

Analyses multivariées avec R

MARBEC, Sète, 22-23 mai 2023

Monique Simier, IRD, UMR MARBEC

Objectifs de la formation

- Apprendre à identifier les **questions** statistiques relevant du domaine de l'analyse multivariée
- Savoir choisir une **méthode** en fonction de la nature des données et des questions posées
- Connaitre les principaux **packages R** permettant de réaliser des analyses multivariées
- Savoir mettre en pratique les analyses factorielles à un tableau de données (**ACP, AFC, ACM**) et à deux tableaux de données (**Analyses Inter/Intra-classes, Analyse Discriminante, Analyses sur Variables Instrumentales, Analyse de Co-inertie**) et interpréter les résultats
- Savoir mettre en pratique les méthodes de classification automatique (**CAH, partitionnement**) et interpréter les résultats
- Avoir un aperçu général des **méthodes d'analyses multi-tableaux**

Plan

- **1. Introduction**
 - Contexte et objectifs de l'analyse multivariée
 - Ordination ou classification ?
 - Ordination basée sur les distances ou sur les valeurs propres (analyses factorielles)
 - Les packages R spécialisés : ade4 et FactoMineR
- **2. Analyses factorielles à un tableau de données, avec les packages ade4 et FactoMineR**
 - 2.1 Analyse en Composantes Principales (ACP) : plusieurs variables quantitatives
 - 2.2 Analyse Factorielle des Correspondances (AFC) : deux variables qualitatives
 - 2.3 Analyse des Correspondances Multiples (ACM) : plusieurs variables qualitatives
- **3. Classification Automatique**
 - 3.1 Classification Ascendante Hiérarchique (CAH) avec la fonction hclust
 - 3.2 Partitionnement avec la fonction kmeans
- **4. Méthodes de couplage de tableaux avec le package ade4**
 - 4.1 Analyses Inter et Intra-classes
 - 4.2 Analyse Discriminante
 - 4.3 Analyses sur variables Instrumentales (ACPVI ou RDA, AFCVI ou CCA)
 - 4.4 Analyse de Co-inertie
- **5. Aperçu des méthodes d'analyse multi-tableaux**
 - 5.1 Analyse Triadique Partielle
 - 5.2 STATIS
 - 5.3 Analyse Factorielle Multiple

1. Introduction

- En statistiques, les données sont des **mesures** effectuées sur des unités d'observation (individus, animaux, objets...).
- La nature de ces mesures peut être :
 - **quantitative** (variable numérique) – exemple : taille, poids...
 - **qualitative** (variable catégorielle ou facteur) – exemple : couleur des yeux ou des cheveux
 - **ordinale** si le classement des modalités a un sens – exemple : notation en A>B>C>D>E
 - **nominale** sinon – exemple : pays, sexe...
- Une même variable peut être traitée comme quantitative ou qualitative – exemple : année
- Une variable quantitative peut être découpée en classes pour être transformée en variable qualitative – exemple : taille => petit/moyen/grand
- On n'utilise pas le même type d'approche statistique suivant le nombre de mesures prises sur chaque unité d'observation:
 - une seule mesure => approche univariée (moyenne, histogramme, camembert...)
 - deux mesures => approche bivariée (graphe en xy, corrélation, régression linéaire, anova...)
 - au-delà => approche **multivariée (c'est le sujet de cette formation !)**

L'approche multivariée

- Dans le contexte multivarié, on considère un **tableau rectangulaire** de données avec :
 - **n lignes** (unités statistiques ou observations ou individus)
 - **p colonnes** (mesures, exprimées par des variables quantitatives ou qualitatives)
- On cherche à synthétiser au mieux l'information contenue dans ce tableau, soit en :
 - Résumant les mesures par un **petit nombre de variables de synthèse ou facteurs** qui retiennent l'essentiel de l'information et permettent des représentations graphiques => **Méthodes Factorielles**
 - Etablissant une **partition de l'ensemble des individus**, hiérarchisée ou non => **Méthodes de Classification Automatique**

