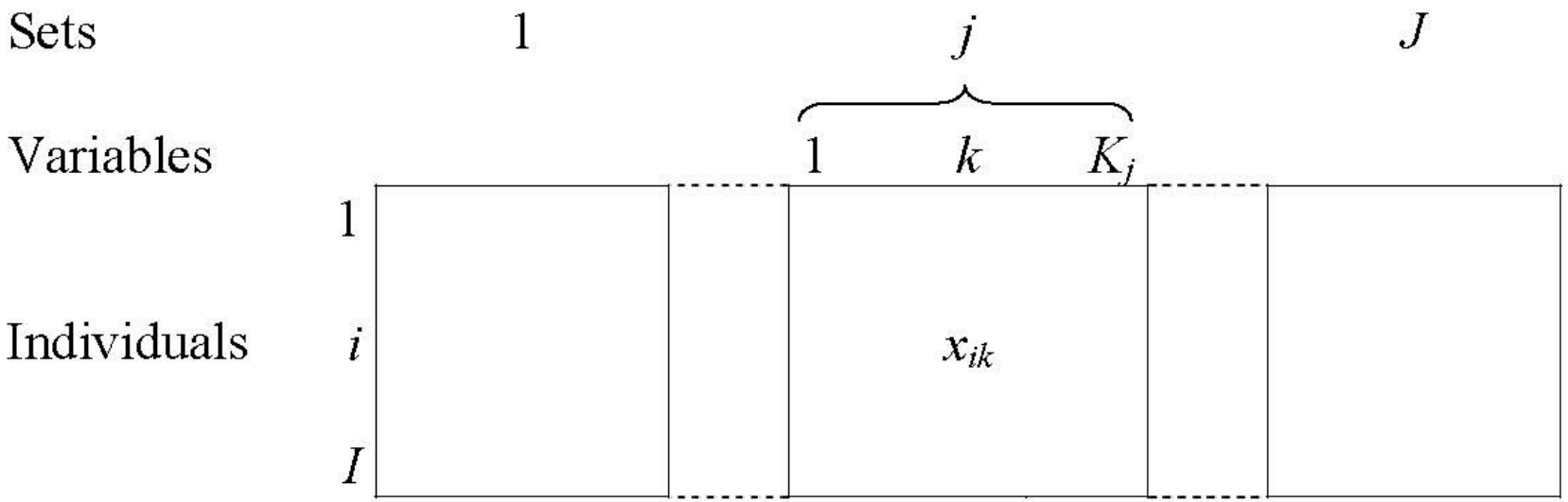
Multiple Factor Analysis

"Doing a data analysis, in good mathematics, is simply searching eigenvectors, all the science of it (the art) is just to find the right matrix to diagonalize"

Benzécri

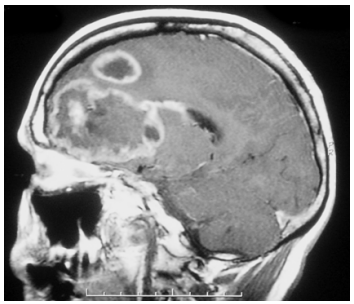
Multiway data set



Examples with **continuous and/or categorical** sets of variables:

- genomic: DNA, protein
- sensory analysis: sensorial, physico-chemical
- survey: student health (addicted consumptions, psychological conditions, sleep, identification, etc.)
- economy: economic indicators for countries by year

Example: gliomas brain tumors



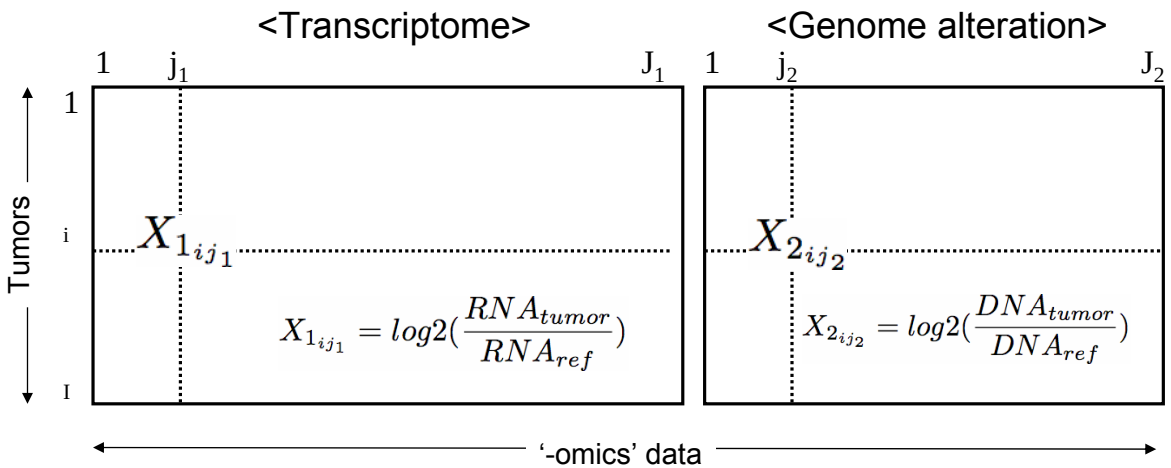
Gliomas: Brain tumors, WHO classification

- astrocytoma (A)..... x5
- oligodendroglioma (O)..... x8
- oligo-astrocytoma (OA)..... x6
- glioblastoma (GBM)..... x24

43 tumor samples

(Bredel *et al.*,2005)

- Transcriptional modification (RNA), microarrays: 489 variables
- Damage to DNA (CGH array): 113 variables



Objectives

- Study the similarities between individuals with respect to all the variables
- Study the linear relationships between variables

$\Rightarrow$ taking into account the structure on the data (balancing the influence of each group)

- Find the common structure with respect to all the groups - highlight the specificities of each group
- Compare the typologies obtained from each group of variables (separate analyses)

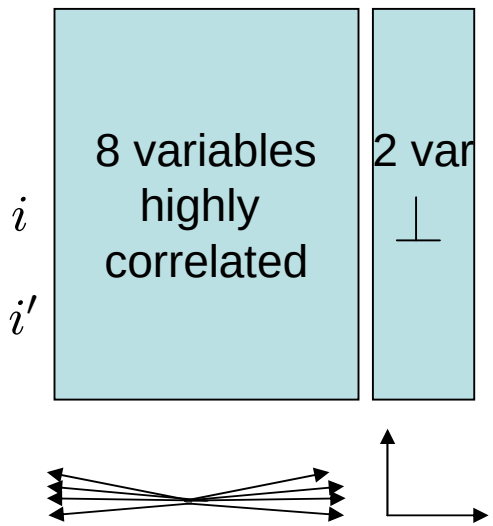
Balancing the groups of variables

MFA is a weighted PCA:

- compute the first eigenvalue λ_1^j of each group of variables
- perform a global PCA on the weighted data table:

$$\left[\frac{X_1}{\sqrt{\lambda_1^1}}; \frac{X_2}{\sqrt{\lambda_1^2}}; \ldots; \frac{X_J}{\sqrt{\lambda_1^J}}\right]$$

$\Rightarrow$ Same idea as in PCA when variables are standardized: variables are weighted to compute distances between individuals i and i'

