Multivariate Statistical Analysis - Exercise Session 6

15.02.2022

- Book "Correspondence analysis in practice" by Michael Greenacre gives an excellent introduction to 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 in your project.

Problem 1: Bivariate correspondence analysis

Install package `ca` if you haven't yet. With package `ca` we can perform correspondence analysis.

```
install.packages("ca")
```

Next, read the data and import package `ca`.

```
library(ca)
science <- read.table("SCIENCEDOCTORATES.txt", header = TRUE, sep = "\t",
                      row.names = 1)
science <- science[-nrow(science), -ncol(science)]
science
```

```
##                Y1960 Y1965 Y1970 Y1971 Y1972 Y1973 Y1974 Y1975
## Engineering      794  2073  3432  3495  3475  3338  3144  2959
## Mathematics      291   685  1222  1236  1281  1222  1196  1149
## Physics          530  1046  1655  1740  1635  1590  1334  1293
## Chemistry       1078  1444  2234  2204  2011  1849  1792  1762
## EarthSciences    253   375   511   550   580   577   570   556
## Biology         1245  1963  3360  3633  3580  3636  3473  3498
## Agriculture      414   576   803   900   855   853   830   904
## Psychology       772   954  1888  2116  2262  2444  2587  2749
## Sociology        162   239   504   583   638   599   645   680
## Economics        341   538   826   791   863   907   833   867
## Anthropology      69    82   217   240   260   324   381   385
## Others           314   502  1079  1392  1500  1609  1531  1550
```

χ^2-test

As a first step, we perform χ^2-test. Null hyphothesis and alternative hyphothesis of χ^2-test are

$$H_0 : \text{Discipline and year are independent,}$$
$$H_1 : \text{Discipline and year are not independent.}$$

First, we perform χ^2-test manually.

```
n <- sum(science)
v1 <- matrix(rowSums(science), ncol = 1)
v2 <- matrix(colSums(science), nrow = 1)
```

```r
# Theoretical frequencies under independence
e <- v1 %*% v2 / n

# Calculate chi-squared statistic
chisq_statistic <- sum((science - e)^2 / e)
chisq_statistic
```

```
## [1] 1686.083
```

```r
# p-value
i <- nrow(science)
j <- ncol(science)
pchisq(chisq_statistic, df = (i - 1) * (j - 1), lower.tail = FALSE)
```

```
## [1] 4.825946e-301
```

Of course, R provides a function for performing χ^2-test.

```r
chisq.test(science)
```

```
##
##  Pearson's Chi-squared test
##
## data:  science
## X-squared = 1686.1, df = 77, p-value < 2.2e-16
```

P-value is really small which suggest that there is dependence between rows and columns.

Correspondence analysis with `ca` package

```r
science_ca <- ca(science)
names(science_ca)
```

```
##  [1] "sv"         "nd"         "rownames"   "rowmass"    "rowdist"
##  [6] "rowinertia" "rowcoord"   "rowsup"     "colnames"   "colmass"
## [11] "coldist"    "colinertia" "colcoord"   "colsup"     "N"
## [16] "call"
```

Now let's go through the relevant parts of the output.

1. `sv`: Singular values of the matrix Z from the lecture slides. Let σ_i be a singular value of Z. Then σ_i^2 is an eigenvalue of the matrix $V = Z^T Z$ and matrix $W = ZZ^T$. Sum of eigenvalues equals to χ^2/n, where χ^2 is the chi-squared test statistic. This value χ^2/n is called *inertia* and it describes how much row/column profiles deviate from their average row/column profile. Eigenvalues also describe how much of the variation is explained by components.

```r
round(sum(science_ca$sv^2), 2) == round(chisq_statistic / n, 2)
```

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

```r
science_ca$sv^2 / sum(science_ca$sv^2)
```

```
## [1] 0.705490525 0.247117034 0.022330556 0.014552053 0.005123493 0.003222065
## [7] 0.002164275
```

2. `rowmass`/`colmass`: Average column profile/row profile

```r
all(round(science_ca$rowmass, 2) == round(rowSums(science) / sum(science), 2))
```