2. Analyses factorielles à un tableau

- Résumer les différentes mesures par un **petit nombre de variables de synthèse**
- Utiliser ces variables de synthèse comme **référentiel** pour représenter graphiquement :
 - les relations entre les **variables** originelles
 - les proximités entre les **individus**
- Cette possibilité d'obtenir une **représentation graphique** de la structure d'un grand jeu de données, autorisant une interprétation intuitive, a fait le succès de ces méthodes.
- Parmi les méthodes factorielles, ou méthodes **d'ordination**, les plus connues sont :
 - l'**Analyse en Composantes Principales (ACP)** quand les variables originelles sont quantitatives
 - l'**Analyse des Correspondances simple (AFC)** pour deux variables qualitatives
 - l'**Analyse des Correspondances multiple (ACM)** pour plusieurs variables qualitatives
 - l'**Analyse Factorielle avec données mixtes** pour un mélange de variables qualitatives et quantitatives
- Il existe aussi des méthodes d'ordination qui au lieu de s'appliquer aux données brutes (variables quantitatives) passent par une matrice de (dis)similarité entre les individus : **Analyse en Coordonnées Principales (PCoA)**, équivalente au Multidimensional Scaling (**MDS**).
- Ces méthodes sont disponibles dans la librairie de base de R, mais il existe deux librairies spécialisées que nous allons utiliser ici : **ade4** (développé à l'université de Lyon I, à l'origine pour les écologistes) et **FactoMineR** (développé à l'université de Rennes). La librairie **vegan** fournit aussi des outils d'ordination plus spécifiquement dédiés à l'écologie des communautés.

Analyses à un tableau: mise en œuvre avec R

Méthode	Variables	R de base	ade4	FactoMineR	vegan
ACP	Plusieurs variables quantitatives	prcomp princomp	dudi.pca	PCA	-
PCoA	Plusieurs variables quantitatives	cmdscale	dudi.pco	-	capscale
AFC	Deux variables qualitatives	corresp	dudi.coa	CA	decorana
ACM	Plusieurs variables qualitatives	mca	dudi.acm	MCA	-
Analyse mixte	Plusieurs variables qualitatives et/ou quantitatives	-	dudi.mix	FAMD	-

D'après analyse-R de Joseph Larmarange (<https://larmarange.github.io/analyse-R>)

dudi pour duality diagram = **schéma de dualité**, théorie qui permet une approche géométrique unifiée des différentes méthodes factorielles. En **rouge** les fonctions abordées dans ce cours.

Plan

- **1. Introduction**
 - Contexte et objectifs de l'analyse multivariée
 - Ordination ou classification ?
 - Ordination basée sur les distances ou sur les valeurs propres (analyses factorielles)
 - Les packages R spécialisés : ade4 et FactoMineR
- **2. Analyses factorielles à un tableau de données, avec les packages ade4 et FactoMineR**
 - 2.1 Analyse en Composantes Principales (ACP) : plusieurs variables quantitatives
 - 2.2 Analyse Factorielle des Correspondances (AFC) : deux variables qualitatives
 - 2.3 Analyse des Correspondances Multiples (ACM) : plusieurs variables qualitatives
- **3. Classification Automatique**
 - 3.1 Classification Ascendante Hiérarchique (CAH) avec la fonction hclust
 - 3.2 Partitionnement avec la fonction kmeans
- **4. Méthodes de couplage de tableaux avec le package ade4**
 - 4.1 Analyses Inter et Intra-classes
 - 4.2 Analyse Discriminante
 - 4.3 Analyses sur variables Instrumentales (ACPVI ou RDA, AFCVI ou CCA)
 - 4.4 Analyse de Co-inertie
- **5. Aperçu des méthodes d'analyse multi-tableaux**
 - 5.1 Analyse Triadique Partielle
 - 5.2 STATIS
 - 5.3 Analyse Factorielle Multiple

2.1 Analyse en Composantes Principales

- Application : tableau dont les variables sont **quantitatives**
- Méthode qui permet de construire une nouvelle variable ressemblant le plus possible à toutes les variables du tableau étudié. On appelle cette variable de synthèse une **composante principale** ou **axe principal**.
- La composante principale a pour propriété d'être de **corrélation maximum avec l'ensemble des variables du tableau étudié**. Plus précisément, la somme des carrés de corrélations de la variable de synthèse avec les variables originelles est maximisée. Ce maximum, qui quantifie son pouvoir de synthèse, est appelé **valeur propre**.
- Une fois la première composante définie, une **seconde composante**, qui est la seconde meilleure synthèse des données, est calculée avec la contrainte d'être **indépendante** de la première. Et ainsi de suite...
- Les axes principaux sont orthogonaux entre eux (car indépendants) et définissent un **espace géométrique**. La corrélation des variables avec les axes permet de les interpréter (cercle des corrélations). Les observations sont projetées sur les axes.
- Les n premiers axes fournissent la meilleure représentation possible du tableau de données. Les valeurs propres permettent de **quantifier la part de la variance totale** (ou inertie) de l'ensemble des variables expliquée par chaque axe.

ACP : en entrée

- Le tableau de données doit être sous forme de **data frame** à n lignes (observations) et p colonnes (variables).
- Toutes les variables doivent être de type **numérique**
- **Les données manquantes ne sont pas permises** : supprimer les observations ou les variables qui en comportent beaucoup, ou les remplacer par la moyenne de la variable.
- Lors de l'analyse, le tableau de données peut être au choix:
 - **normé** (chaque valeur est remplacée par l'écart à la moyenne de la variable et divisée par l'écart-type) : cas le plus classique, indispensable si différentes unités.
 - **centré** (écart à la moyenne, sans division par l'écart-type) : possible si les variables sont toutes dans la même unité (ex: comptage d'espèces)
- Lors de l'analyse, des **pondérations** sont attribuées aux lignes et aux colonnes. Les valeurs par défaut peuvent être modifiées si besoin.