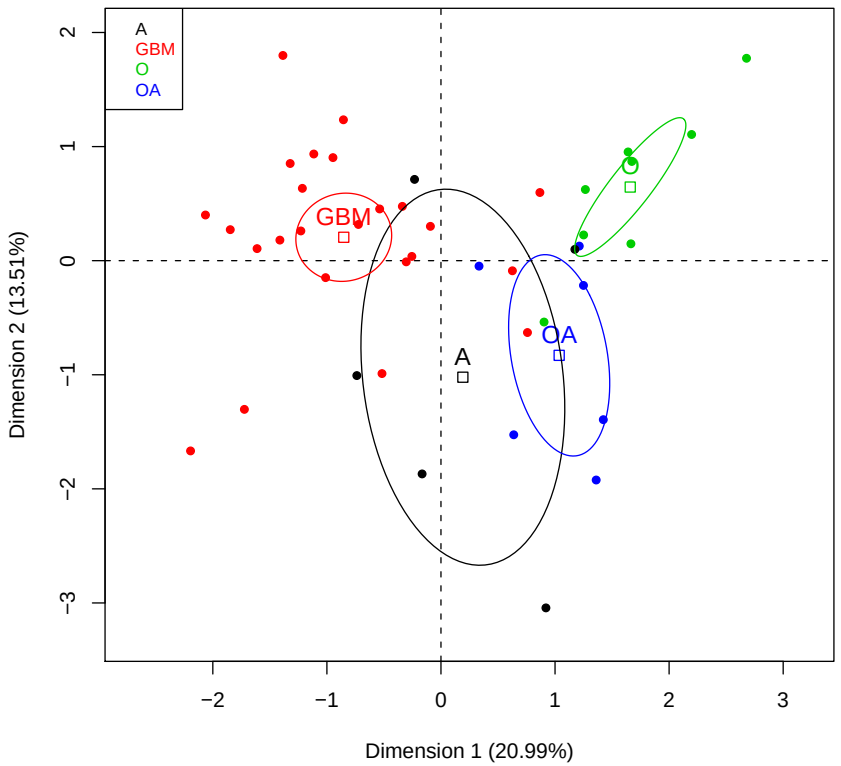


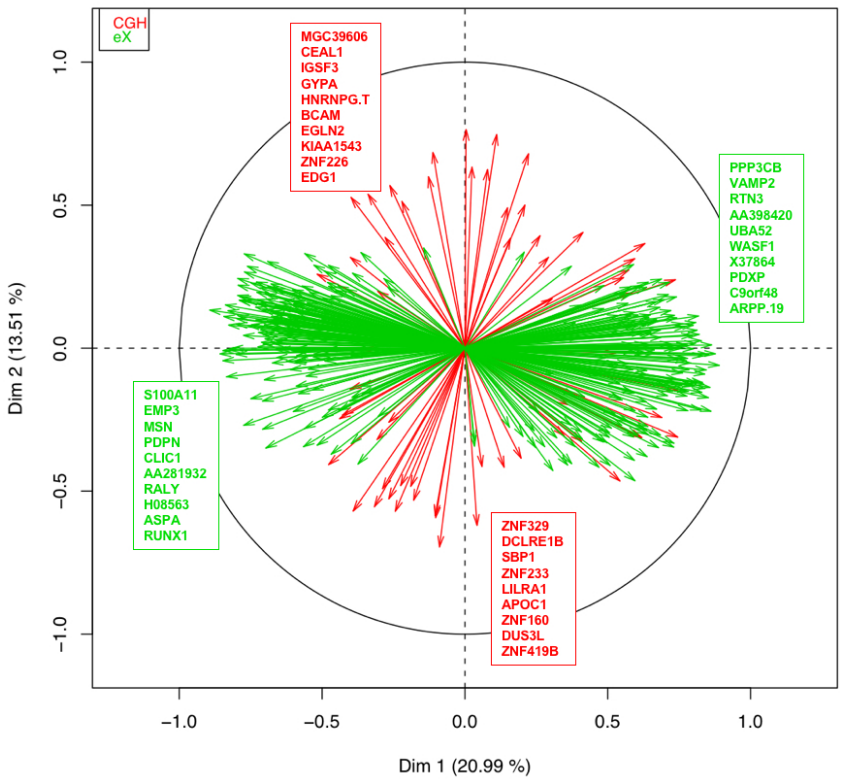
Balancing the groups of variables

This weighting allows that:

- same weight for all the variables of one group: the structure of the group is preserved
- for each group the variance of the main dimension of variability (first eigenvalue) is equal to 1
- no group can generate by itself the first global dimension
- a multidimensional group will contribute to the construction of more dimensions than a one-dimensional group

Individuals and variables representations





Same representations and same interpretation as in PCA

Groups study

$\Rightarrow$ Synthetic comparison of the groups

$\Rightarrow$ Are the relative positions of individuals globally similar from one group to another? Are the partial clouds similar?

$\Rightarrow$ Do the groups bring the same information?

Similarity between two groups

Measure of similarity between groups K_j and K_m:

$$\mathcal{L}_g(K_j, K_m) = \sum_{k \in K_j} \sum_{l \in K_m} cov^2 \left(\frac{x_{.k}}{\lambda_1^k}, \frac{x_{.l}}{\lambda_1^l} \right)$$

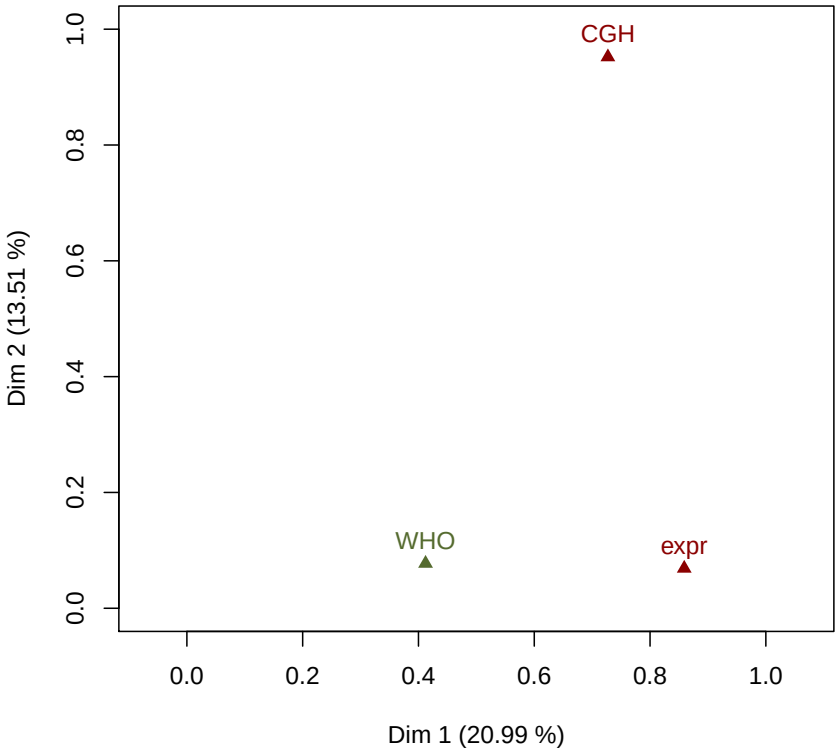
MFA = weighted PCA $\Rightarrow$ first principal component of MFA maximizes

$$\sum_{j=1}^{J} \mathcal{L}_g(v_1, K_j) = \sum_{j=1}^{J} \sum_{k \in K_j} cov^2 \left(\frac{x_{.k}}{\sqrt{\lambda_1^j}}, v_1 \right)$$

Inertia of K_j projected on v_1

Representation of the groups

Group j has the coordinates $(\mathcal{L}_g(v_1, K_j), \mathcal{L}_g(v_2, K_j))$



- 2 groups are all the more close that they induce the same structure
- The 1st dimension is common to all the groups
- 2nd dimension mainly due to CGH

$$0 \leq \mathcal{L}_g(v_1, K_j) = \frac{1}{\lambda_1^j} \underbrace{\sum_{k \in K_j} cov^2(x_{.k}, v_1)}_{\leq \lambda_1^j} \leq 1$$

Numeric indicators

```
> res.mfa$group$Lg
      CGH expr  WHO  MFA
CGH  2.51 0.60 0.46 1.96
expr 0.60 1.10 0.36 1.07
WHO  0.46 0.36 0.50 0.51
MFA  1.96 1.07 0.51 1.91

> res.mfa$group$RV
      CGH expr  WHO  MFA
CGH  1.00 0.36 0.41 0.90
expr 0.36 1.00 0.48 0.74
WHO  0.41 0.48 1.00 0.53
MFA  0.90 0.74 0.53 1.00
```

$$\mathcal{L}_g(K_j, K_j) = \frac{\sum_{k=1}^{K_j}(\lambda_k^j)^2}{(\lambda_1^j)^2} = 1 + \frac{\sum_{k=2}^{K_j}(\lambda_k^j)^2}{(\lambda_1^j)^2}$$

- CGH gives richer description ($\mathcal{L}_g$ greater)
- RV: a standardized $\mathcal{L}_g$
- CGH and expr are not linked (RV=0.36)
- CGH closest to the overall (RV=0.90)

Contribution of each group to each component of the MFA

```
> res.mfa$group$contrib
     Dim.1 Dim.2 Dim.3
CGH   45.8  93.3  78.1
expr  54.2   6.7  21.9
```

- Similar contribution of the 2 groups to the first dimension
- Second dimension only due to CGH

The RV coefficient

$X_{j_{(I\times K_j)}}$ and $X_{m_{(I\times K_m)}}$ not directly comparable
$W_{j_{(I\times I)}} = X_j X_j'$ and $W_{m_{(I\times I)}} = X_m X_m'$ can be compared
Inner product matrices = relative position of the individuals

Covariance between two groups:

$$< W_j, W_m >= \sum_{k\in K_j}\sum_{l\in K_m} cov^2(x_{.k}, x_{.l})$$

Correlation between two groups:

$$RV(K_j, K_m) = \frac{< W_j, W_m >}{\|W_j\| \ \|W_m\|} \qquad 0 \leq RV \leq 1$$

$RV = 0$: variables of K_j are uncorrelated with variables of K_m
$RV = 1$: the two clouds of points are homothetic

