Multiple Correspondence Analysis

Julie Josse, François Husson, Sébastien Lê

Applied Mathematics Department, Agrocampus Ouest

useR-2008
Dortmund, August 11th 2008

MCA deals with which kind of data?

- MCA deals with categorical variables, but continuous variables can also be included in the analysis
- Many examples (almost in survey) and today we illustrate MCA with:
  - Questionnaire: tea consumers' habits
  - Ecological data

Multiple Correspondence Analysis

- Generalization of PCA, Generalization of CA
- Analyse the pattern of relationships of several categorical variables
- Dimensionality reduction, sum-up a data table

$\Rightarrow$ Factorial Analysis: data visualisation with a lot of graphical representations to represent proximities between individuals and proximities between variables

$\Rightarrow$ Pre-processing: MCA before clustering

Tea data

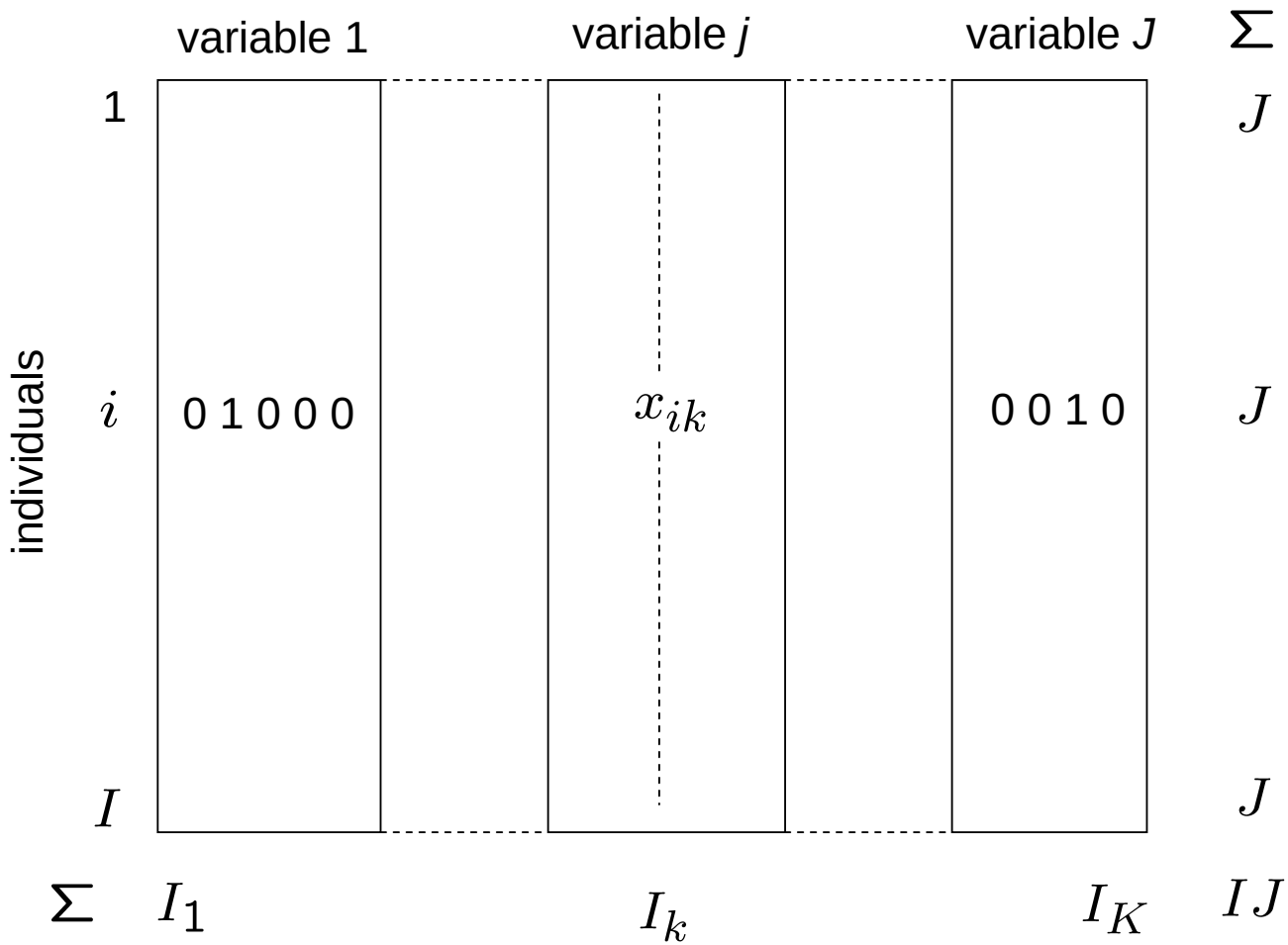
- 300 individuals
- 3 kinds of variables:
  - the way you drink tea (18 questions): kind of tea drunk? How do you drink your tea: with lemon, milk?
  - the product's perception (12 questions): is tea good for health? is it stimulating?
  - personal details (4 questions): sex, age

Problems - objectives

- Individuals study: similarity between individuals (for all the variables) $\longrightarrow$ partition between individuals.
  Individuals are different if they don't take the same levels
- Variables study: find some synthetic variables (continuous variables that sum up categorical variables); link between variables $\Rightarrow$ levels study
- Categories study:
  - two levels of different variables are similar if individuals that take these levels are the same (ex: 65 years and retire)
  - two levels are similar if individuals taking these levels behave the same way, they take the same levels for the other variables (ex: 60 years and 65 years)
- Link between these studies: characterization of the groups of individuals by the levels (ex: executive dynamic women)

