Book (Greenacre 2017) gives an excellent introduction to multiple correspondence analysis (MCA). Book focuses on interpretations and provides many examples. Book also includes an appendix about the package ca, that is used for performing MCA in this course. Of course, it is not necessary to read the book for the course but it can be a good reference if, for example, you decide to use MCA in your project.

Demo Problem 1: Multiple Correspondence Analysis

Install the package `ca` if you haven't yet. The data set `household.txt` is a simplified version of the survey performed by Statistics Finland in 2019. The survey contains answers of Finnish households to questions related to sharing economy (link to the full data set). Below we present the questions and possible answers to each question. Variable names are next to the questions in brackets, and also, codings of the answers are indicated by the integers next to the possible answers.

1. Have you bought/got something from a traditional flea market or auction during the last 12 months? (`flea`)
   - Yes (1)
   - No (2)
2. Have you bought/got something from internet auction during the last 12 months? (`web`)
   - Yes (1)
   - No (2)
3. Have you bought/got something from a recycling group in a social networking service during the last 12 months? (`recycling`)
   - Yes (1)
   - No (2)
4. What is the structure of your household? (`living`)
   - Living alone (1)
   - Living together with a spouse (2)
   - Living together with a spouse and children (3)
   - Living alone with children (4)
5. What is your residential environment? (`environment`)
   - Major city (1)
   - City (2)
   - Town/Coutryside (3)
6. Household net monthly income (`income`)
   - Under 1500 euros (1)
   - 1500-2299 euros (2)
   - 2300-3299 euros (3)
   - 3,300 - 4,499 (4)
   - At least 4500 (5)

Perform correspondence analysis to the data set and interpret the results.

Solution

First, we read the data. The data set has 1229 observations and 6 variables.

```
house <- read.csv("data/household.csv", header = TRUE)
dim(house)
```

```
## [1] 1229    6
```

```
head(house)
```

```
##   flea web recycling living environment income
## 1    2   2         2      3           2      3
```

```
## 2    1   1         1     2           3     2
## 3    1   1         2     2           2     4
## 4    1   2         1     3           1     5
## 5    1   1         2     3           1     5
## 6    2   2         2     2           1     4
```

Now we perform multiple correspondence analysis (MCA) for the data set `house` with the function `mjca` from the package `ca`. There are multiple almost equivalent ways to define MCA. One way to define MCA is that it is CA performed for the *complete disjunctive table* (also called the *indicator matrix*). We can perform this version of MCA by setting `lambda = "indicator"`. Argument `reti` controls whether the complete disjunctive table is returned.

```
house_mca <- ca::mjca(house, lambda = "indicator", reti = TRUE)
```

If you wish you can try to perform bivariate correspondence analysis to the indicator matrix, and see that the results are the same as given by the `house_mca` object.

```
house_ca_ind <- ca::ca(house_mca$indmat)
```

For example, below we check that the column standard coordinates given by the first two components of `house_mca` and `house_ca_ind` are the same. When checking the equality of the coordinates, remember that the results of MCA are unique up to sign changes.

```
abs(round(house_mca$colcoord[, 1:2], 2)) ==
  abs(round(house_ca_ind$colcoord[, 1:2], 2))
```

```
##                 Dim1 Dim2
## flea:1          TRUE TRUE
## flea:2          TRUE TRUE
## web:1           TRUE TRUE
## web:2           TRUE TRUE
## recycling:1     TRUE TRUE
## recycling:2     TRUE TRUE
## living:1        TRUE TRUE
## living:2        TRUE TRUE
## living:3        TRUE TRUE
## living:4        TRUE TRUE
## environment:1 TRUE TRUE
## environment:2 TRUE TRUE
## environment:3 TRUE TRUE
## income:1        TRUE TRUE
## income:2        TRUE TRUE
## income:3        TRUE TRUE
## income:4        TRUE TRUE
## income:5        TRUE TRUE
```

The object returned by the function `mjca` is very similar to the object returned by the function `ca`.

```
names(house_mca)
```

```
##  [1] "sv"         "lambda"     "inertia.e"  "inertia.t"  "inertia.et"
##  [6] "levelnames" "factors"    "levels.n"   "nd"         "nd.max"
## [11] "rownames"   "rowmass"    "rowdist"    "rowinertia" "rowcoord"
## [16] "rowpcoord"  "rowctr"     "rowcor"     "colnames"   "colmass"
## [21] "coldist"    "colinertia" "colcoord"   "colpcoord"  "colctr"
## [26] "colcor"     "colsup"     "subsetcol"  "Burt"       "Burt.upd"
## [31] "subinertia" "JCA.iter"   "indmat"     "call"
```