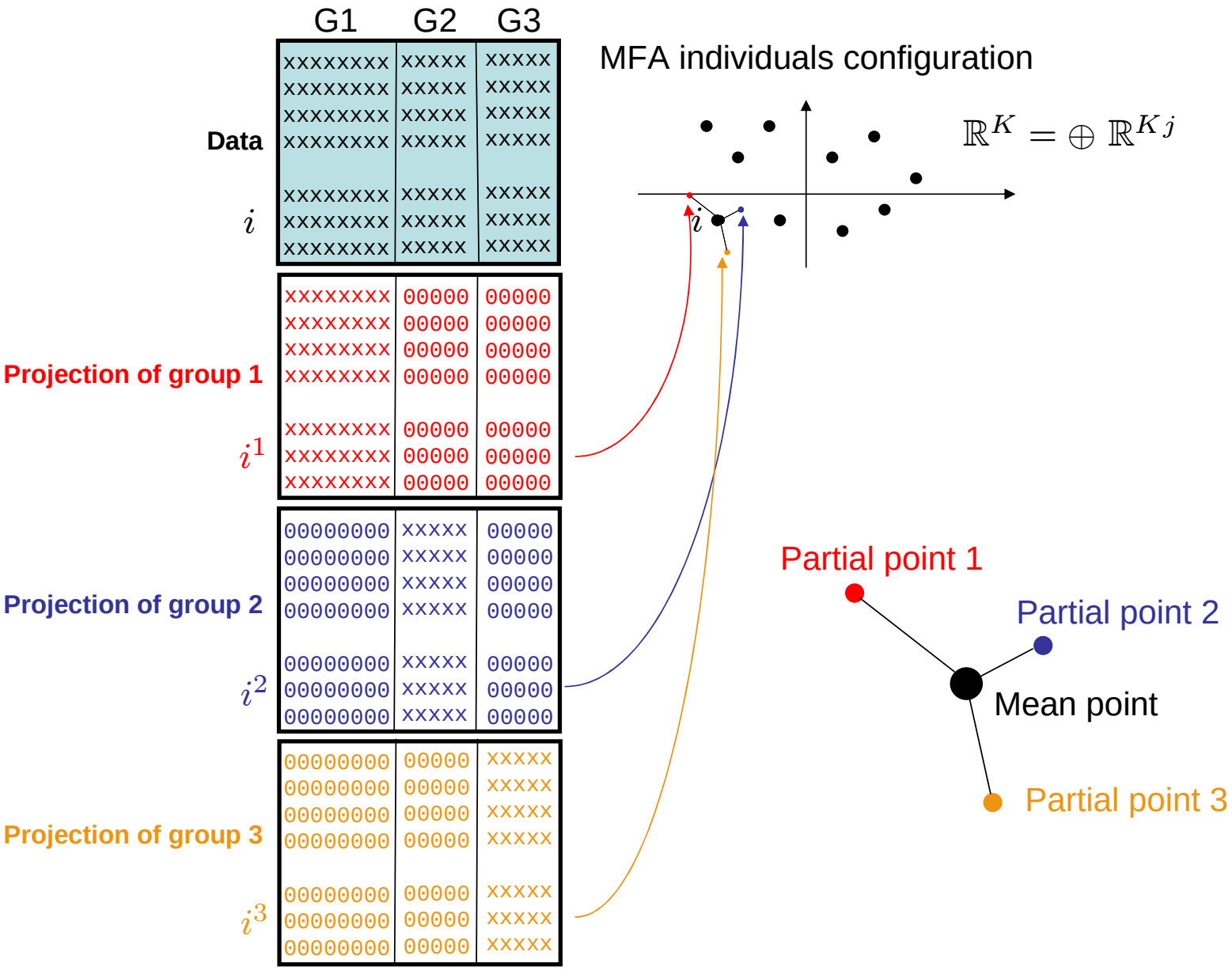
Partial analyses

- Comparison of the groups through the individuals

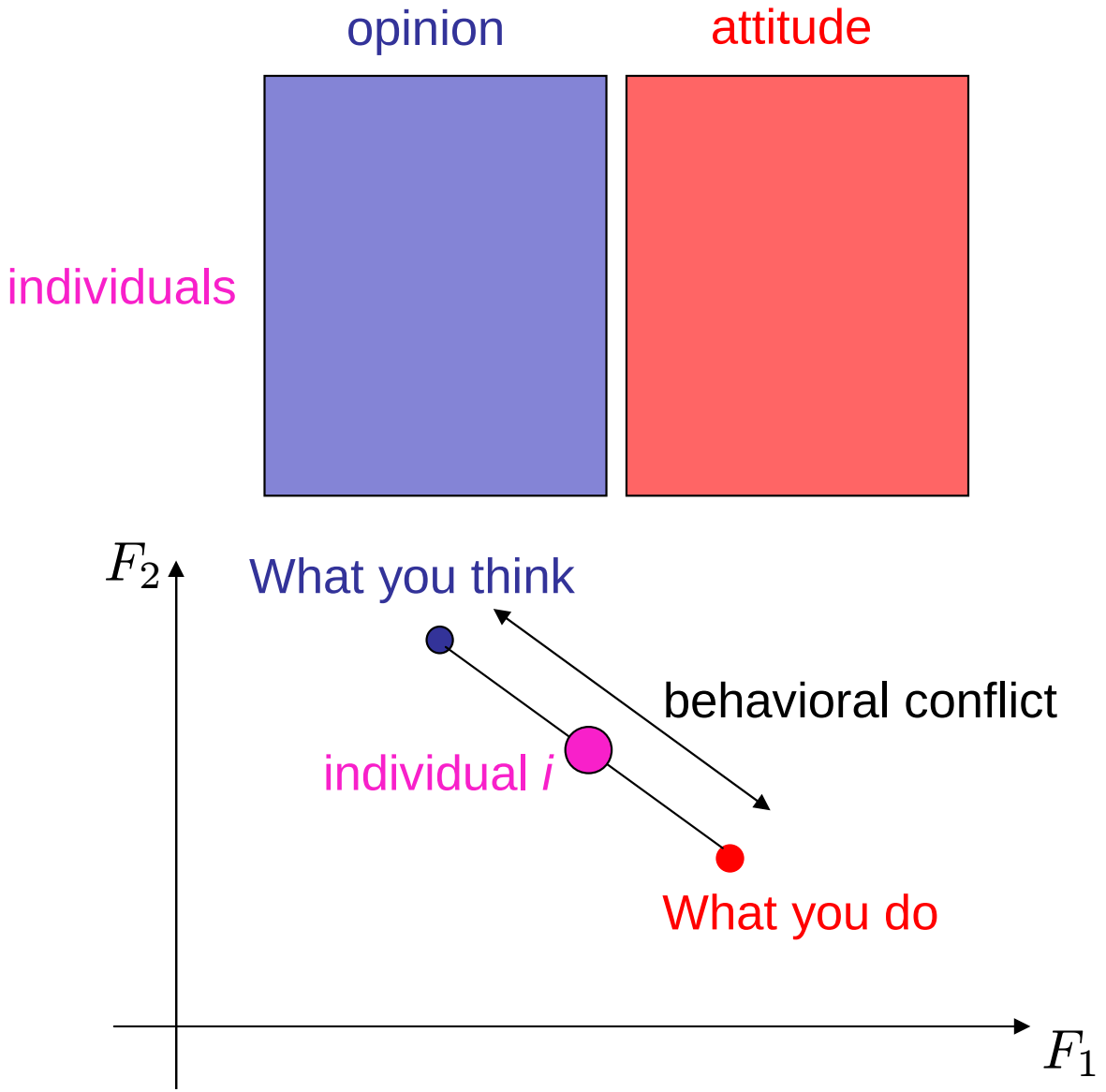
$\Rightarrow$ Comparison of the typologies provided by each group in a common space
$\Rightarrow$ Are there individuals very particular with respect to one group?