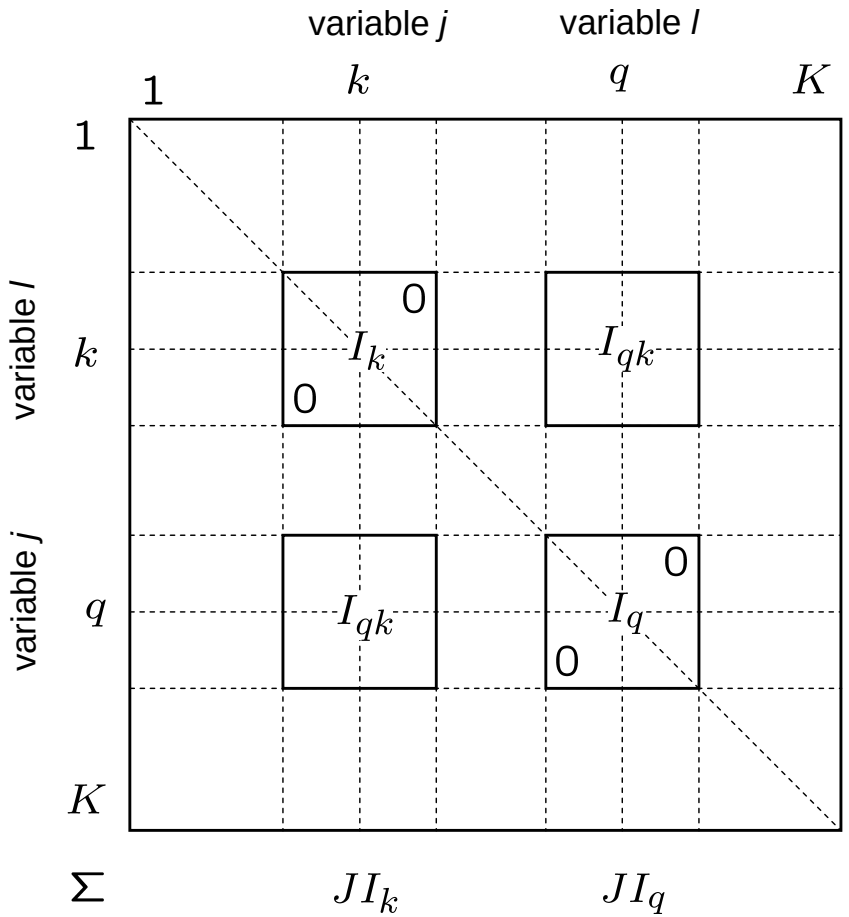
Indicator matrix

- Binary coding of the factors: a factor with K_j levels $\longrightarrow$ K_j columns containing binary values, also called dummy variables



Burt matrix

- This matrix concatenates all two-way cross-tabulations between pairs of variables
- The analysis of the Burt matrix only gives results for the levels



History

At the beginning, when Correspondence Analysis algorithms were available, someone has the idea to use theses algorithms on the Indicator Matrix! You could see the Indicator Matrix (with a lot of imagination!) as a contingency table which cross two categorical variables. This strategy leads to very interesting results: that how is born Multiple Correspondence Analysis. (Lebart)

Construction of the cloud of individuals

$\Rightarrow$ We need a distance between individuals:

- Two individuals take the same levels: distance = 0
- Two individuals take all the categories except one which is uncommon: we want to put it far away
- Two individuals have in common a rare level: they should be closed even if they take different levels for the other variables

$$\begin{aligned} d^2_{i,i'} &= \frac{I}{J}\sum_{k=1}^{K}\frac{1}{I_k}\left(x_{ik}-x_{i'k}\right)^2 \\ &= \sum_{k=1}^{K}\frac{1}{I_k/(IJ)}\left(\frac{x_{ik}/(IJ)}{1/I}-\frac{x_{i'k}/(IJ)}{1/I}\right)^2 \end{aligned}$$

$$d_{\chi^2}(\text{row profile } i, \text{row profile } i') = \sum_{j=1}^{J}\frac{1}{f_{\bullet j}}\left(\frac{f_{ij}}{f_{i\bullet}}-\frac{f_{i'j}}{f_{i'\bullet}}\right)^2$$

Construction of the cloud of levels

$\Rightarrow$ We need a distance between levels:

- Two levels are closed if there is a lot of common individuals which take these levels
- Rare levels are far away from the others

$$
\begin{aligned}
d^2_{k,k'} &= I \sum_{i=1}^{I} \left(\frac{x_{ik}}{I_k} - \frac{x_{ik'}}{I_{k'}} \right)^2 \\
&= \sum_{i=1}^{I} \frac{1}{1/I} \left(\frac{x_{ik}/(IJ)}{I_k/(IJ)} - \frac{x_{ik'}/(IJ)}{I_{k'}/(IJ)} \right)^2
\end{aligned}
$$

$$
d_{\chi^2}(\text{column profile } j, \text{column profile } j') = \sum_{i=1}^{I} \frac{1}{f_{i\bullet}} \left(\frac{f_{ij}}{f_{\bullet j}} - \frac{f_{ij'}}{f_{\bullet j'}} \right)^2
$$

History (later)

MCA has been (re)discovered many times and could be seen under different points of view:

- PCA on a particular data table with particular weights for the variables
- CA on the Indicator Matrix
- CA on the Burt Table

MCA is also known under several different names such as homogeneity analysis

look at your data

```
library(FactoMineR)
data(tea)
summary(tea)
par(ask=T)
for (i in 1:ncol(tea)) barplot(table(tea[,i]))
```

$$\text{Inertia of category } k = \frac{1}{J}\left(1 - \frac{I_k}{I}\right)$$

$\Rightarrow$ How to deal with rare levels?

