Multivariate Statistical Analysis - Exercise Session 7

24.02.2022

- Book "Correspondence analysis in practice" by Michael Greenacre gives an excellent introduction to correspondence analysis/multiple correspondence analysis. Book focuses on interpretations and provides many examples. Book also includes an appendix about the package `ca`. 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 CA/MCA in your project.

Problem 1: Multiple correspondence analysis

Read the data and import package `ca`.

```
library(ca)
tea <- read.table("TEA.txt", header = TRUE, sep = "\t")
dim(tea)
```

```
## [1] 300   6
```

```
head(tea)
```

```
##         Tea   How     how    sugar       where     always
## 1     black alone tea bag    sugar chain store Not.always
## 2     black  milk tea bag No.sugar chain store Not.always
## 3 Earl Grey alone tea bag No.sugar chain store Not.always
## 4 Earl Grey alone tea bag    sugar chain store Not.always
## 5 Earl Grey alone tea bag No.sugar chain store     always
## 6 Earl Grey alone tea bag No.sugar chain store Not.always
```

Now we perform *multiple correspondence analysis* (MCA) for the data set `tea` 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 *complete disjunctive table* (*indicator matrix*). We can perform this version of MCA by setting `lambda = "indicator"`. Argument `reti` controls whether the complete disjunctive table is returned.

```
tea_mca <- mjca(tea, lambda = "indicator", reti = TRUE)
names(tea_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"
```

Explanations for most of the returned objects are already explained on file `6session.pdf`. Objects `Burt`, `Burt.upd`, `subinertia` and `JCA.iter` are related to other definitions of MCA and are not relevant here.

By default `summary(tea_mca)` only gives summary for columns. By setting `rows = TRUE` one can see also summary for rows. More info can be found from the help pages `?summary.mjca`.

```r
s <- summary(tea_mca)
s
```

```
##
## Principal inertias (eigenvalues):
##
##  dim    value      %   cum%   scree plot
##  1      0.279762  15.3  15.3  ****
##  2      0.257748  14.1  29.3  ****
##  3      0.220138  12.0  41.3  ***
##  4      0.187930  10.3  51.6  ***
##  5      0.168765   9.2  60.8  **
##  6      0.163687   8.9  69.7  **
##  7      0.152888   8.3  78.1  **
##  8      0.138387   7.5  85.6  **
##  9      0.115692   6.3  91.9  **
##  10     0.086126   4.7  96.6  *
##  11     0.062211   3.4 100.0  *
##         -------- -----
##  Total: 1.833333 100.0
##
##
## Columns:
##                            name   mass  qlt  inr     k=1 cor ctr     k=2 cor
## 1  |                  Tea:black |   41   72   66 |  -446  65  29 |   143   7
## 2  |              Tea:Earl Grey |  107  135   32 |   250 113  24 |   111  22
## 3  |                  Tea:green |   18  144   74 |  -464  27  14 |  -974 117
## 4  |                  How:alone |  108  118   30 |    22   1   0 |  -251 117
## 5  |                  How:lemon |   18   84   75 |  -682  58  31 |   464  27
## 6  |                   How:milk |   35   43   64 |   331  29  14 |   229  14
## 7  |                  How:other |    5  144   82 |  -289   3   1 |  2141 142
## 8  |                how:tea bag |   94  639   45 |   616 497 128 |  -329 142
## 9  |     how:tea bag+unpackaged |   52  519   66 |  -371  63  26 |  1001 457
## 10 |             how:unpackaged |   20  667   92 | -1943 515 270 | -1057 152
## 11 |             sugar:No.sugar |   86   62   41 |  -238  60  17 |    40   2
## 12 |                sugar:sugar |   81   62   44 |   254  60  19 |   -42   2
## 13 |          where:chain store |  107  715   38 |   533 506 108 |  -343 209
## 14 | where:chain store+tea shop |   43  705   75 |  -481  81  36 |  1333 624
## 15 |             where:tea shop |   17  699   95 | -2164 520 279 | -1269 179
## 16 |              always:always |   57   13   53 |  -109   6   2 |   118   7
## 17 |          always:Not.always |  109   13   28 |    57   6   1 |   -62   7
##    ctr
## 1    3 |
## 2    5 |
## 3   67 |
## 4   27 |
## 5   15 |
## 6    7 |
## 7   89 |
## 8   40 |
## 9  203 |
## 10  87 |
## 11   1 |
## 12   1 |
```

```
## 13  49 |
## 14 299 |
## 15 104 |
## 16   3 |
## 17   2 |
```