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

```
all(round(science_ca$colmass, 2) == round(colSums(science) / sum(science), 2))
```

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

3. `rowdist`/`coldist` and `rowinertia`/`colinertia`: χ^2 distances between rows/columns to average row/column profiles (centroid) are given by `rowdist`/`coldist`. Large χ^2 distance indicates that corresponding row/column deviates from the average profile. Below we calculate χ^2 distance between first row profile and average row profile manually. Also, we show that inertia is weighted sum of squared χ^2 distances.

$$\frac{\chi^2}{n} = \sum_i \underbrace{(i\text{th row mass}) \times (\chi^2 \text{ distance from } i\text{th profile to centroid})^2}_{i\text{th row inertia}}.$$

```
# Chi-squared distances
science_ca$rowdist
```

```
##  [1] 0.08952388 0.07049341 0.11028846 0.16957446 0.08827503 0.01235856
##  [7] 0.10393910 0.13970921 0.12706869 0.06346696 0.27380295 0.16970149
```

```
rowprof <- proportions(as.matrix(science), 1)
sqrt(sum((rowprof[1, ] - science_ca$colmass)^2 / science_ca$colmass))
```

```
## [1] 0.08952388
```

```
# Inertia is equal to the sum of row inertias
sum(science_ca$rowinertia)
```

```
## [1] 0.0131821
```

```
chisq_statistic / sum(science)
```

```
## [1] 0.0131821
```

```
# Row inertias are weighted chi-squared distances
science_ca$rowmass * science_ca$rowdist^2
```

```
##  [1] 1.422986e-03 3.217643e-04 1.029233e-03 3.231501e-03 2.419862e-04
##  [6] 2.912176e-05 5.181770e-04 2.406814e-03 5.112553e-04 1.878816e-04
## [11] 1.147611e-03 2.133771e-03
```

```
science_ca$rowinertia
```

```
##  [1] 1.422986e-03 3.217643e-04 1.029233e-03 3.231501e-03 2.419862e-04
##  [6] 2.912176e-05 5.181770e-04 2.406814e-03 5.112553e-04 1.878816e-04
## [11] 1.147611e-03 2.133771e-03
```

4. `rowcoord`/`colcoord`: Standardized row and column coordinates or scores.

Correspondence analysis manually

First let us form the matrix Z, V and W from the lecture slides.

```
# Observed frequencies
f <- science / n

# Theoretical relative frequencies under independence
f1 <- matrix(rowSums(f), ncol = 1)
f2 <- matrix(colSums(f), nrow = 1)
e <- f1 %*% f2
```

```r
# Matrices Z, V and W
z <- (f - e) / sqrt(e)
z <- as.matrix(z)
v <- t(z) %*% z
w <- z %*% t(z)

# Save nonzero eigenvalues, and eigenvectors of V and W
princip_inertia <- matrix(eigen(v)$values[1:7], nrow = 1)
u1 <- eigen(v)$vectors
u2 <- eigen(w)$vectors
```

Then let us check that eigenvalues of V and W correspond to principal inertias.

```r
eigen(v)$values
```

```
## [1]  9.299848e-03  3.257522e-03  2.943637e-04  1.918266e-04  6.753840e-05
## [6]  4.247359e-05  2.852970e-05 -7.500213e-19
```

```r
eigen(w)$values
```

```
##  [1]  9.299848e-03  3.257522e-03  2.943637e-04  1.918266e-04  6.753840e-05
##  [6]  4.247359e-05  2.852970e-05  4.122024e-19  4.981311e-20  1.933726e-21
## [11] -2.545800e-19 -1.930096e-18
```

```r
science_ca$sv^2
```

```
## [1] 9.299848e-03 3.257522e-03 2.943637e-04 1.918266e-04 6.753840e-05
## [6] 4.247359e-05 2.852970e-05
```