ACP : en sortie

- Les **p valeurs propres**, dont on retient les 2, 3 ou 4 premières au vu de l'éboulis des valeurs propres => calcul du pourcentage représenté par chaque axe.
- Les **coordonnées des variables sur les axes** (=corrélations si ACP normée) => représentation graphique des variables (cercle des corrélations si ACP normée)
- Les **coordonnées des observations sur ces mêmes axes** => représentation graphique (ou nuage) des observations.
- Les **contributions** des variables / individus aux axes
- La **qualité de la représentation** des variables sur les axes, mesurée par le cosinus² de l'angle entre position variable / individu et axe
- Dans les représentations graphiques, on croise les axes 2 à 2. Pour le plan défini par les axes 1 et 2, il est d'usage de mettre **l'axe 1 horizontal et l'axe 2 vertical**. Idem pour les plans 3-4, etc.

Analyse en Composantes Principales avec la fonction **dudi.pca()** d'ade4

`dudi.pca(df, row.w = rep(1, nrow(df))/nrow(df), col.w = rep(1, ncol(df)),
center = TRUE, scale = TRUE, scannf = TRUE, nf = 2)`

df	un “data frame” à n lignes (unités statistiques) et p colonnes (variables numériques)
row.w	un vecteur optionnel des poids des lignes (par défaut, poids uniformes : 1/n)
col.w	un vecteur optionnel des poids des colonnes (par défaut, poids unitaire : 1)
center	option de centrage : une valeur logique (TRUE/FALSE) ou un vecteur numérique <ul style="list-style-type: none">• si TRUE, centrage par la moyenne• si FALSE pas de centrage• si vecteur numérique, sa longueur doit être égale au nombre de colonnes du data frame et donner les valeurs centrales pour chaque variable
scale	une valeur logique (TRUE/FALSE) indiquant si les variables doivent être normées
scannf	une valeur logique (TRUE/FALSE) indiquant si le diagramme des valeurs propres doit être affiché
nf	si scannf =FALSE, un entier indiquant le nombre d’axes à conserver

Les données Méaudret (librairie ade4)

Données de la thèse de D. Pegaz-Maucet (1980)

Impact d'une perturbation d'origine organique sur la dérive des macro-invertébrés benthiques d'un cours d'eau (Méaudret, Vercors).

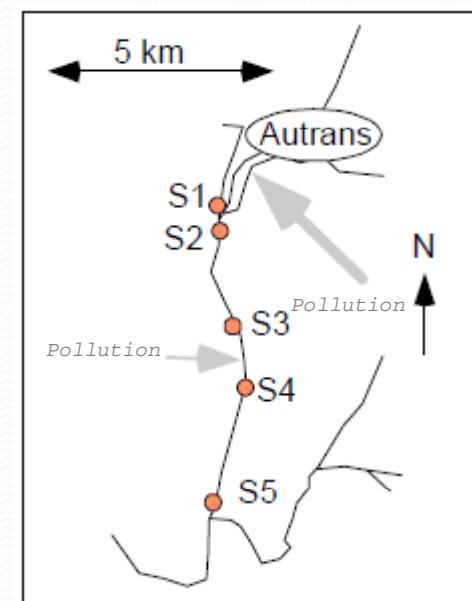
```
library(ade4)
data(meaudret)
meaudret
```

meaudret est une liste de 4 composantes:

- env est un data frame avec 20 échantillons et 9 variables.
- fau est un data frame avec 20 échantillons et 13 espèces d'Epheméroptères.
- design est un data frame avec 20 échantillons et 2 facteurs.
 - season est un facteur à 4 modalités = saisons.
 - site est un facteur à 5 modalités = sites sur la rivière Meaudret.
- spe.names est un vecteur de type "character" contenant les noms des 13 espèces de fau.

Code des variables

- 1- Temp : Température (°C)
- 2- Flow : Débit (l/s)
- 3- pH : pH
- 4- Cond : Conductivité (mmho/cm)
- 5- Bdo5 : DBO5 (mg/l oxygène)
- 6- Oxyd : Oxydabilité (idem)
- 7- Ammo : Ammoniaque (mg/l)
- 8- Nitr : Nitrates (mg/l)
- 9- Phos : Orthophosphates (mg/l)



