Analyse de données avec R
Complémentarité des méthodes d'analyse factorielle et de classification

François Husson & Julie Josse

Laboratoire de mathématiques appliquées
Agrocampus Rennes
husson@agrocampus-ouest.fr

SFdS SOCIÉTÉ FRANÇAISE DE STATISTIQUE Marseille - 2010

FACTOMINER

Plan

- Qu'est-ce que  et pourquoi FACTOMINER  ?
- Les méthodes classiques et les aides à l'interprétation
- Traitement de données multi-tableaux
- Complémentarité analyse factorielle - classification
- Interface graphique

Qu'est-ce que  ?

- Logiciel gratuit de plus en plus utilisé
- Disponible sous Windows, Mac, Linux
- Permettant d'utiliser et de fabriquer des librairies de fonctions pointues (plus de 2500 librairies)
- Évolution rapide du logiciel, disponibilité immédiate de nouvelles méthodes ou méthodologies
- Graphes et sorties facilement exportables
- Possibilité de modifier / sélectionner les données facilement
- Possibilité de combiner programmation et utilisation de fonctions pré-définies

Pourquoi FACTOMINER ?

Pour pouvoir faire de l'analyse de données

- en utilisant un point de vue géométrique permettant de dessiner des graphes
- en ayant la possibilité d'ajouter de l'information supplémentaire
- sur de nouvelles méthodes (prenant en compte différentes structures sur les données)
- à l'aide d'une interface graphique simple d'utilisation et orientée vers l'utilisateur

Les méthodes factorielles

Différentes méthodes pour différents formats de données

Données	Méthode	Fct.
Variables quantitatives	An. en Composantes Principales	PCA
Table de contingence	An. des Correspondances	CA
Variables qualitatives	An. des Correspondances Multiples	MCA
Variables quantitatives et qualitatives	An. Factorielle de données mixtes	AFDM
Groupes de variables	An. Factorielle Multiple	MFA
Hiérarchie sur les variables	An. Factorielle Multiple Hiérarchique	HMFA
Groupes d'individus	An. Factorielle Multiple Duale	DMFA

Les méthodes de classification

Méthodes de classification	Fonction
Classification ascendante hiérarchique	HCPC
K-means	kmeans

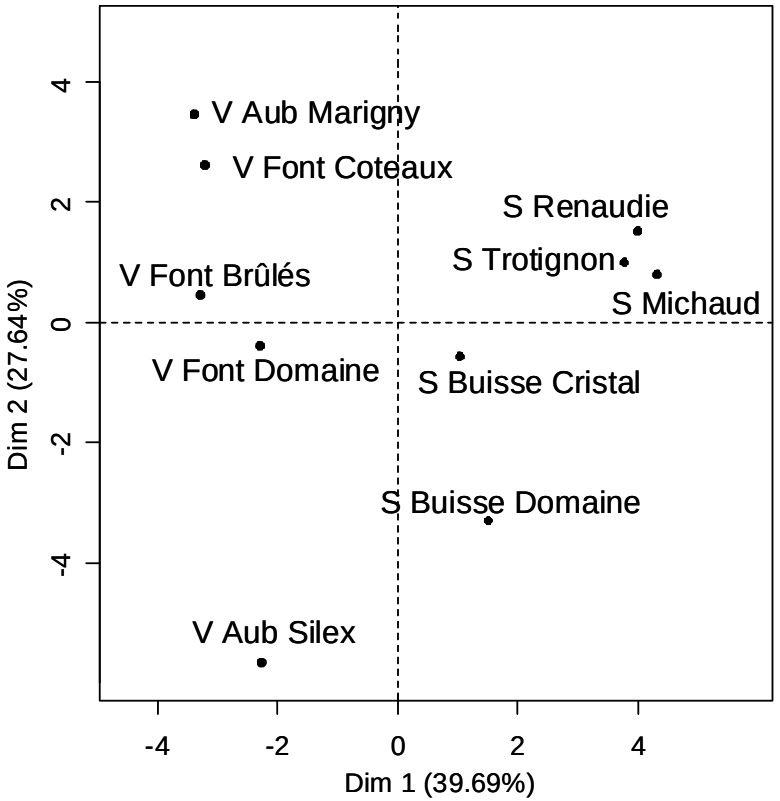
Méthodes outils	Fonction
Description d'une variable quantitative par des variables quantitatives et/ou qualitatives (ex. d'une dimension)	condes, dimdesc
Description d'une variable qualitative par des variables quantitatives et/ou qualitatives (ex. d'une classe)	catdes
Construction d'un tableau de données textuel	textuel

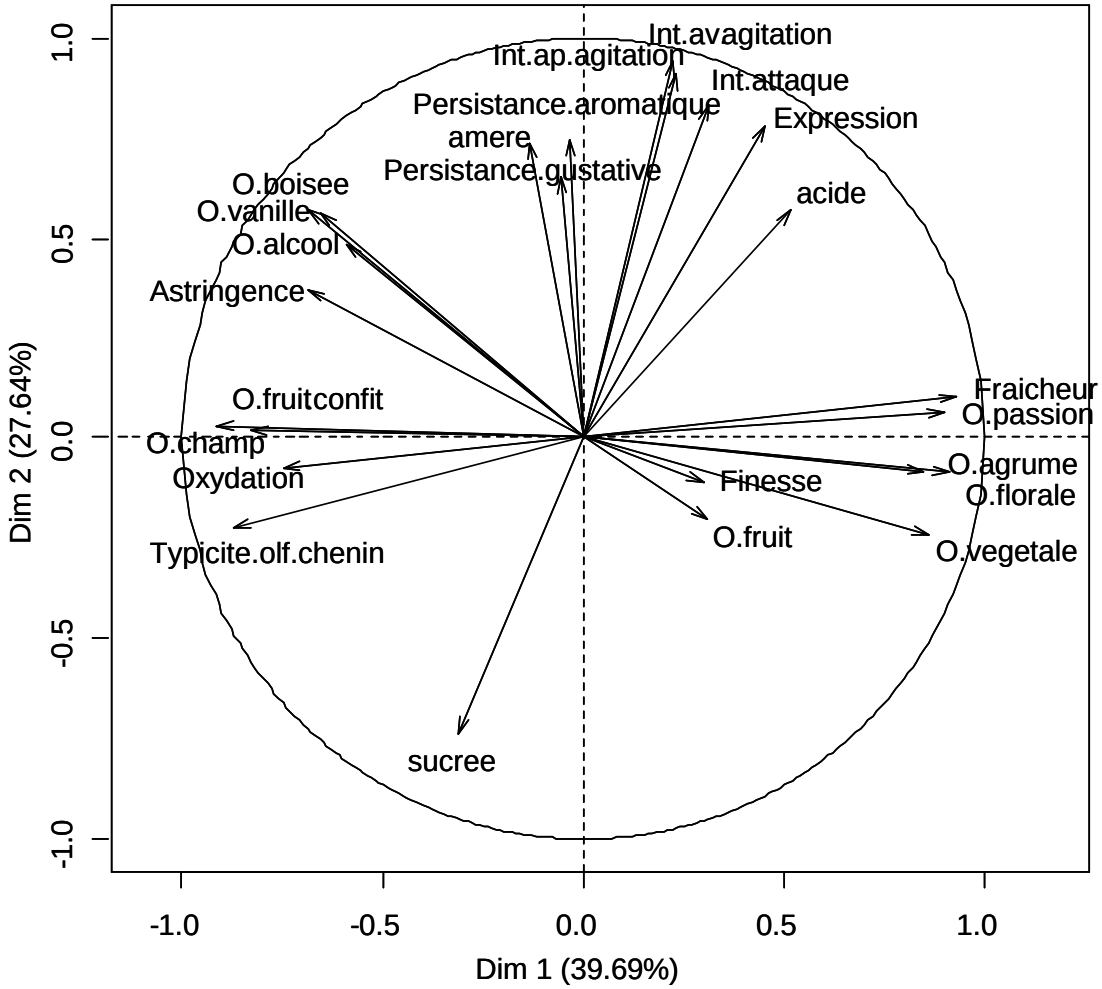
Exemple en ACP

Les données : 10 vins blancs (5 Sauvignons, 5 Vouvrays) évalués par des experts selon 27 variables sensorielles

	Int.av.agitation	Int.ap.agitation	Expression	O.fruit	O.passion	O.agrume	O.fruit.confit	...	Astringence	Fraicheur	Oxydation	Finesse	Persistance.gustative	Persistance.aromatique	Appréciation.hédonique	Appréciation.olfactive	cepage
S Michaud	7.8	8.0	7.1	4.3	2.4	5.7	0.7		1.4	6.6	0.4	5.3	7.1	6.7	5.0	6.0	Sauvignon
S Renaudie	7.1	7.6	7.0	4.4	3.1	5.3	0.7		2.3	6.6	0.3	5.1	7.2	6.6	5.5	5.4	Sauvignon
S Trotignon	7.1	7.4	7.1	5.1	4.0	5.3	1.0		2.4	6.9	0.4	5.3	6.1	6.1	5.5	5.0	Sauvignon
S Buisse Domaine	5.5	6.3	5.4	4.3	2.4	3.6	1.5		3.0	6.3	0.4	4.6	4.9	5.1	4.6	5.3	Sauvignon
S Buisse Cristal	6.0	6.5	6.2	5.6	3.1	3.5	3.0		3.1	6.4	1.0	5.6	6.1	5.1	5.0	6.1	Sauvignon
V Aub Silex	3.9	4.3	4.4	3.9	0.7	3.3	2.9		2.4	4.7	1.9	5.6	5.9	5.6	5.5	5.0	Vouvray
V Aub Marigny	7.8	8.0	6.0	2.1	0.7	1.0	3.3		4.0	4.9	1.5	4.1	6.3	6.7	4.1	5.1	Vouvray
V Font Domaine	6.3	6.8	6.0	5.1	0.5	2.5	4.5		2.5	5.3	2.8	4.2	6.7	6.3	5.1	4.4	Vouvray
V Font Brûlés	6.8	7.5	6.7	5.1	0.8	3.8	4.7		3.1	4.3	4.0	4.1	7.0	6.1	6.4	4.4	Vouvray
V Font Coteaux	7.1	7.3	6.7	4.1	0.9	2.7	3.6		4.3	5.3	1.2	6.0	7.3	6.6	5.7	6.0	Vouvray