Notice that V and W have the same nonzero eigenvalues. Next let us form matrices R and C from lecture slides

```r
# Matrix R
one <- matrix(rep(1, nrow(science)), ncol = 1)
shifting <- one %*% sqrt(f2)
scaling <- f1 %*% sqrt(f2)
r <- f / scaling - shifting
r <- as.matrix(r)

# Matrix C
one <- matrix(rep(1, ncol(science)), nrow = 1)
shifting <- sqrt(f1) %*% one
scaling <- sqrt(f1) %*% f2
c <- f / scaling - shifting
c <- as.matrix(c)
```

Now we can calculate row and column principal coordinates (scores).

```r
# Row principal coordinates
rowcoord <- r %*% u1
#omit the dimension corresponding to zero eigenvalue
rowcoord <- rowcoord[, 1:7]

# Column principal coordinates
colcoord <- t(c) %*% u2
# Omit dimensions corresponding to zero eigenvalues
colcoord <- colcoord[, 1:7]
```

Below we calculate so called standardized coordinates. These coordinates correspond to ones returned by function `ca`.

```
# Standardized rowcoordinates
stand <- matrix(rep(1, nrow(science), ncol = 1)) %*% princip_inertia
standrowcoord <- rowcoord / sqrt(stand)

standrowcoord[, 1]
```

```
##   Engineering   Mathematics       Physics     Chemistry EarthSciences
##  -0.322438878   0.077507427  -1.011402351  -1.585369258  -0.425654104
##       Biology   Agriculture    Psychology     Sociology     Economics
##  -0.008109185  -0.570837241   1.321263122   1.254532552  -0.233745056
##  Anthropology        Others
##   2.719309874   1.683657551
```

```
science_ca$rowcoord[, 1]
```

```
##   Engineering   Mathematics       Physics     Chemistry EarthSciences
##   0.322438878  -0.077507427   1.011402351   1.585369258   0.425654104
##       Biology   Agriculture    Psychology     Sociology     Economics
##   0.008109185   0.570837241  -1.321263122  -1.254532552   0.233745056
##  Anthropology        Others
##  -2.719309874  -1.683657551
```

```
# Standardized colcoordinates
stand2 <- matrix(rep(1, ncol(science), ncol = 1)) %*% princip_inertia
standcolcoord <- colcoord / sqrt(stand2)

standcolcoord[, 1]
```

```
##      Y1960      Y1965      Y1970      Y1971      Y1972      Y1973      Y1974
## -1.6106625 -1.9439873 -0.9299411 -0.3975732  0.1056064  0.5886455  1.0136835
##      Y1975
##  1.2385154
```

```
science_ca$colcoord[, 1]
```

```
##      Y1960      Y1965      Y1970      Y1971      Y1972      Y1973      Y1974
##  1.6106625  1.9439873  0.9299411  0.3975732 -0.1056064 -0.5886455 -1.0136835
##      Y1975
## -1.2385154
```

Interpretation and `summary(ca)`

Summary gives some additional information such as quality of representation and contributions of modalities to each axis. Below we go through some relevant parts of the summary.

```
summary(science_ca)
```

```
##
## Principal inertias (eigenvalues):
##
##  dim    value      %   cum%   scree plot
##  1      0.009300  70.5  70.5  ******************
##  2      0.003258  24.7  95.3  ******
##  3      0.000294   2.2  97.5  *
```

```
##  4      0.000192   1.5  98.9
##  5      6.8e-050   0.5  99.5
##  6      4.2e-050   0.3  99.8
##  7      2.9e-050   0.2 100.0
##         -------- -----
##  Total: 0.013182 100.0
##
##
## Rows:
##      name   mass  qlt  inr    k=1 cor ctr    k=2 cor ctr
## 1  | Engn |  178  971  108 |   31 121  18 |  -83 851 372 |
## 2  | Mthm |   65  928   24 |   -7  11   0 |  -67 916  91 |
## 3  | Phys |   85  956   78 |   98 782  87 |  -46 174  55 |
## 4  | Chms |  112  985  245 |  153 813 282 |   70 172 171 |
## 5  | ErtS |   31  736   18 |   41 216   6 |   64 520  39 |
## 6  | Blgy |  191  485    2 |    1   4   0 |    9 481   4 |
## 7  | Agrc |   48  907   39 |   55 281  16 |   82 627 100 |
## 8  | Psyc |  123  993  183 | -127 832 215 |   56 161 119 |
## 9  | Sclg |   32  907   39 | -121 906  50 |    1   0   0 |
## 10 | Ecnm |   47  498   14 |   23 126   3 |   39 371  21 |
## 11 | Anth |   15  949   87 | -262 917 113 |   49  32  11 |
## 12 | Othr |   74  942  162 | -162 915 210 |  -28  27  18 |
##
## Columns:
##      name   mass  qlt  inr    k=1 cor ctr    k=2 cor ctr
## 1 | Y1960 |   49  995  270 |  155 332 127 |  220 663 725 |
## 2 | Y1965 |   82  961  229 |  187 952 310 |  -18   9   9 |
## 3 | Y1970 |  139  929  107 |   90 789 120 |  -38 139  60 |
## 4 | Y1971 |  148  744   37 |   38 449  23 |  -31 295  44 |
## 5 | Y1972 |  148  755   19 |  -10  62   2 |  -34 694  53 |
## 6 | Y1973 |  148  830   48 |  -57 758  51 |  -17  72  14 |
## 7 | Y1974 |  143  962  110 |  -98 946 147 |   13  16   7 |
## 8 | Y1975 |  143  983  180 | -119 862 220 |   45 121  88 |
```