- Comparison of the separate PCA

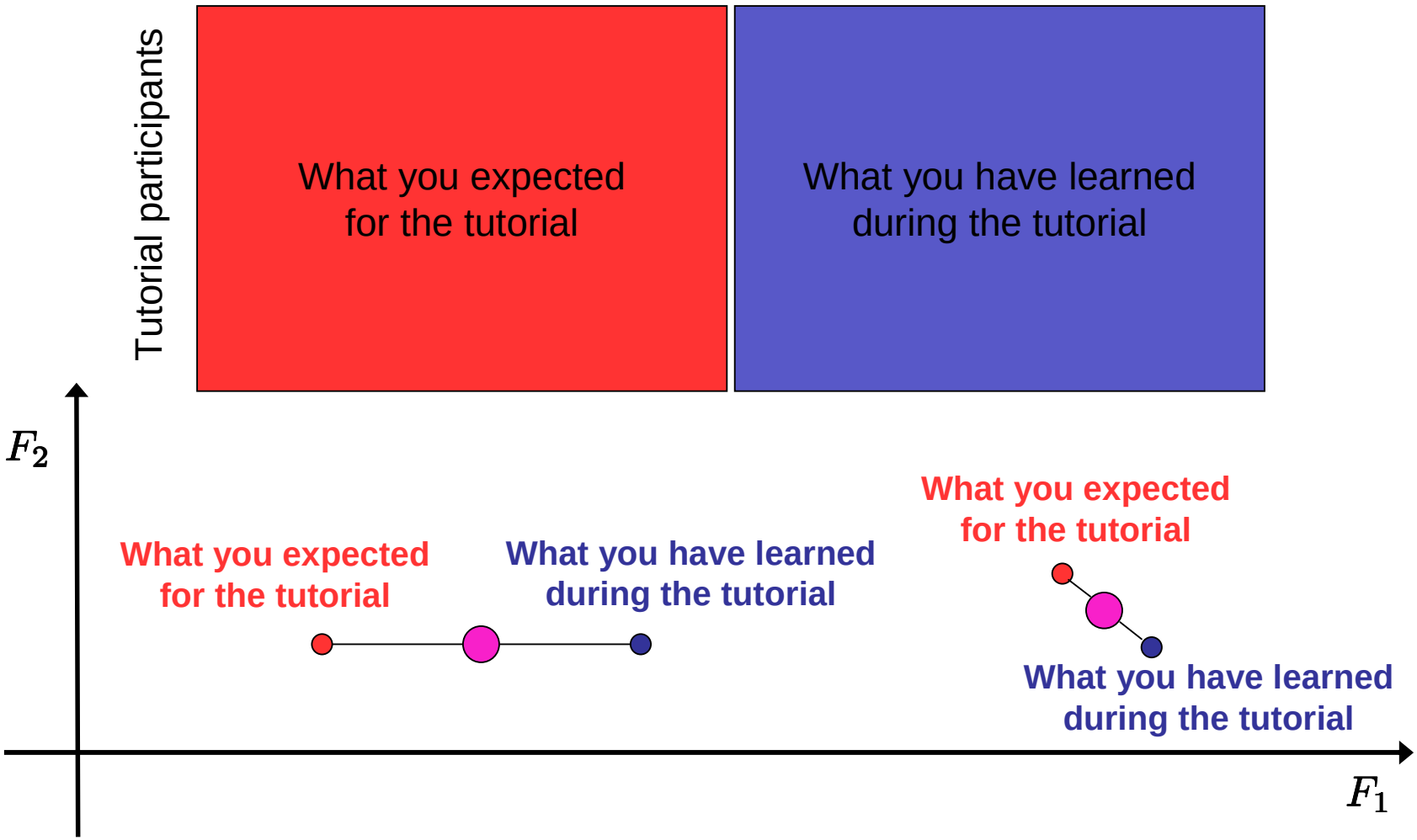
Projection of partial points



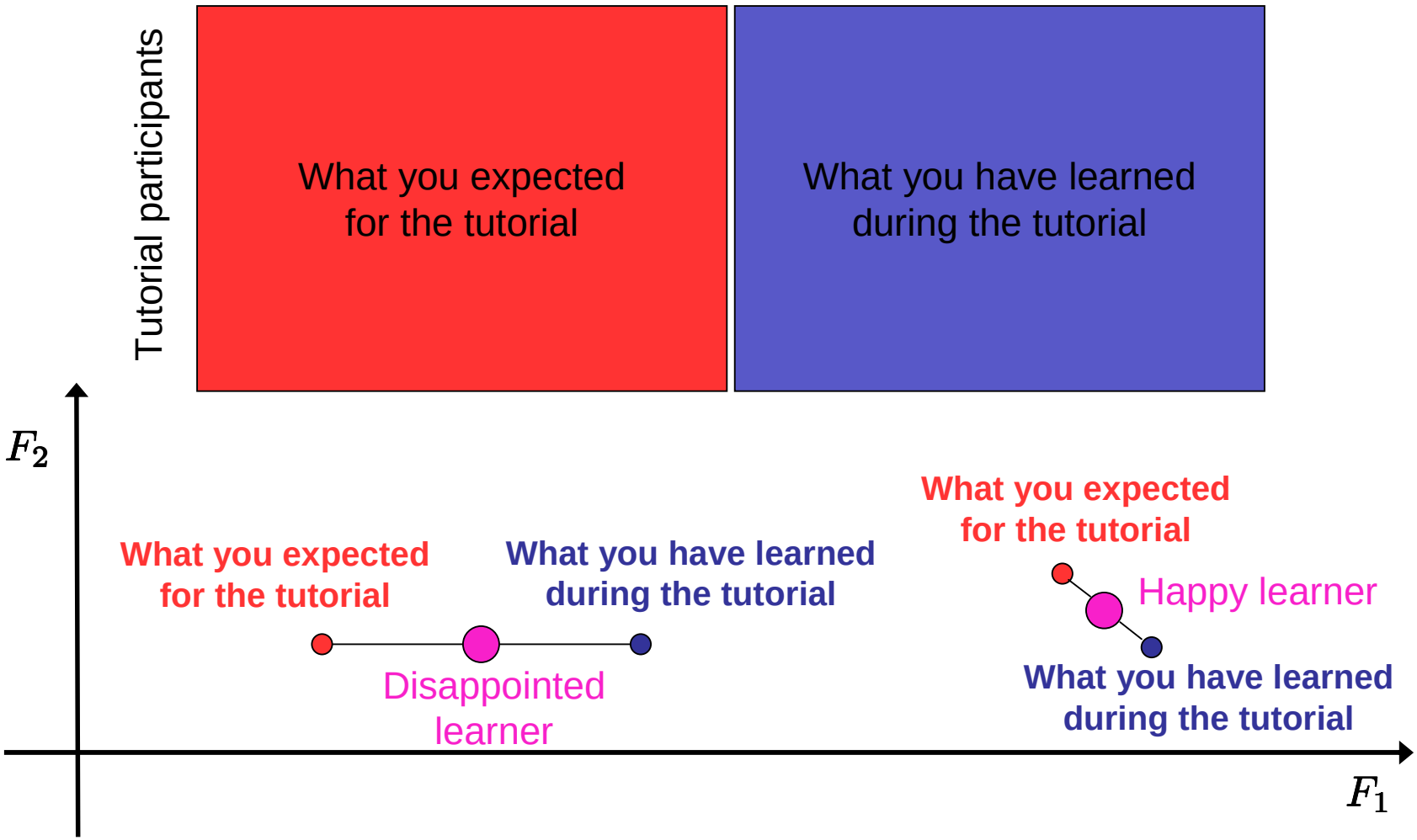
Partial points



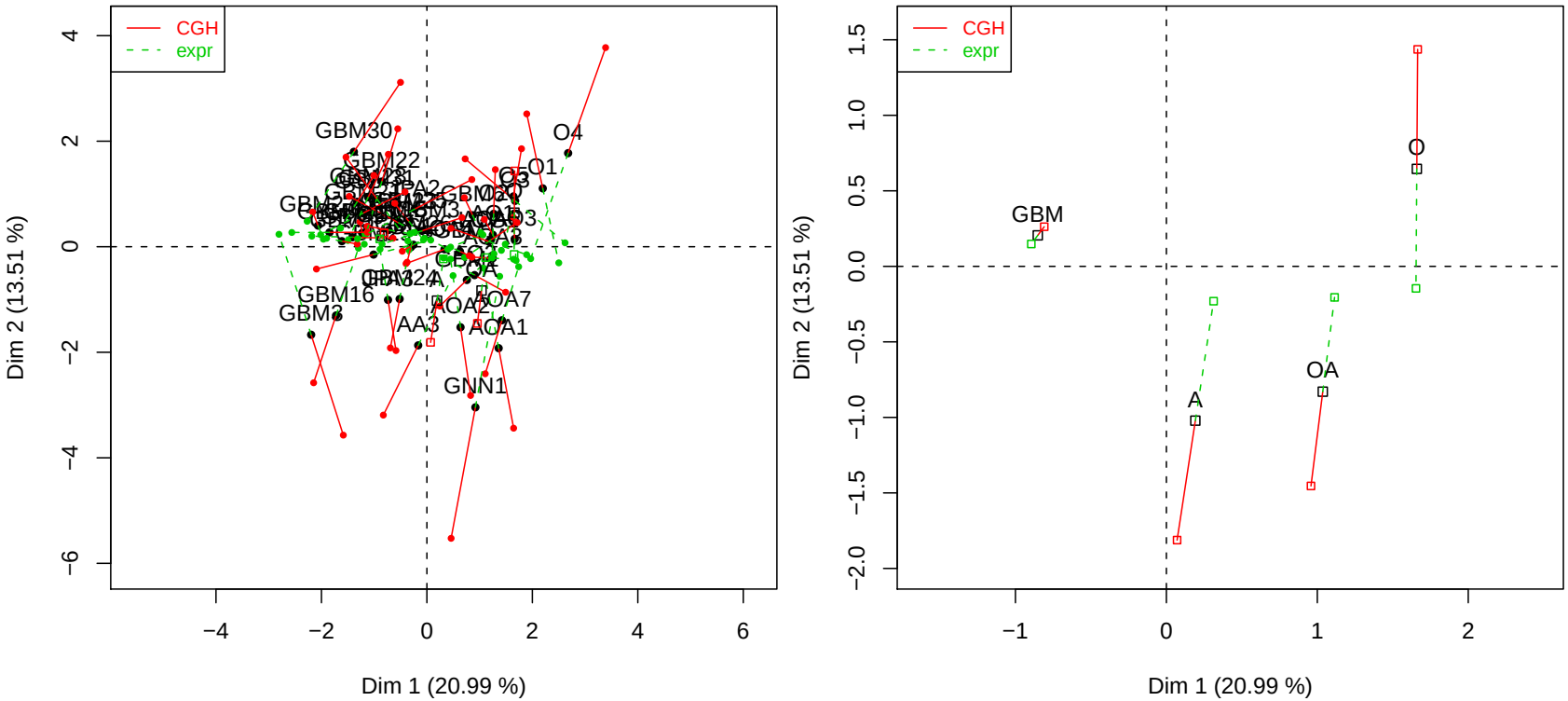
Partial points



Partial points



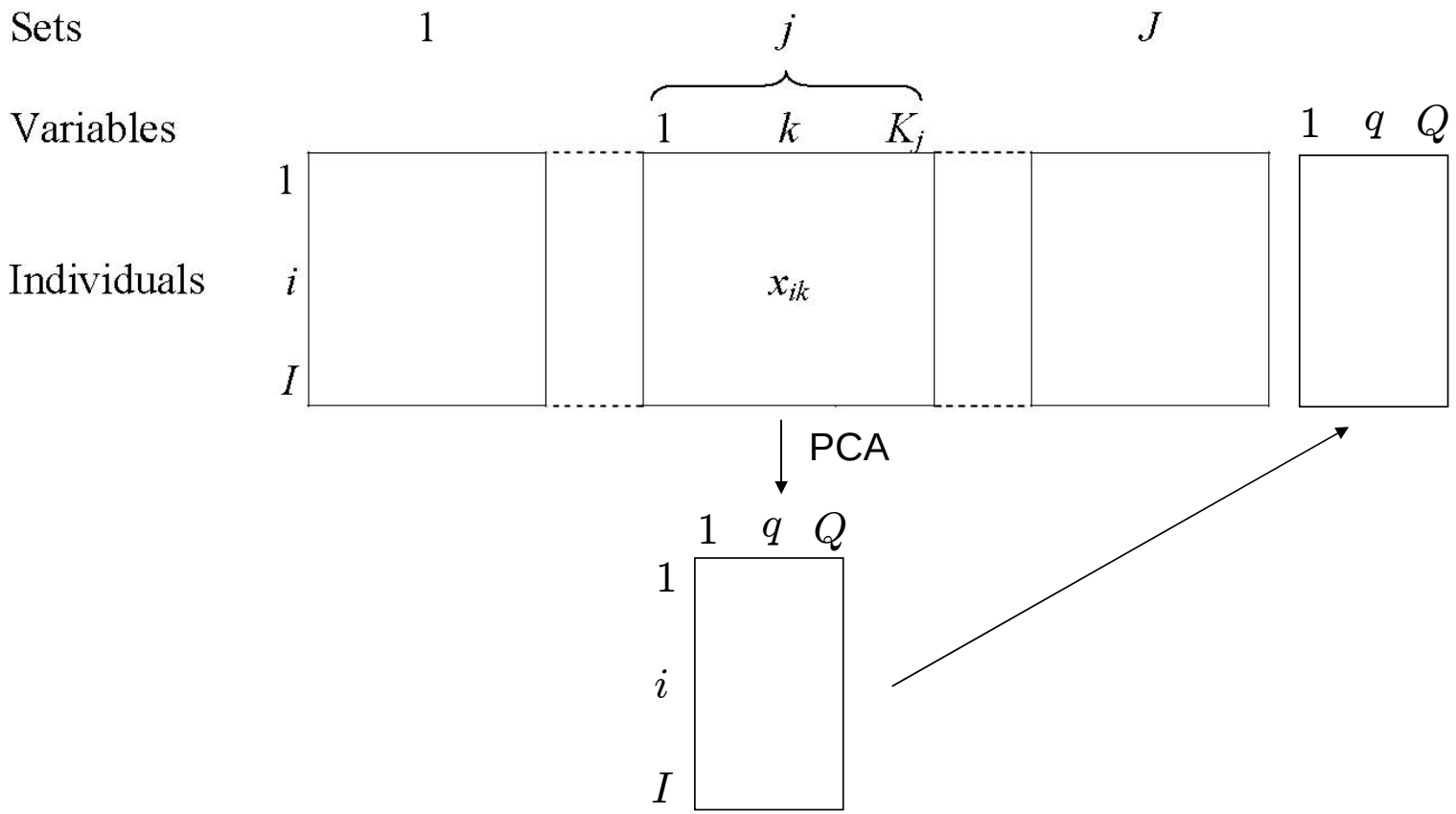
Representation of the partial points



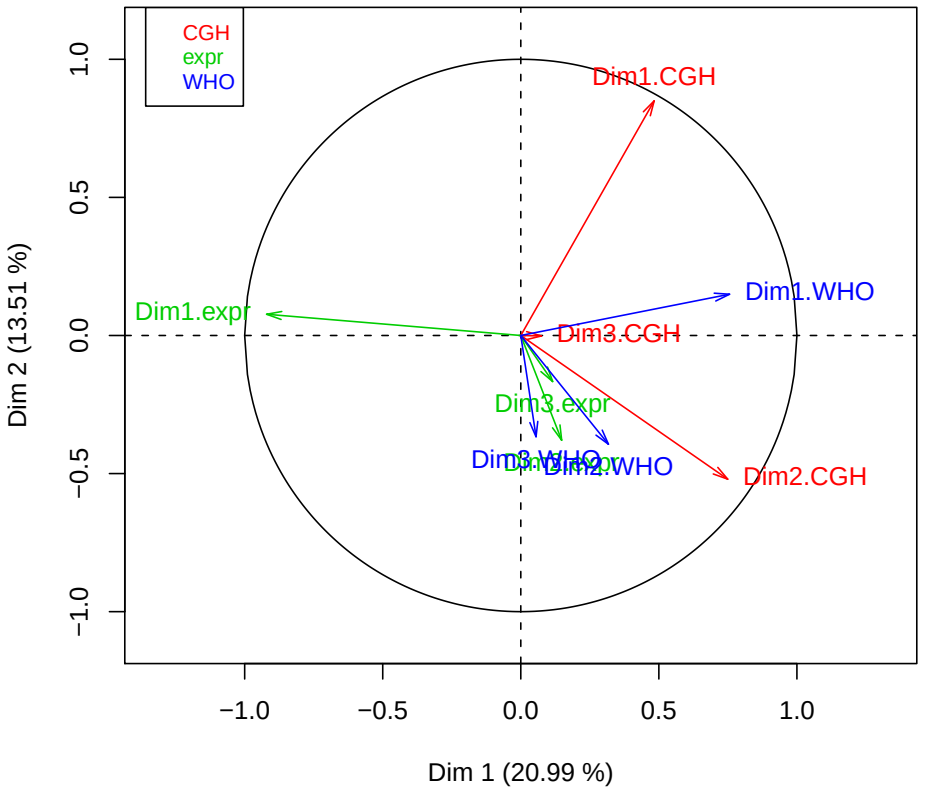
- an individual is at the barycentre of its partial points
- an individual is all the more "homogeneous" that its superposed representations are close (res.mfaindwithin.inertia)

Representation of the partial components

