Practicals 2 : CA and MCA

1 Preliminaries

The goals of this practical session are the construction of the correspondence analysis (CA) that we have introduced during the lectures and the introduction of the multiple correspondence analysis (MCA).

To make easier the recovery of the data and to have at our disposal some graphical commands in the sequel, we start by loading the script *tp2.R*,

```
source("http://www.math.univ-toulouse.fr/~xgendre/ens/m2se/tp2.R")
```

2 Studies and socio-professional categories (CA)

We begin with the data used in the work of the french sociologist Pierre Bourdieu,

```
T <- DataBourdieu()
```

The contingency table `T` contains the numbers of students involved in the poll. On the lines, we have the several socio-professional categories for the head of the family (`EAG` - farm-owner, `SAG` - farm-worker, `PT` - boss, `PLCS` - private profession and senior executive, `CM` - executive, `EMP` - employee, `OUV` - worker et `AUT` - other). In the columns, we have the field of the studies (`DR` - law, `SCE` - economics, `LET` - literatur, `SC` - sciences, `MD` - medicine or dental, `PH` - pharmacy, `PD` - multidisciplinary et `IUT` - university institute of technology).

The first step consists in computing the line and column profiles associated to this data set. We also give a name to the rows to make easier the next graphical representations.

```
P1 <- diag(1/rowSums(T)) %*% T
rownames(P1) <- rownames(T)
P2 <- diag(1/colSums(T)) %*% t(T)
rownames(P2) <- colnames(T)
```

We compute the PCA of the line profiles as we did in Practicals 1. We get the weight matrix $W_1 = \mathrm{diag}(n_{1\cdot}/n, \ldots, n_{p\cdot}/n)$ and the matrix $M_1 = \mathrm{diag}(n/n_{\cdot 1}, \ldots, n/n_{\cdot q})$. Execute the following commands and explain why what they do is equivalent to this PCA.

```
n <- sum(rowSums(T))
M1half <- diag(sqrt(n/colSums(T)))
X <- P1 %*% M1half
ACP1 <- eigen(t(X) %*% diag(rowSums(T)/n) %*% X)
```

What is the value of `ACP1$values[1]` ? Remind why it is important to follow the following steps,

```
k <- min(nrow(T), ncol(T))
ACP1$values <- ACP1$values[2:k]
ACP1$vectors <- ACP1$vectors[, 2:k]
C1 <- X %*% ACP1$vectors
```

With the help of the transition formulae, we can directly obtain the PCA of the column profiles. Thus, we get the principal component matrix associated to this second PCA.

```
C2 <- P2 %*% C1 %*% diag(1/sqrt(ACP1$values))
```

Compute the inertia explained by the principal plane and the quality of the representations of each socio-professional category and kind of studies. Comment these values. The graphical representation of the CA can be done with the help of the imported command `PlotAFC`,

```
PlotAFC(C1, C2)
```

Comment and interpret these results.

A library for the CA

There exist several tools to compute and represent the CA. Among them, the R package `ca` is often used and is easy to handle. If this package is not installed on your computer, you can get it with the command `install.packages("ca")`. Thus, the following commands lead to the results of the CA and to the graphics,

```
library(ca)
AFC <- ca(as.table(T))
print(AFC)
summary(AFC)
plot(AFC)
```

3 French presidential elections of 1995 (CA)

The second example of CA that we consider in these practicals is related to the results of the first round of the French presidential elections of 1995. For each departement (the two Corsican departments are aggregated), we have the number of entries, the number of voters, the number of votes cast and the number of votes received by each candidate. Load these data set and briefly glance them.

```
data <- DataElection1995()
```

The contingency table to study is given by the number of votes received by each candidate in each department.

```
T <- as.matrix(data[, 4:12])
```

Compute the CA and comment the obtained results. Pay attention to the particular case of the Vendée (department 85). Check the line `data[85,]` to help you to understand. Compute again the CA without this department using the contingency table given by `as.matrix(data[-85,4:12])`.

4 MCA : a bit of theory

The multiple correspondence analysis (MCA) is the generalization of the CA for a number of qualitative variables higher than two. We assume that we have at our disposal $m > 2$ qualitative variables $x^1, \ldots, x^m$ that we observe on n individuals. For each $k \in \{1, \ldots, m\}$, the variable x^k can take c_k distinct values denoted by $1, \ldots, c_k$ and c denotes the total number of modes,

$$c = \sum_{k=1}^{m} c_k \ .$$

For each $i \in \{1,\dots,n\}$, $x_i^k \in \{1,\dots,c_k\}$ is the value of the variable x^k observed for the i-th individual.

For each mode of each variable, we consider the associated indicator vector, *i.e.* for each $k \in \{1,\dots,m\}$ and each $\ell \in \{1,\dots,c_k\}$,

$$\mathbb{I}_{m\ell} = \begin{pmatrix} \mathbf{1}_{x_1^k=\ell} \\ \vdots \\ \mathbf{1}_{x_n^k=\ell} \end{pmatrix} \in \{0,1\}^n \ .$$

If you put these vectors in the columns of a matrix X of size $n \times c$, you get the **complete disjunctive table**,

$$X = \left[\mathbb{I}_{11} \cdots \mathbb{I}_{1c_1} \mathbb{I}_{21} \cdots \mathbb{I}_{2c_2} \cdots \mathbb{I}_{m1} \cdots \mathbb{I}_{mc_m}\right] \ .$$

Finally, we define the main object for the MCA, the **Burt table** $\mathcal{B}$, defined as the symmetric version of X,

$$\mathcal{B} = {}^tXX \ .$$

It is possible to see $\mathcal{B} = [B_{kk'}]_{1\leqslant k,k'\leqslant m}$ as a matrix of $m \times m$ blocks such that

- if $k \neq k'$ then $B_{kk'}$ is the contingency table relative to the variables x^k et $x^{k'}$,
- if $k = k'$ then B_{kk} is a diagonal matrix that contains the marginal totals of the variable x^k, denoted by $n_1^k,\dots,n_{c_k}^k$.

To compute the MCA, we have to consider the PCA of the line profiles given by the Burt table,

$$P = \frac{1}{m} D\mathcal{B}$$

where $D = \mathrm{diag}(1/n_1^1,\dots,1/n_{c_1}^1,1/n_1^2,\dots,1/n_{c_2}^2,\dots,1/n_1^m,\dots,1/n_{c_m}^m)$. The weight matrix is

$$W = \frac{1}{nm} D^{-1}$$

and the metric is given by

$$M = nmD \ .$$

Note that, by symmetry of $\mathcal{B}$, it is not necessary to consider the column profiles. Moreover, with the MCA, only some eigenvalues will be taken in consideration. Thus, it won't be possible to measure the global quality of the representation only with the eigenvalues. Nevertheless, we still can use the contributions of the modes to the inertia of the axes to interpret them.

5 Breast cancer (MCA)

To illustrate the MCA, we consider a data set that comes from a medical study about the breast cancer in three hospitals (Tokyo, Boston and Glamorgan). The patients are distributed among some groups according to their age ("< 50", "$50-69$" and "> 70"), the survival ("Oui" and "Non"), the size of the inflammation ("Petite" and "Grande") and its nature ("Maligne" and "Bénine"). The raw data can be obtained with the following command,

```
data <- DataCancer()
```

If you want, you can try to compute the Burt table directly from `data`. More simply, you can get this table with the help of the command

```
B <- DataCancerBurt()
```

Compute the matrices W and M and, using `B`, do the MCA. Plot the modes in the principal plane. Comment and interpret these results.