- Delete individuals (not a good idea!)
- Group levels
- "Ventilation": allocate at random

Define active variables

18 variables	4 variables	12 variables
Which kind of tea do you drink?	Who are you?	Which adjectives do you associate to the tea?

- Active variables: the way you drink tea
- Supplementary: the others

$\Rightarrow$ How to deal with continuous variable?

- Supplementary information: projected on the dimensions and calculate the correlation with each dimension
- Active Information: cut the variables in classes.

```
res.mca=MCA(tea,quanti.sup=19,quali.sup=20:36)
```

Graph of the individuals

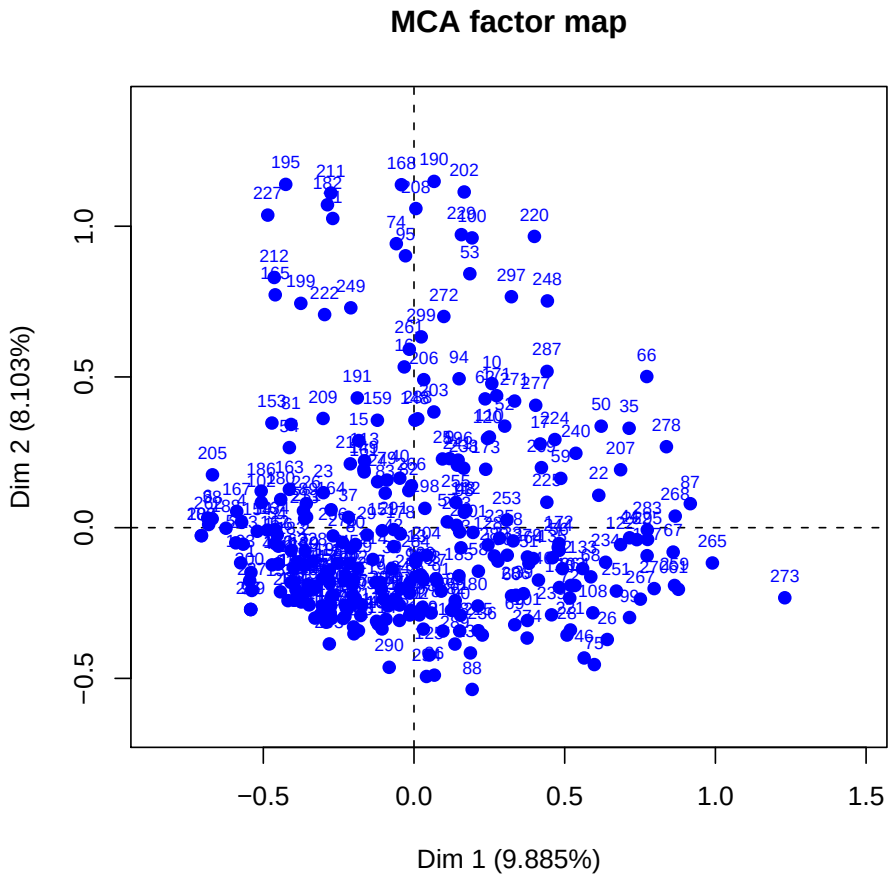
```
plot(res.mca,invisible=c("var","quali.sup","quanti.sup"))
```

Distance between individual i and the barycenter:

$$d(x_{i.}, g)^2 = \frac{I}{J} \sum_{j=1}^{J} \sum_{k=1}^{K} \frac{x_{ik}}{I_k} - 1$$

Distance between individuals:

$$d_(x_{i.}, x_{i'.})^2 = \frac{I}{J} \sum_{k=1}^{K} \frac{(x_{ik} - x_{i'k})^2}{I_k}$$



Graph of the active levels

```
plot(res.mca,invisible=c("ind","quali.sup","quanti.sup"))
```

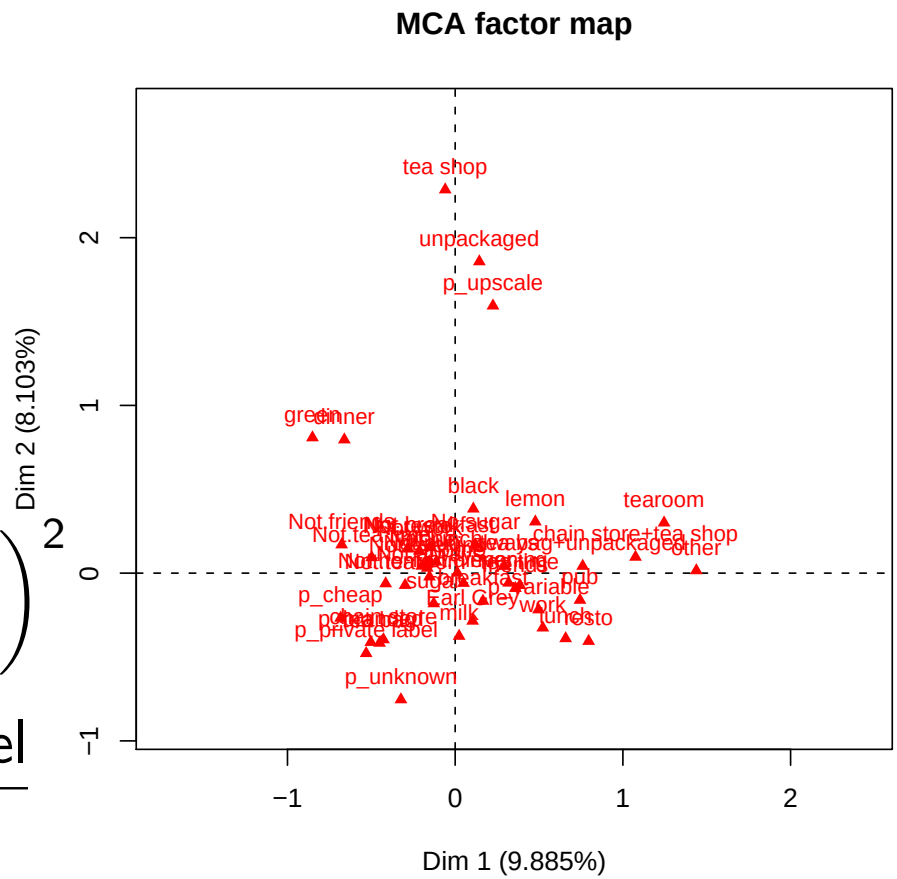
Distance between levels and the barycenter:

