

L'analyse de données avec FactoMineR : les nouveautés

Gestion des données manquantes - module graphique - aides

François Husson & Julie Josse

Laboratoire de mathématiques appliquées, Agrocampus Rennes

Rencontres R, Lyon, juin 2013

FactoMineR en quelques mots

- propose des méthodes d'analyses factorielles et de classification
- de nombreux indicateurs (qualité de représentation, contribution, description automatique des axes, ...)
- possibilité d'ajouter des éléments supplémentaires
- interface graphique (en français et en anglais)
- gestion des données manquantes (avec le package missMDA)
- module graphique
- aides à l'utilisateur (site internet, vidéos, livres)

FactoMineR en quelques mots

Différentes méthodes pour différents formats de données :

Données	Méthodes	Fonction
Variables quantitatives	An. en composantes principales	PCA
Table de contingence	An. des correspondances	CA
Variables qualitatives	An. des correspondances multiples	MCA
Données mixtes	An. factorielle de données mixtes	FAMD
Groupes de variables	An. factorielle multiple	MFA
Hierarchie sur les variables	An. factorielle multiple hiérarchique	HMFA
Groupes d'individus	An. factorielle multiple duale	DMFA

Méthodes de classification et méthodes outils complémentaires :

Méthodes	Fonction
Classification ascendante hiérarchique	HCPC
Description d'une variable qualitative (ex. var. de classe)	catdes
Description d'une variable quantitative (ex. d'une dimension)	condes, dimdesc
Construction d'un tableau de données textuel	textual

Gestion des données manquantes avec le package missMDA

① Imputation des données manquantes par ACP itérative

② ACP sur le tableau complété

⇒ Fournit les axes et composantes principales (mieux que Nipals)

⇒ Fournit une imputation du jeu de données

⇒ Possible pour l'ACM, l'AFDM et l'AFM

Gestion des données manquantes : exemple en ACP

```

> library(missMDA)
> data(orange)
> nb <- estim_ncpPCA(orange,ncp.max=5)           ## Estime le nb de dimensions
> comp <- imputePCA(orange,ncp=nb,scale=TRUE)    ## Complète le tableau
> res.pca <- PCA(comp$completeObs)              ## Effectue l'ACP

```

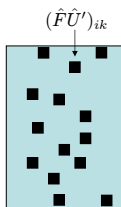
```

> orange                                     > comp$completeObs
  Sweet Acid Bitter Pulp Typicity          Sweet Acid Bitter Pulp Typicity
1  NA   NA  2.83   NA   5.21           5.54 4.13   2.83 5.89   5.21
2 5.46 4.13  3.54 4.62   4.46           5.46 4.13   3.54 4.62   4.46
3  NA  4.29  3.17 6.25   5.17           5.45 4.29   3.17 6.25   5.17
4 4.17 6.75   NA 1.42   3.42           4.17 6.75   4.73 1.42   3.42
...
5  NA   NA   NA 7.33   5.25           5.71 3.87   2.80 7.33   5.25
6 4.88 5.29  4.17 1.50   3.50           4.88 5.29   4.17 1.50   3.50

```

Imputation multiple en ACP

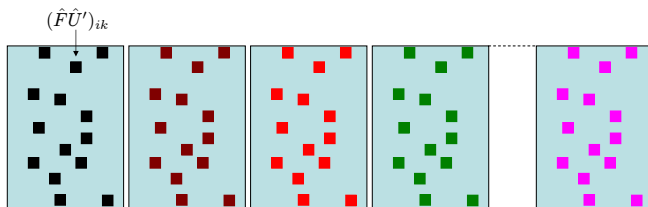
⇒ ACP itérative : une méthode d'imputation simple



⇒ Une valeur unique ne peut pas refléter la variabilité de prédiction

Imputation multiple en ACP

⇒ ACP itérative : une méthode d'imputation simple



⇒ Une valeur unique ne peut pas refléter la variabilité de prédiction

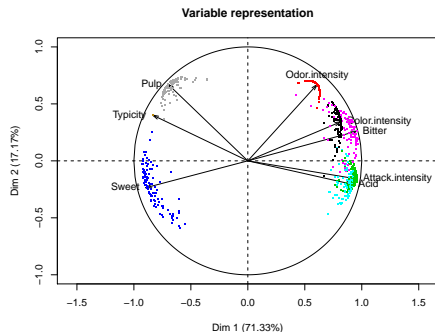
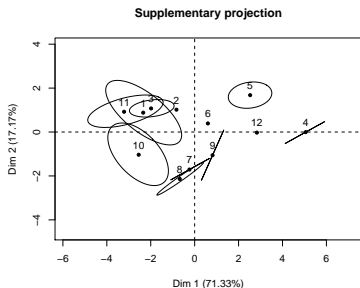
⇒ Imputation multiple : générer plusieurs valeurs plausibles pour chaque valeur manquantes