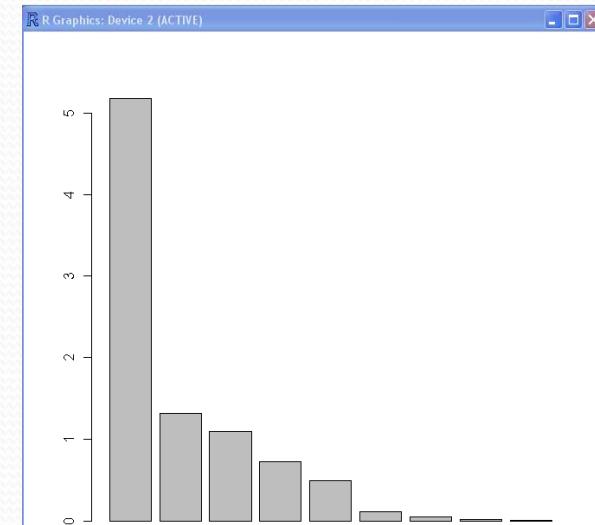
Etude 1: ACP normée sur le tableau Méaudret environnement

```
library(ade4)
data(meaudret)
meaudret
$env
  Temp Flow  pH Cond Bdo5 Oxyd  Ammo Nitr Phos
sp_1    10    41 8.5   295   2.3   1.4   0.12   3.4  0.11
sp_2    11   158 8.3   315   7.6   3.3   2.85   2.7  1.50
...
wi_4     3   498 8.3   330   4.8   1.6   1.04   4.4  0.82
wi_5     2   390 8.2   330   1.7   1.2   0.56   5.0  0.60

mil <- meaudret$env

# ACP normée par défaut
pca1 <- dudi.pca(mil)
Select the number of axes: 5

pca1 <- dudi.pca(df = mil, scannf = F, nf = 5)
```



ACP normée sur le tableau Méaudret environnement

```
pca1
Duality diagramm
class: pca dudi
$call: dudi.pca(df = mil, scannf = F, nf = 5)

$nf: 5 axis-components saved
$rank: 9
eigen values: 5.175 1.32 1.093 0.7321 0.4902 ...
  vector length mode    content
1 $cw     9    numeric column weights ← Poids des colonnes = variables
2 $lw    20    numeric row weights  ← Poids des lignes = individus
3 $eig    9    numeric eigen values ← Valeurs propres

  data.frame nrow ncol content
1 $stab      20    9    modified array ← Tableau centré et réduit
2 $li       20    5    row coordinates ← Coordonnées des lignes
3 $l1       20    5    row normed scores ← Coord. lignes / val. propre
4 $co        9    5    column coordinates ← Coordonnées des colonnes
5 $c1        9    5    column normed scores ← Coord. colonnes / val. propre
other elements: cent norm
```

ACP normée sur le tableau Méaudret environnement

```
pve <- 100 *pca1$eig / sum(pca1$eig) # Pourcentages d'inertie des axes
print(pve, digits=2)
[1] 57.50 14.67 12.15 8.13 5.45 1.22 0.59 0.22 0.07

cpve <- cumsum(pve) # Pourcentages cumulés
print(cpve, digits=3)
[1] 57.5 72.2 84.3 92.5 97.9 99.1 99.7 99.9 100.0

pca1$co # Coordonnées des colonnes (=variables)
          Comp1        Comp2        Comp3        Comp4        Comp5
Temp    0.1054115 -0.32090155  0.83747545  0.42252564  0.057462741
Flow   -0.2727582 -0.47432815 -0.58116930  0.60159191 -0.005507834
...
pca1$li # Coordonnées des lignes (=échantillons)
          Axis1        Axis2        Axis3        Axis4        Axis5
sp_1   -1.75632208 -0.4635311  1.1291841 -1.12513814  0.34323496
sp_2    0.03943889 -1.0741475  0.6037606 -0.32789822  0.34618838
...
```

Statistiques d'inertie

```
inertia.dudi(pca1, row.inertia=TRUE, col.inertia=TRUE)
```

Decomposition of total inertia:

	inertia	cum	cum(%)
Ax1	5.174737	5.175	57.50
Ax2	1.320419	6.495	72.17
...			
Ax9	0.006316	9.000	100.00

Row contributions (%):

sp_1	sp_2	sp_3	sp_4	sp_5	su_1	su_2	su_3	su_4	su_5	au_1	au_2
3.323	1.052	2.780	3.459	3.563	2.395	13.126	7.333	3.480	2.724	3.748	26.875
au_3	au_4	au_5	wi_1	wi_2	wi_3	wi_4	wi_5				
4.749	3.318	4.100	3.046	1.089	1.974	4.620	3.245				

Row absolute contributions (%):

	Axis1	Axis2	Axis3	Axis4	Axis5
sp_1	2.980507	0.8136	5.8308	8.64576	1.20162
sp_2	0.001503	4.3690	1.6670	0.73429	1.22239
...					
wi_5	1.647894	1.0203	12.7976	2.18989	7.45691

Signed row relative contributions:

