Correspondence Analysis

Julie Josse, François Husson, Sébastien Lê

Applied Mathematics Department, Agrocampus Ouest

useR-2008
Dortmund, August 11th 2008

History

- Theoretical principles: Fisher (1940)
- Correspondence Analysis has been actively developed in 1965 ... in Rennes!
- JP. Benzécri: mathematician and linguist
- PhD thesis of his student B. Escofier : Correspondence Analysis

- The beginning of the "French school"

CA in the R packages

- anacor (de Leeuw and mair)
- ca (Nenadic and Greenacre)
- ade4 (Chessel)
- vegan (Dixon)
- homals (de Leeuw)
- FactoMineR (Husson *et al.*)

Data, examples

- Two categorical variables $\Rightarrow$ contingency table. Symmetric role of the rows and the columns

- Examples:
  - examples where a χ^2 test can be applied
  - text-mining: number of times the word i is in the text j
  - solutions (acid, bitter, etc.) - answers (acid, bitter, etc.): number of persons who answer j for the stimulus i
  - perfumes - descriptors: number of times the descriptor j is used to describe the perfume i

Notations

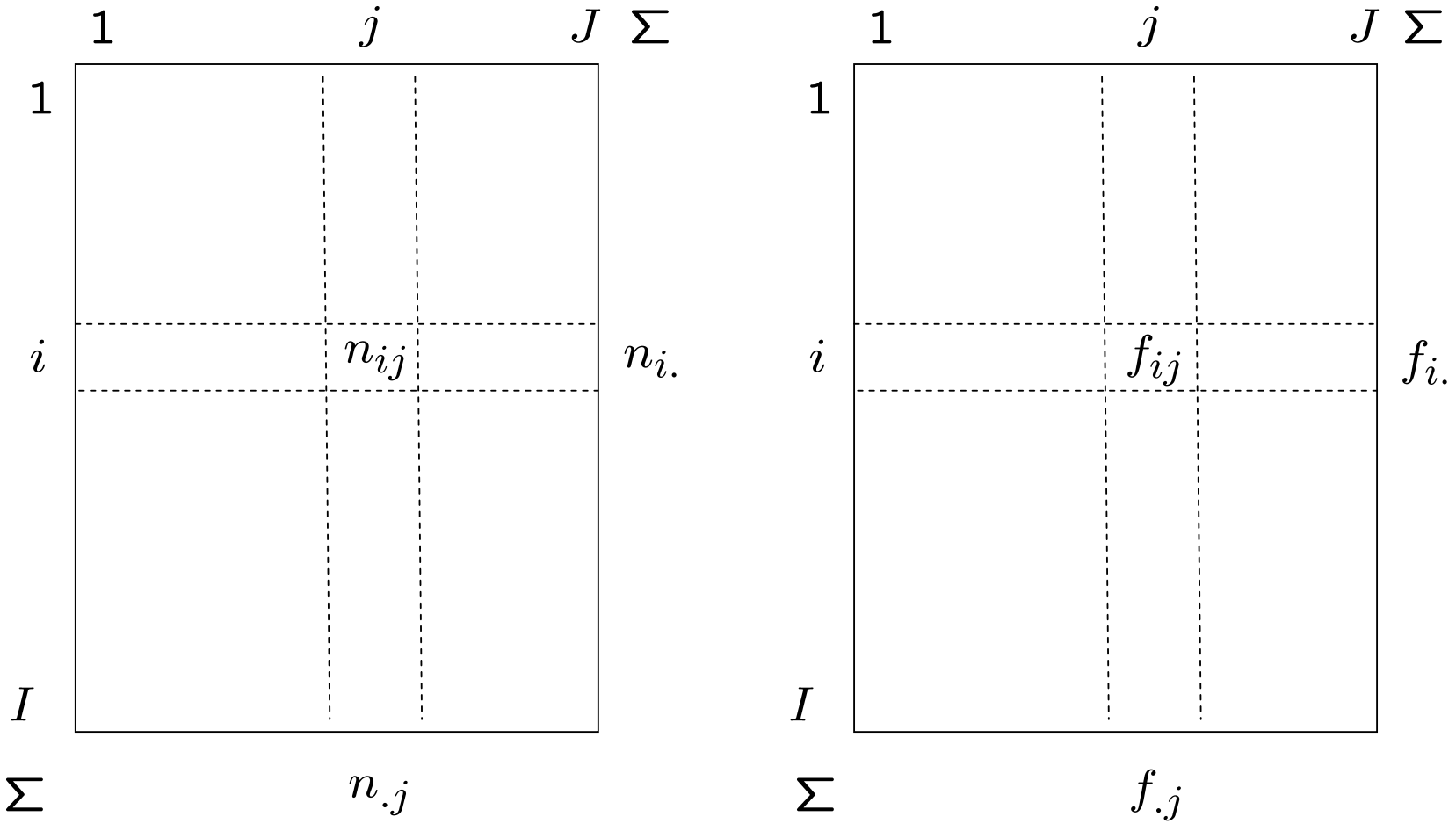


Figure: Data table in CA.