Visualisation de l'incertitude liée aux données manquantes

```

> library(missMDA)
> mi <- MIPCA(orange, scale = TRUE, ncp=2)
> mi$res.MI          ## sortie pour les tableaux imputés
> plot(mi)

```



Gestion des données manquantes : exemple en ACM

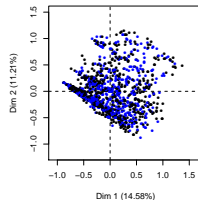
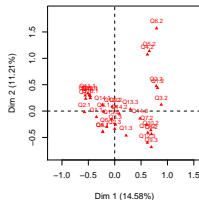
```

> library(missMDA)
> data(vnf)
> nb <- estim_ncpMCA(vnf,ncp.max=5)      ## Estime le nb de dimensions
> imp <- imputeMCA(vnf, ncp=nb)         ## Complète le tableau disjonctif
> res <- MCA(vnf,tab.disj=imp$tab.disj)  ## ACM utilisant tab disj complété

```

	V1	V2	V3	...	V14
ind 1	a	NA	g	...	u
ind 2	NA	f	g	...	u
ind 3	a	e	h	...	v
ind 4	a	e	h	...	v
ind 5	b	f	h	...	u
ind 6	c	f	h	...	u
ind 7	c	f	NA	...	v
...
ind 1232	c	f	h	...	v

	V1_a	V1_b	V1_c	V2_e	V2_f	V3_g	V3_h	...
ind 1	1	0	0	0.71	0.29	1	0	...
ind 2	0.12	0.29	0.59	0	1	1	0	...
ind 3	1	0	0	1	0	0	1	...
ind 4	1	0	0	1	0	0	1	...
ind 5	0	1	0	0	1	0	1	...
ind 6	0	0	1	0	1	0	1	...
ind 7	0	0	1	0	1	0.37	0.63	...
...
ind 1232	0	0	1	0	1	0	1	...



⇒ Même principe avec FAMD et MFA

Nouveautés dans le module graphique

```
> library(FactoMineR)
> data(decathlon)
> res.pca <- PCA(decathlon, quanti.sup = 11:12, quali.sup=13)
> summary(res.pca, nbelements=2, ncp=3)      ## fonction summary.PCA
```

```
Call:      PCA(decathlon, quanti.sup = 11:12, quali.sup = 13)
```

Eigenvalues

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5	Dim.6	Dim.7	Dim.8	Dim.9	Dim.10
Variance	3.272	1.737	1.405	1.057	0.685	0.599	0.451	0.397	0.215	0.182
% of var.	32.719	17.371	14.049	10.569	6.848	5.993	4.512	3.969	2.148	1.822
Cumulative % of var.	32.719	50.090	64.140	74.708	81.556	87.548	92.061	96.030	98.178	100.000

Individuals (the 2 first)

	Dist	Dim.1	ctr	cos2	Dim.2	ctr	cos2	Dim.3	ctr	cos2	
SEBRLE	2.369	0.792	0.467	0.112	0.772	0.836	0.106	0.827	1.187	0.122	
CLAY	3.507	1.235	1.137	0.124	0.575	0.464	0.027	2.141	7.960	0.373	

Variables (the 2 first)

	Dim.1	ctr	cos2	Dim.2	ctr	cos2	Dim.3	ctr	cos2	
100m	-0.775	18.344	0.600	0.187	2.016	0.035	-0.184	2.420	0.034	
Long.jump	0.742	16.822	0.550	-0.345	6.869	0.119	0.182	2.363	0.033	