	Axis1	Axis2	Axis3	Axis4	Axis5
sp_1	-51.56561	-3.5918	21.3148	-21.1623	1.969404
sp_2	0.08213	-60.9232	19.2479	-5.6772	6.328193
...					
wi_5	-29.19876	-4.6131	-47.9120	5.4897	-12.516741

Cumulative sum of row relative contributions (%):

	Axis1	Axis1:2	Axis1:3	Axis1:4	Axis1:5	Axis6:9
sp_1	51.56561	55.157	76.47	97.63	99.60	0.39607
sp_2	0.08213	61.005	80.25	85.93	92.26	7.74138
...						
wi_5	29.19876	33.812	81.72	87.21	99.73	0.26967

Statistiques d'inertie (suite)

Column contributions (%):

Temp	Flow	pH	Cond	Bdo5	Oxyd	Ammo	Nitr	Phos
11.11	11.11	11.11	11.11	11.11	11.11	11.11	11.11	11.11

Column absolute contributions (%):

	Axis1	Axis2	Axis3	Axis4	Axis5
Temp	0.2147	7.7989	64.14674	24.38529	0.673577
Flow	1.4377	17.0391	30.89127	49.43400	0.006188
pH	12.1670	0.1213	0.52821	0.68544	70.788084
Cond	15.9889	2.6373	3.44540	1.32741	2.654485
Bdo5	18.0025	0.8780	0.43463	0.44751	7.473441
Oxyd	16.3878	3.7632	0.25313	1.58254	10.761163
Ammo	18.7344	0.2698	0.07917	0.03637	0.348308
Nitr	0.6966	62.2135	0.04641	16.94420	2.149243
Phos	16.3704	5.2789	0.17503	5.15725	5.145510

Signed column relative contributions:

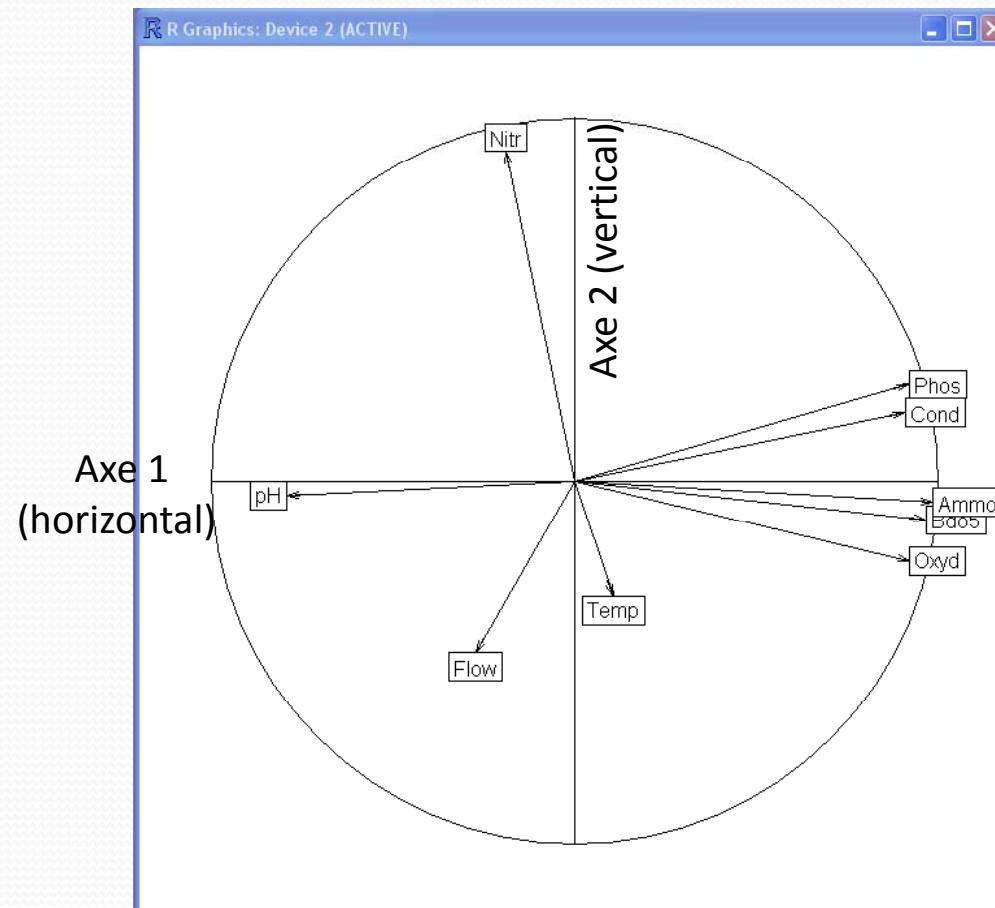
	Axis1	Axis2	Axis3	Axis4	Axis5
Temp	1.111	-10.2978	70.13651	17.85279	0.330197
Flow	-7.440	-22.4987	-33.77578	36.19128	-0.003034
pH	-62.961	-0.1602	-0.57754	-0.50182	34.701288
Cond	82.738	3.4824	-3.76711	0.97182	-1.301265
Bdo5	93.158	-1.1593	-0.47521	-0.32763	3.663583
Oxyd	84.802	-4.9691	-0.27677	-1.15860	5.275270
Ammo	96.946	-0.3562	0.08657	0.02662	0.170746
Nitr	-3.605	82.1478	0.05074	12.40507	1.053589
Phos	84.712	6.9704	-0.19138	3.77569	2.522400