Exemple en ACP

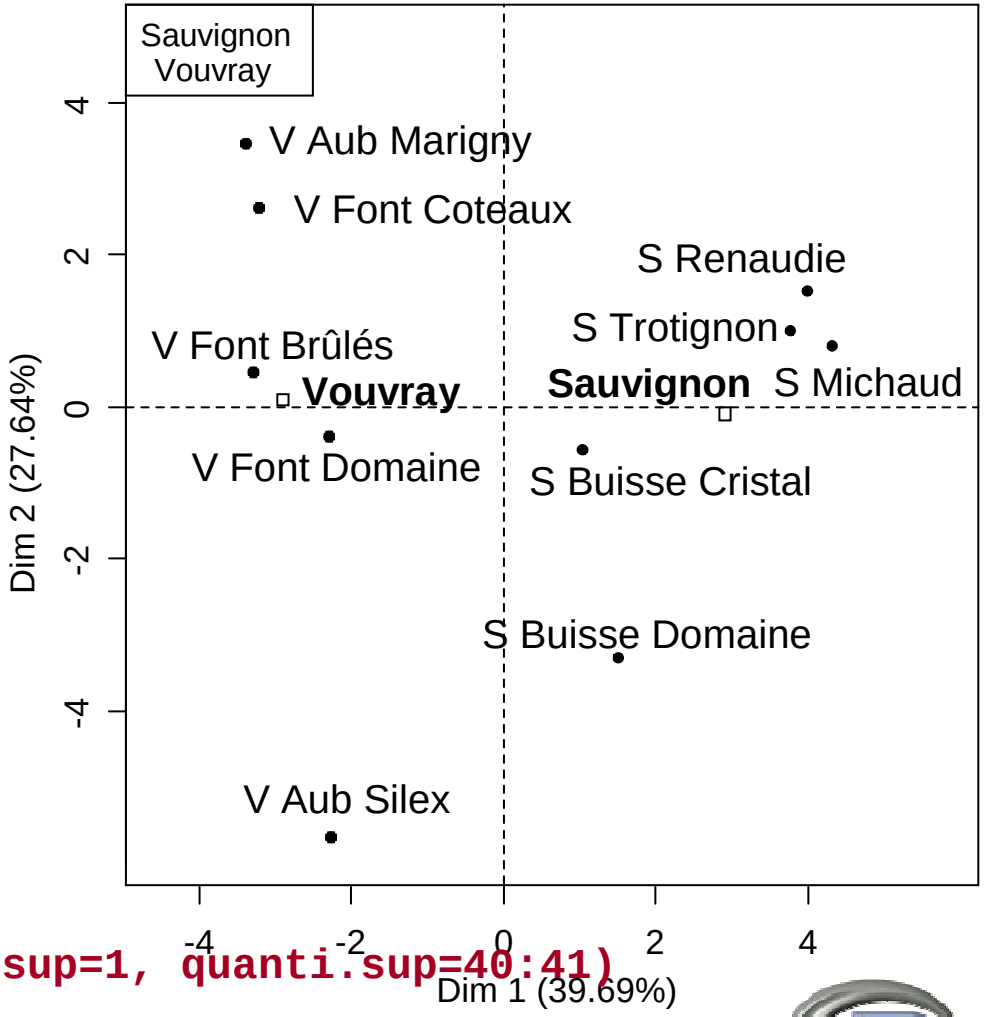





```
> resultat <- PCA(donnees, quali.sup=1, quanti.sup=40:41)
```

Exemple en ACP

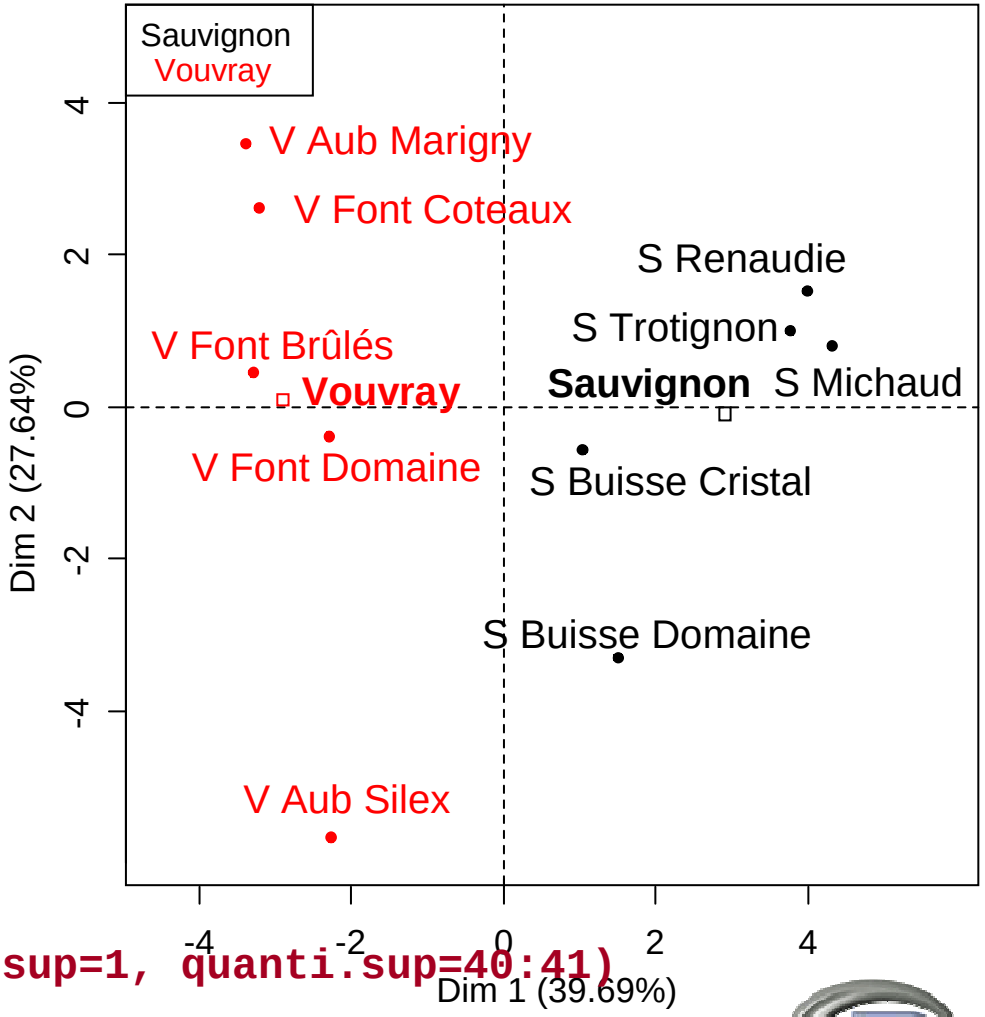
- Introduction d'information supplémentaire :
  - Individus supp.
  - Variables qualitatives supp.




```
> resultat <- PCA(donnees, quali.sup=1, quanti.sup=40:41)
```

Exemple en ACP

- Introduction d'information supplémentaire :
  - Individus supp.
  - Variables qualitatives supp.
- Graphes enrichis :
  - en coloriant les individus en fonction d'information supplémentaire

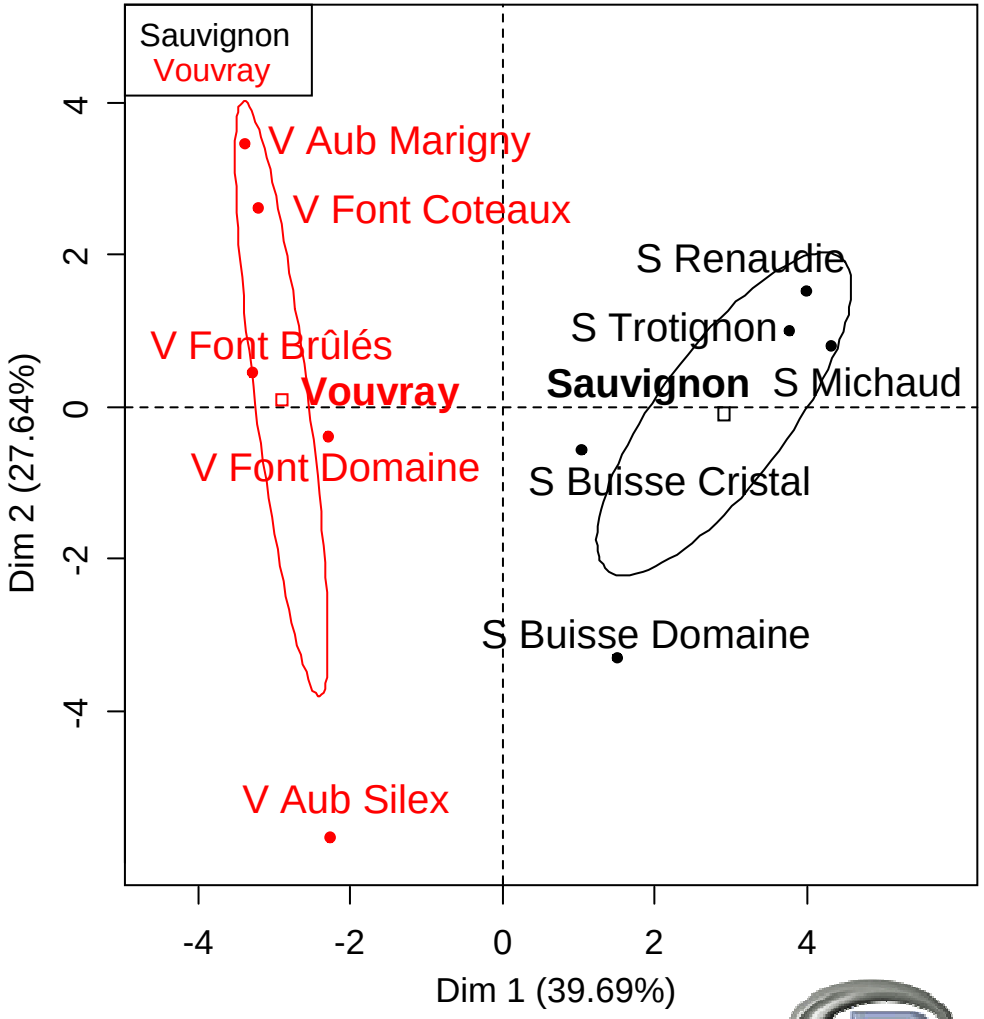



```
> resultat <- PCA(donnees, quali.sup=1, quanti.sup=40:41)
> plot(resultat, habillage=1)
```