Figure 1 shows that only 29.3% of variation is explained by the first two components. Nevertheless, we proceed to analyze the first two components.

```
barplot(s$scree[, 3], ylim = c(0, 20), names.arg = paste("PC", 1:11), las = 2,
        xlab = "Component", ylab = "% of variation explained", col = "skyblue")
```

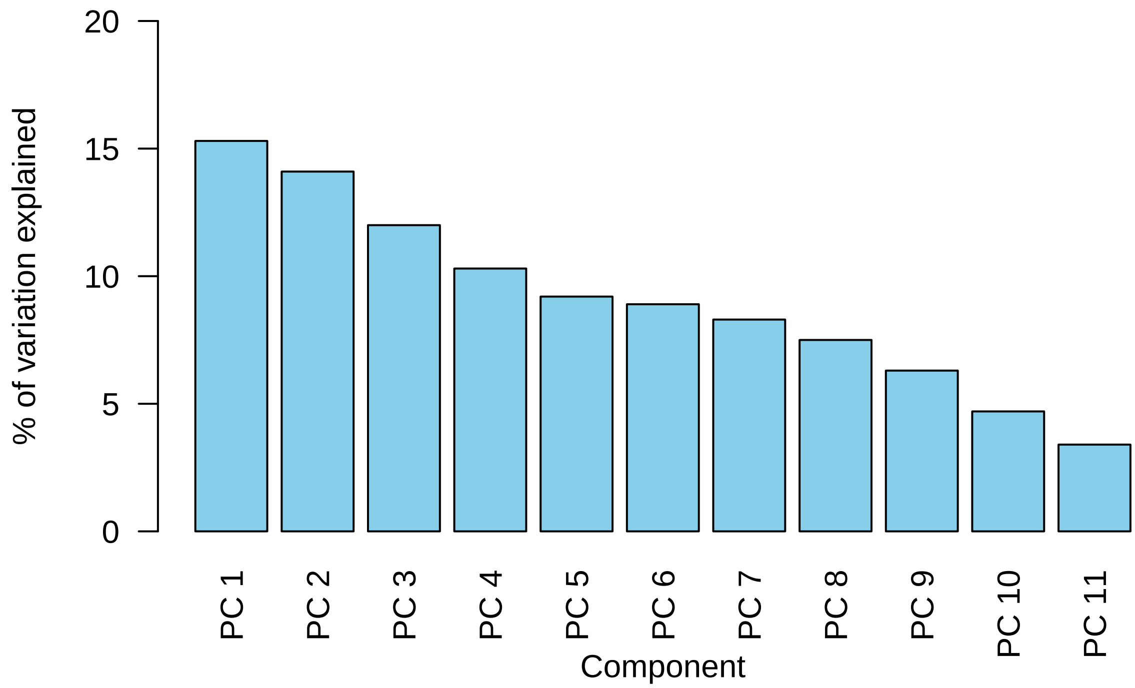

Figure 1: Scree plot.