Cumulative sum of column relative contributions (%):

	Axis1	Axis1:2	Axis1:3	Axis1:4	Axis1:5	Axis6:9
Temp	1.111	11.41	81.55	99.40	99.73	0.27156
Flow	7.440	29.94	63.71	99.91	99.91	0.09149
pH	62.961	63.12	63.70	64.20	98.90	1.09795
Cond	82.738	86.22	89.99	90.96	92.26	7.73930
Bdo5	93.158	94.32	94.79	95.12	98.78	1.21587
Oxyd	84.802	89.77	90.05	91.21	96.48	3.51793
Ammo	96.946	97.30	97.39	97.42	97.59	2.41418
Nitr	3.605	85.75	85.80	98.21	99.26	0.73819
Phos	84.712	91.68	91.87	95.65	98.17	1.82771

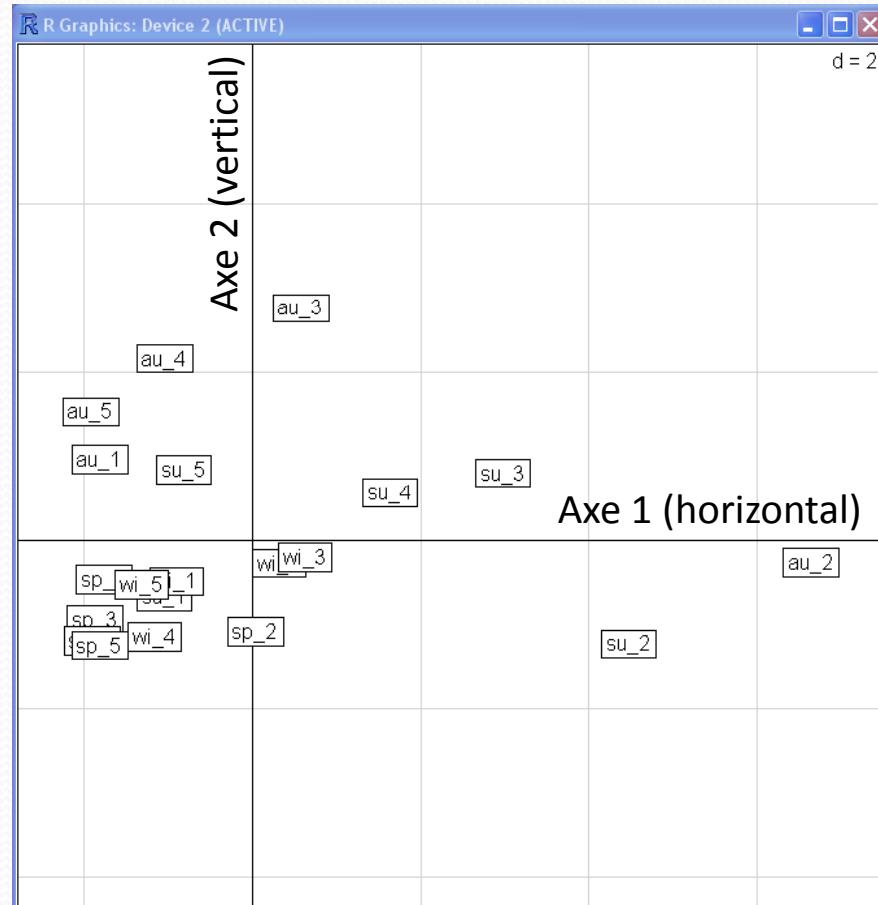
Cercle de corrélations des variables

```
# Cercle de corrélations
# Paramétrage par défaut
s.corcircle(
  pcal$co,
  xax = 1,
  yax = 2,
  label = row.names(pcal$co),
  clabel = 1,
  grid = FALSE,
  sub = "",
  csub = 1,
  possub = "bottomleft",
  cgrid = 0,
  fullcircle = TRUE,
  box = FALSE,
  add.plot = FALSE)
```