By default `summary.mjca` gives the summary for the columns only. Motivation for this is that in MCA often the relations between different variables are of interest. By setting `rows = TRUE` one can see the summary for the rows as well. For details, see the help pages `?ca::summary.mjca`.

```
house_summary <- summary(house_mca)
house_summary
```

```
##
## Principal inertias (eigenvalues):
##
##  dim    value      %   cum%   scree plot
##  1      0.327300  16.4  16.4  ****
##  2      0.251808  12.6  29.0  ***
##  3      0.198050   9.9  38.9  **
##  4      0.177587   8.9  47.7  **
##  5      0.172507   8.6  56.4  **
##  6      0.161890   8.1  64.5  **
##  7      0.153648   7.7  72.1  **
##  8      0.143797   7.2  79.3  **
##  9      0.141449   7.1  86.4  **
##  10     0.111484   5.6  92.0  *
##  11     0.096487   4.8  96.8  *
##  12     0.063992   3.2 100.0  *
##         -------- -----
##  Total: 2.000000 100.0
##
##
## Columns:
##               name   mass  qlt  inr    k=1 cor ctr    k=2 cor ctr
## 1  |        flea:1 |  104  429   34 |  442 321  62 |  256 108  27 |
## 2  |        flea:2 |   63  429   55 | -726 321 102 | -421 108  44 |
## 3  |         web:1 |   79  564   50 |  704 445 119 |  363 119  41 |
## 4  |         web:2 |   88  564   45 | -633 445 107 | -327 119  37 |
## 5  |   recycling:1 |   59  571   63 |  966 508 168 |  340  63  27 |
## 6  |   recycling:2 |  108  571   34 | -526 508  91 | -185  63  15 |
## 7  |      living:1 |   39  687   75 | -967 289 112 | 1136 398 201 |
## 8  |      living:2 |   64  300   48 | -173  19   6 | -668 281 114 |
## 9  |      living:3 |   56  389   61 |  857 368 125 | -204  21   9 |
## 10 |      living:4 |    7   83   71 |  206   2   1 | 1327  81  51 |
## 11 | environment:1 |   78  105   40 |  -40   1   0 |  344 104  37 |
## 12 | environment:2 |   39   26   55 |  -45   1   0 | -292  26  13 |
## 13 | environment:3 |   50   45   52 |   97   4   1 | -310  41  19 |
## 14 |      income:1 |   22  508   76 | -834 103  46 | 1649 404 233 |
## 15 |      income:2 |   30   94   63 | -419  38  16 |  508  56  30 |
## 16 |      income:3 |   37   14   57 |  -74   2   1 | -204  12   6 |
## 17 |      income:4 |   39  144   59 |  435  59  23 | -525  85  43 |
## 18 |      income:5 |   39  152   61 |  414  52  20 | -576 100  51 |
```

Summary for MCA is very similar to the summary of the bivariate correspondence analysis. Again, `qlt` gives quality of representation of modalities with respect to the first and second components, `inr` gives column inertias, `k=i` gives column principal coordinates, `cor` gives qualities of representation with respect ro the component `k=i`, and `ctr` gives contribution of the component `k=i` to different modalities.

Figure 1 shows that only 29% of variation is explained by the first two components. Nevertheless, we proceed to analyze the first two components.

```
barplot(house_summary$scree[, 3], ylim = c(0, 20),
        names.arg = paste("PC", 1:12), las = 2, xlab = "Component",
        ylab = "% of variation explained", col = "skyblue")
```