Supplementary continuous variables

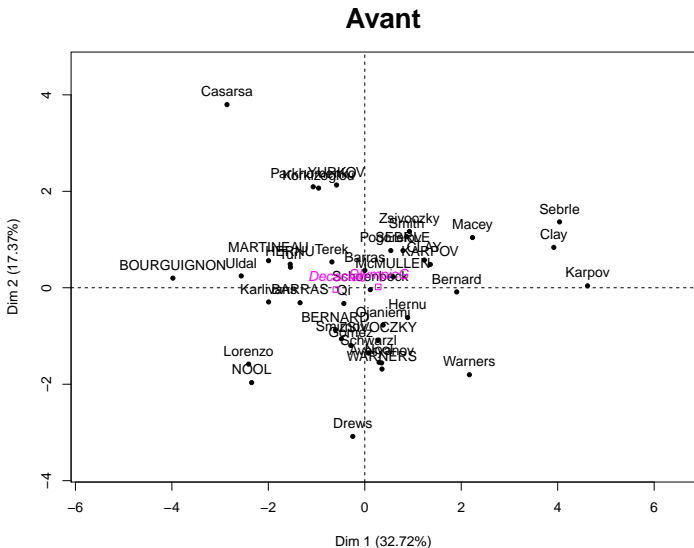
	Dim.1	cos2	Dim.2	cos2	Dim.3	cos2	
Rank	-0.671	0.450	0.051	0.003	-0.058	0.003	
Points	0.956	0.914	-0.017	0.000	-0.066	0.004	

Supplementary categories

	Dist	Dim.1	cos2	v.test	Dim.2	cos2	v.test	Dim.3	cos2	v.test	
Decastar	0.946	-0.600	0.403	-1.430	-0.038	0.002	-0.123	0.289	0.093	1.050	
OlympicG	0.439	0.279	0.403	1.430	0.017	0.002	0.123	-0.134	0.093	-1.050	

Graphe des individus : AVANT

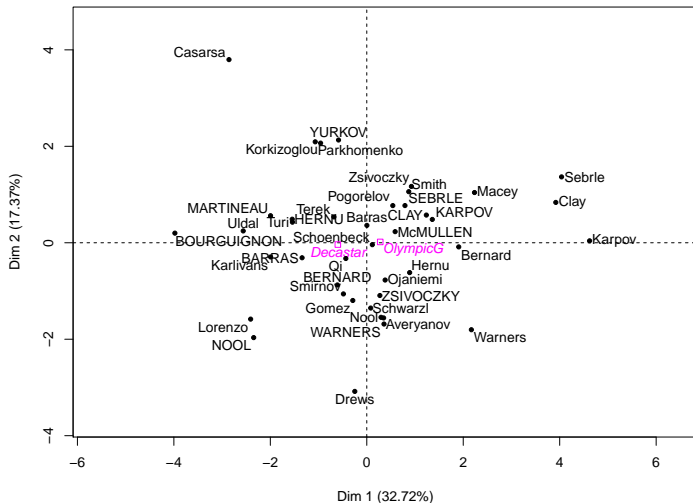
```
> plot(res.pca, autoLab="no")
```



Graphe des individus : APRES

```
> plot(res.pca, autoLab="auto") ## si <50 éléments = yes, sinon no
```

Après

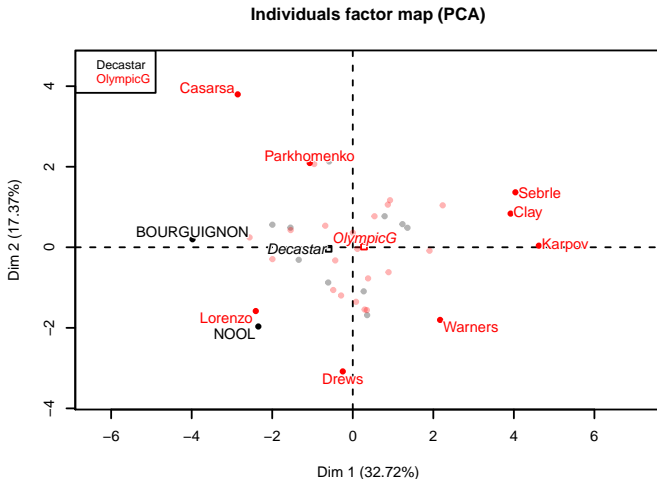


Nouveau module graphique

- Concerne toutes les fonctions graphiques `plot.PCA`, `plot.MCA`, `plot.CA`, `plot.FAMD`, `plot.MFA`
- La fonction `autoLab` positionne les libellés de façon optimale
 - placement des libellés sur l'extérieur du graphique
 - calcul du taux de recouvrement des libellés
 - algorithme itératif minimisant le taux de recouvrement
- Quelques astuces complémentaires :
 - sélectionner des éléments
 - réduire la taille des caractères (`cex = 0.7`)
 - agrandir la fenêtre graphique
 - mettre une ombre sous les libellés
 - relancer la fonction \implies graphe légèrement différent