Notations

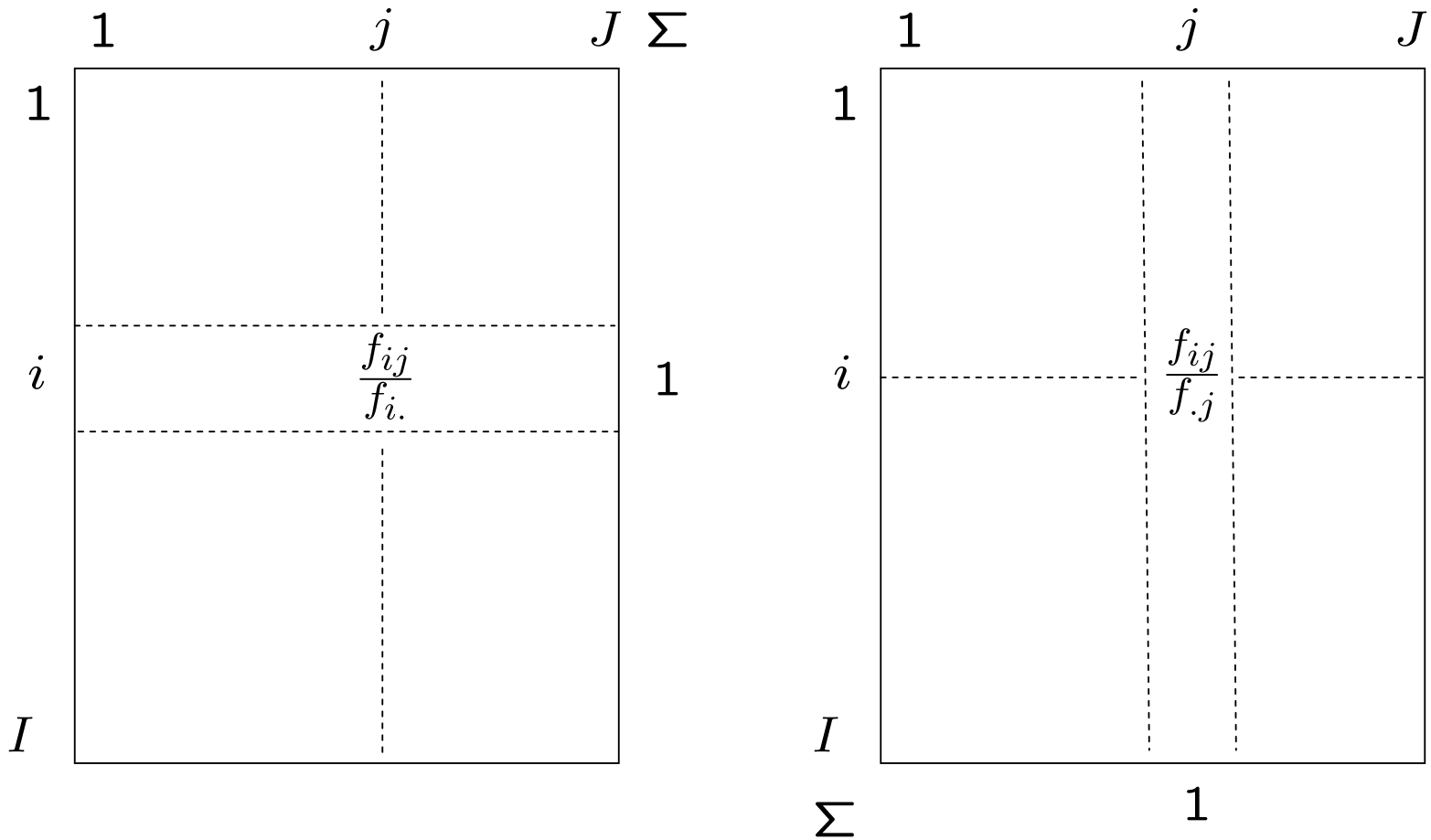


Figure: Row profile and column profile.

Aim

- Rows typology
- Columns typology
- Relationship between these two typologies

$\Rightarrow$ Study the relationship (the correspondence) between the two variables, the gap to independence

$\Rightarrow$ Visualize the association between levels

Example

12 perfumes described by 39 words:

	floral	fruity	strong	soft	light	...
Angel	2	11	18	3	1	...
Aromatics Elixir	2	3	29	2	0	...
Chanel 5	5	0	19	3	1	...
Cinéma	14	14	3	12	9	...
Coco Mademoiselle	10	10	6	10	7	...
.....	.	.	.	.	.	

Intensity of the relationship

- Chi-square:

$$\begin{aligned}\chi^2 &= \sum_{ij} \frac{(n_{ij} - n_{i.}n_{.j}/n)^2}{n_{i.}n_{.j}/n}, \\ &= n\sum_{ij} \frac{(f_{ij} - f_{i.}f_{.j})^2}{f_{i.}f_{.j}}, \\ &= n\phi^2.\end{aligned}$$

  $\Rightarrow \phi^2$ is the intensity of the relationship

- Chi-square test (Pearson):