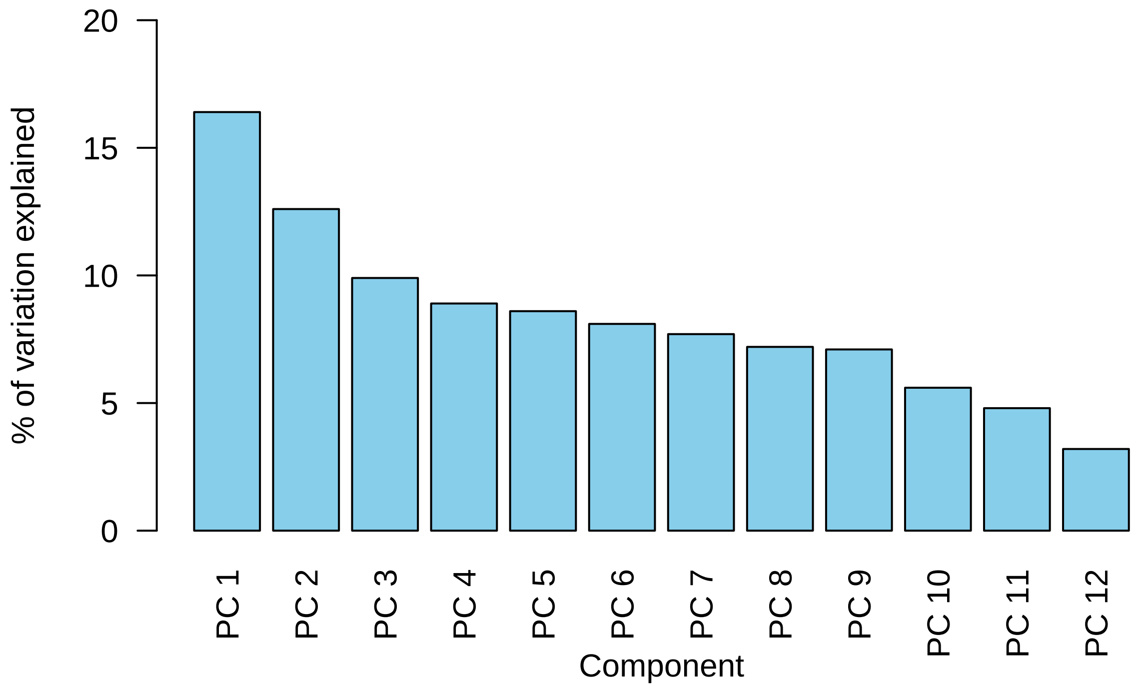

Figure 1: Scree plot.

By default `plot.mjca` plots only the column scores. By modifying argument `what` one can specify whether row/column scores are plotted. Again, for more information see help pages `?ca::plot.mjca`.

```
plot(house_mca, arrows = c(TRUE, TRUE))
```

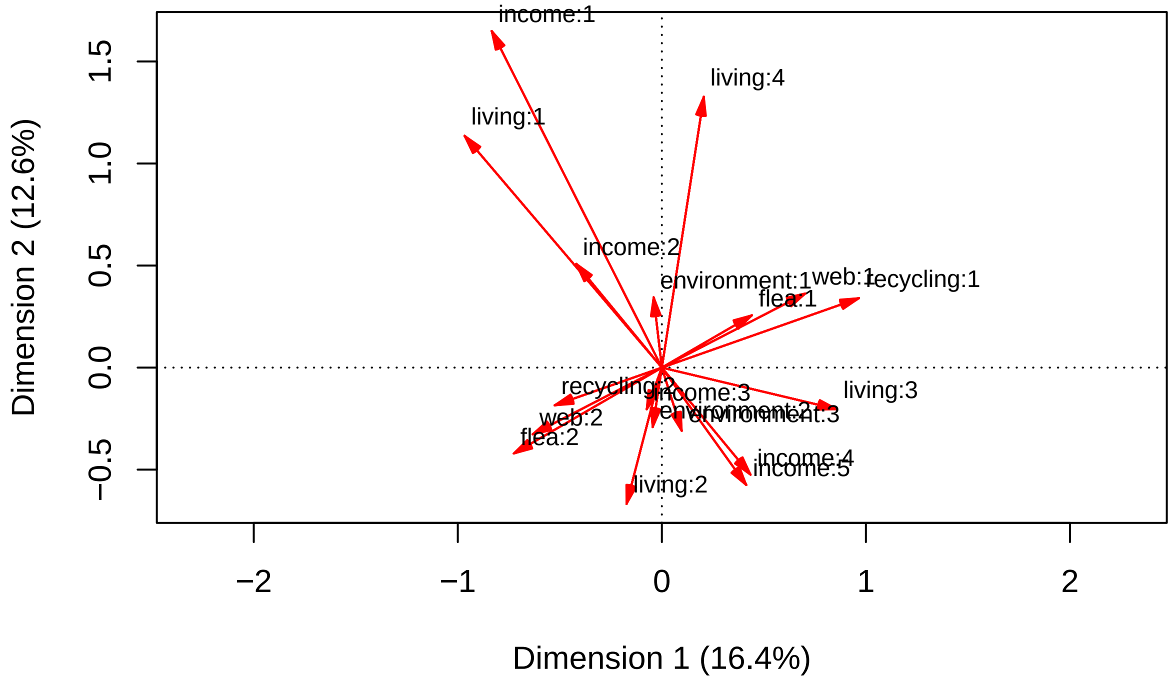

Figure 2: First two column principal coordinates.

Denote principal column coordinate corresponding to the component h and modality l of variable p by $\psi_{h,pl}$. From the relation

$$d_{p_1l_1,p_2l_2} \approx 1 + \sum_{h=1}^{2} \psi_{h,p_1l_1}\psi_{h,p_2l_2}$$

we get interpretation for Figure 2:

- angle between modalities less than 90 degrees = attraction,
- angle between modalities more than 90 degrees = repulsion and
- angle between modalities 90 degrees = independent.