$$d(x_{.k}, g)^2 = \frac{I}{I_k} - 1$$

Distance between levels:

$$d(x_{.k}, x_{.k'})^2 = I \sum_{i=1}^{I} \left(\frac{x_{ik}}{I_k} - \frac{x_{ik'}}{I_{k'}} \right)^2$$

$$= \frac{\text{nb of ind taking only one level}}{I_k I'_k}$$



Transition Formulae

$$\boxed{F_s(i) = \frac{1}{\sqrt{\lambda_s}} \sum_k \frac{x_{ik}}{J} G_s(k)}$$

$\Rightarrow$ Individual i is (up to $\frac{1}{\sqrt{\lambda_s}}$) at the barycenter its levels

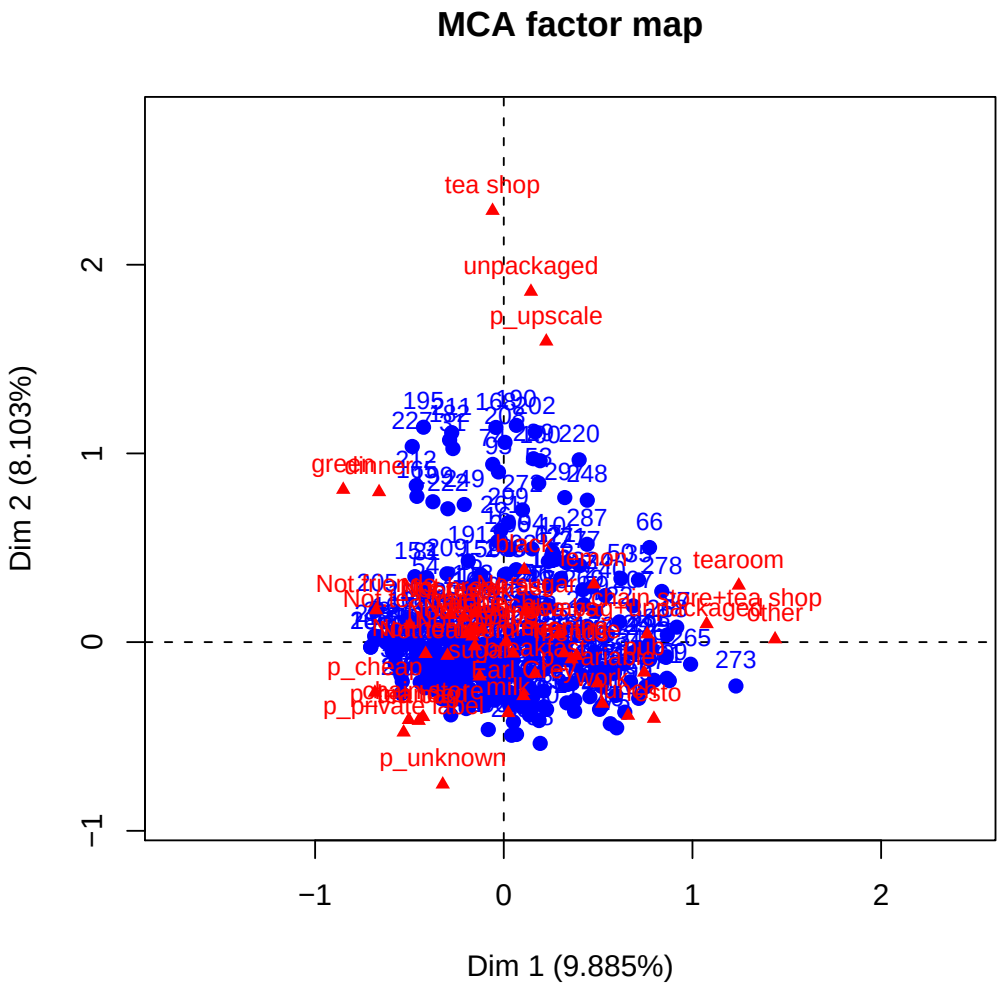
$$\boxed{G_s(k) = \frac{1}{\sqrt{\lambda_s}} \sum_i \frac{x_{ik}}{I_k} F_s(i)}$$

$\Rightarrow$ Level k is (up to $\frac{1}{\sqrt{\lambda_s}}$) at the barycenter of the individuals who take this level