$$\chi^2_{\text{obs}} \sim \chi^2_{(I-1)\times(J-1)}$$

  $\chi^2_{\text{obs}} = 615.8$, (p-value = 1.7e-56) $\Rightarrow$ Highly significant

Nature of the relationship

- Contribution to the chi-square:

$$x_{ij}^2 = \frac{(n_{ij} - n_{i.}n_{.j}/n)^2}{n_{i.}n_{.j}/n}$$

$\Rightarrow$ Contribution of each cell, contribution of each row, contribution of each column?

- Residuals (positive or negative association):

$$x_{ij} = \frac{(n_{ij} - n_{i.}n_{.j}/n)}{\sqrt{n_{i.}n_{.j}/n}}$$

$\Rightarrow$ CA: visualize the residuals matrix X (the gap to independence)
$\Rightarrow$ As usual, the association structure of X is revealed using the SVD

Total inertia

$$\text{Total inertia} = \phi^2 = \text{trace}(XX') = \frac{\chi^2}{n}$$

$$\phi^2 = \frac{\chi^2}{n} = \sum_{ij} \frac{(f_{ij} - f_{i.}f_{.j})^2}{f_{i.}f_{.j}}$$

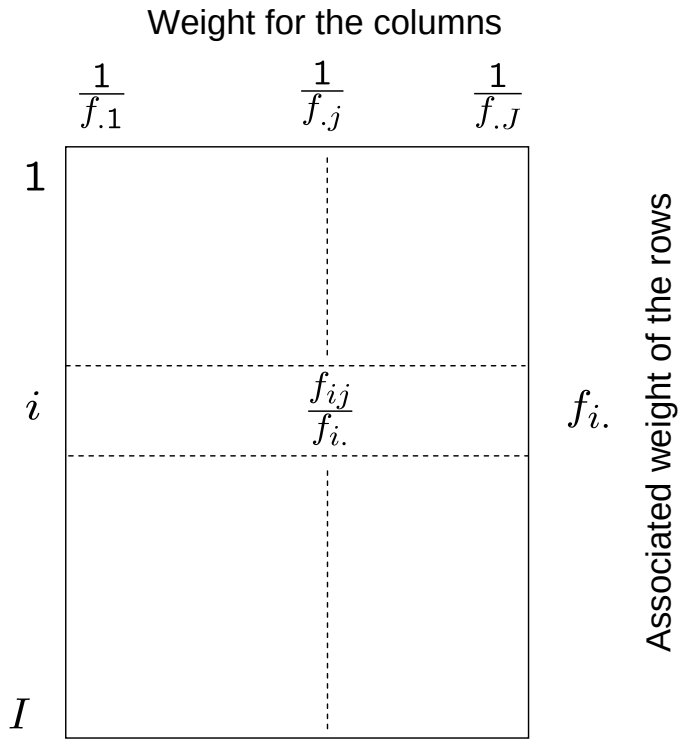
$$\phi^2 = \sum_{ij} f_{.j} \frac{\left(\frac{f_{ij}}{f_{.j}} - f_{i.}\right)^2}{f_{i.}}$$