Also, proximity of the column profiles hints that the χ^2-distances are small between the profiles (assuming good quality of representation). Thus, similar modalities should be close to each other.

Interpretations from the biplot are only valid when modalities are represented well by the components. Thus, we could modify Figure 2 in such a way that point size represents quality of representation of the corresponding modality.

```r
# Function for scaling values from 0 to 1 (this is for visualization purposes):
normalize <- function(x) {
  (x - min(x)) / (max(x) - min(x))
}

# Generate the scatter plot. Point size is now scaled according to qlt:
qlt <- house_summary$columns[, 3]
house_covariates <- house_mca$colpcoord[, 1:2]
plot(house_covariates, pch = 21, asp = 1, bg = "red", cex = normalize(qlt) + 1,
```

```
     xlab = paste0("Dimension 1", " (", house_summary$scree[1, 3], "%", ")"),
     ylab = paste0("Dimension 2", " (", house_summary$scree[2, 3], "%", ")"))

# Add arrows. Slight transparency is added to increase visibility.
arrows(rep(0, 17), rep(0, 17), house_covariates[, 1], house_covariates[, 2],
       length = 0, col = rgb(1, 0, 0, 0.25))

# "Cross-hair" is added, i.e., dotted lines crossing x and y axis at 0.
abline(h = 0, v = 0, lty = 3)

# Add variable:category names to the plot.
text(house_covariates, house_mca$levelnames, pos = 2, cex = 0.5)
```

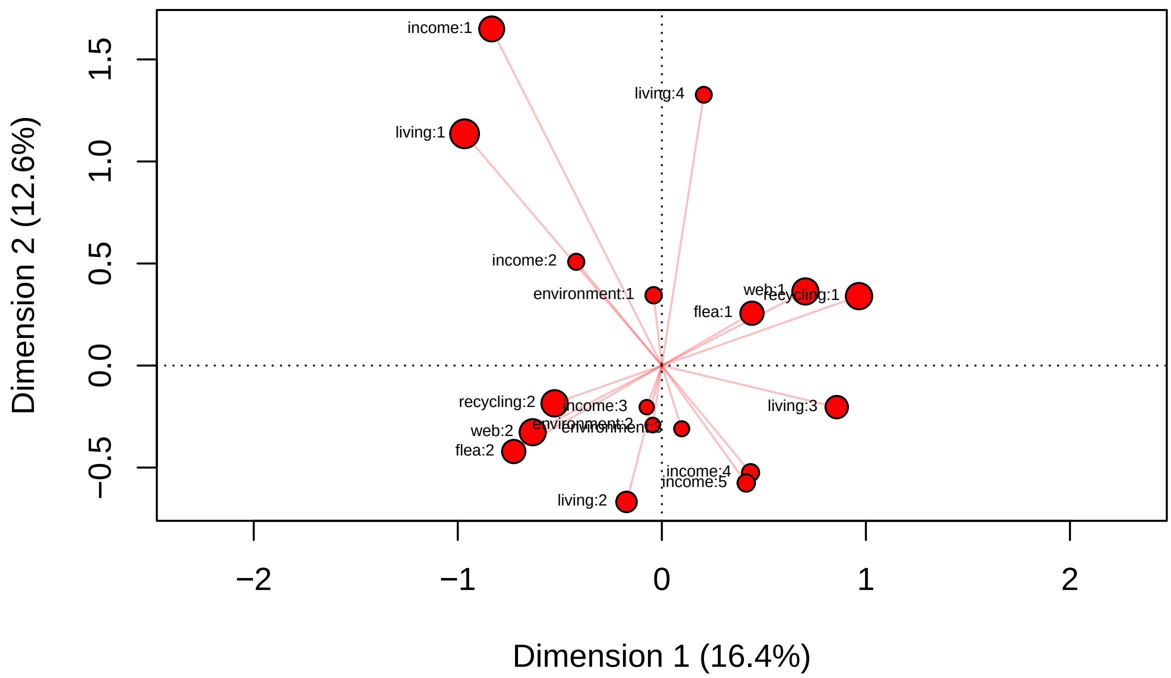

Figure 3: First two column principal coordinates. Point sizes are scaled according to quality of representation.

For example, following interpretations can be made for the variables from Figures 2 and 3:

- If a household has a habit of buying items from one kind of auction (internet, flea market or recycling group), then most likely they also buy items from another kind of auction. On the other hand, if a household does not buy anything from one type of auction then probably they do not buy items from another type of auction.
- Single living people tend to have low income (as expected since the income accounts for the whole household).