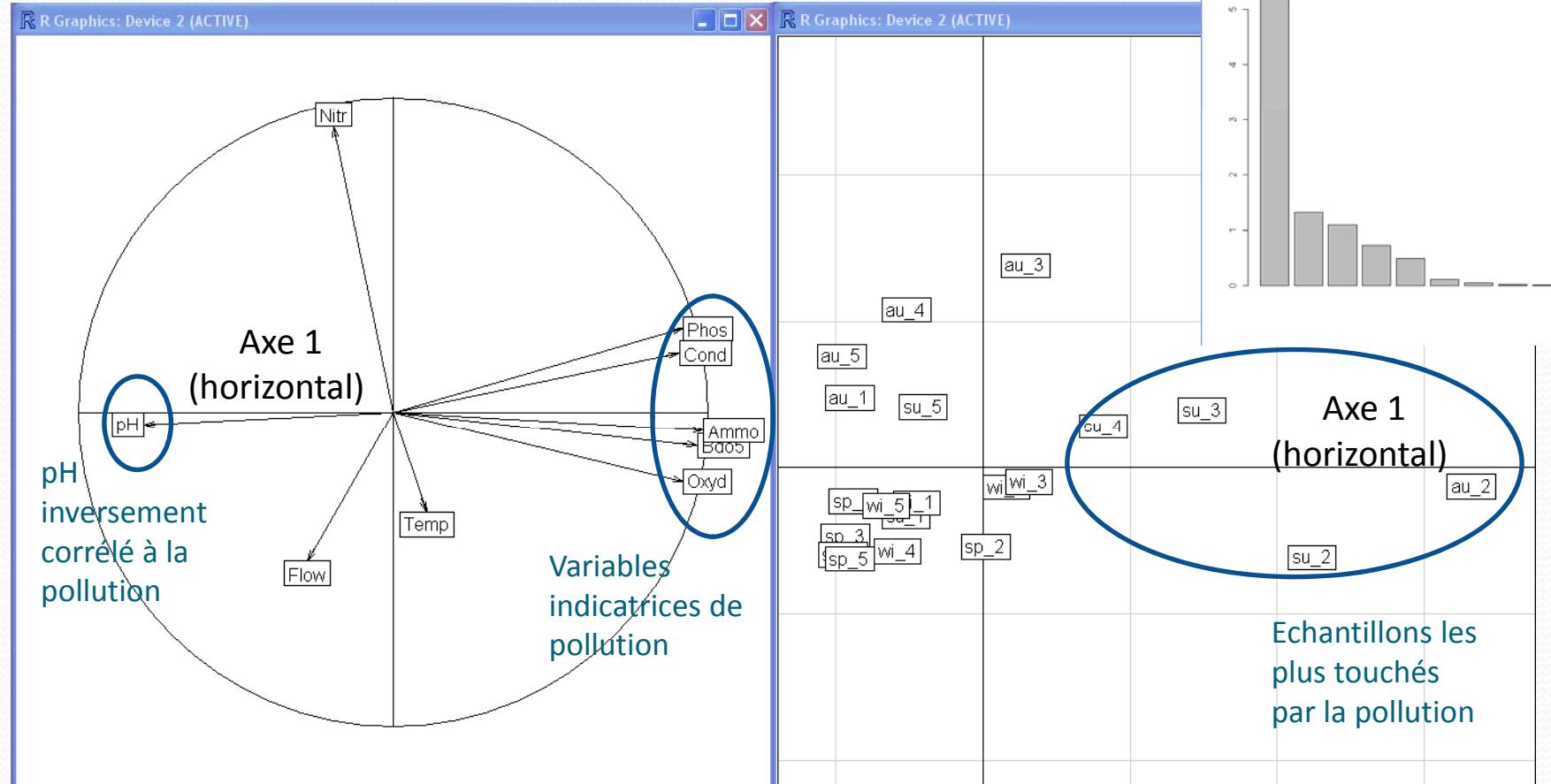


Nuage des individus

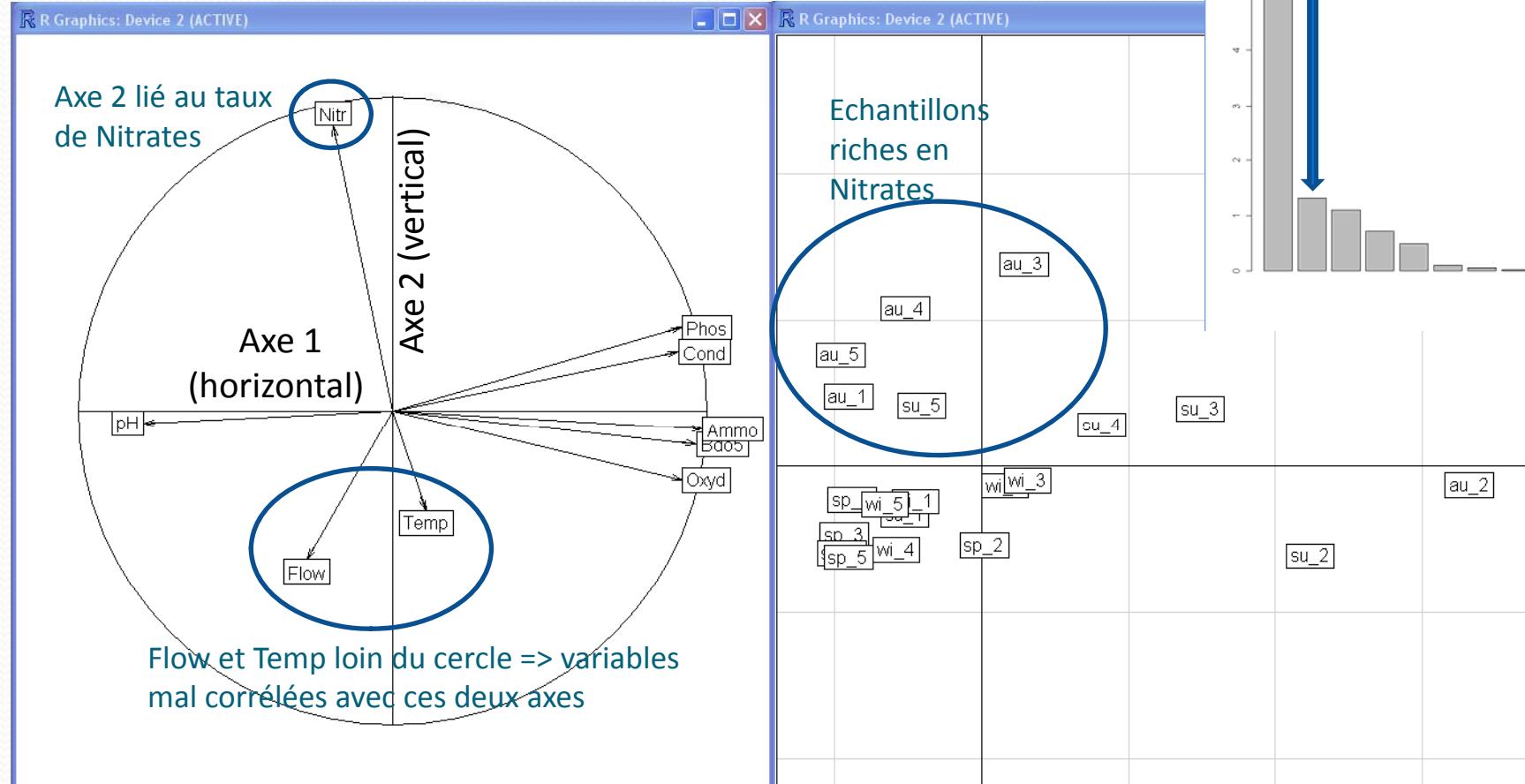
```
# Nuage des individus
# Valeurs par défaut
s.label(pca1$li,
xax = 1, yax = 2,
label = row.names(pca1$li),
clabel = 1,
pch = 20,
cpoint = 0,
boxes = TRUE,
neig = NULL, cneig = 2,
xlim = NULL, ylim = NULL,