$\Rightarrow$ Possibility to simultaneously represent the two representations

Superimposed representation

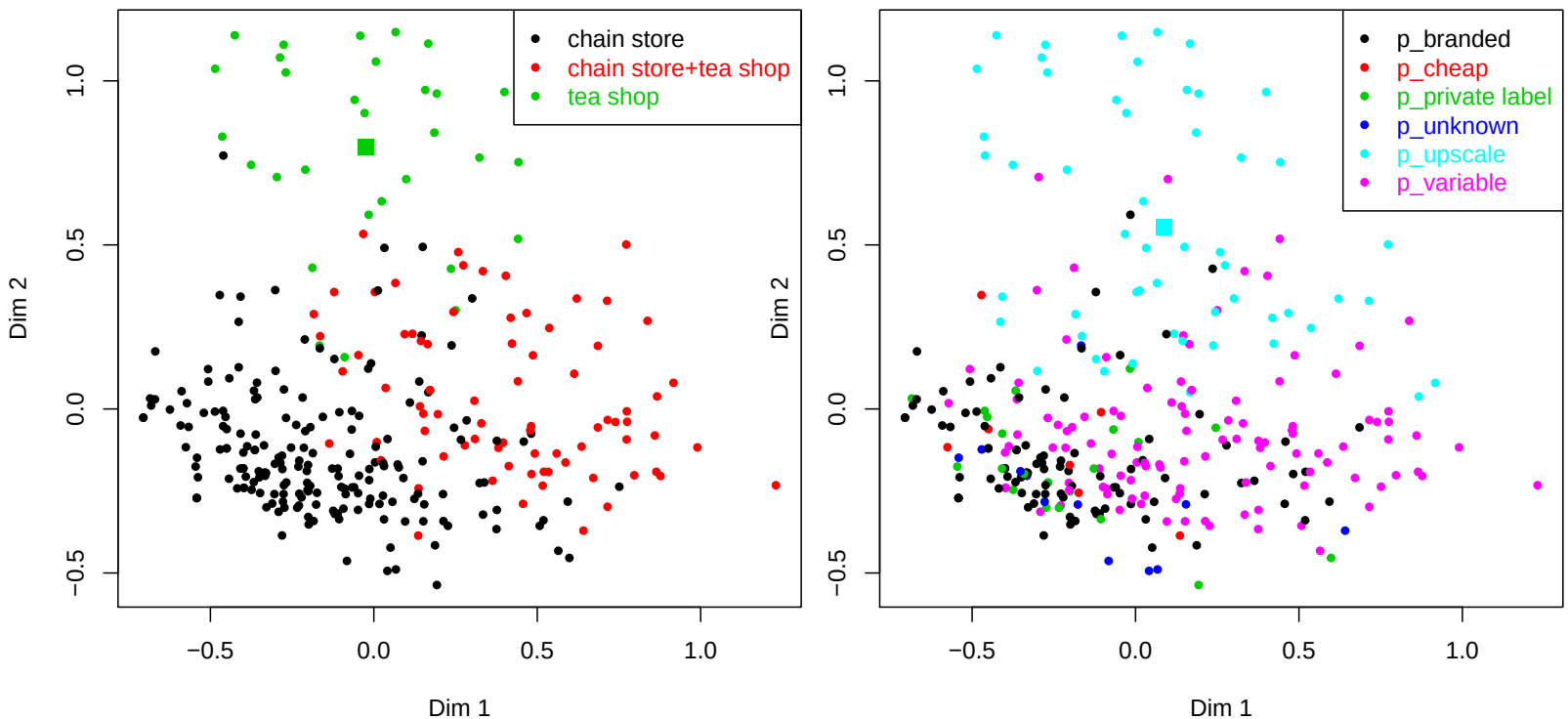
```
plot(res.mca,invisible=c("quali.sup","quanti.sup"))
```



Interpretation of the location of the levels

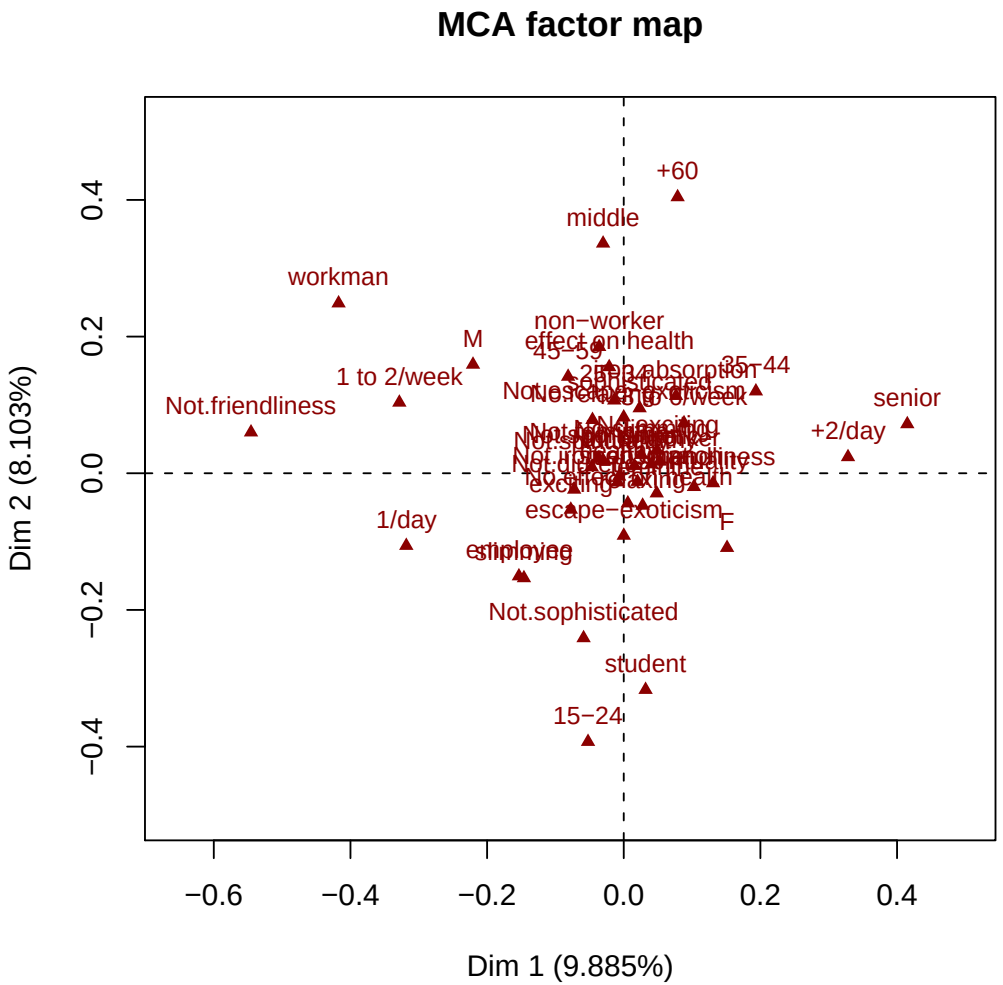
```
plot(res.mca$ind$coord[,1:2],col=as.numeric(tea[,17]),pch=20)
legend("topright",legend=levels(tea[,17]),text.col=1:3,col=1:3)
aa=by(res.mca$ind$coord,tea[,17],FUN=mean)[[3]][1:2]
points(aa[1],aa[2],col=3,pch=15,cex=1.5)
x11()
plot(res.mca$ind$coord[,1:2],col=as.numeric(tea[,18]),pch=20)
legend("topright",legend=levels(tea[,18]),text.col=1:6,col=1:6)
bb=by(res.mca$ind$coord,tea[,18],FUN=mean)[[5]][1:2]
points(bb[1],bb[2],col=5,pch=15,cex=1.5)
```