Rows can be analyzed similarly to columns. Denote principal row coordinate corresponding to component h and household i by $\phi_{h,i}$. From relation

$$d_{i_1,i_2} \approx 1 + \sum_{h=1}^{2} \phi_{h,i_1}\phi_{h,i_2}$$

we get interpretation for Figure 4:

- angle between households less than 90 degrees = similar profiles and
- angle between households more than 90 degrees = profiles differ.

Also, proximity of the row profiles hints that the χ^2-distances are small (assuming good quality of representation). Thus, similar households should be clustered together.

For the sake of clarity, observation labels are dropped from Figure 4 and instead of arrows we have points. Indeed, Figure 4 shows that households are in five clusters. However, clusters are not determined by one single variable. For example, one can check that clusters do not correspond to modalities of `income`, even though `income` has exactly five different modalities.

```
plot(house_mca, arrows = c(FALSE, FALSE), what = c("all", "none"),
     labels = c(0, 0))
```

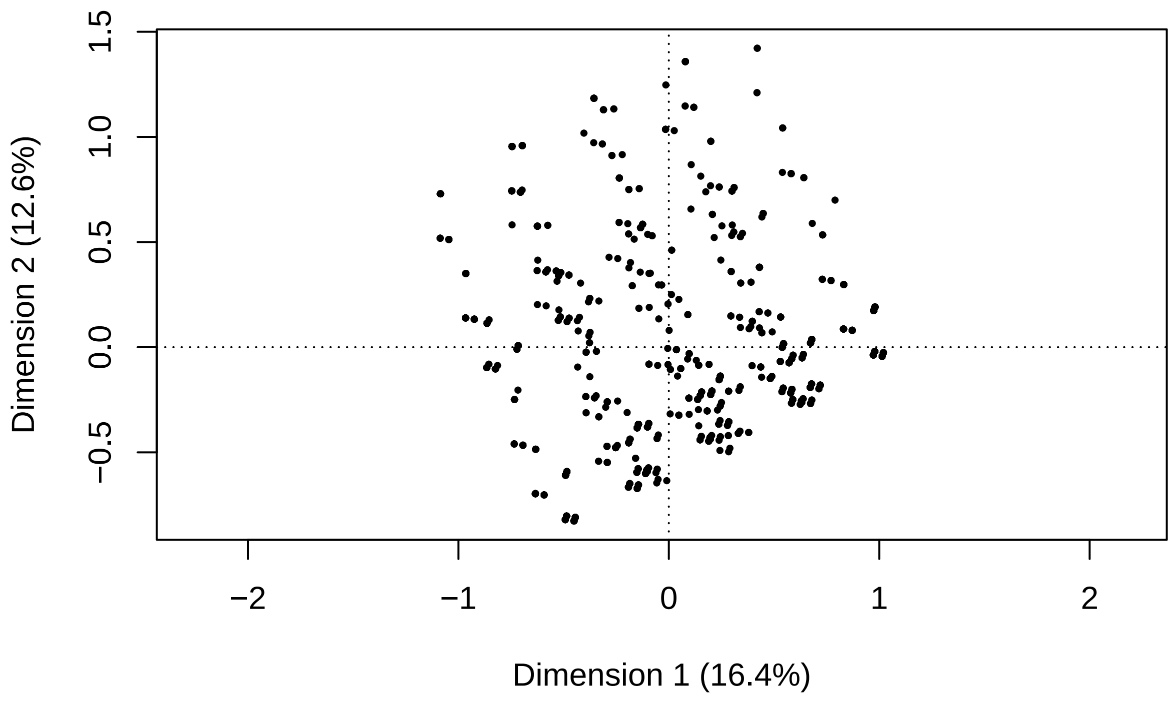

Figure 4: First two row principal coordinates.

Denote standard row coordinate corresponding to component h and household i by $\hat{\phi}_{h,i}$. Lastly, relation

$$d_{i,pl} \approx 1 + \sum_{h=1}^{2} \hat{\phi}_{h,i}\psi_{h,pl}$$

gives interpretation for Figure 5. Notice that since columns are in principal coordinates, we can also interpret angles and distances between columns in Figure 5 similarly as in Figure 2.

```
plot(house_mca, arrows = c(FALSE, TRUE), what = c("all", "all"),
     map = "colprincipal", labels = c(0, 2), cex = 0.5)
```