Similarly:

$$\phi^2 = \frac{\chi^2}{n} = \sum_{ij} f_{i.} \frac{\left(\frac{f_{ij}}{f_{i.}} - f_{.j}\right)^2}{f_{.j}}$$

Total inertia explained via the profile

For the row profile:



$$\sum_i f_{i.} \frac{f_{ij}}{f_{i.}} = f_{.j}$$

$$\phi^2 = \sum_{ij} f_{i.} \frac{\left(\frac{f_{ij}}{f_{i.}} - f_{.j}\right)^2}{f_{.j}}$$

$\Rightarrow$ Total inertia = weighted sum of squared distances of the rows profile to the average profile the weight of the row profile is its mass $f_{i.}$ and the squared distance is an Euclidean distance where each squared difference is divided by the corresponding average value $f_{.j}$

χ^2-distance

- the row i is a point in $\mathbb{R}^J$ (with the weight $f_{i.}$)

$$d^2(i,l) = \sum_j \frac{1}{f_{.j}} \left(\frac{f_{ij}}{f_{i.}} - \frac{f_{lj}}{f_{l.}} \right)^2$$

- the column j is a point in $\mathbb{R}^I$ (with the weight $f_{.j}$)

$$d^2(j,h) = \sum_i \frac{1}{f_{i.}} \left(\frac{f_{ij}}{f_{.j}} - \frac{f_{ih}}{f_{.h}} \right)^2$$

These χ^2-distances enjoy good properties (distributional equivalence principle)

CA: different formulations

- In factorial analysis such as PCA, we are looking for dimension which represent in the better way the variability between individuals (i.e the distance to the barycenter), in CA we are looking for dimensions which better represent the gap to independence.

- Classical presentation of CA: two weighted PCA on "row profile" and on "column profil"