FACTOMINER

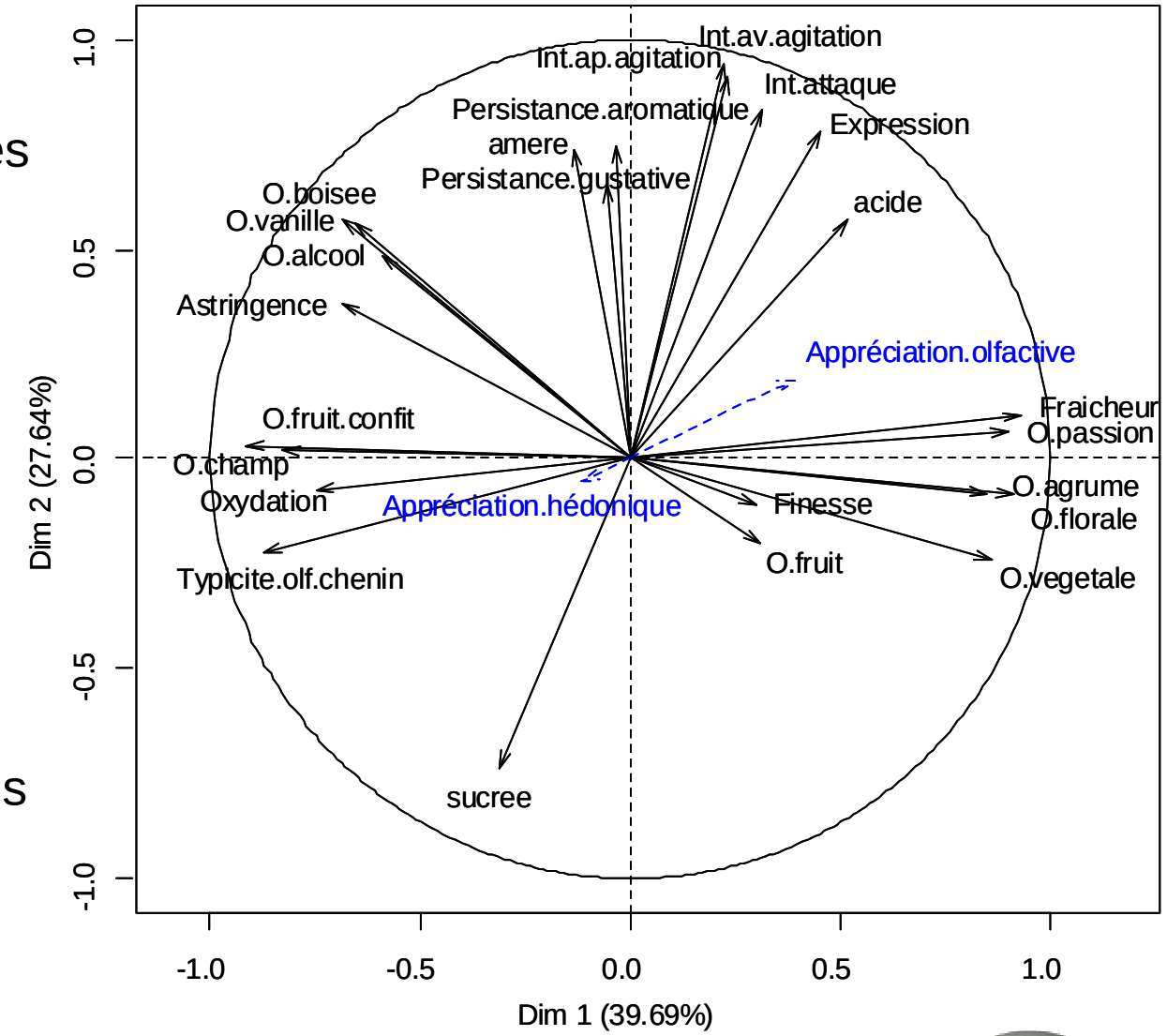
Exemple en ACP

- Introduction d'information supplémentaire :
  - Individus supp.
  - Variables qualitatives supp.
- Graphes enrichis :
  - en coloriant les individus en fonction d'information supplémentaire
  - en construisant des ellipses de confiance autour des modalités



Exemple en ACP

- Graphes enrichis :
  - variables représentées en fonction de leur qualité de représentation
- Introduction d'information supplémentaire :
  - Variables quantitatives supplémentaires




```
> resultat <- PCA(donnees, quali.sup=1, quanti.sup=40:41)
```

Exemple en ACP

- Des indicateurs sont disponibles :
  - Contribution des individus et des variables à la construction des axes
  - Qualité de représentation des individus et des variables

```
> $eig  eigenvalue   inertia   cumulative inertia
comp 1       10.36      41.44                41.44

> $ind         coord          contrib          cos2
             Dim.1 Dim.2    Dim.1 Dim.2    Dim.1 Dim.2
S Michaud    -4.38  0.93    18.49  1.30     0.69  0.03

> $var         coord          contrib          cos2
             Dim.1 Dim.2    Dim.1 Dim.2    Dim.1 Dim.2
O.agrume     -0.84 -0.06     6.82  0.05     0.71  0.00

> $quanti.sup
> $quali.sup coord              cos2          v-test
             Dim.1 Dim.2    Dim.1 Dim.2    Dim.1 Dim.2
Sauvignon    -3.03 -0.02     0.98     0    -2.82 -0.03
```

```
> resultat
```

Description des dimensions

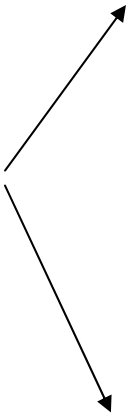
➢ **Par des variables quantitatives :**

• Calcul du coefficient de corrélation entre chaque variable et les coordonnées des individus sur l'axe *s*

• Tri des coefficients de corrélation

• Les coefficients significativement différents de 0 sont fournis

Meilleures variables pour décrire la première dimension



```
$Dim.1
$Dim.1$quanti
                     correlation p.value
A.alcool                    0.92 1.5e-04
O.fruit.confit              0.91 2.6e-04
Typicite.olf.chenin         0.86 1.3e-03
O.champ                     0.84 2.5e-03
Oxydation                   0.75 1.3e-02

O.vegetale                 -0.87 1.2e-03
O.passion                  -0.90 4.3e-04
O.florale                  -0.90 3.3e-04
Fraicheur                  -0.93 7.5e-05
```

```
> dimdesc(resultat)
```

Description des dimensions

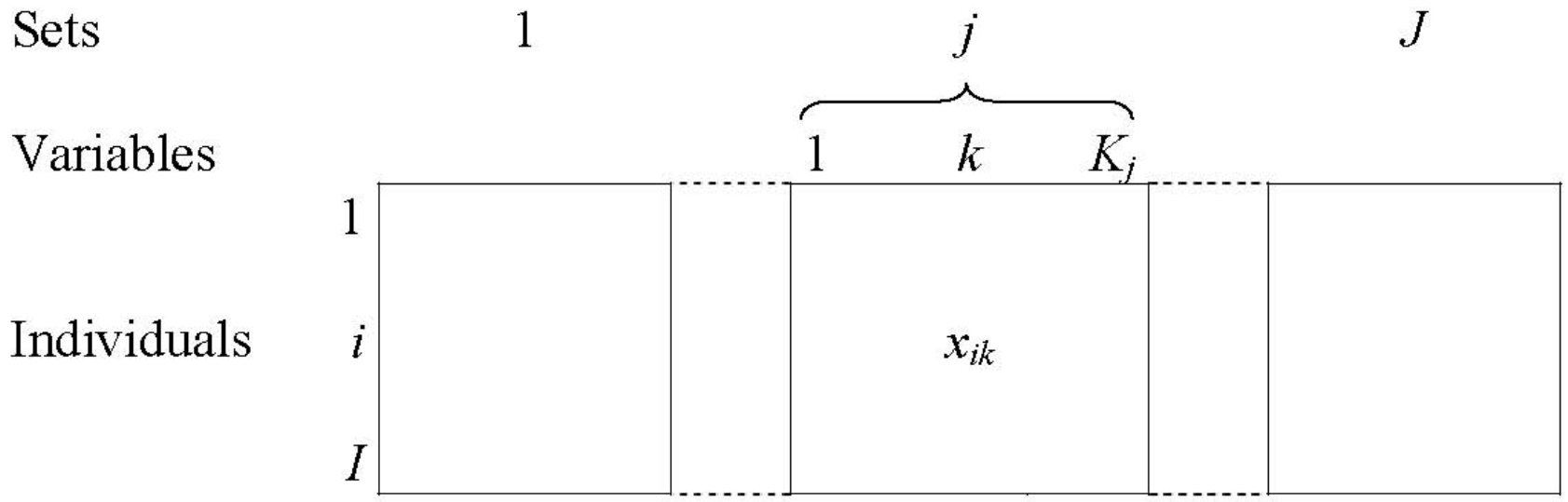
➢ Par les variables qualitatives :

• Réalisation d'une analyse de variance avec les coordonnées des individus sur l'axe en fonction de la variable qualitative

• Réalisation d'un test *T* par modalité

```
$Dim.1$quali
          R2 p.value
cepage 0.89 4.9e-05

$Dim.1$category
           Estimate p.value
Vouvray        3.00 4.9e-05
Sauvignon     -3.00 4.9e-05
```

Groupes de variables (AFM)



Les groupes de variables peuvent être quantitatifs et/ou qualitatifs

Objectifs : - équilibrer l'influence de chaque groupe de variables

- étudier le lien entre groupes de variables
- fournir des graphes spécifiques :

représentation partielle (individus vus par un groupe)

groupes de variables

Exemple d'AFM : les données

Les données : un panel d'experts + 1 panel étudiant + 1 panel conso + données d'appréciation + 1 variable qualitative (cépage)