Interpretation of the location of the levels



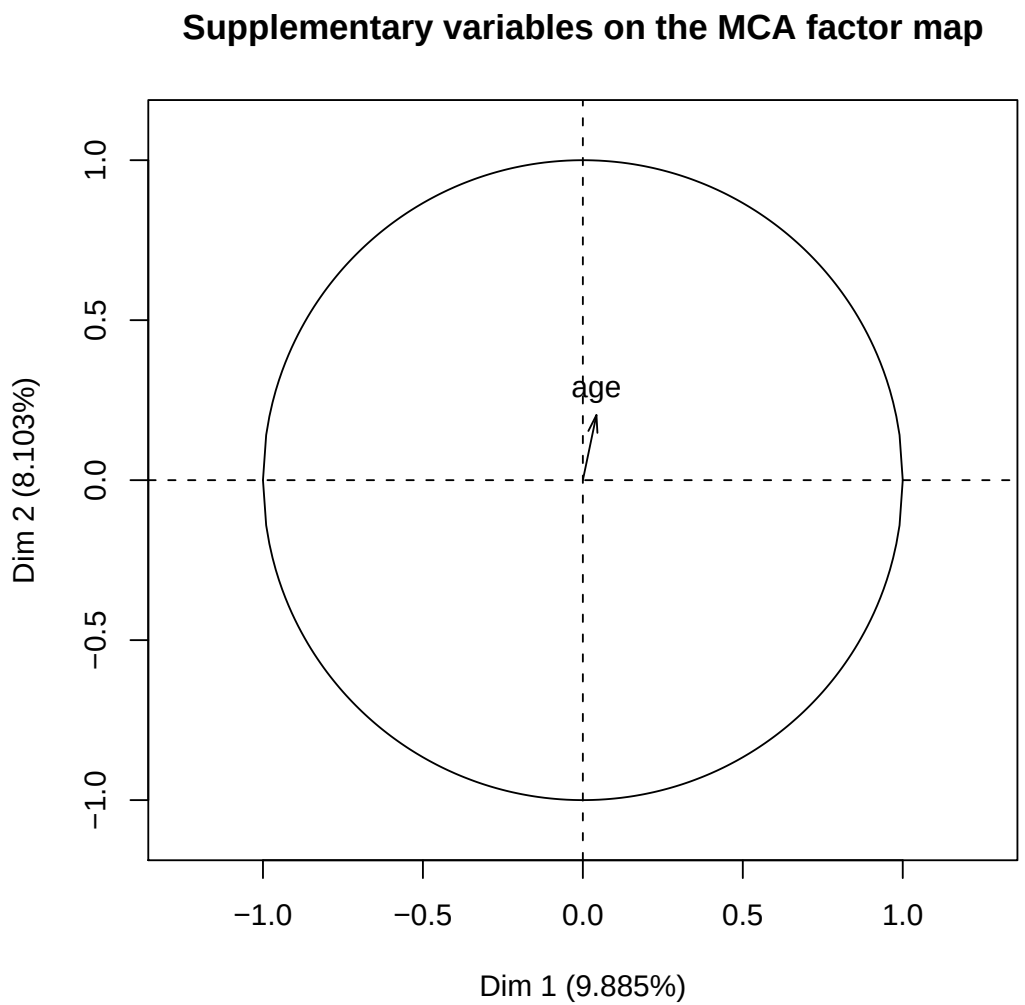
Supplementary levels representation

```
plot(res.mca,invisible=c("var","ind","quanti.sup"),cex=0.8)
```



Continuous supplementary variable

```
plot(res.mca,invisible=c("var","ind","quali.sup"),cex=0.8)
```