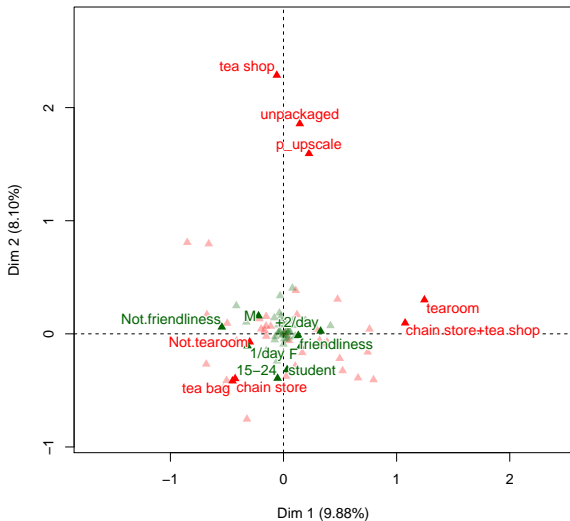
Sélection des individus par leur qualité de représentation

```
> plot(res.pca, habillage="Competition", select="cos2 0.6")
```



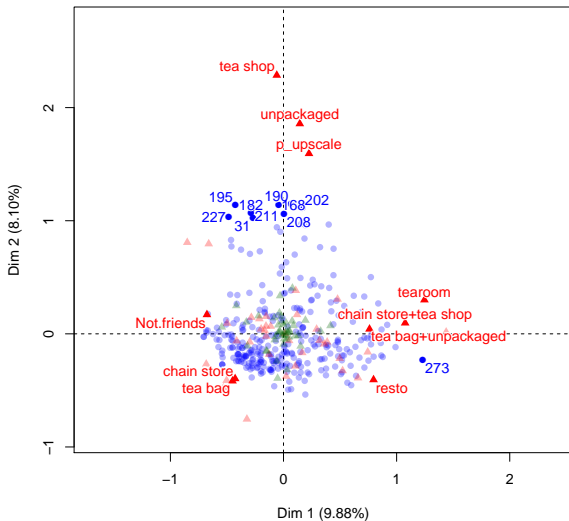
Sélection des modalités par leur qualité de représentation

```
> plot(res.mca, invisible="ind", selectMod="cos2 8")
```



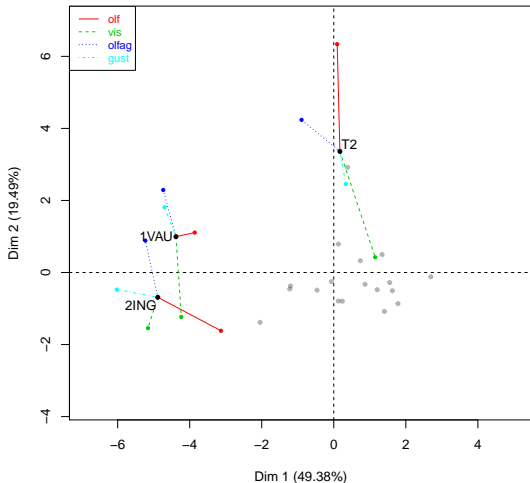
Sélection des éléments par leur contribution

```
> plot(res.mca,select="contrib 10",selectMod="contrib 10",shadow=TRUE)
```



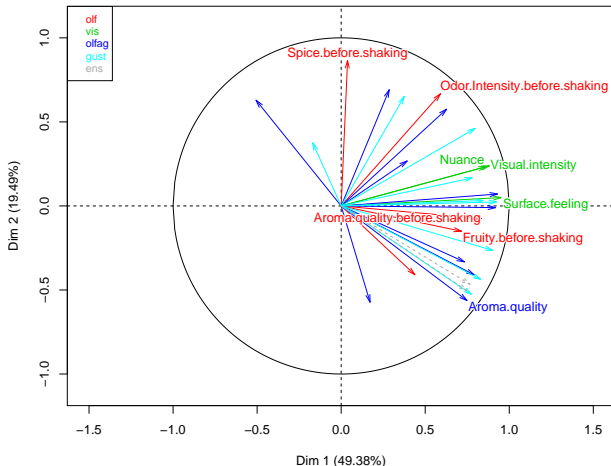
Sélection des éléments par une liste

```
> liste <- c("1VAU","2ING","T2")  
> plot(res.mfa, select=liste, partial="all", habillage="group")
```



Sélection des variables par leur contribution

```
> plot(res,choix="var", select="contrib 8", habillage="group",  
unselect=0, shadow=TRUE)
```



Nouveau module graphique

- Principaux arguments :

```
autoLab = "auto" ## position optimale des libellés si nb éléments <50
shadowtext = TRUE ## ombre sous le libellé
invisible=c("ind","ind.sup") ## rend invisibles certains éléments
```