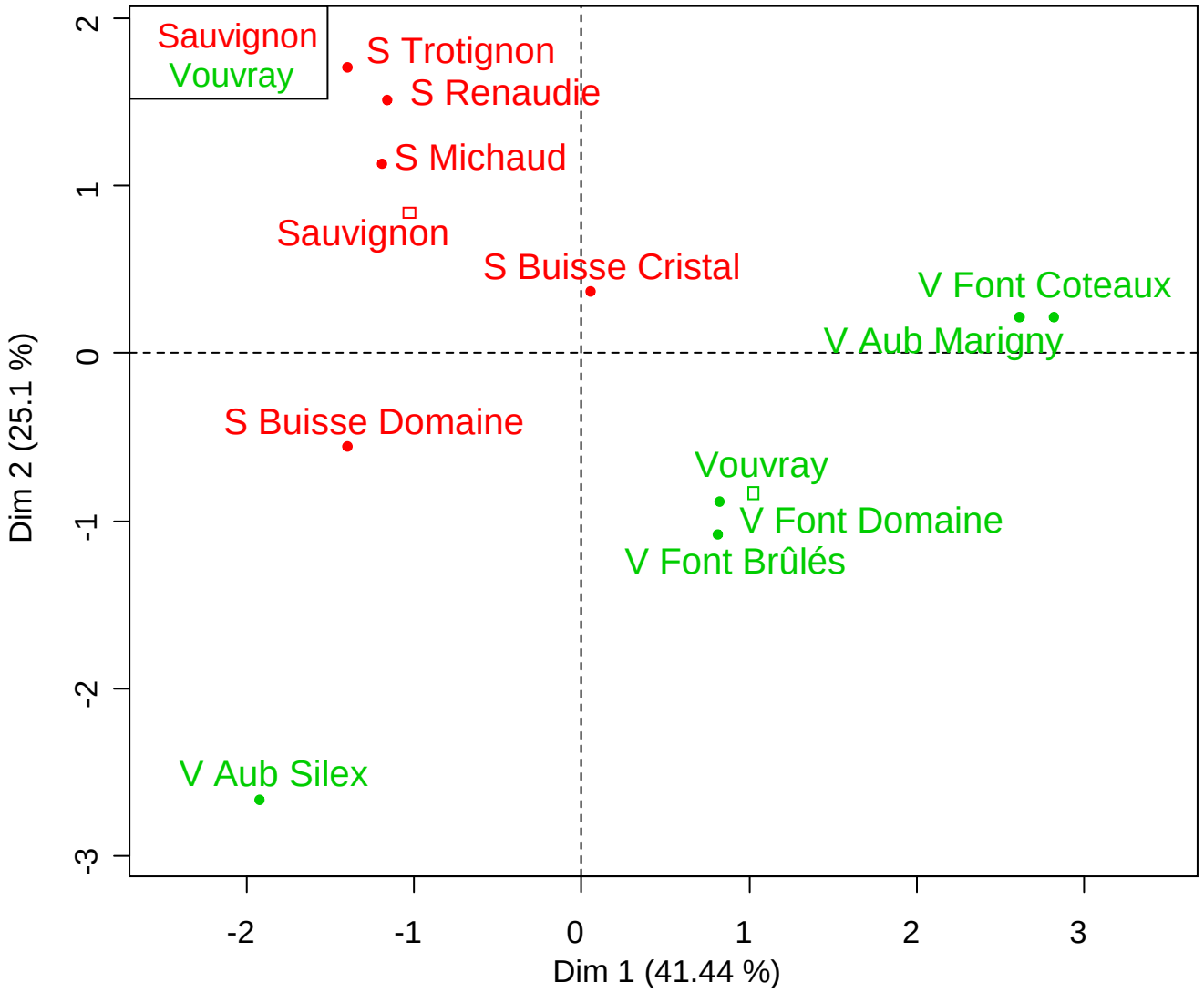
Groupes de variables quantitatives

Qlt

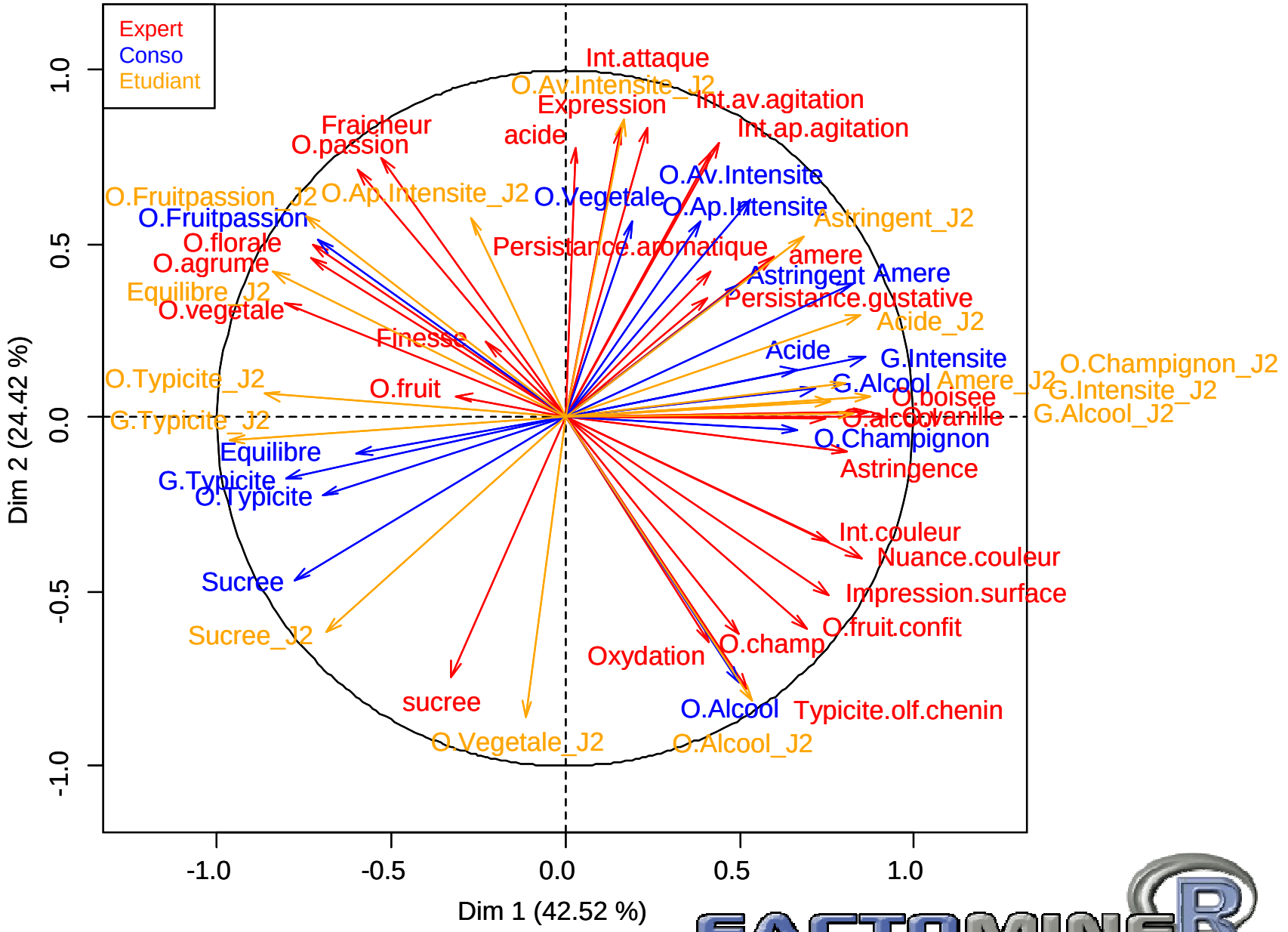
	Expert (27)	Conso (15)	Etudiant (15)	Appréciation (60)	Cépage (1)
Vin 1					
Vin 2					
…					
Vin 10					

```
> resAFM <- MFA(don.comp, group=c(27,15,15,60,1),
    type=c(rep("s",4),"n"), num.group.sup=c(4:5),
    name.group=c("Expert","Conso","Etudiant","Appréciation","Cépage"))
```