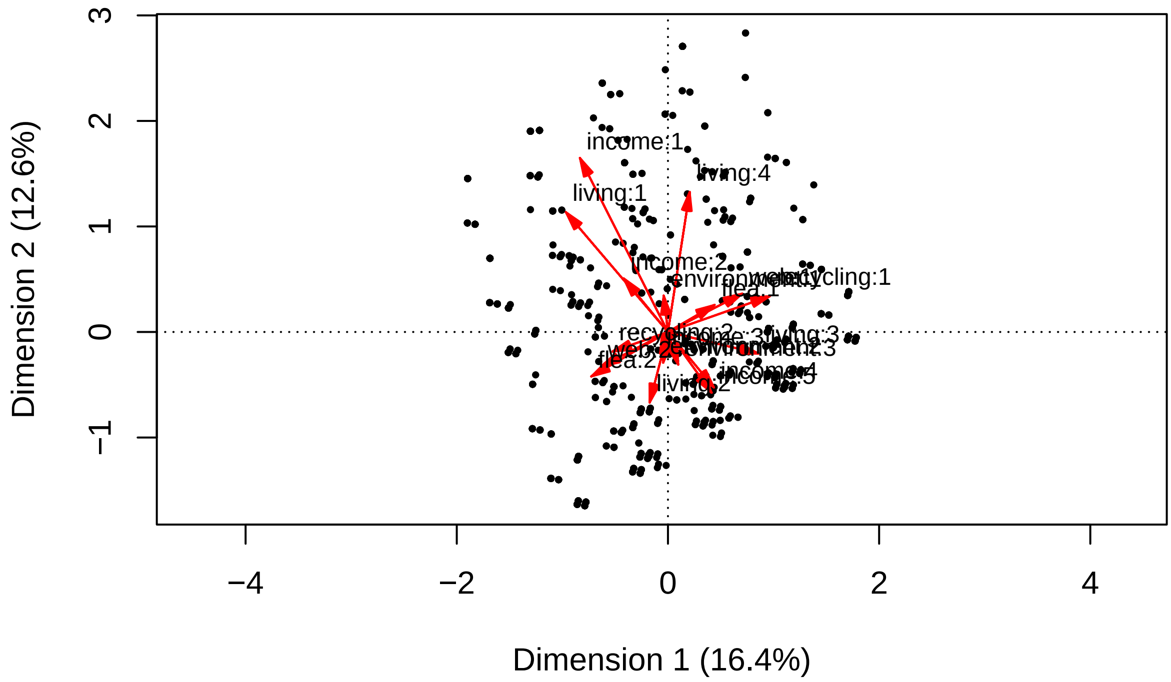

Figure 5: Columns in principal coordinates and rows in standard coordinates.

However, Figure 5 is hard to interpret since labels get on top of each other. Thus, we can make a similar plot manually. From Figure 6 attractions between columns and rows are more visible.

```
# Row standard coordinates
plot(house_mca$rowcoord[, 1:2], cex = 0.5, pch = 4, asp = 1,
     col = rgb(0, 0, 1, 0.1),
     xlab = paste0("Dimension 1", " (", house_summary$scree[1, 3], "%", ")"),
     ylab = paste0("Dimension 2", " (", house_summary$scree[2, 3], "%", ")"))

# Generate the scatter plot. Point size is now scaled according to qlt.
points(house_covariates, pch = 21, bg = "red", cex = normalize(qlt) + 1)

# Add arrows. Slight transparency is added to increase visibility.
arrows(rep(0, 17), rep(0, 17), house_covariates[, 1], house_covariates[, 2],
       length = 0, col = rgb(1, 0, 0, 0.25))

# "Cross-hair" is added, i.e., dotted lines crossing x and y axis at 0.
abline(h = 0, v = 0, lty = 3)

# Add variable:category names to the plot. Now we only use first letters of the
# variables so that the plot does not become too stuffed.
first <- substr(house_mca$levelnames, 1, 1)
```

```r
last <- substr(house_mca$levelnames, nchar(house_mca$levelnames),
               nchar(house_mca$levelnames))
labels <- paste0(first, last)
text(house_covariates, labels, pos = 3, cex = 0.5)
```