- Sélection des éléments par :

```
select = 1:4 ## 4 premiers indiv
select = c("i1","i3") ## liste d'indiv
select = "cos2 5" ## 5 indiv les mieux représentés
select = "cos2 0.6" ## indiv avec cos2 > 0.6
select = "coord 3" ## 3 indiv avec coordonnées les + grandes
selectMod = "contrib 8" ## 8 modalités avec contributions les + grandes
unselect = 0.5 ## transparence des éléments non sélectionnés
unselect = 0 ## éléments non sélectionnés de même couleur
unselect = 1 ## éléments non sélectionnés non dessinés
unselect = "grey30" ## éléments non sélectionnés en gris
```

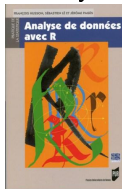
Aides à l'utilisateur

- Un menu déroulant en français
 - > `source("http://factominer.free.fr/install-facto-fr.r")`
- Un site internet : <http://factominer.free.fr>
- Un Google group pour poser des questions
<https://groups.google.com/group/factominer-users/>
- Des jeux de données avec les lignes de code (cliquer ici)
- Des livres :

Statistique avec R (3^e ed.)



Analyse de données avec R



Aides à l'utilisateur : des vidéos sur Youtube

- <https://www.youtube.com/HussonFrancois>
- une playlist de 14 vidéos en français
- une playlist de 10 vidéos en anglais

Analyse de données avec FactoMineR



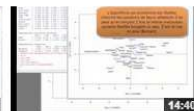
Démarche en analyse multivariée

de François Husson
188 vues



Faire une ACP avec le menu déroulant de FactoMineR

de François Husson
270 vues



Analyse en Composantes Principales (ACP) avec R e...

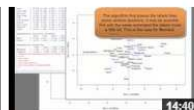
de François Husson
1 361 vues

Exploratory multivariate analysis with R and FactoMineR



Methodology in multivariate exploratory data analysis

de François Husson
197 vues



Principal component analysis (PCA) with R

de François Husson
538 vues



Correspondence analysis with R and FactoMineR

de François Husson
103 vues

Merci de votre attention

On aurait pu parler ...

... du menu déroulant en français ...

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

The screenshot shows the R Commander interface. The 'FactoMineR' menu is open, displaying a list of analysis methods in French. The 'Analyse en Composantes Principales (ACP)' option is highlighted. In the foreground, the 'Analyse en Composantes Principales (ACP)' dialog box is open. It features a list of variables to select, with '100m', 'Long jump', 'Shot.put', 'High.jump', '400m', '110m.hurdle', 'Discus', 'Pole.vault', 'Javeline', and '1500m' listed. Below the list are three tabs: 'Sélection de facteurs illustratifs', 'Sélection de variables illustratives', and 'Sélectionner les individus illustratifs'. There are also buttons for 'Options graphiques', 'Sorties', and 'Réinitialiser'. A section titled 'Options générales' contains fields for 'Nom de l'objet résultat' (set to 'res'), 'Nombre de dimensions' (set to 5), 'Réduire les variables' (checked), and 'Sorties graphiques : choix des dimensions' (set to 1 and 2). At the bottom, there is a 'Réaliser une classification après l'ACP' checkbox and an 'Appliquer' button. The bottom of the dialog has 'OK', 'Annuler', and 'Aide' buttons.

R Commander

Fichier Edition Données Statistiques Graphes Modèles Distributions Outils Aide FactoMineR

Importer des données depuis un fichier texte

Analyse en Composantes Principales (ACP)

Analyse Factorielle des Correspondances (AFC)

Analyse des Correspondances Multiples (ACM)

Analyse Factorielle Multiple (AFM)

Analyse Factorielle Multiple Hiérarchique (AFMH)

Analyse Factorielle Multiple Duale (AFMD)

Analyse Factorielle de Données Mixtes (AFDM)

Analyse Procustéenne Généralisée (APG)

Nuage de points avec variables additionnelles

Description des modalités

Classification Hiérarchique sur Composantes Principales (HCPC)

ACP

Analyse en Composantes Principales (ACP)

Sélectionner les variables actives (par défaut, toutes les variables sont actives)

100m
Long jump
Shot.put
High.jump
400m
110m.hurdle
Discus
Pole.vault
Javeline
1500m

Sélection de facteurs illustratifs Sélection de variables illustratives Sélectionner les individus illustratifs

Options graphiques Sorties Réinitialiser

Options générales

Nom de l'objet résultat : res

Nombre de dimensions : 5

Réduire les variables :

Sorties graphiques : choix des dimensions 1 2

Réaliser une classification après l'ACP

Appliquer

OK Annuler Aide

on.PCA a 41 lignes et 13 colonnes.
on a 41 lignes et 13 colonnes.