1. `ctr`: Contributions of modalities to the construction of the ith CA axis. Below we calculate contribution of modality `Engineering` to the second principal axis.

```
1000 * f1[1] * (rowcoord[1, 2])^2 / princip_inertia[2]
```

```
## Engineering
##    371.6586
```

2. `cor` and `qlt`: Column `cor` gives QLT of rows/columns with respect to ith axis. However, column `qlt` gives quality of representation with respect to the plane spanned by the first two axes. Below we calculate quality of representation of row `Engineering` by second CA axis.

```
1000 * (rowcoord[1, 2])^2 / science_ca$rowdist[1]^2
```

```
## Engineering
##    850.8066
```

3. `k=i`: Principal row/column coordinates.
4. `inr`: Row/column inertias divided by the total inertia.

Function `plot.ca` provides many options for scaling.

1. `map == "symmetric"`: Here one *cannot* make any conclusions based on the distances between columns

and rows. However, row-to-row distances approximate χ^2 distances between rows. Also, row-to-origin distances approximate χ^2 distances between corresponding row and the average row profile. Same interpretations can be made for col-to-col distances.

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

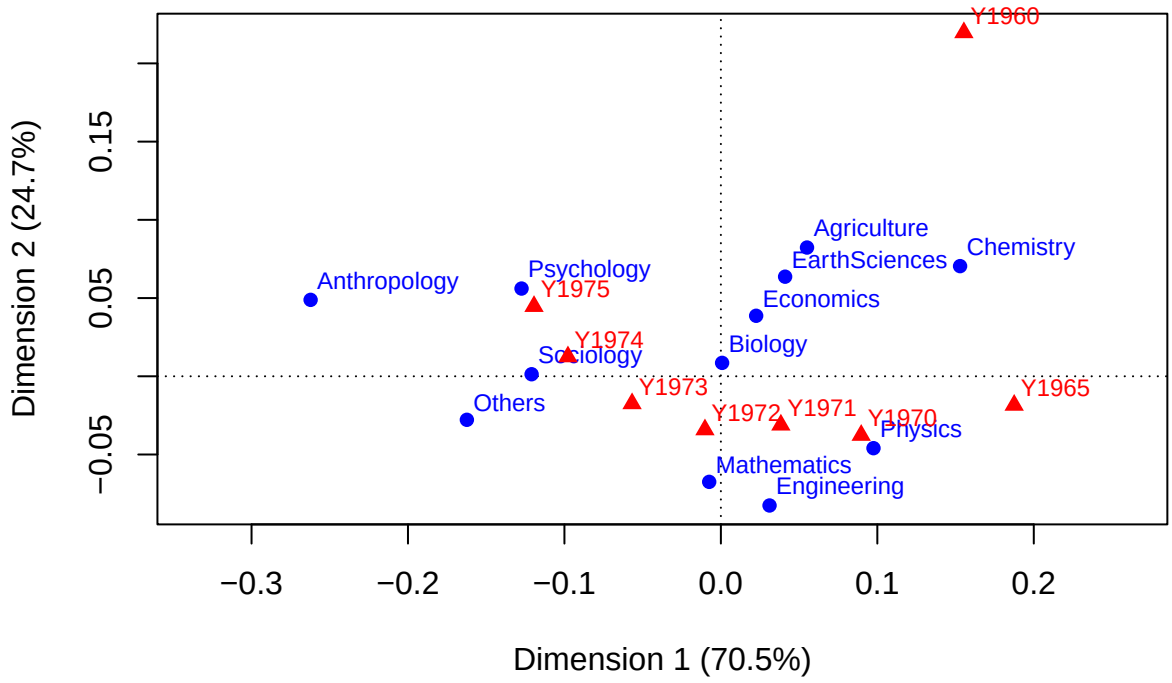

Figure 1: `map == "symmetric"`

2. `map == "rowprincipal"`: Here rows are in principal coordinates (representing row profiles) and columns are in standard coordinates (representing column vertices). Here row-to-row distances approximate χ^2 distances between rows. Distances between rows and columns describe row profiles. Closer a profile is to vertex, the higher its profile is for that category, assuming the first two axes account for large proportion of inertia.

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

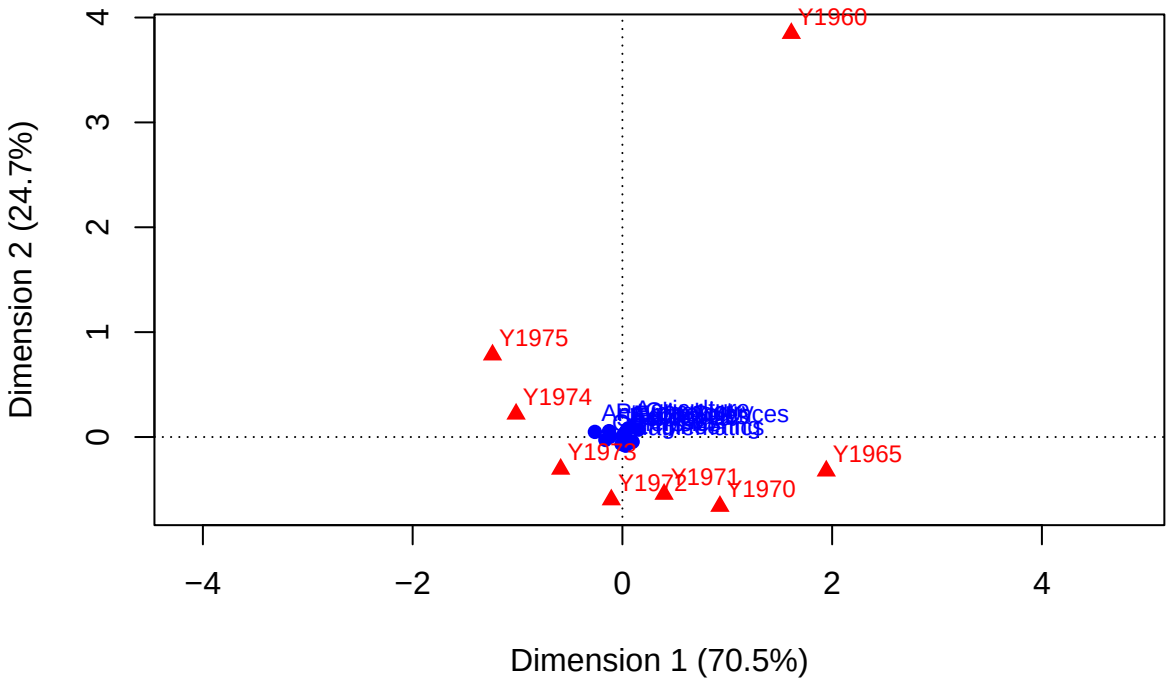

Figure 2: `map == "rowprincipal"`

As Figure 2 shows, the disadvantage of `rowprincipal/colprincipal` plot is that profiles points tend to get clustered in the middle (especially when inertia is small) and the visualization can be hard to read.