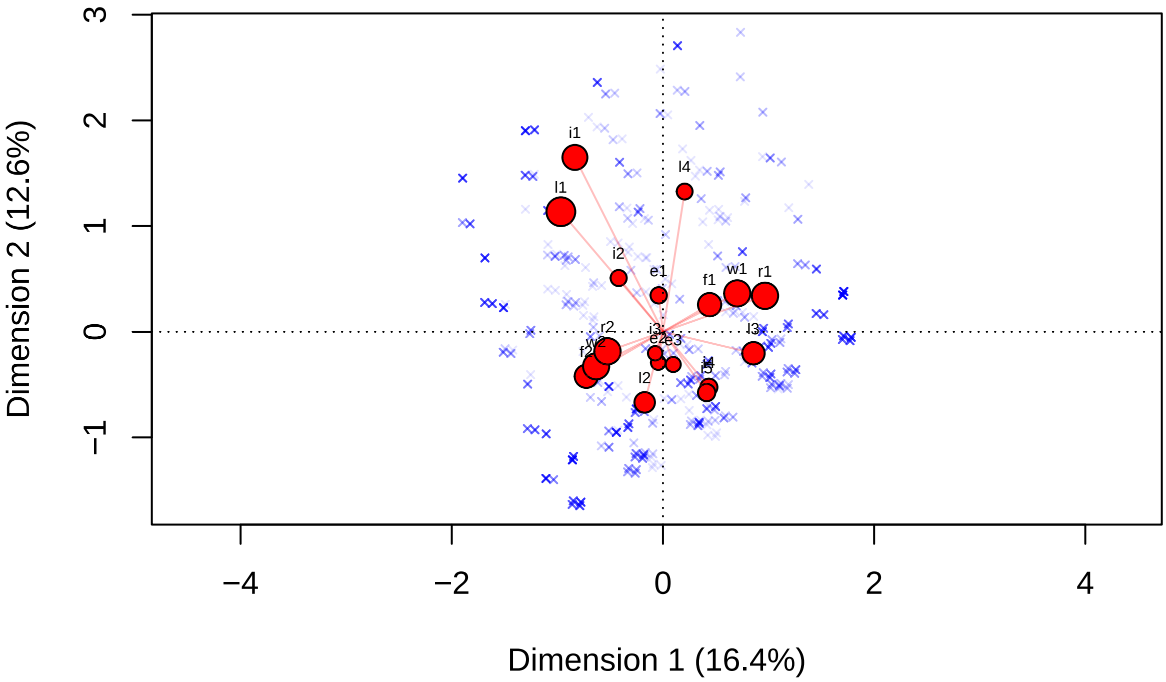

Figure 6: Columns in principal coordinates and rows in standard coordinates. For clarity we use shortened labels for modalities and points corresponding to households are transparent. Size of the points corresponding to modalities are scaled according to quality of representation.

Demo Problem 2: The Trace of Matrix V

Let V be the matrix defined as in lecture slides 7. Show that

$$\text{Trace}(V) = \frac{K}{P} - 1.$$

Message of the result: Total inertia in MCA does not depend on the data, but only on the number of modalities K and number of variables P.

Solution

First, let us review some notation

$n =$ sample size, $\quad K =$ total number of modalities, $\quad K_p =$ number of modalities of pth variable,

$P =$ number of qualitive variables, $\quad n_{pl} =$ number of individuals having modality l of variable Y_p,

$$x_{ipl} = \begin{cases} 1, & \text{if individual } i \text{ has modality } l \text{ of variable } Y_p \\ 0, & \text{otherwise} \end{cases}.$$

Notice that we have

$$\sum_{p=1}^{P}\sum_{l=1}^{K_p} x_{ipl} = P, \quad \sum_{i=1}^{n} x_{ipl} = n_{pl} \quad \text{and}$$
$$\sum_{p=1}^{P}\sum_{l=1}^{K_p} n_{pl} = nP.$$

Above relations will be useful in the proof. Remember also that matrix $T \in \mathbb{R}^{n\times K}$ is defined as

$$T = \begin{pmatrix} t_{1,1} & \cdots & t_{1,K} \\ \vdots & \ddots & \vdots \\ t_{n,1} & \cdots & t_{n,K} \end{pmatrix}, \quad \text{where} \quad t_{i,pl} = \frac{x_{ipl} - n_{pl}/n}{\sqrt{Pn_{pl}}}.$$

We have that $V = T^\intercal T$ and

$$T^\intercal = \begin{pmatrix} t_{1,1} & \cdots & t_{n,1} \\ \vdots & \ddots & \vdots \\ t_{1,K} & \cdots & t_{n,K} \end{pmatrix}.$$

Thus,

$$\operatorname{diag}(V) = \operatorname{diag}(T^\intercal T) = \begin{pmatrix} t_{1,1}^2 + t_{2,1}^2 + \cdots + t_{n,1}^2 \\ t_{1,2}^2 + t_{2,2}^2 + \cdots + t_{n,2}^2 \\ \vdots \\ t_{1,K}^2 + t_{2,K}^2 + \cdots + t_{n,K}^2 \end{pmatrix}.$$