... du site internet en français ...

Accueil Méthodes Classiques Méthodes Avancées Interface Les + Excel F.A.Q. Docs Contact

FACTOMINE^R

> Nouveautés



Analyse de données avec FactoMineR
8 vidéos | Il y a 2 semaines

Vidéos sur l'utilisation de FactoMineR pour faire une ACP, AFC, ACM, analyse factorielle multiple, classification ascendante hiérarchique

La version 1.24 de FactoMineR propose un nouveau module graphique qui "optimise" la position des libellés pour éviter qu'ils se chevauchent, qui permet de sélectionner les éléments que l'on souhaite visualiser, etc.

Quatre reviewings sur le livre Analyse de données avec R sont disponibles à l'adresse suivante. Pour voir le reviewing complet de Gary Evans pour Journal of Statistical Software.

English Version

Version française

> Top Menu

Accueil

Méthodes Classiques

Méthodes Avancées

Interface

Les Plus de Facto

FactoMineR et Excel

F.A.Q.

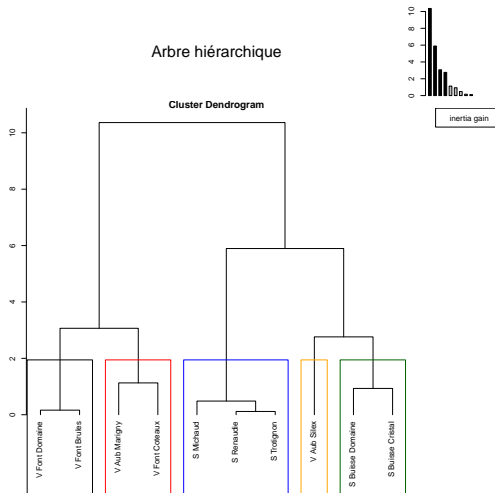
Documents

Contacts

> Liens utiles

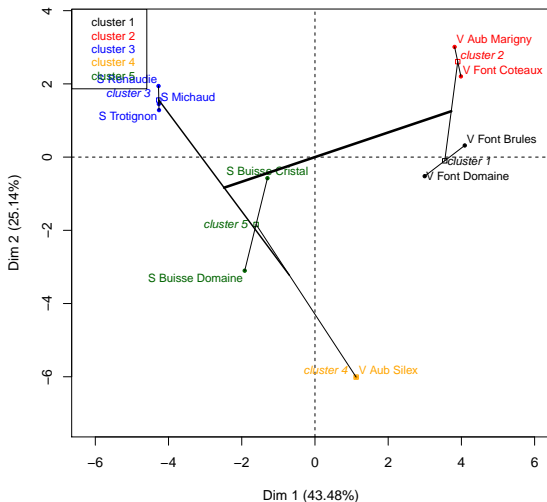
Département de
Mathématiques

... de classification hiérarchique ...



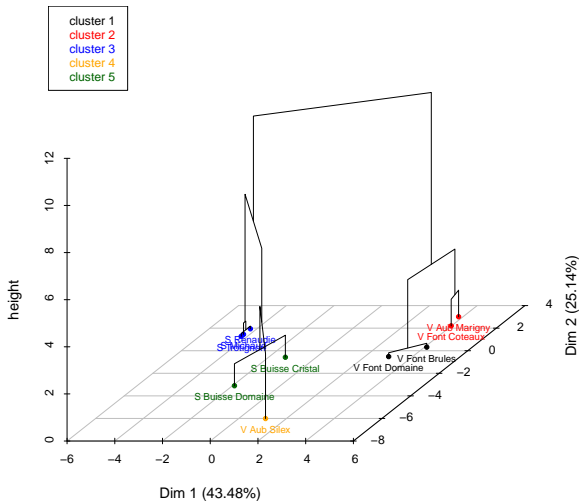
```
> res.hcpc <- HCPC(res.pca)
> plot(res.hcpc, choice="tree")
```

... de représentation factorielle et d'arbre vu de dessus ...