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

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

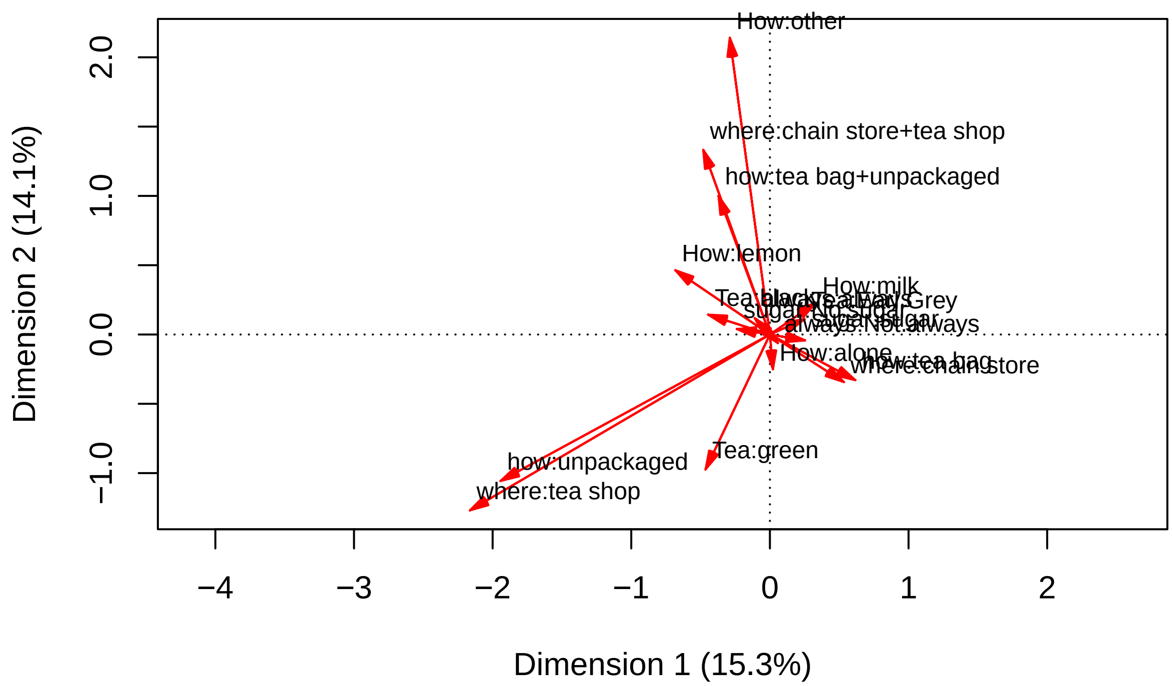

Figure 2: First two column principal coordinates.

From relation

$$d_{p_1 l_1, p_2 l_2} \approx 1 + \sum_{h=1}^{2} \psi_{h,p_1 l_1} \psi_{h,p_2 l_2}$$

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.

Remember that interpretations are only valid when modalities are represented well in two dimensions. Thus we could modify Figure 2 in such a way that point size represents quality of representation of corresponding modality.

```
# 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 <- s$columns[, 3]
tea_covariates <- tea_mca$colpcoord[, 1:2]
plot(tea_covariates, xlim = c(-2.5, 1), ylim = c(-1.5, 2.5), pch = 21,
     bg = "red", cex = normalize(qlt) + 1,
```

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

# Add arrows. Slight transparency is added to increase visibility.
arrows(rep(0, 17), rep(0, 17), tea_covariates[, 1], tea_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(tea_covariates, tea_mca$levelnames, pos = 2, cex = 0.75)
```

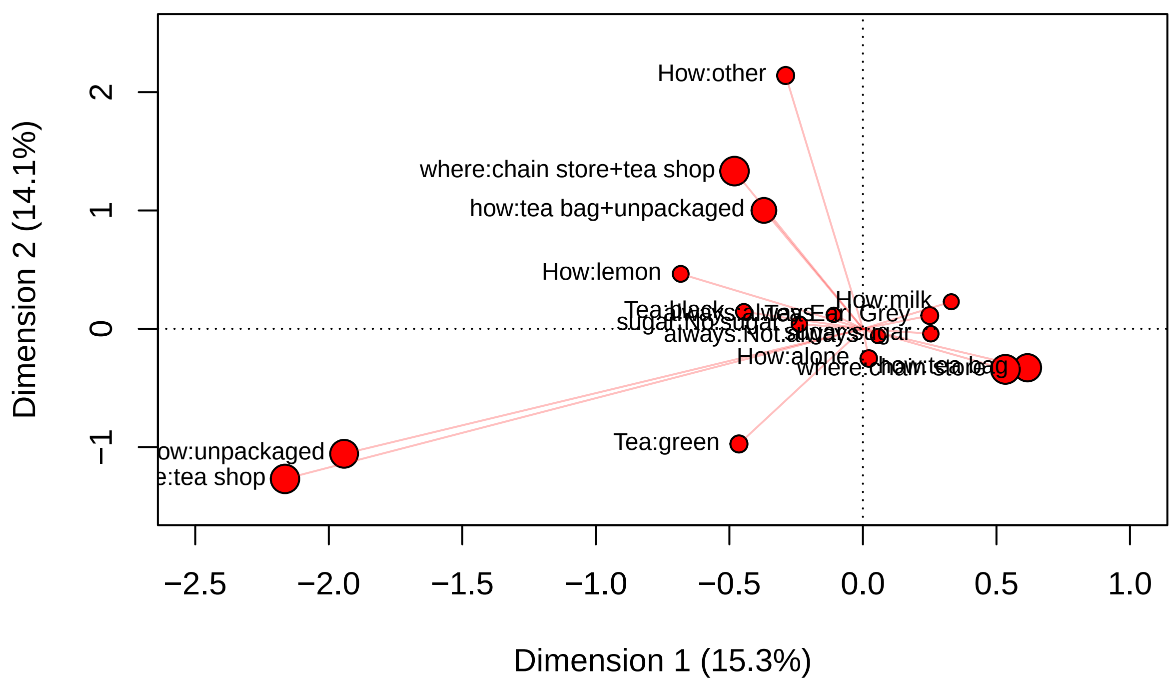

Figure 3: First two column principal coordinates. Point sizes are scaled according to quality of representation.

Figure 4 illustrates that MCA is just CA performed for complete disjunctive table. That is, Figures 2 and 4 are identical.