Then,

$$\begin{aligned}
\operatorname{Trace}(V) &= \sum_{m=1}^{K}\sum_{i=1}^{n} t_{i,m}^2 = \sum_{i=1}^{n}\sum_{m=1}^{K} t_{i,m}^2 = \sum_{i=1}^{n}\sum_{p=1}^{P}\sum_{l=1}^{K_p} t_{i,pl}^2 \\
&= \sum_{i=1}^{n}\sum_{p=1}^{P}\sum_{l=1}^{K_p} \left(\frac{x_{ipl} - n_{pl}/n}{\sqrt{Pn_{pl}}}\right)^2 = \sum_{i=1}^{n}\sum_{p=1}^{P}\sum_{l=1}^{K_p} \left(\frac{x_{ipl}^2 - 2x_{ipl}\frac{n_{pl}}{n} + \frac{n_{pl}^2}{n^2}}{Pn_{pl}}\right) \\
&= \frac{1}{P}\sum_{i=1}^{n}\sum_{p=1}^{P}\sum_{l=1}^{K_p} \left(\frac{x_{ipl}^2}{n_{pl}} - 2\frac{x_{ipl}}{n} + \frac{n_{pl}}{n^2}\right).
\end{aligned}$$

Consider the terms of the sum separately. For the second term, we have

$$\frac{1}{P}\sum_{i=1}^{n}\sum_{p=1}^{P}\sum_{l=1}^{K_p} \left(-2\frac{x_{ipl}}{n}\right) = \frac{-2}{Pn}\sum_{i=1}^{n}\sum_{p=1}^{P}\sum_{l=1}^{K_p} x_{ipl} = \frac{-2}{Pn} nP = -2.$$

Likewise, for the third term we have

$$\frac{1}{P}\sum_{i=1}^{n}\sum_{p=1}^{P}\sum_{l=1}^{K_p} \frac{n_{pl}}{n^2} = \frac{1}{Pn^2}\sum_{i=1}^{n}\sum_{p=1}^{P}\sum_{l=1}^{K_p} n_{pl} = \frac{1}{Pn^2}\sum_{i=1}^{n} nP = 1.$$

The first term is the most difficult one here. Note that $x_{ipl} = x_{ipl}^2$, since $x_{ipl} \in \{0, 1\}$. By opening the sums we get

$$
\begin{aligned}
\frac{1}{P}\sum_{i=1}^{n}\sum_{p=1}^{P}\sum_{l=1}^{K_p}\frac{x_{ipl}}{n_{pl}} &= \frac{1}{P}\sum_{i=1}^{n}\sum_{p=1}^{P}\left(\frac{x_{ip1}}{n_{p1}}+\frac{x_{ip2}}{n_{p2}}+\cdots+\frac{x_{ipK_p}}{n_{pK_p}}\right)\\
&= \frac{1}{P}\sum_{i=1}^{n}\left(\frac{x_{i11}}{n_{11}}+\frac{x_{i12}}{n_{12}}+\cdots+\frac{x_{i1K_1}}{n_{1K_1}}+\frac{x_{i21}}{n_{21}}+\cdots+\frac{x_{iPK_P}}{n_{PK_P}}\right)\\
&= \frac{1}{P}\left(\frac{1}{n_{11}}\sum_{i=1}^{n}x_{i11}+\frac{1}{n_{12}}\sum_{i=1}^{n}x_{i12}+\cdots+\frac{1}{n_{PK_P}}\sum_{i=1}^{n}x_{iPK_P}\right)\\
&= \frac{1}{P}\left(\frac{n_{11}}{n_{11}}+\frac{n_{12}}{n_{12}}+\cdots+\frac{n_{PK_P}}{n_{PK_P}}\right) = \frac{K}{P}.
\end{aligned}
$$

By combining all the terms we get

$$
\mathrm{Trace}(V) = \frac{K}{P} - 2 + 1 = \frac{K}{P} - 1.
$$

Homework Problem 1: Multiple Correspondence Analysis

The data set `attitudes.txt` contains the attitudes of 871 individuals towards science and the environment. Each category contains five possible answers (strongly agree, somewhat agree, neither agree nor disagree, somewhat disagree, strongly disagree, coded as 1 to 5). The questions are:

A) We believe too often in science, and not enough in feelings and faith.

B) Overall, modern science does more harm than good.

C) Any change humans cause in nature – no matter how scientific – is likely to make things worse.

D) Modern science will solve our environmental problems with little change to our way of life.

In addition, the data set contains three demographic variables (sex, age and education). Variables age and education have 6 categories. For the variable age, 1 indicates that the individual belongs to the youngest age group. Likewise, for the variable education, 1 indicates the lowest level of education. Furthermore, the individuals are categorized as either 1 = male or 2 = female. Perform MCA using the indicator matrix. That is, remember to set value of the argument `lambda` to `"indicator"` in the function `mjca`. Provide the requested answers/figures.

a) Find a combination of two MCA components that explain as much of the variation as possible. What is the combination and how much of the total variation is explained by these two components?

b) Produce the MCA graph with respect to the components chosen in a).

c) What is the relationship between education and strong opinions (strongly agree/strongly disagree) in this data set? Justify!

References

Greenacre, Michael (2017). *Correspondence analysis in practice.* CRC press.