Do the separate analyses give similar dimensions as MFA?



Representation of the partial components



- The first dimension of each group is well projected
- CGH has same dimensions as MFA

Use of biological knowledge

Genes can be grouped by gene ontology (GO) biological process

GO:0006928
cell motility

ANXA1
CALD1
EGFR
ENPP2
FN1
FPRL2
LSP1
MSN
PDPN
PLAUR
PRSS3
SAA2
SPINT2
TNFRSF12A
VEGF
WASF1
YARS

GO:0009966
regulation of signal transduction

CASP1
EDG2
F2R
HCLS1
HMOX1
IGFBP3
IQSEC1
LYN
MALT1
TCF7L1
TNFAIP3
TRIO
VEGF
YWHAG
YWHAH

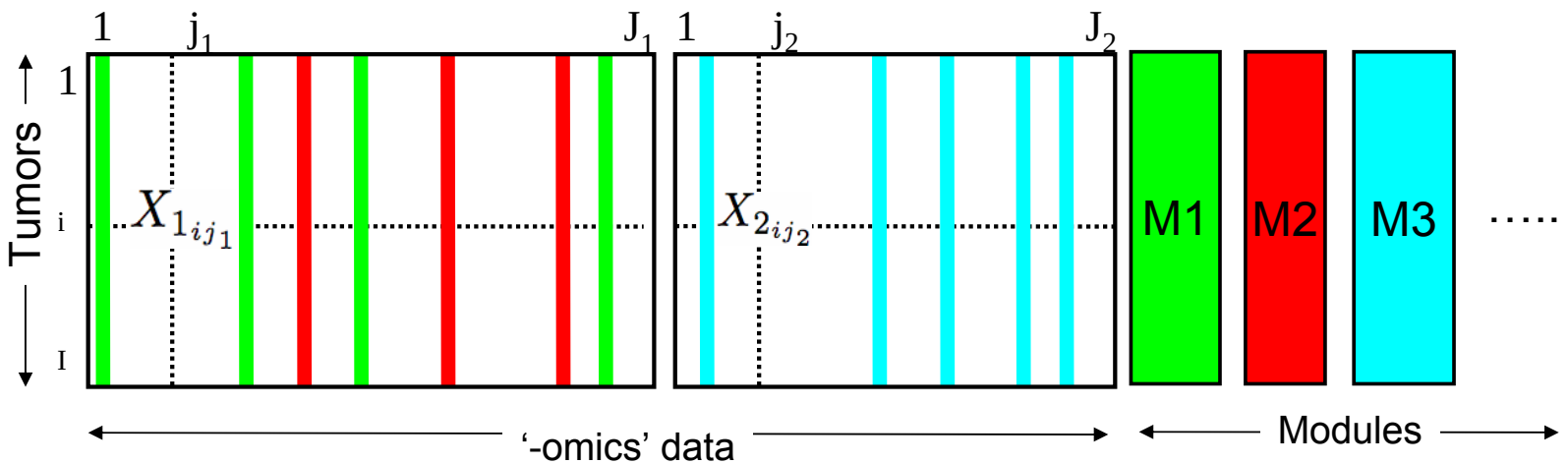
GO:0052276
chromosome organisation and biogenesis

