=T, add=T)
```



From all analyses, provided in Task 3, we can conclude that MCA is a powerful tool for investigating missing and “middle” categories. Both categories involve quite big amount of uncertainty, because of which most of the other methods just ignore these cases. Here we have seen how the information from both categories could be utilized applying MCA approach.

Used and useful links

Package ‘ca’

Oleg Nenadic and Michael Greenacre, Computation of Multiple Correspondence Analysis, with code in R

Michael Greenacre, Biplots in practise

Multiple Correspondence Analysis Essentials: Interpretation and application to investigate the associations between categories of multiple qualitative variables - R software and data mining

Mike Bendixen, A Practical Guide to the Use of Correspondence Analysis in Marketing Research, Marketing Bulletin, 2003, 14, Technical Note 2.

[An Example R Markdown] (<http://www.statpower.net/Content/310/R%20Stuff/SampleMarkdown.html>)

Writing Mathematic Fomulars in Markdown