`map == "rowgreen"`/`map == "rowgab"`: Rows are in principal coordinates and columns are in standard coordinates multiplied by the square root of the mass/the mass of the corresponding point. Again, row-to-row distances approximate χ^2 distances between rows. There is no interpretation for row-to-column distances. However, under 90 degree angle between row and column implies attraction between modalities and over 90 degree angles imply repulsion between modalities, assuming good quality of display. Point of options `"rowgreen"`/`"rowgab"` is to scale vertices in such a way that biplot becomes more readable compared to `rowprincipal`. Note that with the option `rowprincipal` angles can be interpreted similarly as with options `"rowgreen"`/`"rowgab"`.

```
plot(science_ca, map = "rowgreen", arrows = c(FALSE, TRUE))
```

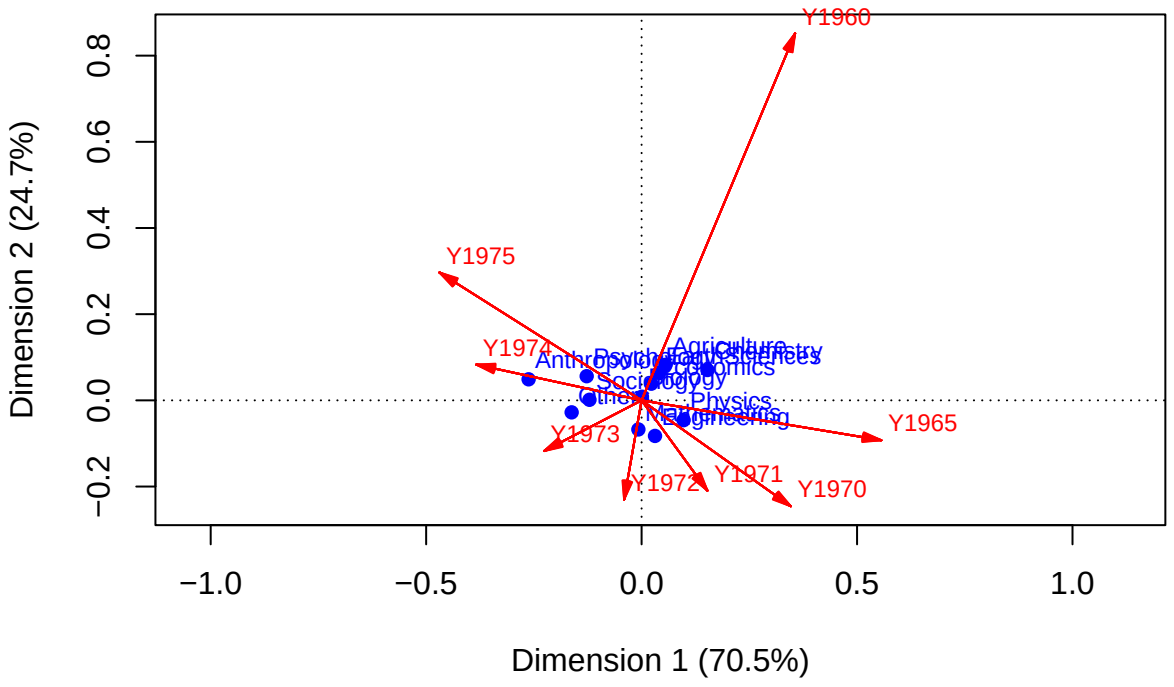

Figure 3: `map == "rowgreen"`

```
plot(science_ca, map = "rowgab", arrows = c(FALSE, TRUE))
```