CBX6
NUSAP1
PCOLN3
PTTG1
SUV39H1
TCF7L1
TSPYL1

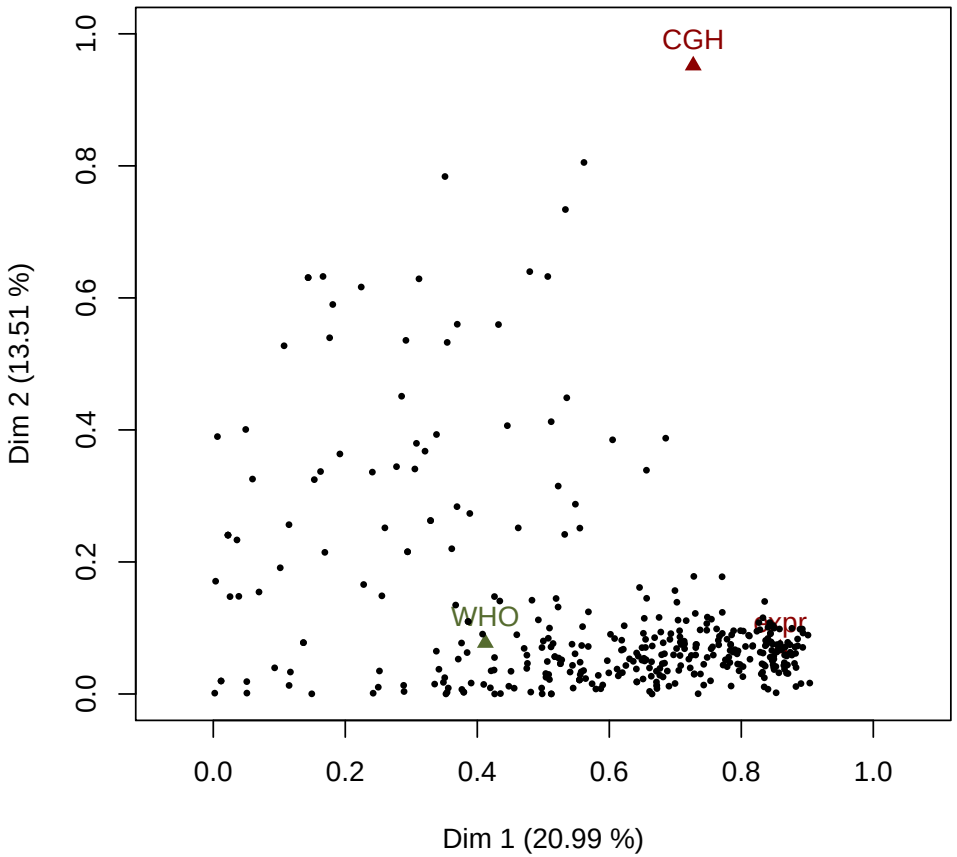
Use of biological knowledge

- Biological processes considered as supplementary groups of variables

Modular approach



Use of biological knowledge



Many biological processes induce the same structure on the individuals than MFA

Back to the wine example!

Continuous variables

Categorical

	Expert (27)	Consumer (15)	Student (15)	Preference (60)	Label (1)
wine 1					
wine 2					
...					
wine 10					

Objectives:

- How are the products described by the panels?
- Do the panels describe the products in a same way? Is there a specific description done by one panel?

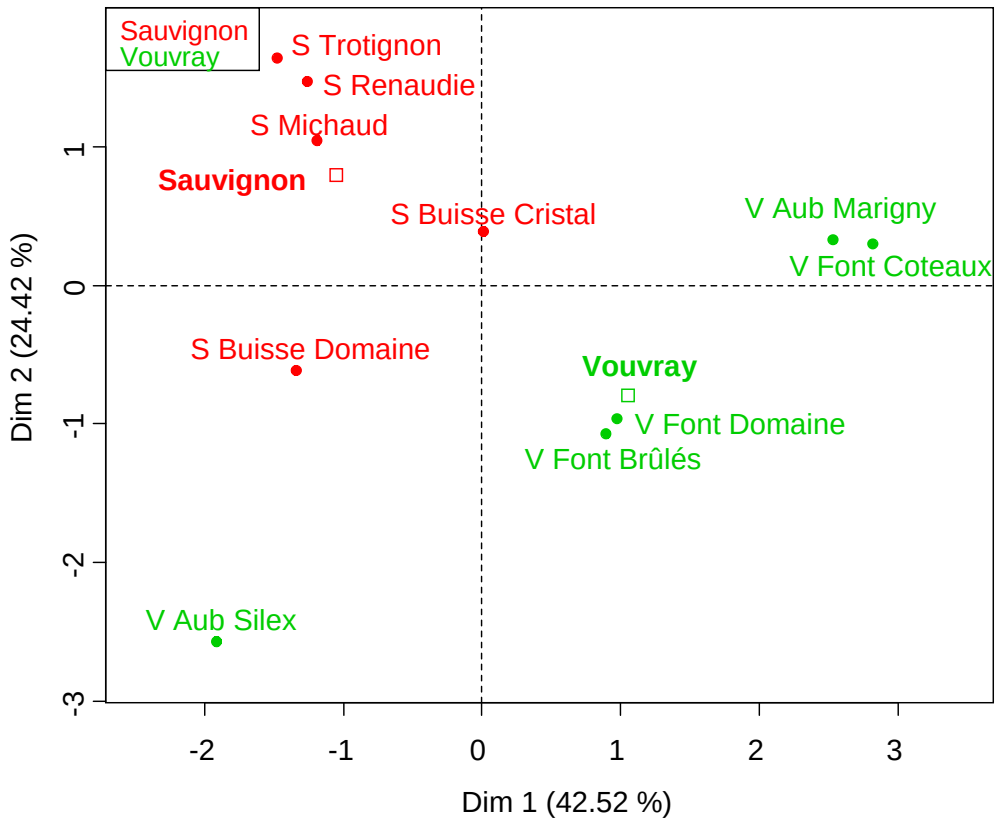
Practice with R

1. Define groups of active and supplementary variables
2. Scale or not the variables
3. Perform MFA
4. Choose the number of dimensions to interpret
5. Simultaneously interpret the individuals and variables graphs
6. Study the groups of variables
7. Study the partial representations
8. Use indicators to enrich the interpretation

Practice with R

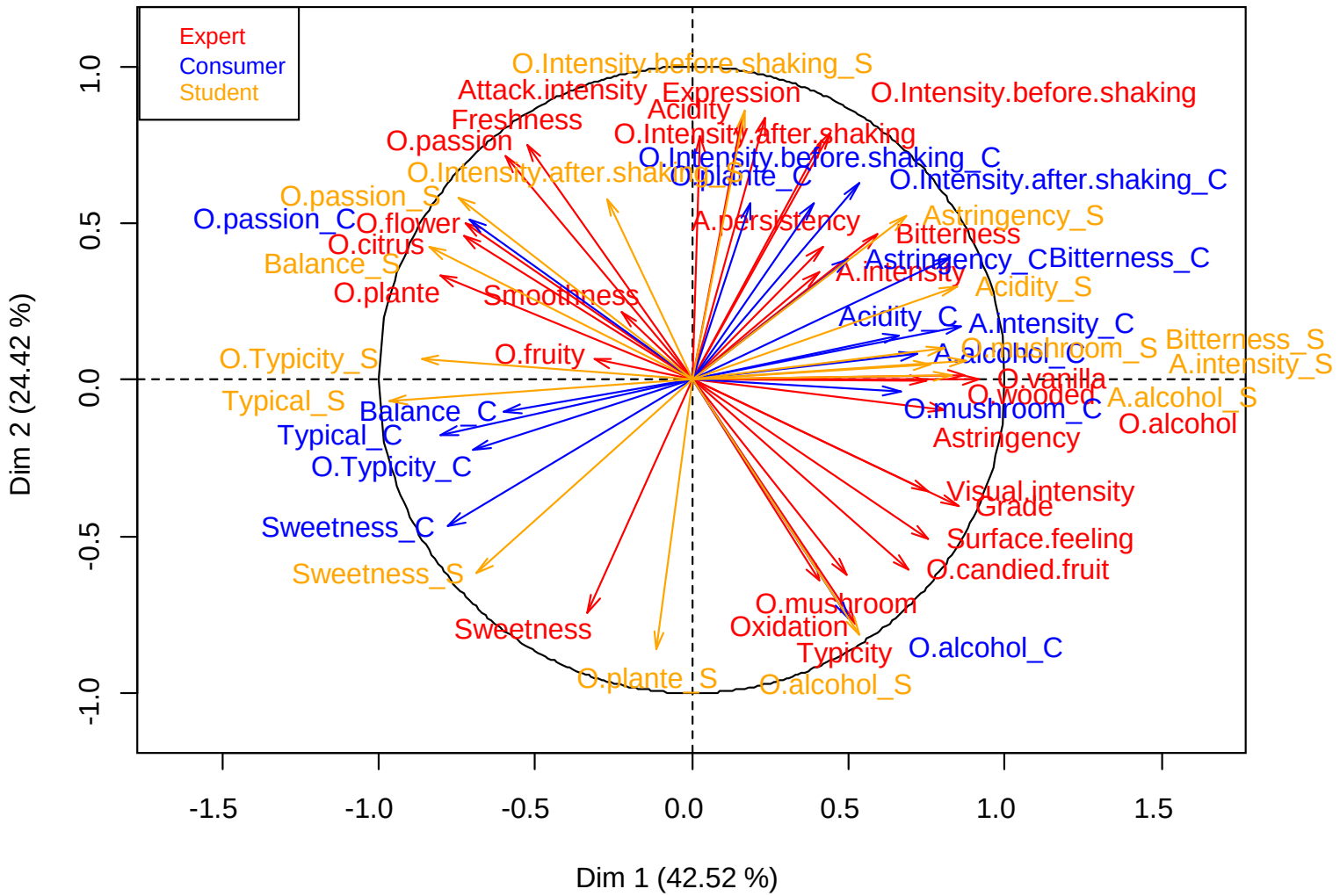
```
library(FactoMineR)
Expert <- read.table("http://factominer.free.fr/useR2010/Expert_wine.csv",
   header=TRUE, sep=";", row.names=1)
Consu <- read.table(".../Consumer_wine.csv",header=T,sep=";",row.names=1)
Stud <- read.table(".../Student_wine.csv",header=T,sep=";",row.names=1)
Pref <- read.table(".../Pref_wine.csv",header=T,sep=";",row.names=1)

palette(c("black","red","blue","orange","darkgreen","maroon","darkviolet"))
complet <- cbind.data.frame(Expert[,1:28],Consu[,2:16],Stud[,2:16],Pref)
res.mfa <- MFA(complet,group=c(1,27,15,15,60),type=c("n",rep("s",4)),
   num.group.sup=c(1,5),graph=FALSE,
   name.group=c("Label","Expert","Consumer","Student","Preference"))
plot(res.mfa,choix="group",palette=palette())
plot(res.mfa,choix="var",invisible="sup",hab="group",palette=palette())
plot(res.mfa,choix="var",invisible="actif",lab.var=FALSE,palette=palette())
plot(res.mfa,choix="ind",partial="all",habillage="group",palette=palette())
plot(res.mfa,choix="axes",habillage="group",palette=palette())
dimdesc(res.mfa)
write.infile(res.pca,file="my_FactoMineR_results.csv") #to export a list
```