Description of the dimensions

By qualitative variables (F-test), categories (t-test) and quantitative variables (correlation)

```
dimdesc(res.mca)
```

```
$`Dim 1`$quali
                  P-value
where        1.255462e-35
tearoom      6.082138e-32
how          1.273180e-23
friends      8.616289e-20
resto        2.319804e-18
tea.time     1.652462e-15
price        4.050469e-14
pub          5.846592e-12
work         3.000872e-09
How          4.796010e-07
Tea          8.970954e-07
lunch        1.570629e-06
frequency    1.849071e-06
friendliness 2.706357e-06
evening      5.586801e-05

$`Dim 1`$category
                      Estimate      P-value
tearoom              0.2973107 6.082138e-32
chain store+tea shop 0.3385378 1.755544e-25
friends              0.1995083 8.616289e-20
resto                0.2080260 2.319804e-18
tea time             0.1701136 1.652462e-15
tea bag+unpackaged   0.2345703 6.851637e-14
pub                  0.1813713 5.846592e-12
work                 0.1417041 3.000872e-09
Not.work            -0.1417041 3.000872e-09
green               -0.2456910 6.935593e-10
Not.pub             -0.1813713 5.846592e-12
Not.tea time        -0.1701136 1.652462e-15
tea bag             -0.2318245 3.979797e-16
Not.resto           -0.2080260 2.319804e-18
chain store         -0.2401244 1.254499e-18
Not.friends         -0.1995083 8.616289e-20
Not.tearoom         -0.2973107 6.082138e-32
```

Inertia

- Variable Inertia:

$$\text{Inertia } (j) = \sum_{k=1}^{K_j} \text{Inertia } (k) = \sum_{k=1}^{K_j} \frac{1}{J}\left(1 - \frac{I_k}{I}\right) = \frac{K_j - 1}{J}$$

$\Rightarrow$ The inertia is large when the variable has many levels

Remark: should we use variables with equal number of levels? No: it doesn't matter because the projected inertia of each variable on each axis is bounded by $1/J$.

- Total Inertia:

$$\text{total inertia} = \frac{K}{J} - 1$$

Choice of the number of dimensions to interpret

```
res.mca$eig
```

- Bar plot: difficult to make a choice
- $K - J$ non-null eigenvalues, $\sum \lambda_s = \frac{K}{J} - 1$
  Average of an eigenvalue: $\frac{1}{K-J} \times \sum_s \lambda_s = \frac{1}{J}$ $\Rightarrow$ a rule consists to keep the eigenvalues greater than $1/J$
- Bootstrap confidence interval

Why percentages of inertia are small?

- Individuals are in $\mathbb{R}^{K-J} \Rightarrow$ Generally, the percentage of variance explained by the first axis is small
- Maximum percentage for one dimension:

$$\begin{aligned} \frac{\lambda_s}{\sum \lambda_s} \times 100 &\leq \frac{1}{\frac{K-J}{J}} \times 100 \\ &\leq \frac{J}{K-J} \times 100 \end{aligned}$$

With $K = 100$, $J = 10$: $\lambda_s \leq 11.1$ %

```
aa=as.factor(rep(1:10,each=100))
bb=cbind.data.frame(aa,aa,aa,aa,aa,aa,aa,aa,aa,aa)
colnames(bb)=paste("a",1:10,sep="")
res=MCA(bb)
res$eig[1:10,]
```

Why the percentages of inertia are small?

Moreover, the percentages are pessimistic!
If the percentages of inertia are calculated from the Burt table analysis:

```
burt=t(tab.disjonctif(tea[,1:18]))%*%tab.disjonctif(tea[,1:18])
res.burt=CA(burt)
res.burt$eig[1:10,]
```

34.8 % explained by the two first axes instead of 19 % (with exactly the same representations for the levels)

$\Rightarrow$ Benzécri and Greenacre noticed that this percentage is optimistic and proposed coefficient to adjust the inertia

Helps to interpret

- Contribution and $\cos^2$ for the individuals and the levels

```
res.mca$ind$contrib
res.mca$ind$cos2
res.mca$var$contrib
res.mca$var$cos2
```

$\Rightarrow$ Extreme levels do not necessarily contribute the most (it depends on the frequencies)
$\Rightarrow$ Cos^2 are very small... but it was awaited since inertia is small

- Variable contribution: $CTR(j) = \sum_k CTR(k)$
- Remark:

$$\eta^2(F_s, j) = \frac{CTR(j)}{J\lambda_s}$$

Remarks

- Return to your data with contingency tables $\longrightarrow$ correspondence analysis
- Non-linear relationships can be highlighted
- Gutman effect
- MCA as a pre-processing for clustering

A pre-processing for Clustering

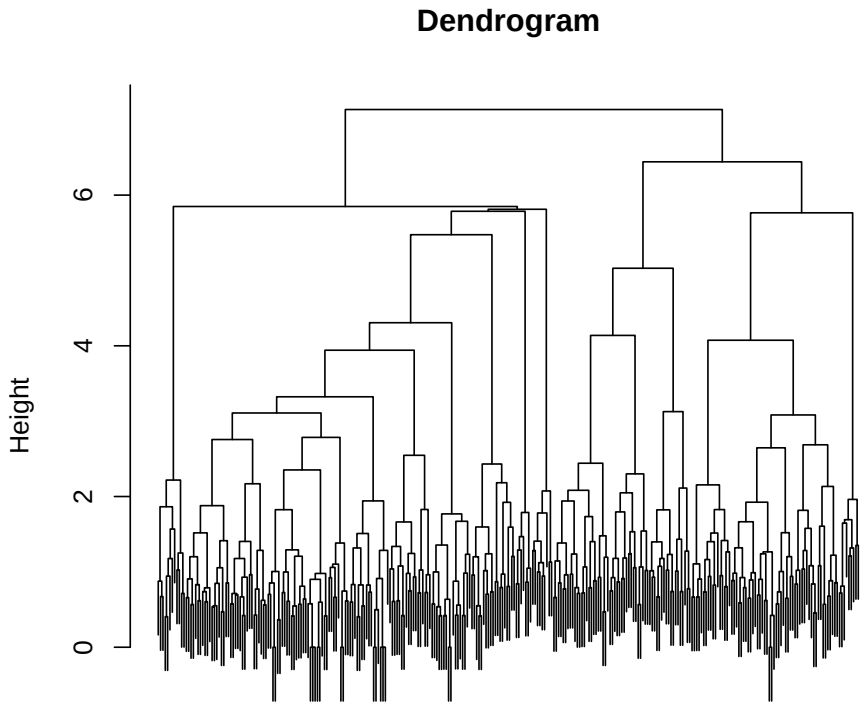
- Transformation of the data: categorical $\longrightarrow$ continuous
- Principal components (individuals coordinates) are synthetic variables: the most linked to the other variables: $\text{argmax}_v \frac{1}{J} \sum_j \eta^2(v,j) = F_s$
- "Denoising": retain only 95% of the information