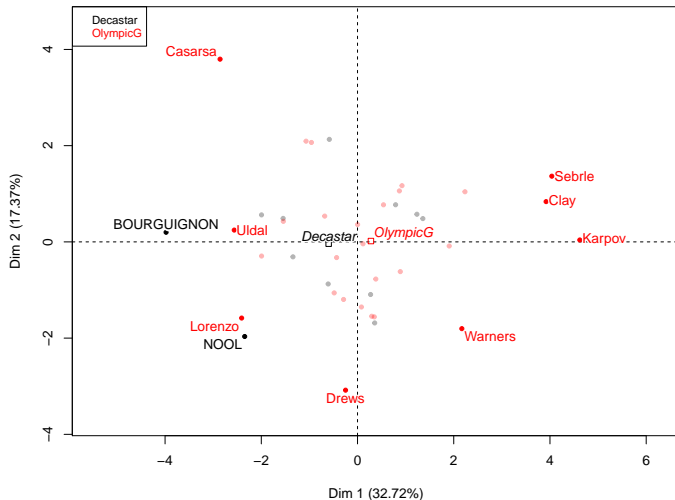
```
> plot(res.hcpc, choice="map")
```

... de représentation factorielle et d'arbre en 3D ...



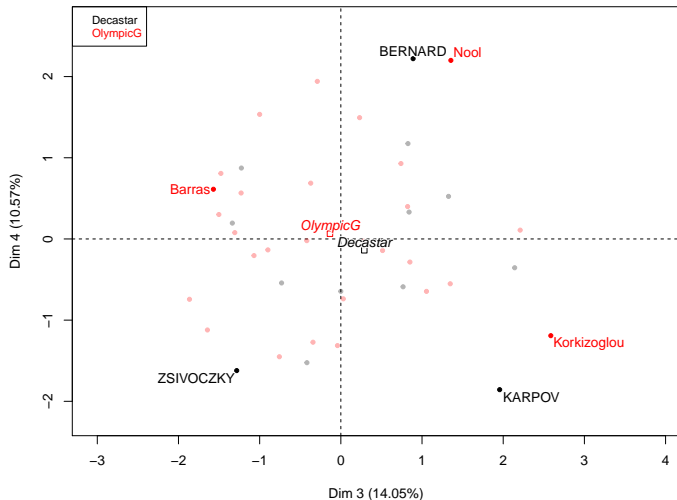
```
> plot(res.hcpc)
```

... de sélection en ACP ...



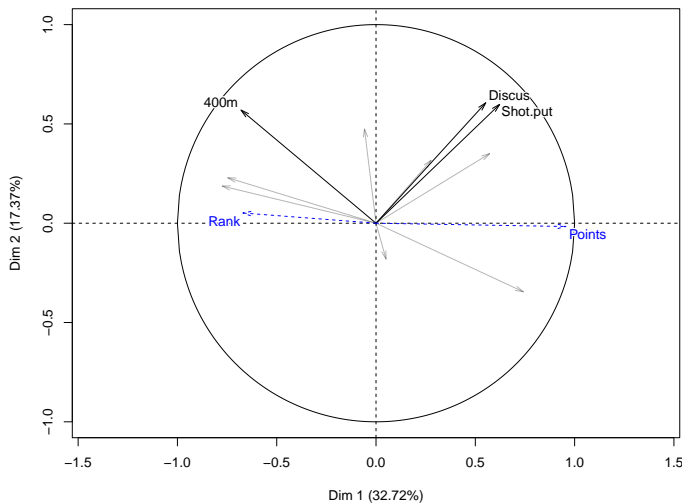
```
> data(decathlon)
> res.pca <- PCA(decathlon, quanti.sup = 11:12, quali.sup=13)
> plot(res.pca, select="contrib 10", shadow = TRUE, habillage = "Competition")
```

... de sélection d'individus contribuant au plan 3-4 ...



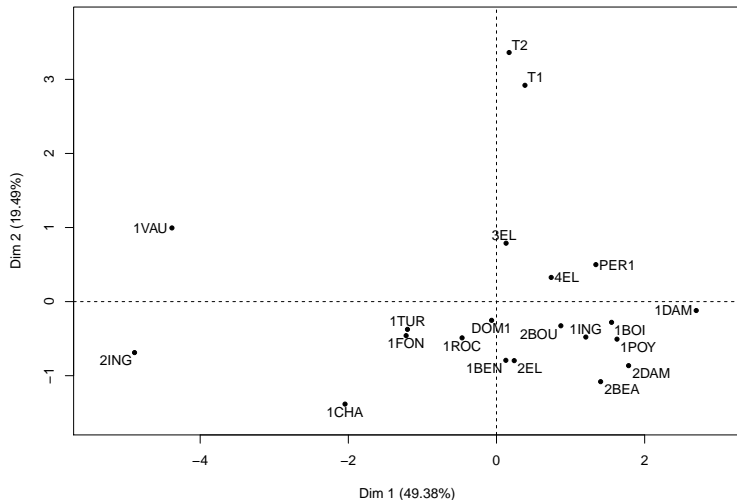
```
> plot(res.pca, select="cos2 0.5", shadow = TRUE,  
      habillage = "Competition", axes = 3:4)
```

... de sélection de variables bien projetées ...