Representation of the individuals

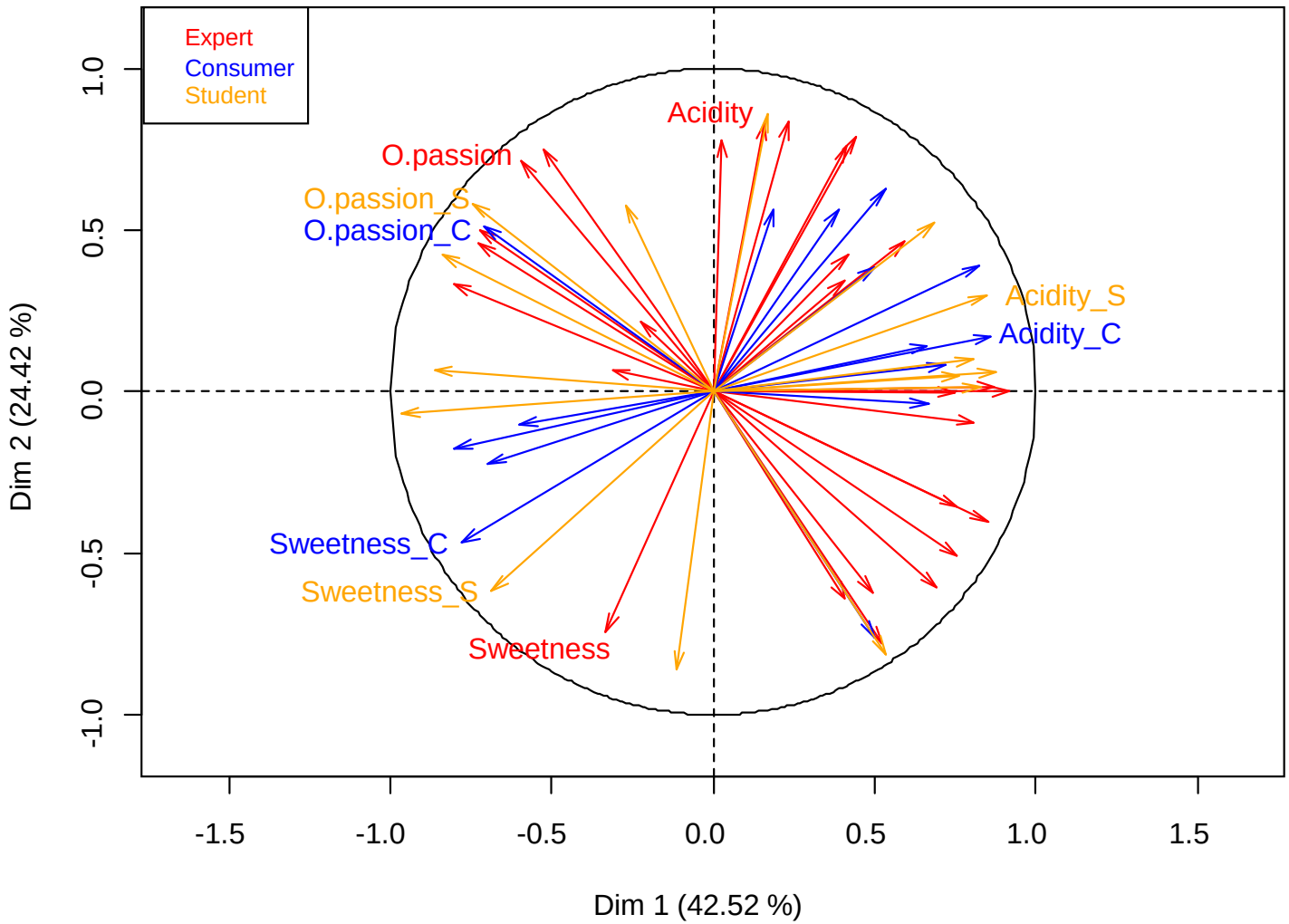


- The two labels are well separated
- Vouvray are sensorially more different
- Several groups of wines, ...