```
plot(ca(tea_mca$indmat), arrows = c(FALSE, TRUE), what = c("none", "all"))
```

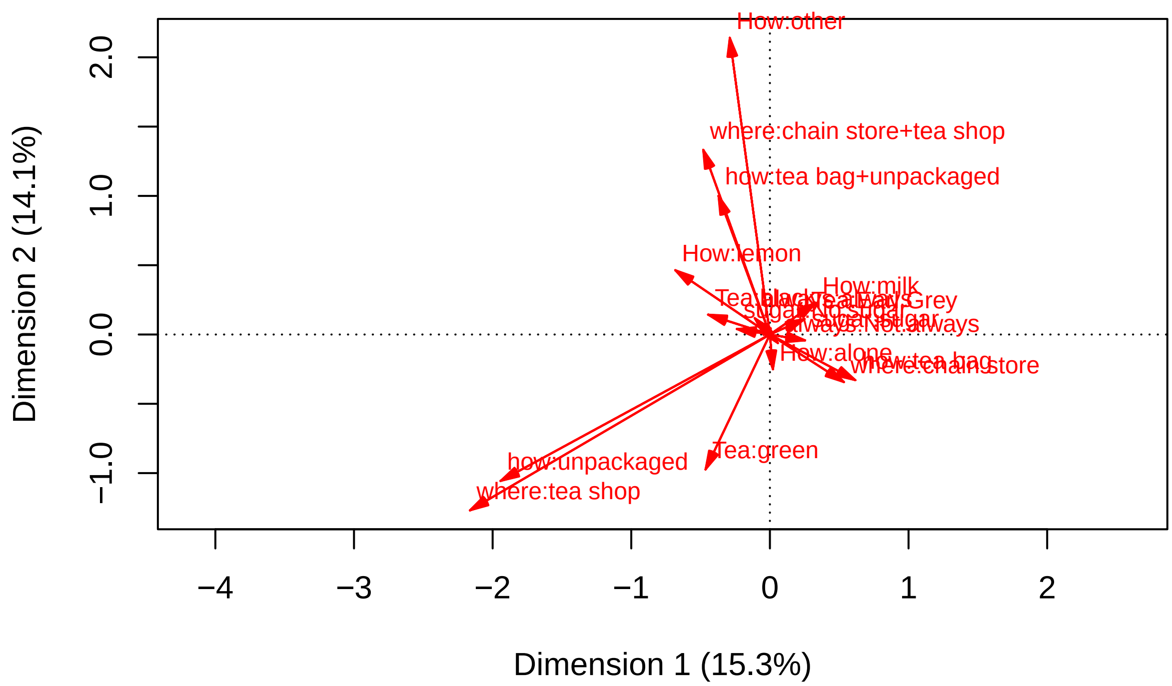

Figure 4: Correspondence analysis performed for complete disjunctive table.

Rows can be analyzed similarly to columns. 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 5:

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

For the sake of clarity, observation labels are dropped from Figure 5 and instead of arrows we have points.

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

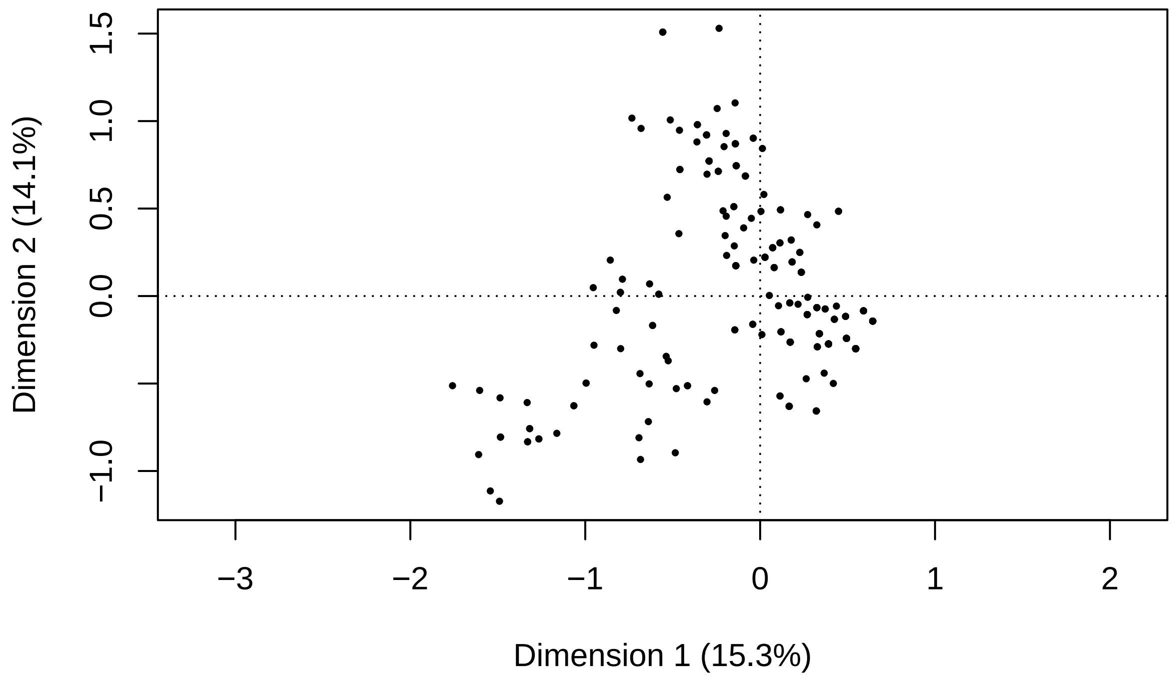

Figure 5: First two row principal coordinates.

Lastly, relation

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

gives interpretation for Figure 6. Notice that since columns are in principal coordinates we can also interpret angles between columns in Figure 6 as in Figure 2.

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

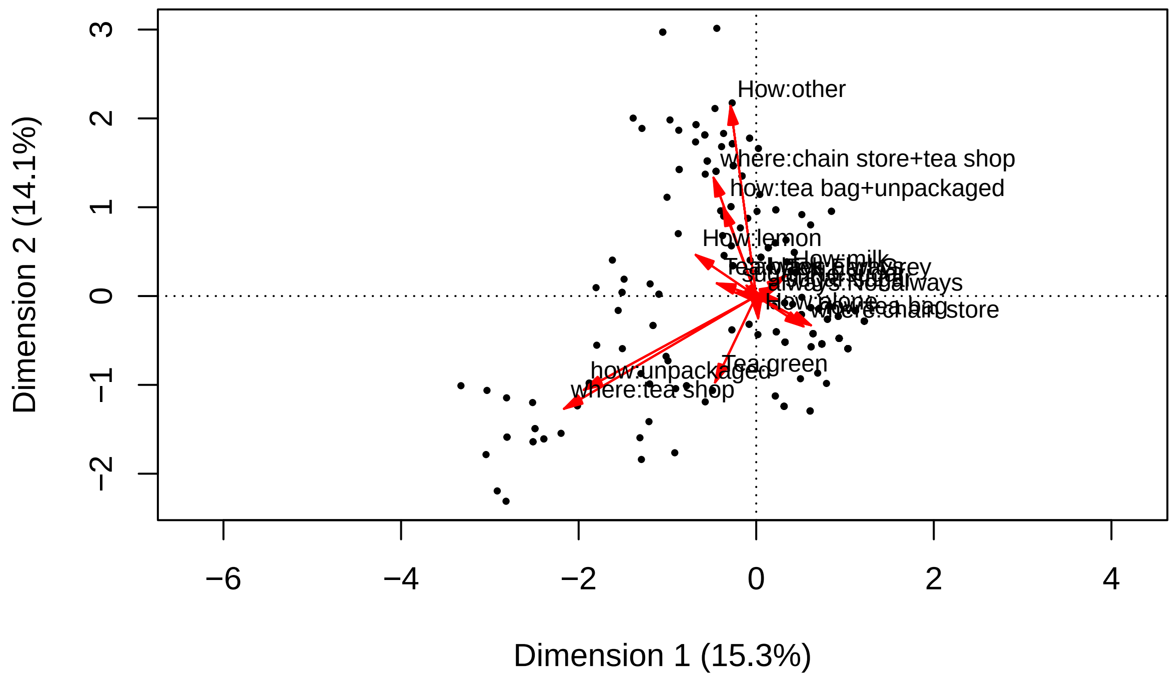

Figure 6: Columns in principal coordinates and rows in standard coordinates.

Problem 2: The trace of matrix V

First, let us review some notation

$$n = \text{sample size}, \quad K = \text{total number of modalities}, \quad K_p = \text{number of modalities of } p\text{th variable},$$
$$P = \text{number of qualitive variables}, \quad n_{pl} = \text{number of individuals having modality } l \text{ of variable } Y_p,$$
$$x_{ipl} = \begin{cases} 1, & \text{if individual } i \text{ has modality } l \text{ of variable } Y_p \\ 0, & \text{otherwise} \end{cases}.$$

Notice that we have

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

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

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

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

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

Thus

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

Then,

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

Then 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

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