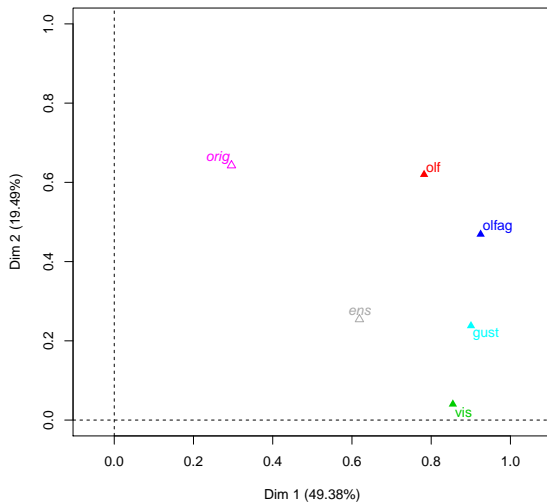
```
> plot(res.pca, choix="var",select="cos2 3", shadow=TRUE)
```


... d'AFM ...



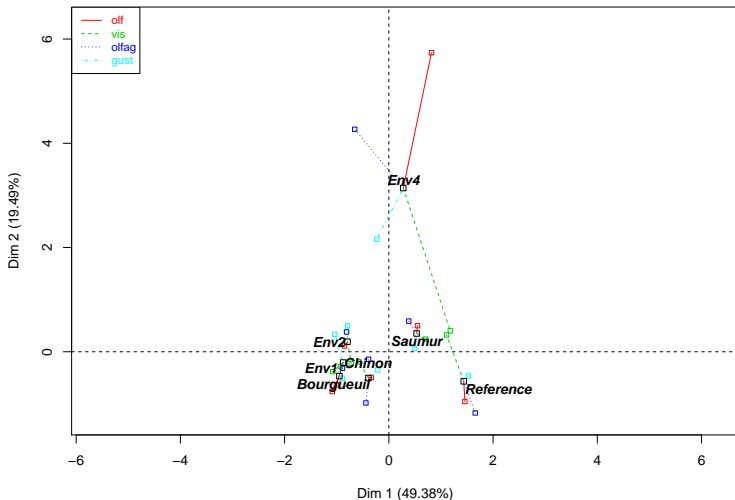
```
> res <- MFA(wine, group=c(2,5,3,10,9,2), type=c("n",rep("s",5)),
+   ncp=5, name.group=c("orig","olf","vis","olfag","gust","ens"),
+   num.group.sup=c(1,6))
> plot(res, invisible="quali",habillage="none")
```

... de représentation des groupes en AFM ...



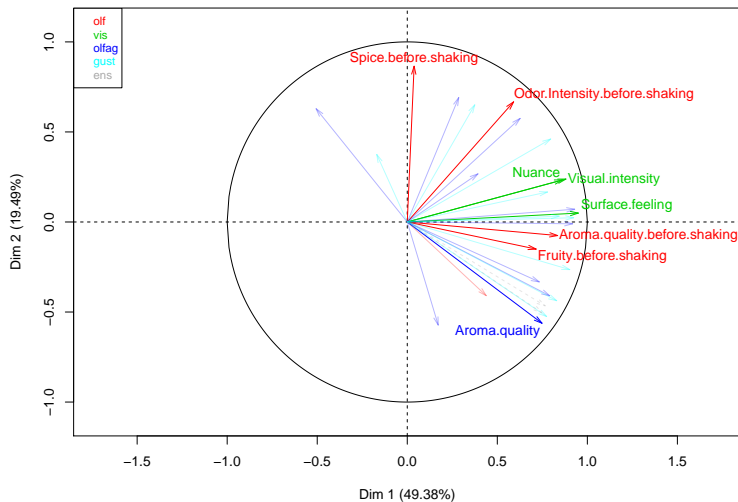
```
> plot(res, choix="group",habillage="group")
```

... de représentation des modalités moyennes et partielles ...



```
> plot(res, invisible="ind", partial="all", habillage="group")
```

... de sélection de variables en AFM ...



```
> plot(res,choix="var",hab="group",select="contrib 8")
```

... etc.