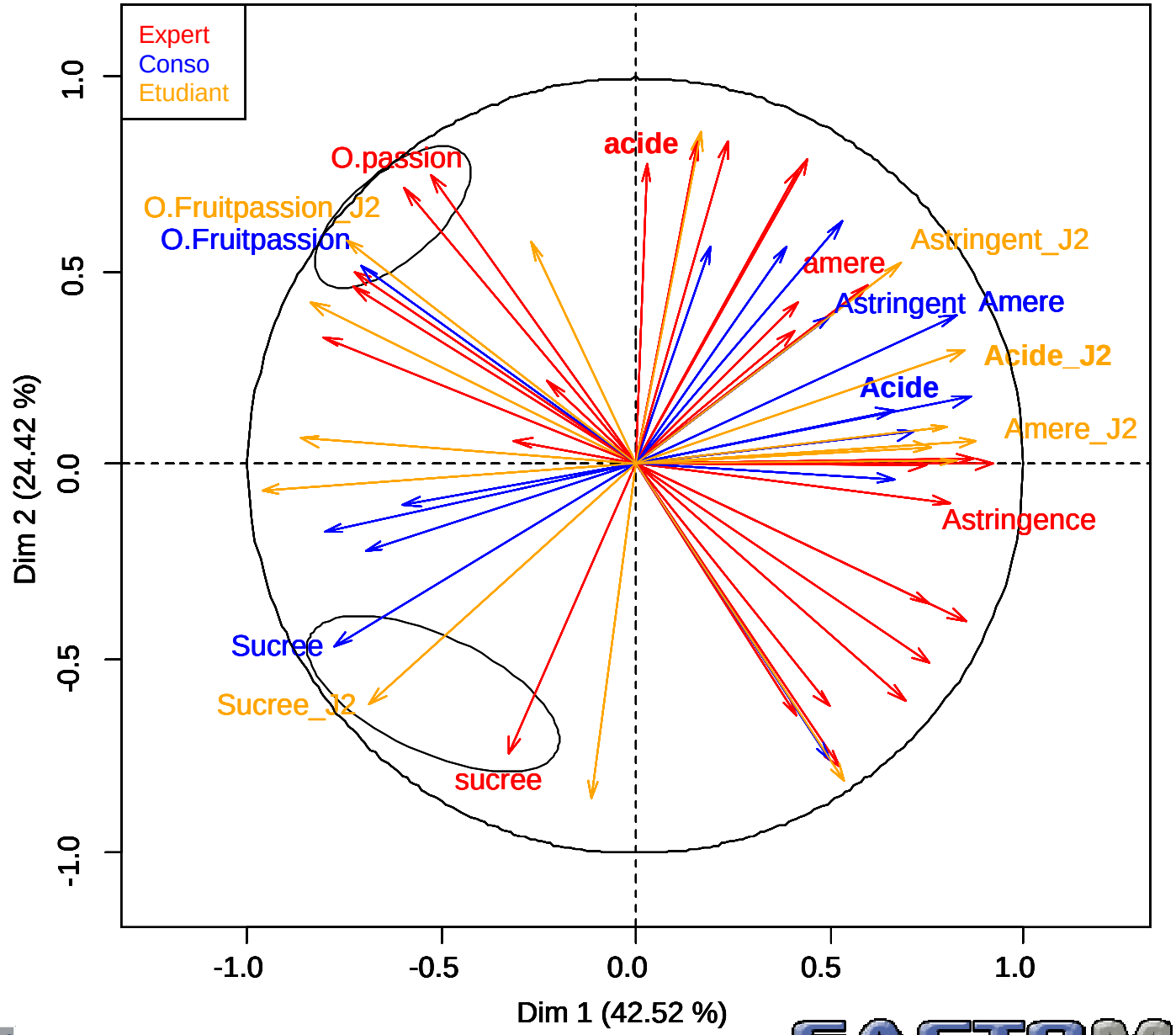
Exemple d'AFM: représentation des individus



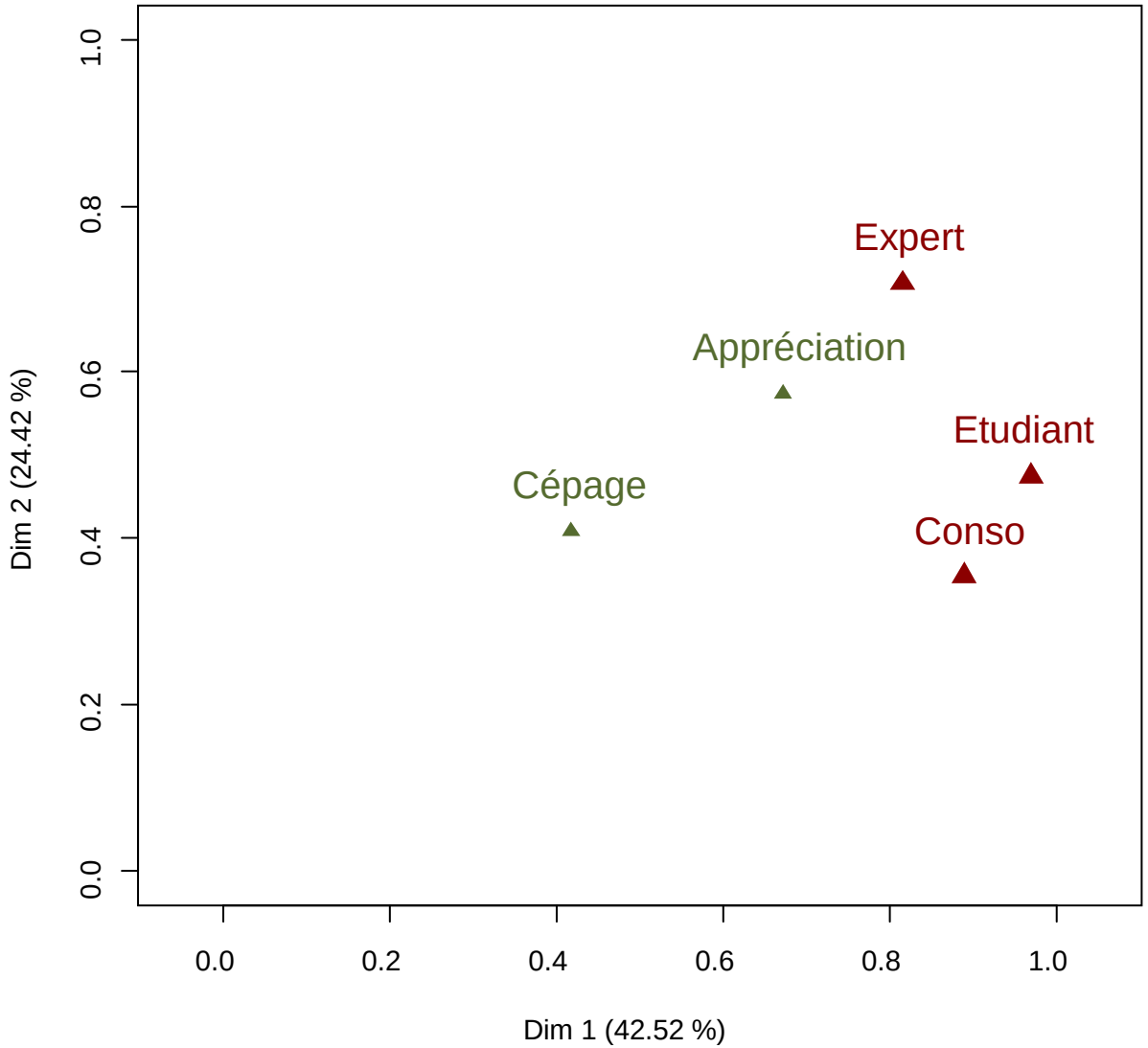
Exemple d’ AFM : représentation des variables



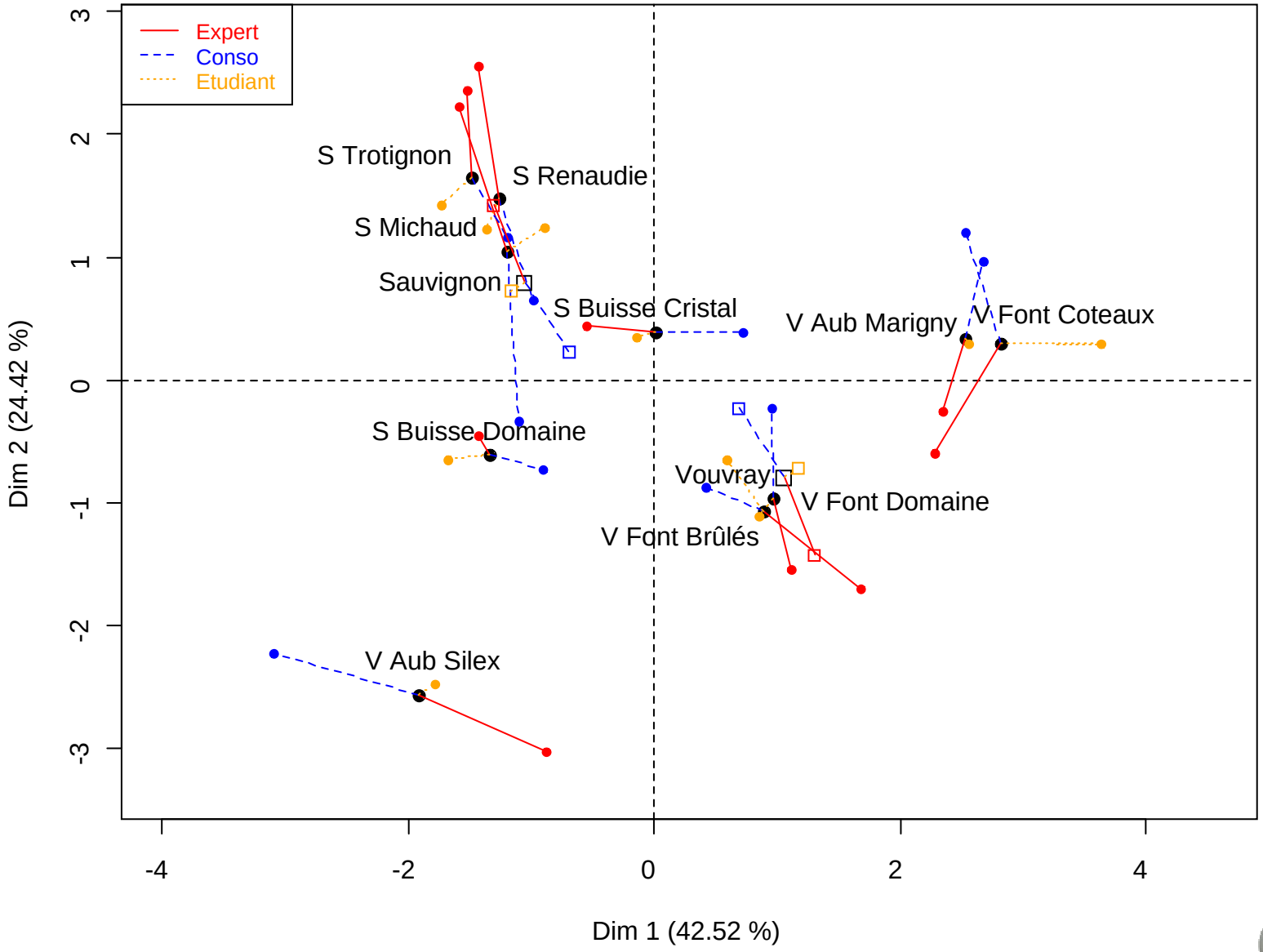
Exemple d’ AFM : représentation des variables