- Canonical Analysis on the two indicator matrices

Link between the two representations: transition formulae

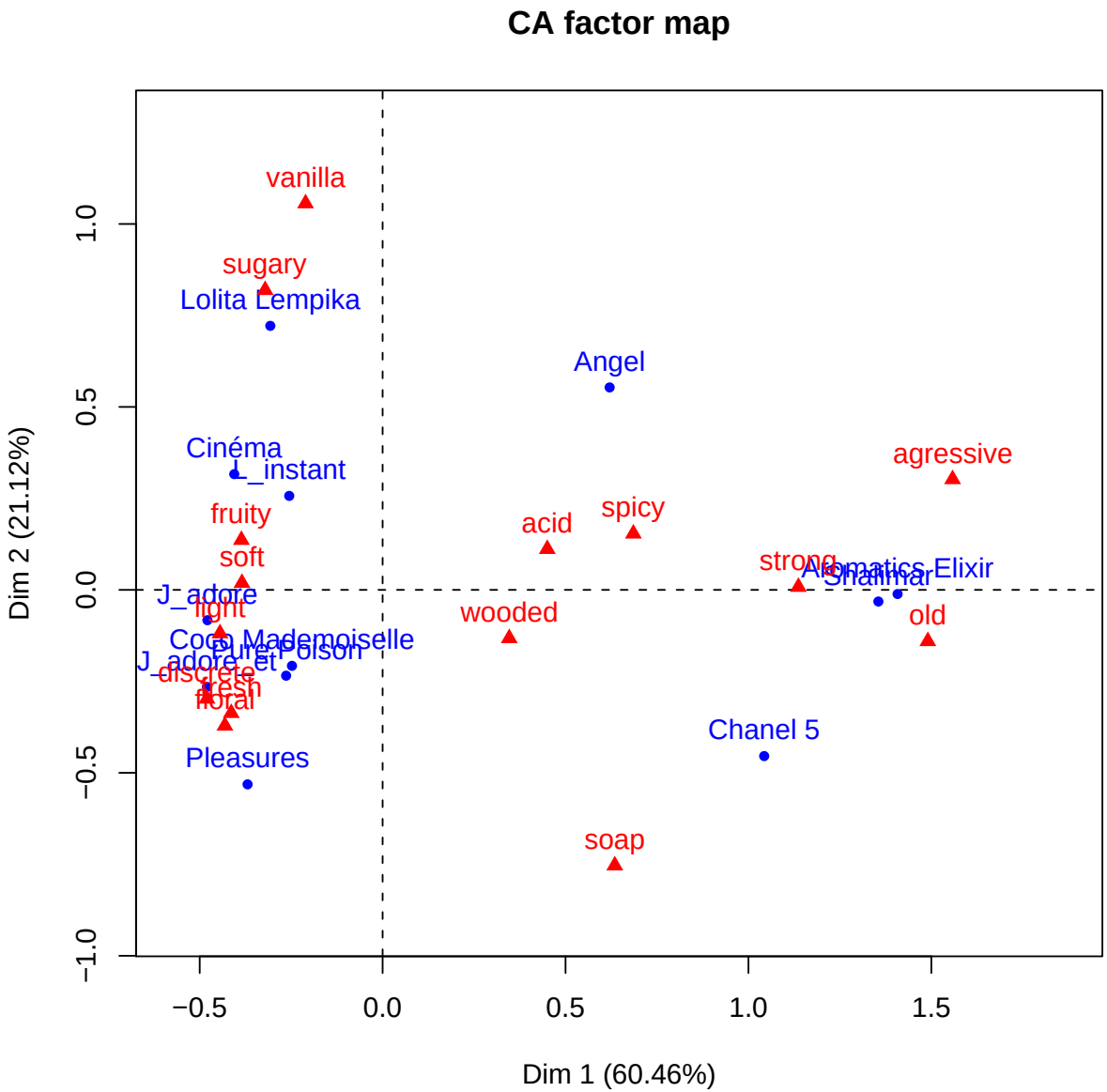
$$\boxed{F_s(i) = \frac{1}{\sqrt{\lambda_s}} \sum_j \frac{f_{ij}}{f_{i.}} G_s(j)}$$

$\Rightarrow$ Row i is at the barycenter of the weighted columns (with a scale $1/\sqrt{\lambda_s}$)