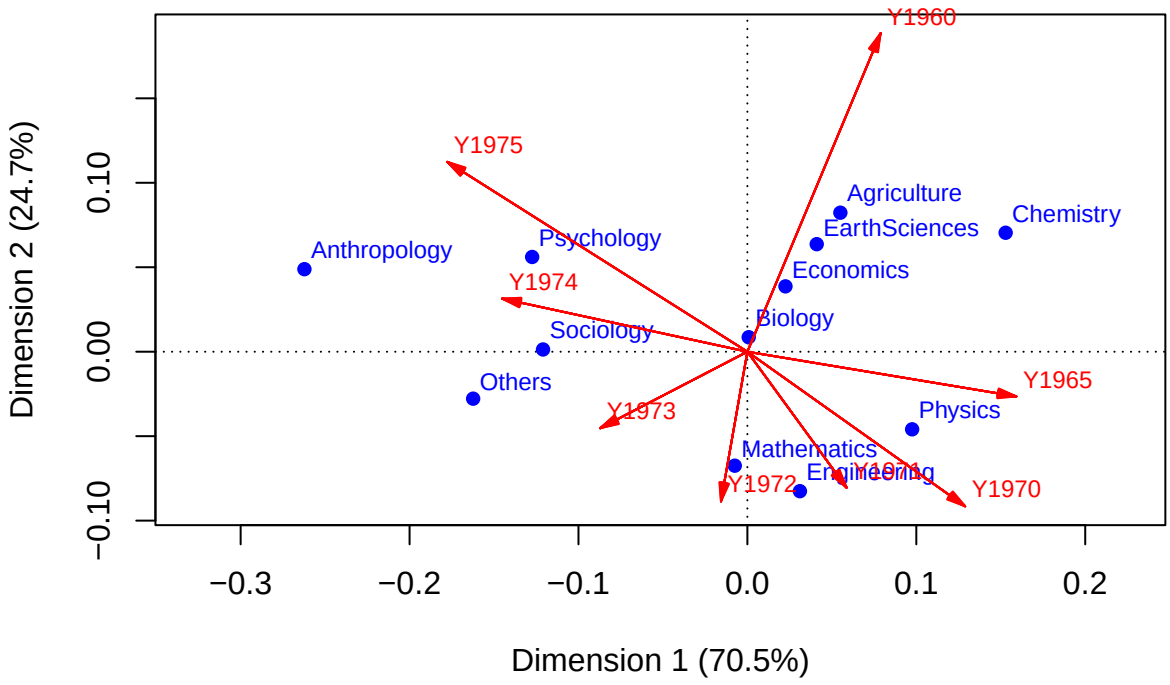

Figure 4: `map == "rowgab"`

`colprincipal`, `colgab`, `colgreen` can be interpreted similarly as `rowprincipal`, `rowgab`, `rowgreen` respectively, but from the viewpoint of column profiles.

From above plots one can try to interpret two first CA axes:

- 1st axis: Hard sciences vs. soft sciences
- 2nd axis: Formula heavy sciences vs. experimental sciences

Problem 2: Association between column and row profiles

First, we show that Z^TZ and ZZ^T have the same eigenvalues. From the definition of an eigenvalue and -vector we get

$$Vv_i = Z^TZv_i = \lambda_i v_i \tag{1}$$

and

$$Ww_i = ZZ^Tw_i = \mu_i w_i. \tag{2}$$

Multiply Equation (1) with Z from the left side and note that $V = Z^TZ$,

$$\Rightarrow ZVv_i = Z\underbrace{Z^TZv_i}_{=\lambda_i v_i} = \lambda_i Zv_i.$$

Thus

$$\begin{aligned} &ZZ^T(Zv_i) = \lambda_i(Zv_i) \\ &\Rightarrow ZZ^Tv_i^* = \lambda_i v_i^*, \quad \text{where } v_i^* = Zv_i. \end{aligned}$$

Hereby, λ_i is the eigenvalue of $ZZ^T = W$ with eigenvector Zv_i. The squared length of the eigenvector is given by

$$\|Zv_i\|_2^2 = (Zv_i)^T(Zv_i) = v_i^T \underbrace{Z^T Zv_i}_{=\lambda_i v_i} = \lambda_i v_i^T v_i = \lambda_i.$$

Thus

$$w_i = \frac{1}{\|Zv_i\|_2} Zv_i = \frac{1}{\sqrt{\lambda_i}} Zv_i.$$

The same proof goes to the other direction also. Just start by multiplying Equation of (2) by Z^T and proceed similarly.