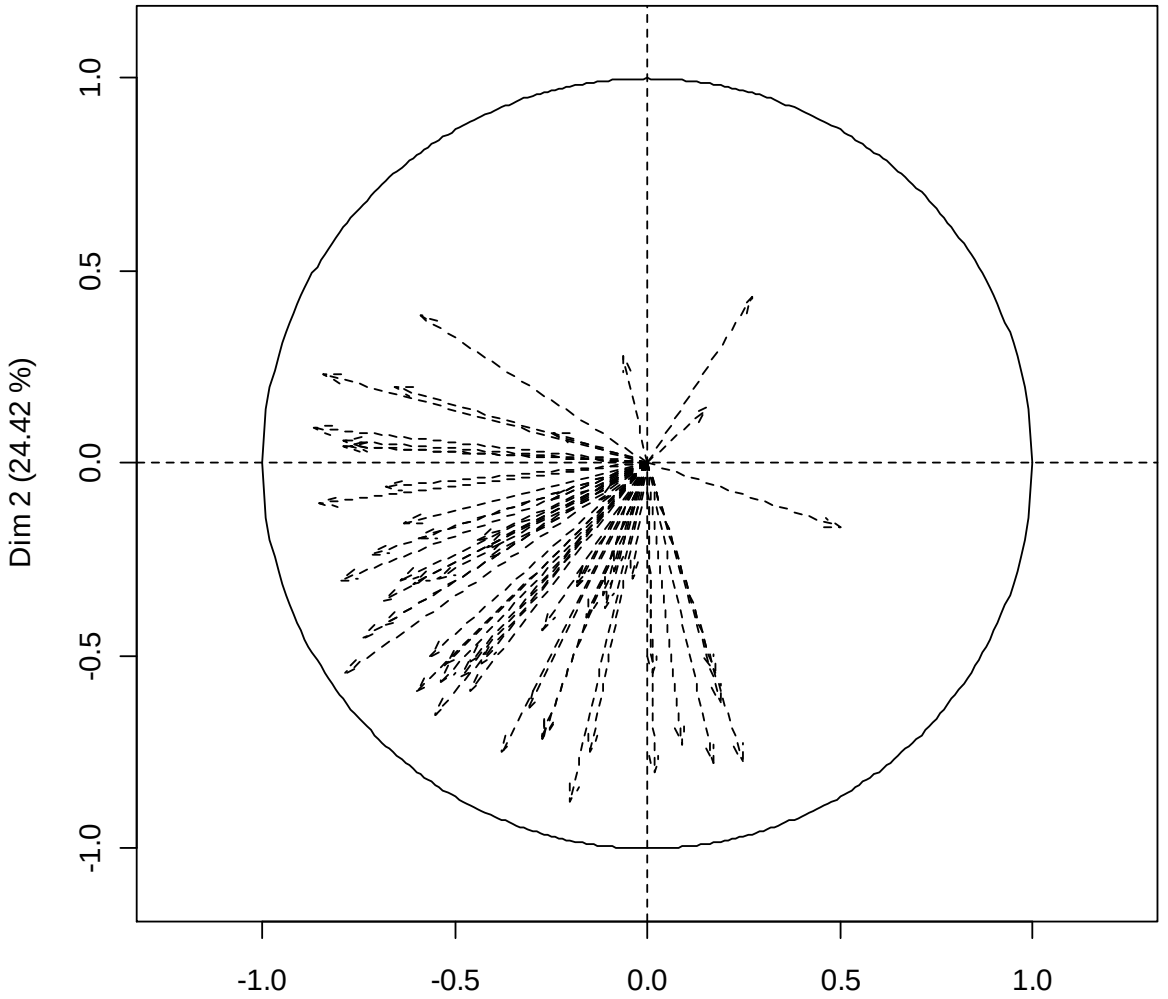
Exemple d'AFM : représentation des groupes



Exemple d'AFM : représentation des points partiels



Exemple d'AFM : représentation des préférences

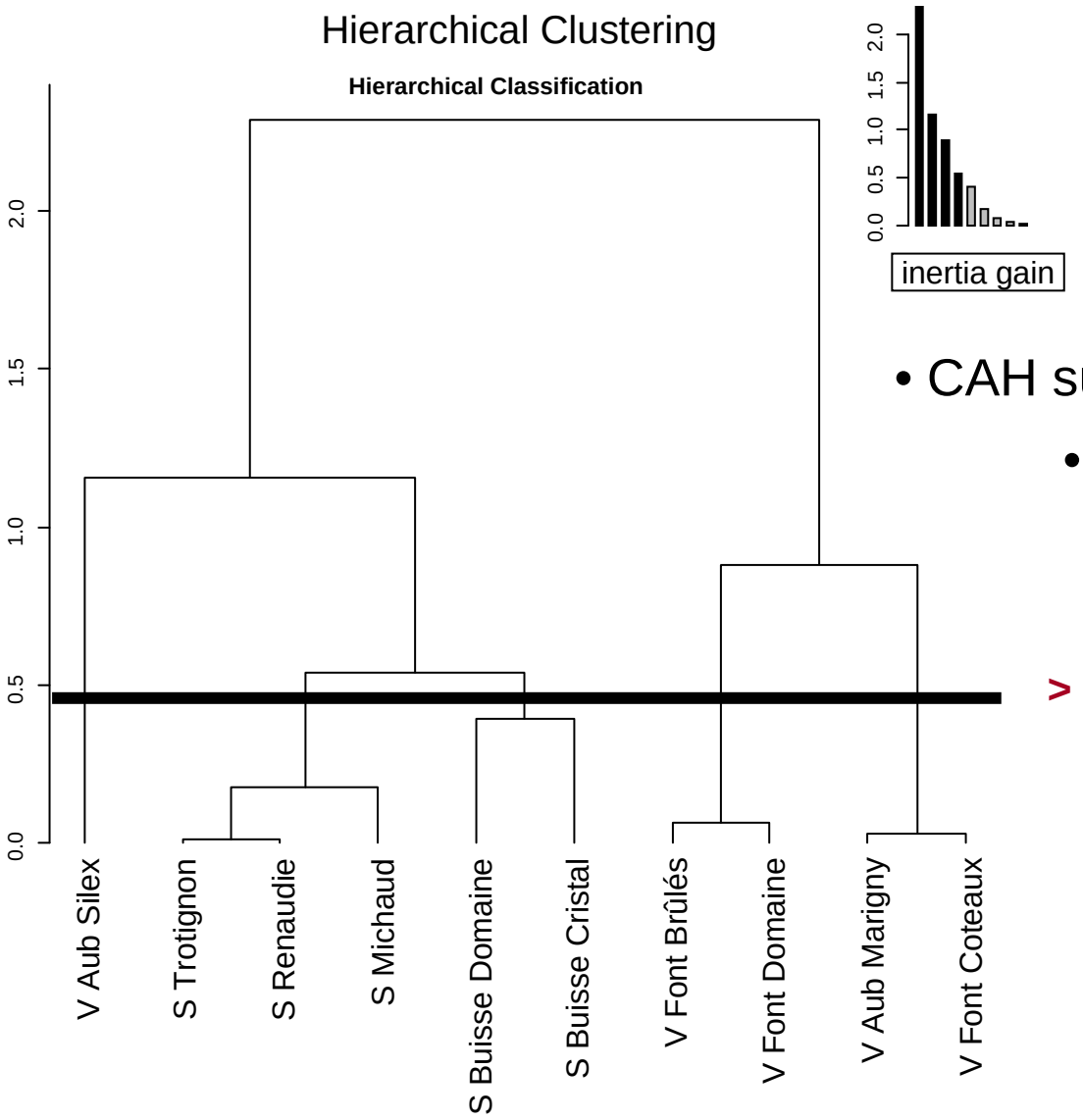



```
> resAFM <- MFA(don.comp, group=c(27,15,15,60,1),
    type=c(rep("s",4),"n"),num.group.sup=c(4:5),
    name.group=c("Expert","Jury 1","Jury 2","Appréciation","Cépage"))
```

Complémentarité analyse factorielle - classification

- Analyse factorielle comme prétraitement avant classification :
  - transformation des variables qualitative en quantitative (ACM)
  - équilibre de l'influence des variables (AFM ou AFMH)
  - suppression des dernières dimensions pour rendre la classification plus robuste
- Complémentarité graphique :
  - vision continue (axes factoriels) et discontinue (classes)
  - arbre hiérarchique donne une idée de l'information portée par les autres dimensions de l'analyse factorielle