- Clustering on the individuals coordinates (with variance λ_s)
  $\Rightarrow$ hierarchical clustering with ward criteria (based on inertia)

- Classification (Fisher Linear Discriminant) on the individuals coordinates (with variance λ_s)

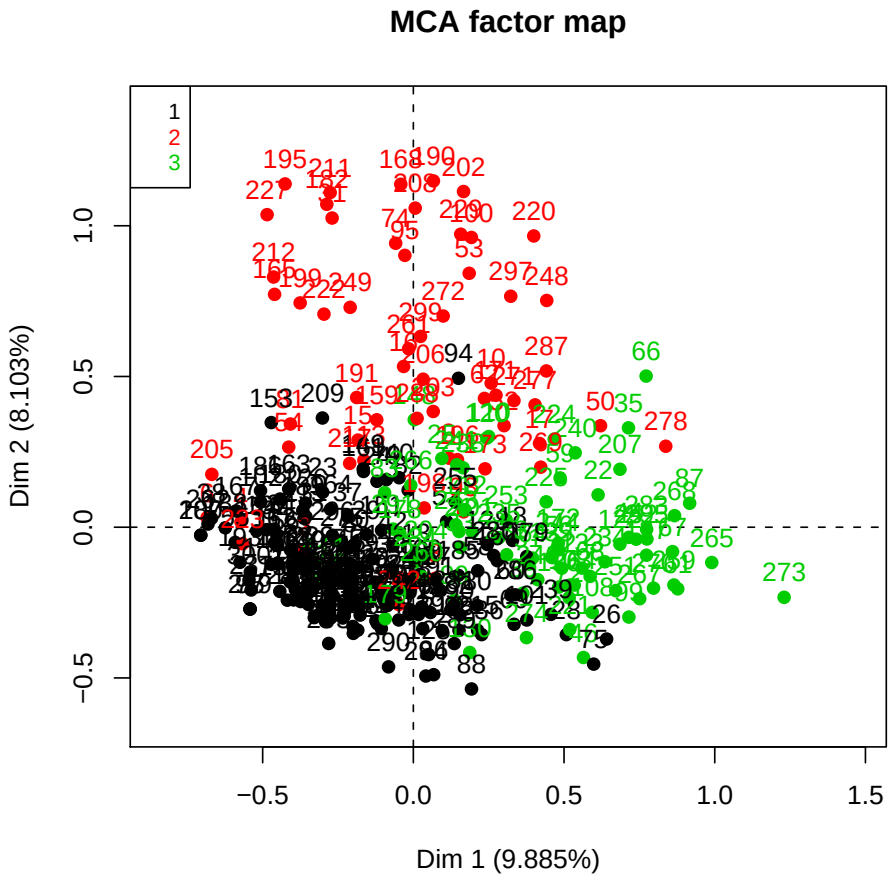
Hierarchical Clustering

```
res.mca=MCA(tea,quanti.sup=19,quali.sup=20:36,ncp=20,graph=F)
library(cluster)
classif = agnes(res.mca$ind$coord,method="ward")
plot(classif,main="Dendrogram",ask=F,which.plots=2,labels=FALSE)
```



Represent the clusters on your factorial map

```
clust = cutree(classif,k=3)
tea.comp = cbind.data.frame(tea,res.mca$ind$coord[,1:3],factor(clust))
res.aux=MCA(tea.comp,quanti.sup=c(19,37:39),quali.sup=c(20:36,40),graph=F)
plot(res.aux,invisible=c("quali.sup","var","quanti.sup"),habillage=40)
```



Describe each cluster

```
catdes(tea.comp,ncol(tea.comp))
```

```
 $test.chi
            P.value df                   P.value df
where   2.316552e-49  4     tearoom 1.025632e-09  2
how     3.592323e-35  4     dinner  3.874810e-09  2
price   1.142914e-31 10     friends 1.859075e-06  2
How     5.884403e-10  6

$category$`2`
                     Cla/Mod   Mod/Cla    Global      p.value    V-test
price=p_upscale   0.73584906 0.6610169 0.1766667 1.467589e-22  9.702737
where=tea shop    0.90000000 0.4576271 0.1000000 6.532117e-19  8.805180
how=unpackaged    0.77777778 0.4745763 0.1200000 3.184180e-16  8.082056
dinner=dinner     0.71428571 0.2542373 0.0700000 1.094845e-07  5.182468
Tea=Earl Grey     0.11917098 0.3898305 0.6433333 8.396333e-06 -4.303756
how=tea bag       0.09411765 0.2711864 0.5666667 3.162813e-07 -4.981001
dinner=Not.dinner 0.15770609 0.7457627 0.9300000 1.094845e-07 -5.182468
where=chain store 0.08854167 0.2881356 0.6400000 7.994965e-10 -6.034051

$quanti$`2`
        v.test Mean in category Overall mean sd in category Overall sd
Dim.2 12.92675        0.5267543 6.280824e-17      0.3746555  0.3486355
```