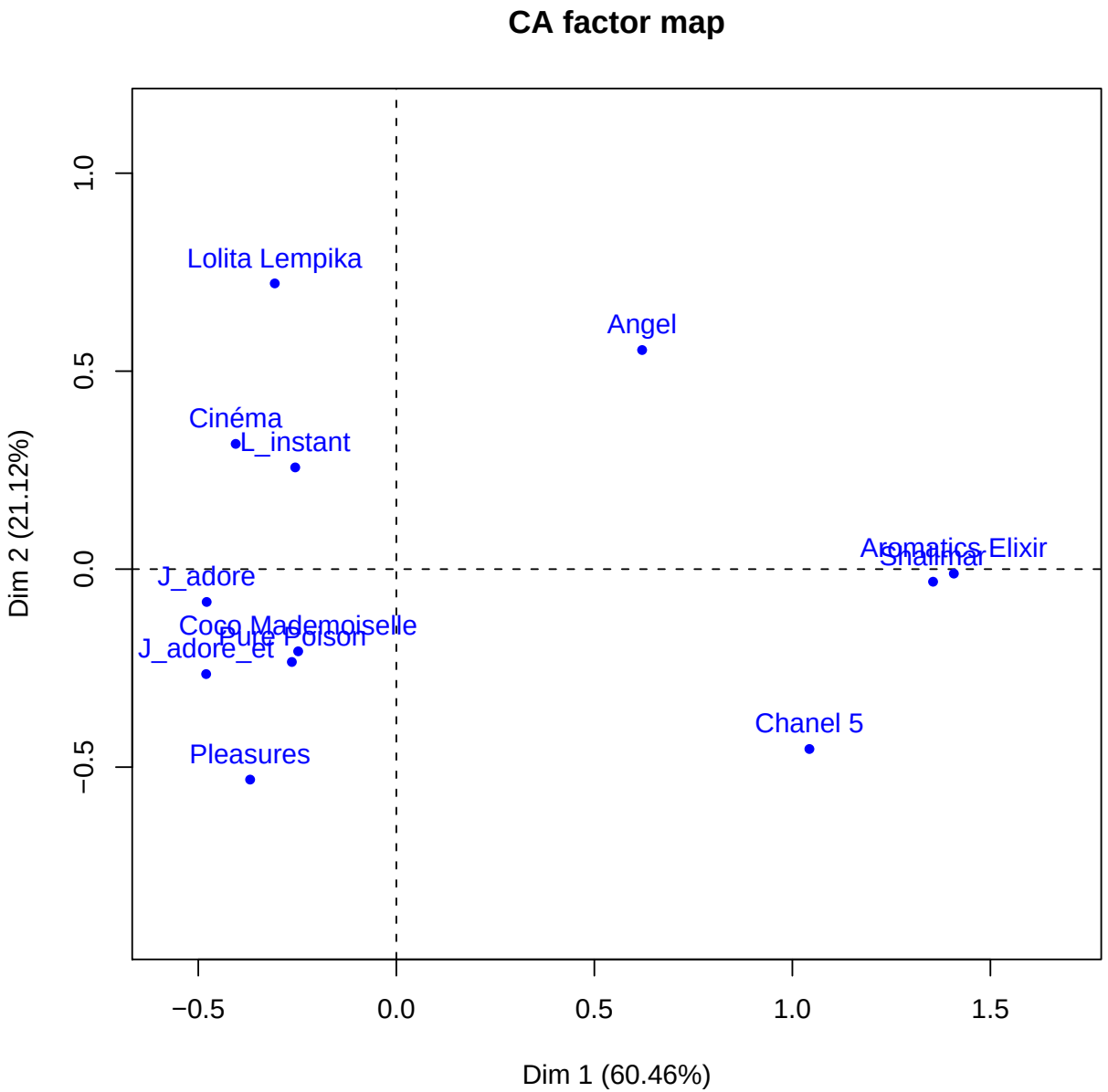
$$\boxed{G_s(k) = \frac{1}{\sqrt{\lambda_s}} \sum_i \frac{f_{ij}}{f_{.j}} F_s(i)}$$

$\Rightarrow$ Column k is at the barycenter of the weighted rows (with a scale $1/\sqrt{\lambda_s}$)