Exemple de classification : arbre hiérarchique

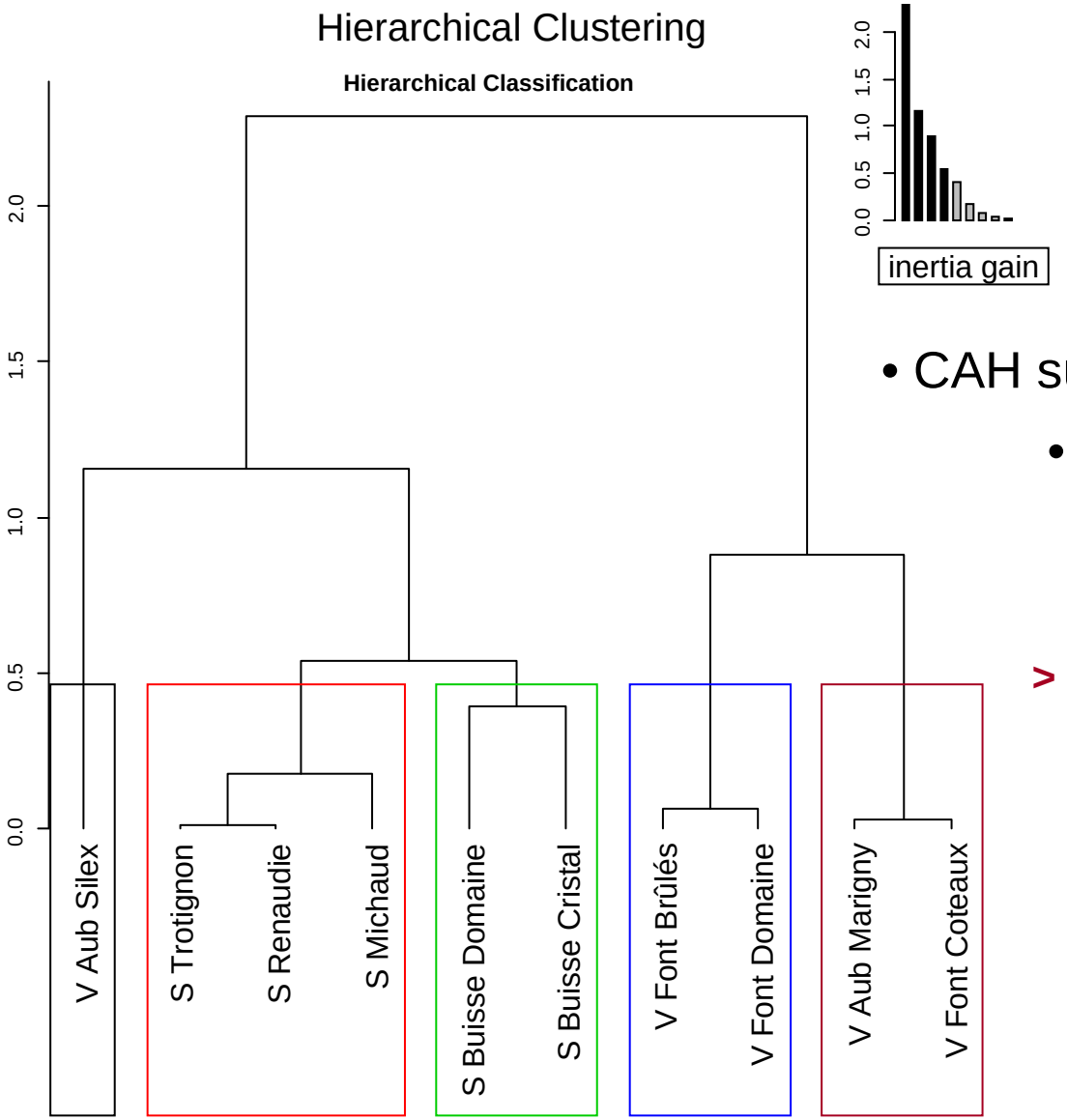


- CAH sur les coordonnées de l'AFM
  - choix du nombre de classes

```
> resClassif <- HCPC(resAFM)
```


FACTOMINER

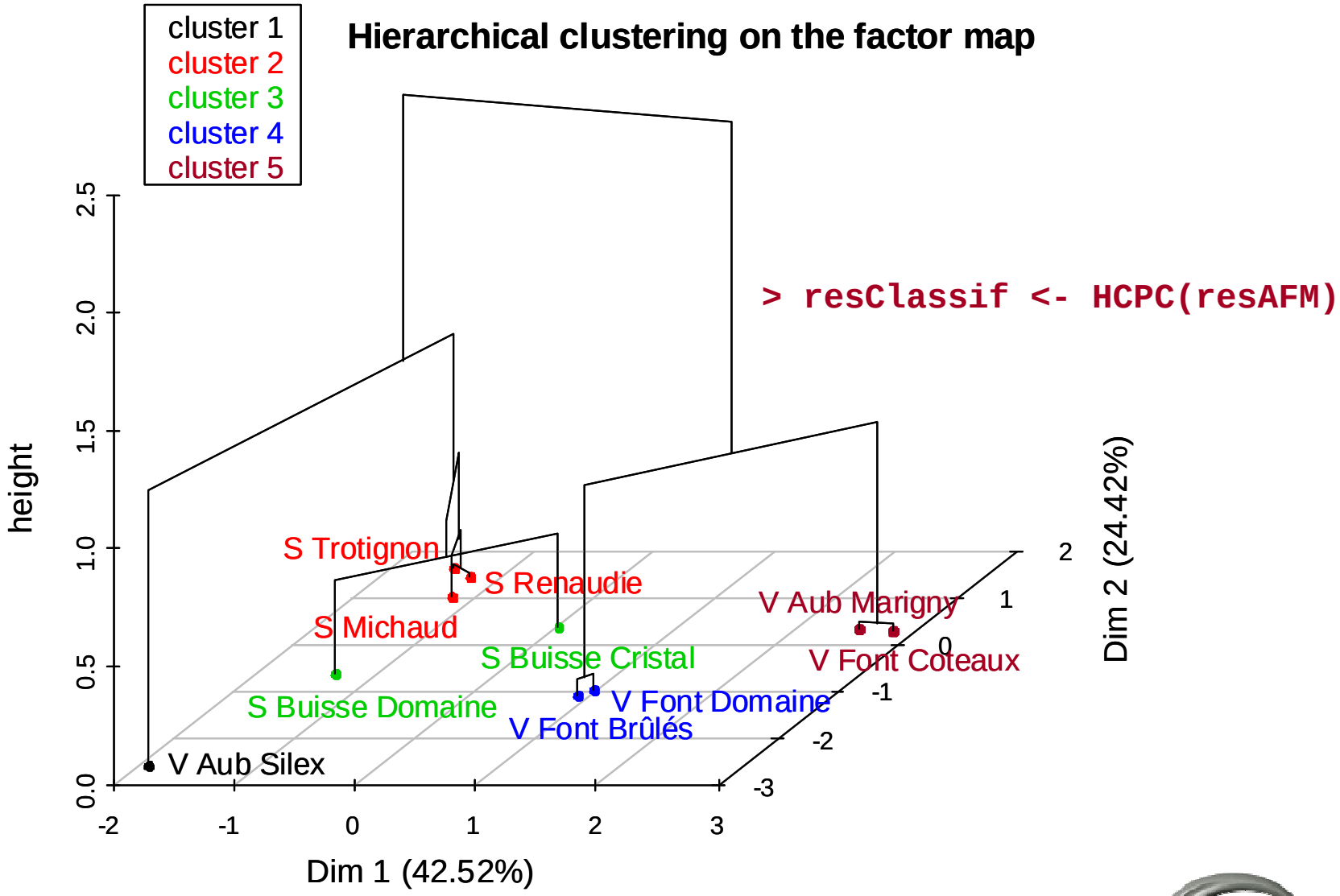
Exemple de classification : arbre hiérarchique



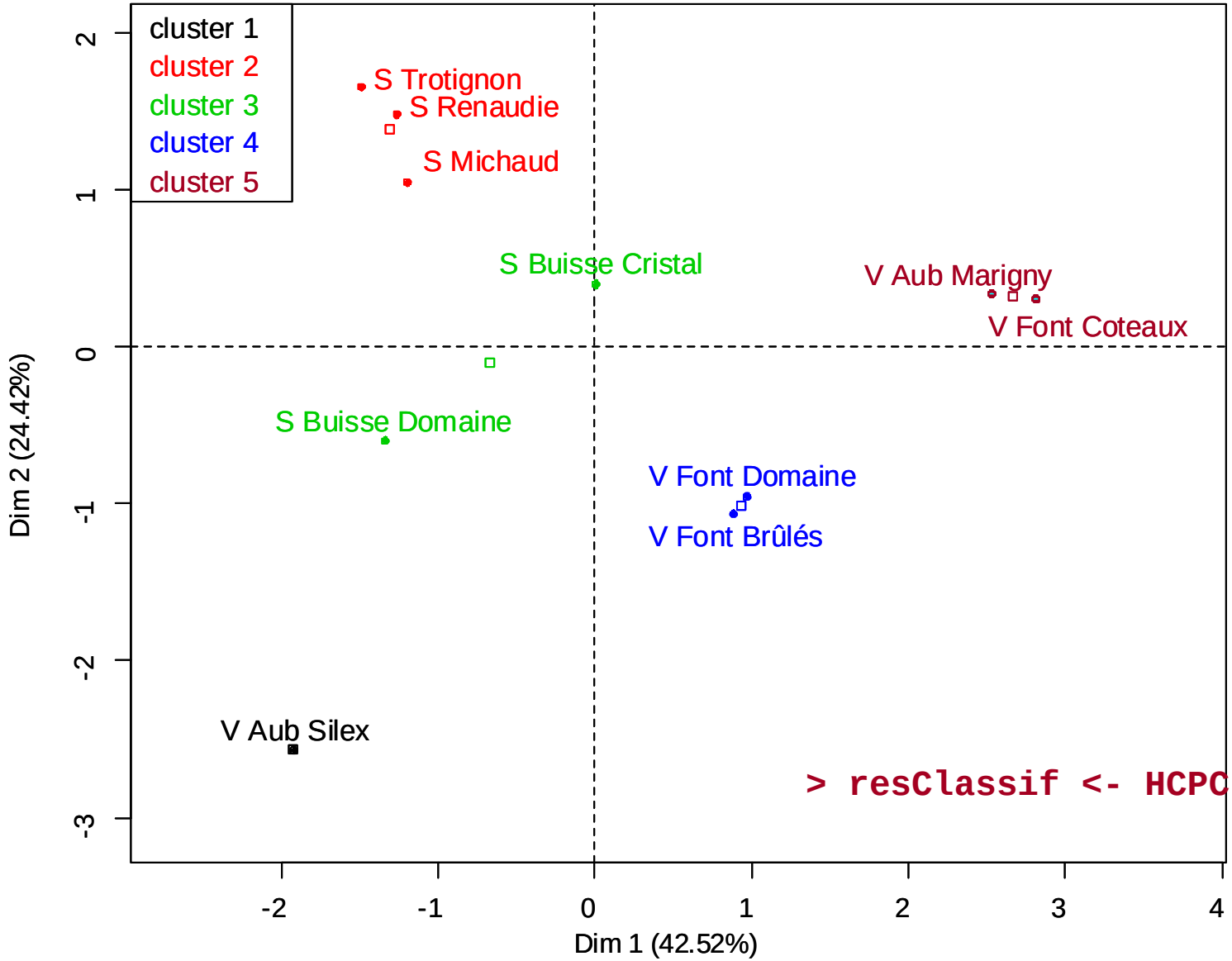
- CAH sur les coordonnées de l'AFM
  - choix du nombre de classes
    - K-Means avec 5 groupes

```
> resClassif <- HCPC(resAFM)
```

Exemple de classification : arbre en 3D



Exemple de classification : partition




```
> resClassif <- HCPC(resAFM)
```

Exemple : description des classes

```
> resClassif <- HCPC(resAFM)
> resClassif
```

```
> $test.chi
$test.chi2
          p.value df
cepage     0.0404  4
```

```
> $category$classe3
                  Cla/Mod Mod/Cla Global   p.value    v.test
cepage=Sauvignon      60     100     50 0.1666667  1.382994
cepage=Vouvray         0       0     50 0.1666667 -1.382994
```