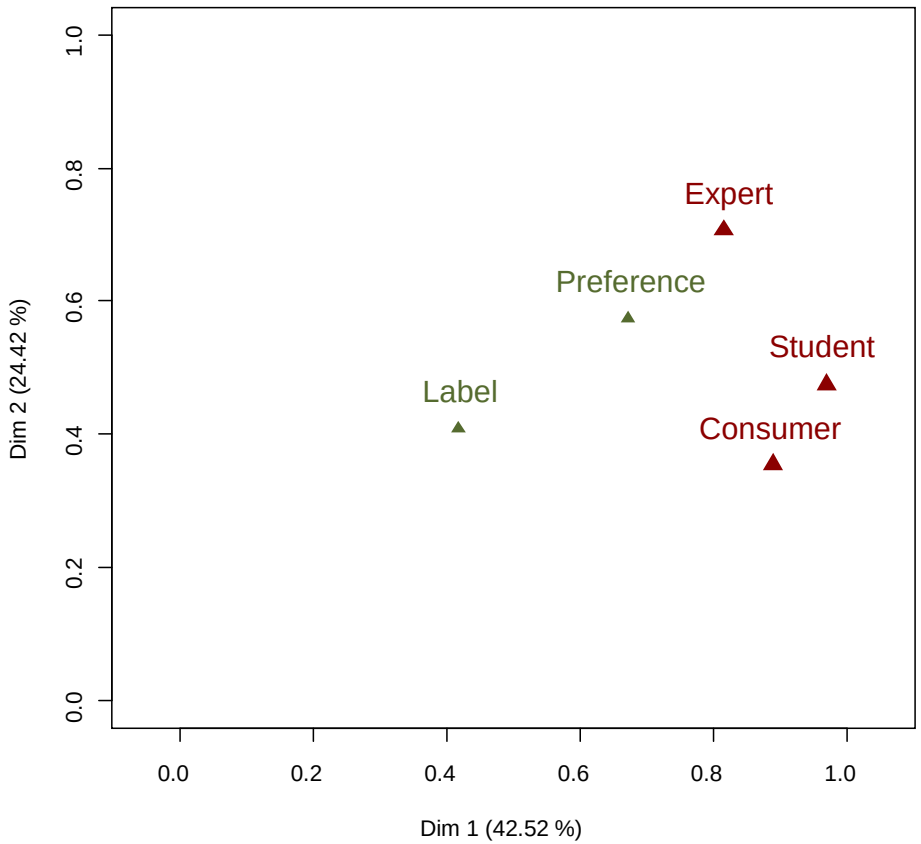
Representation of the active variables



Representation of the active variables

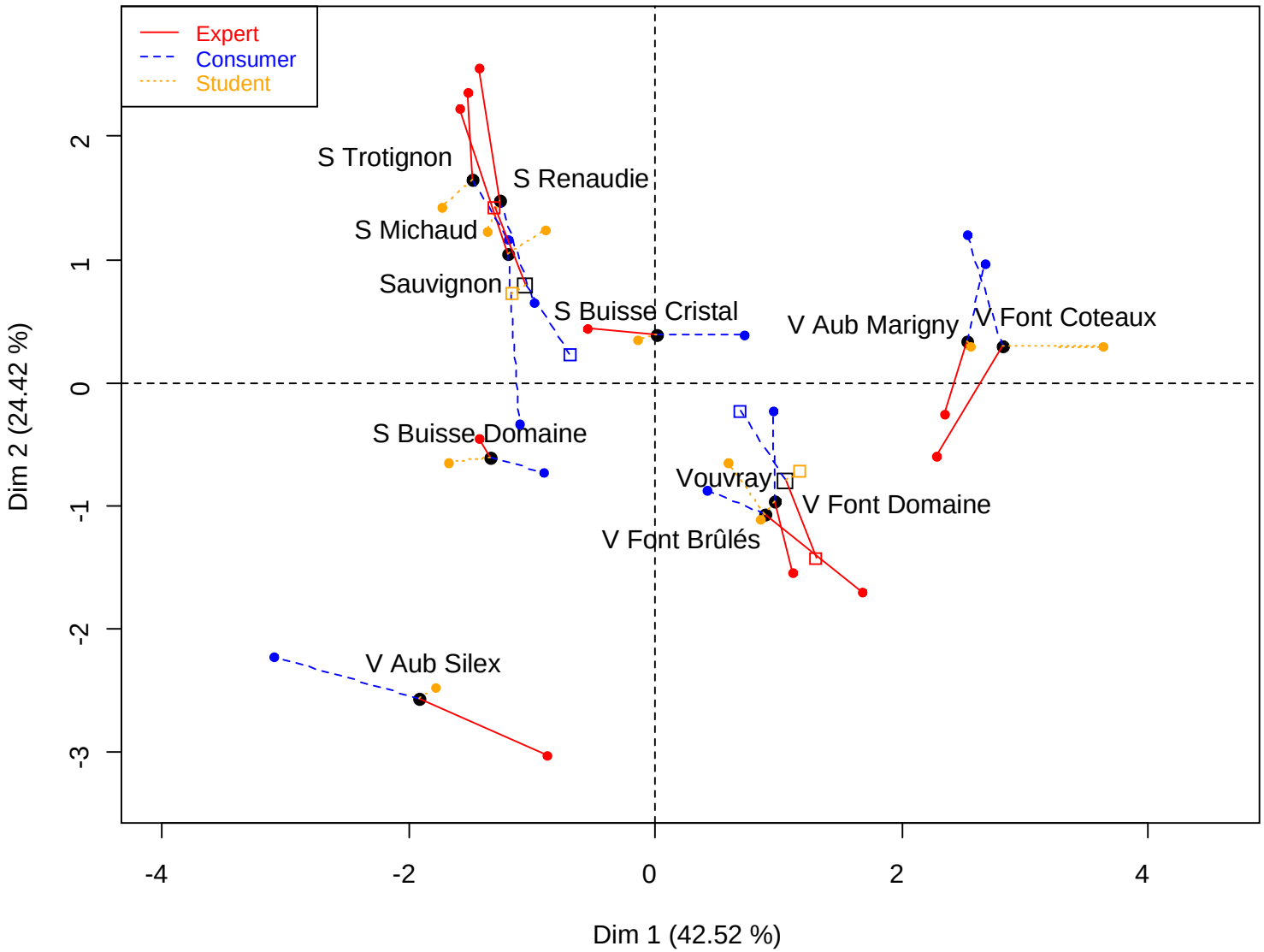


Representation of the groups

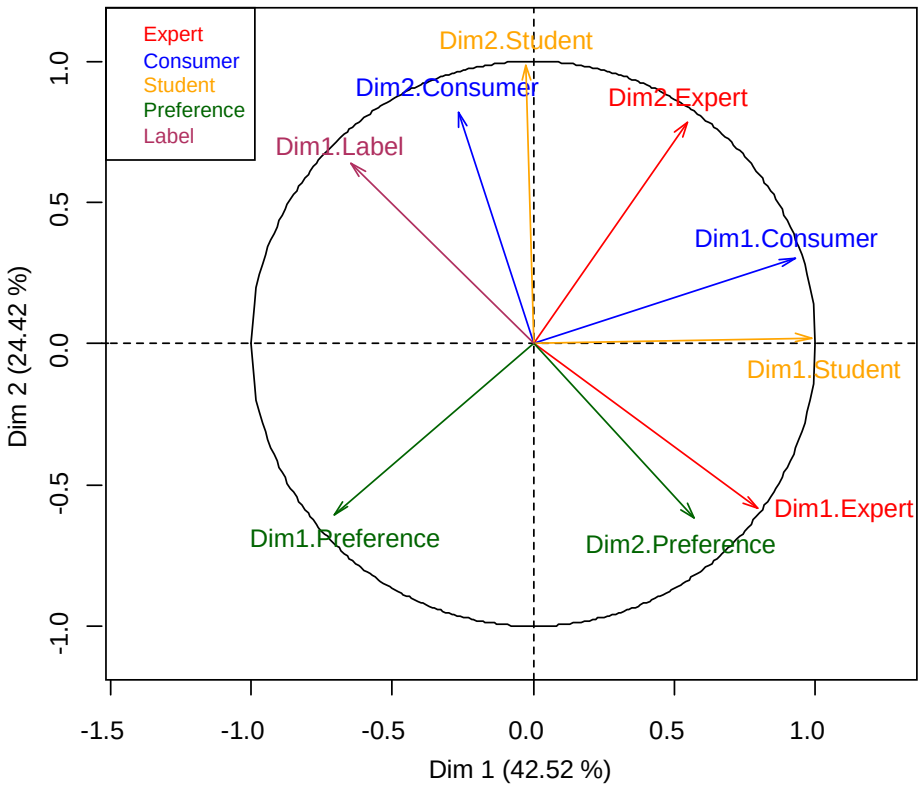


- 2 groups are all the more close that they induce the same structure
- The 1st dimension is common to all the panels
- 2nd dimension mainly due to the experts
- Preference linked to sensory description

Representation of the partial points

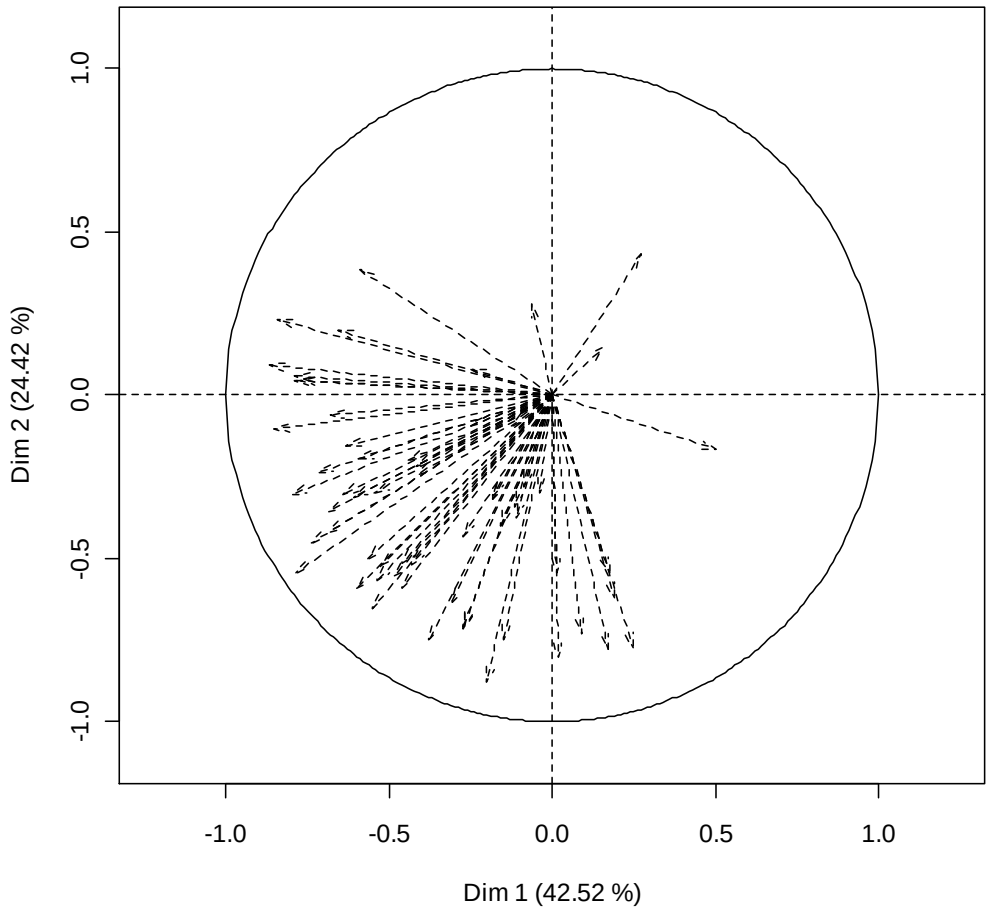


Representation of the partial dimensions



- The two first dimensions of each group are well projected
- Consumer has same dimensions as MFA

Representation of supplementary continuous variables



Preferences are linked to sensory description

The favourite wine is *Vouvray Aubussière Silex*

Helps to interpret

- Contribution of each group of variables to each component of the MFA

```
> res.mfa$group$contrib
          Dim.1 Dim.2 Dim.3
Expert     30.5  46.0  33.7
Consumer   33.2  23.1  31.2
Student    36.3  30.9  35.1
```

  - Similar contribution of the 3 groups to the first dimension
  - Second dimension mainly due to the expert

- Correlation between the global cloud and each partial cloud

```
> res.mfa$group$correlation
          Dim.1 Dim.2 Dim.3
Expert     0.95  0.95  0.96
Consumer   0.95  0.83  0.87
Student    0.99  0.99  0.84
```

First components are highly linked to the 3 groups: the 3 clouds of points are nearly homothetic

Similarity measures between groups

```
> res.mfa$group$Lg
           Expert Consumer Student Preference Label  MFA
Expert       1.45     0.94    1.17       1.01  0.89 1.33
Consumer     0.94     1.25    1.04       1.11  0.28 1.21
Student      1.17     1.04    1.29       1.03  0.62 1.31
Preference   1.01     1.11    1.03       1.47  0.37 1.18
Label        0.89     0.28    0.62       0.37  1.00 0.67
MFA          1.33     1.21    1.31       1.18  0.67 1.44

> res.mfa$group$RV
           Expert Consumer Student Preference Label  MFA
Expert       1.00     0.70    0.85       0.69  0.74 0.92
Consumer     0.70     1.00    0.82       0.82  0.25 0.90
Student      0.85     0.82    1.00       0.75  0.55 0.96
Preference   0.69     0.82    0.75       1.00  0.31 0.81
Label        0.74     0.25    0.55       0.31  1.00 0.56
MFA          0.92     0.90    0.96       0.81  0.56 1.00
```

- Expert gives a richer description ($\mathcal{L}_g$ greater)
- Groups Student and Expert are linked (RV = 0.85)
- Group Student is the closest to the overall (RV = 0.96)

To go further

- Mixed data: MFA with 1 group = 1 variable
  if there are only continuous variables, PCA is recovered; if there are only categorical variables, MCA is recovered
  a specific function: AFDM
- MFA used for methodological purposes:
  - comparison of coding (continuous or categorical)
  - comparison between preprocessing (standardized PCA and unstandardized PCA)
  - comparison of results from different analyses
- Hierarchical Multiple Factor Analysis
  Takes into account a hierarchy on the variables: variables are grouped and subgrouped (like in questionnaires structured in topics and subtopics)