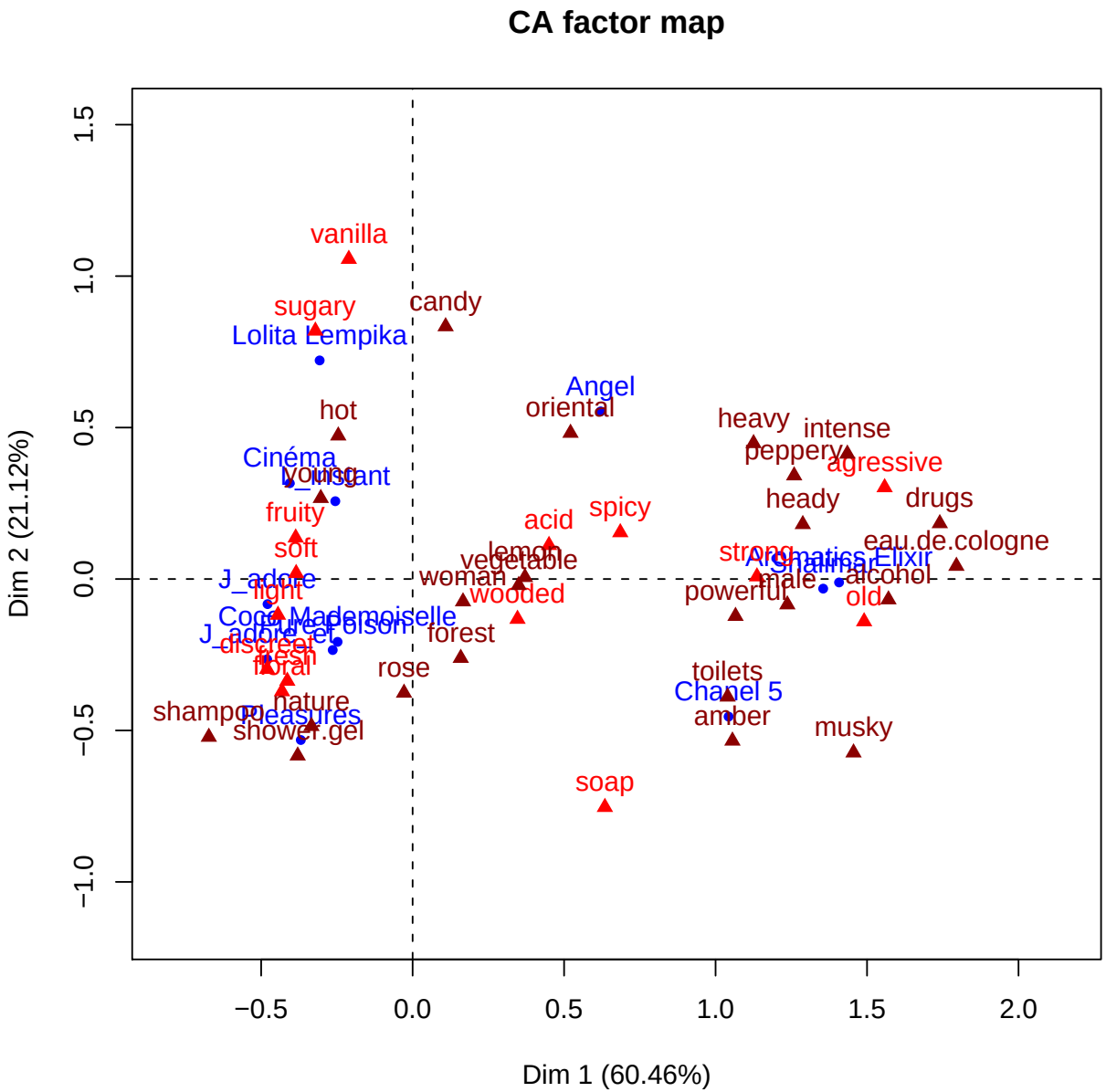
Graphical representation



Graphical representation



Graphical representation



Graphical representation in CA: remarks

- The barycenter represents the independence
- The distance between levels of a same variable can be interpreted
- Representation provided are pseudo-barycentric (dilatation): transition formulae
- It is not possible to interpret the distance between levels of the two variables but ...
- ... it is at a weighted barycenter of all the levels

Helps to interpret

- Supplementary informations can be added (zero weight)!
- Percentage of variance for each axis: information brought by the dimension $\frac{\lambda_s}{\sum_s \lambda_s}$, but, it is interesting to have a look at the eigenvalues.
- λ_s always smaller than 1; the value 1 is obtained for exclusive associations.

```
library(FactoMineR)
don=diag(5)
a=CA(don)
a$eig

don=matrix(1,5,5)
a=CA(don)
a$eig
```

Helps to interpret

- The maximum number of axes is $\min(I, J) - 1$
- Quality of the representation: $\cos^2$
- Contribution:
  - $\frac{\text{inertia of a point}}{\text{total inertia}} = \frac{f_{i.} F_s(i)^2}{\lambda_s}$. Be careful, extreme points are not those which contribute the most to the dimension

Practice

```
library(FactoMineR)
perfume = read.table("perfume.txt",header=T,sep="\t",row.names=1)
res.ca = CA(perfume,col.sup=16:39)

plot(res.ca,invisible="row")
plot(res.ca,invisible=c("col","col.sup"))

res.ca$eig
barplot(res.ca$eig[,1],main="Eigenvalues",names.arg=1:nrow(res.ca$eig))
res.ca$row$coord
res.ca$row$cos2
res.ca$row$contrib
res.ca$col$coord
res.ca$col$cos2
res.ca$col$contrib
```