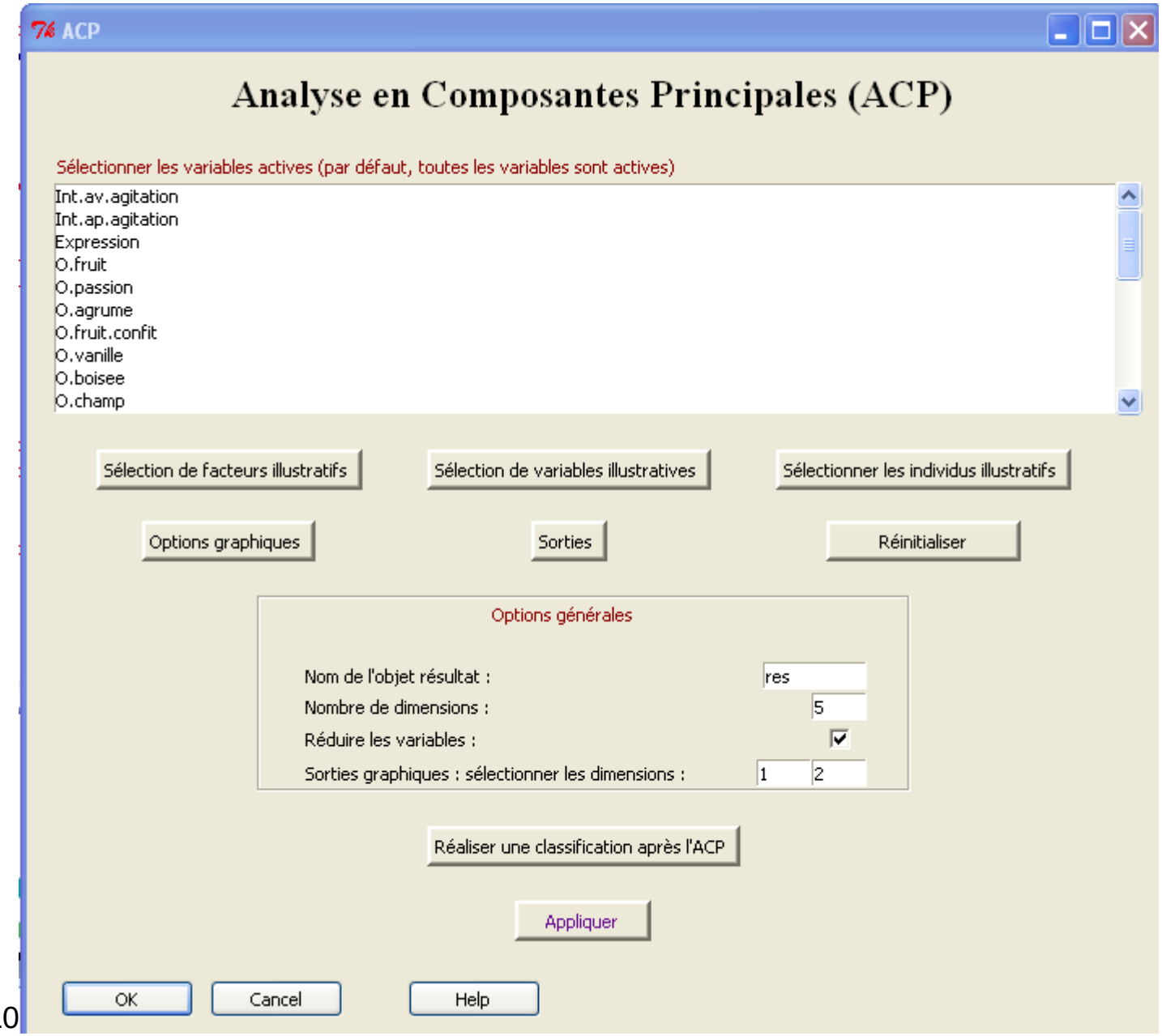
```
> $quanti$classe5
                v.test Mean in  Overall    sd in  Overall p.value
                       category    mean category       sd
O.boisee          2.95     7.50    2.14     0.25     2.73    0.00
O.vanille         2.91     5.98    1.98     0.31     2.06    0.00
O.Champignon_J2   2.87     5.33    3.28     0.47     1.07    0.00
G.Intensite_J2    2.52     6.25    5.50     0.17     0.45    0.01
Amere_J2          2.49     4.69    3.94     0.25     0.45    0.01

O.Typicite_J2    -2.46     4.06    5.08     0.05     0.62    0.01
G.Typicite_J2    -2.62     3.79    4.96     0.24     0.67    0.01
```

L'interface graphique

Menu déroulant de FactoMineR

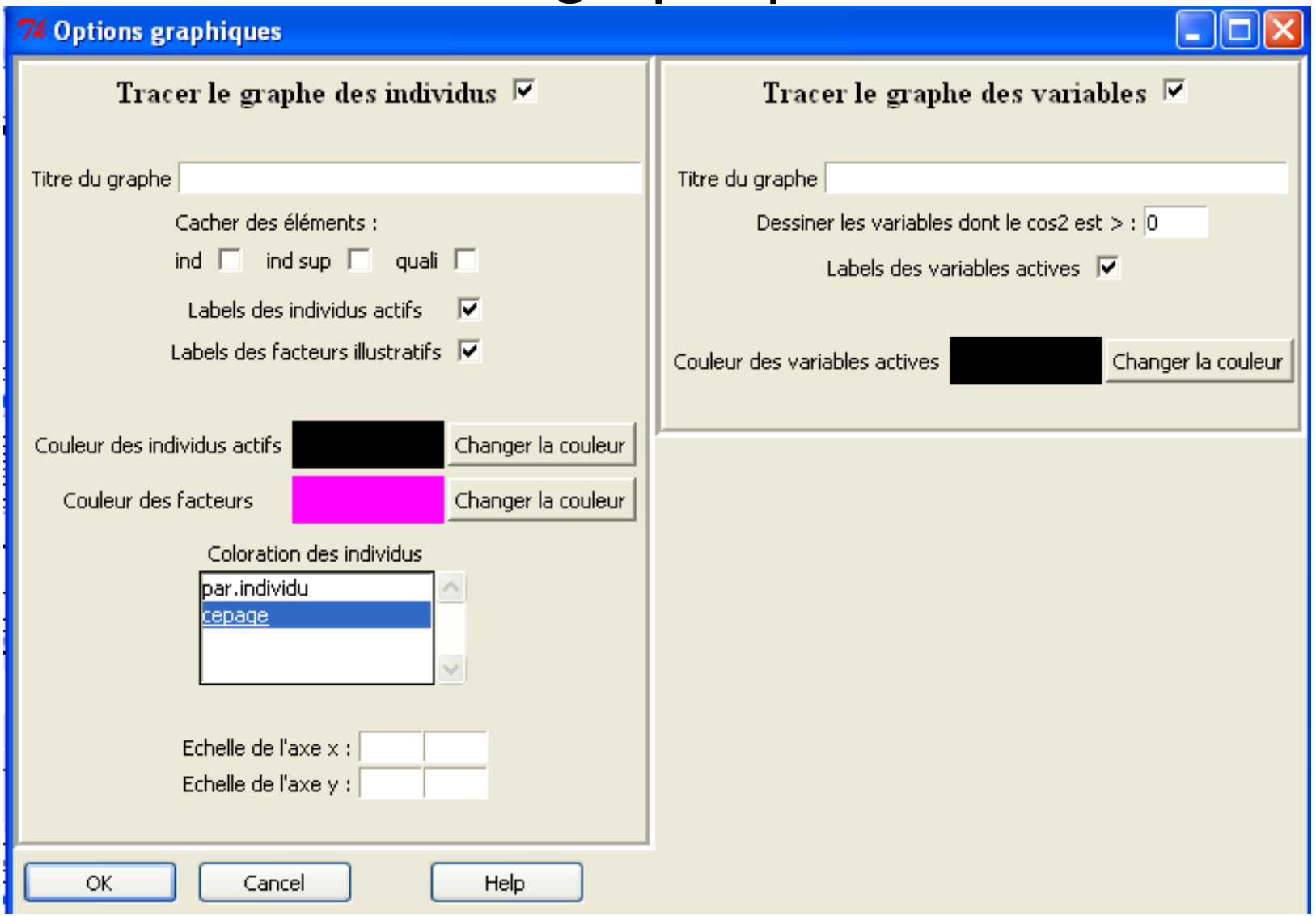
L'interface graphique

Fenêtre principale de l'ACP



L'interface graphique

Options graphiques



Les autres packages

ade4 : équipe de Lyon; package complet, orienté données écologiques

ca : Analyse des correspondances simple et multiple (Greenacre)

homals : homogeneity analysis (De Leeuw)

MASS : analyse des correspondances

cluster : classification

hopach : classification hiérarchique

dynGraph : graphiques dynamiques

missMDA : gestion de données manquantes pour ACP ou ACM

http://cran.r-project.org/web/views/Multivariate.html

http://cran.r-project.org/web/views/Cluster.html

Conclusion

Pour les chercheurs, les utilisateurs et les étudiants : contient des méthodes classiques mais aussi des méthodes avancées

FactoMineR est disponible sur le site officiel de R (CRAN)

L'interface graphique peut être chargée très facilement :

```
source("http://factominer.free.fr/install-facto-fr.r")
```

Un site web est dédié à cette librairie : http://factominer.free.fr

Jeux de données : http://www.agrocampus-ouest.fr/math/husson

Gestion des données manquantes *via* le package missMDA

Conclusion

- *Presses Universitaires de Rennes*

Statistiques avec R (2008, 2010)
Cornillon, Guyader, Husson, Jégou, Josse, Kloareg, Matzner-Lober, Rouvière

Analyse de données avec R (2009)
Husson, Lê, Pagès

- *Nouvelle collection SpringR*

FACTOMINER

L’équipe de FactoMineR

(lors d’une évaluation sensorielle ou d’une réflexion autour du package ?)