grid = TRUE,
addaxes = TRUE,
cgrid = 1,
include.origin = TRUE,
origin = c(0,0),
sub = "", csub = 1.25,
possub = "bottomleft",
pixmap = NULL,
contour = NULL, area = NULL,
add.plot = FALSE)
```



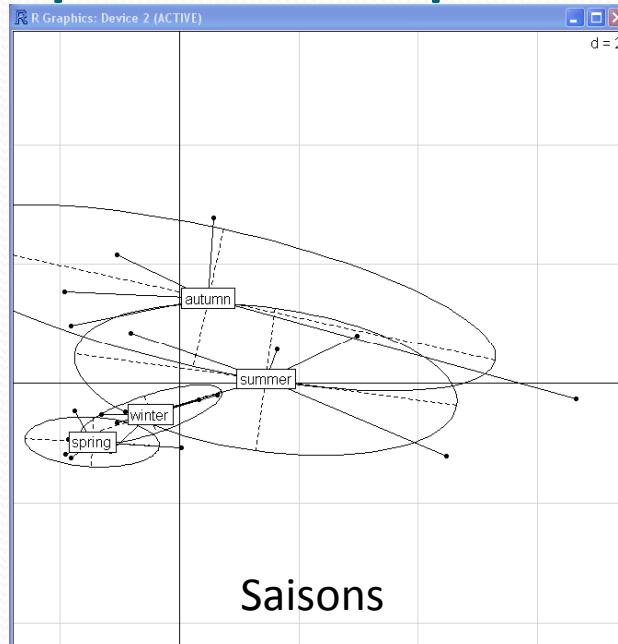
Interprétation de l'ACP normée



Interprétation de l'ACP normée

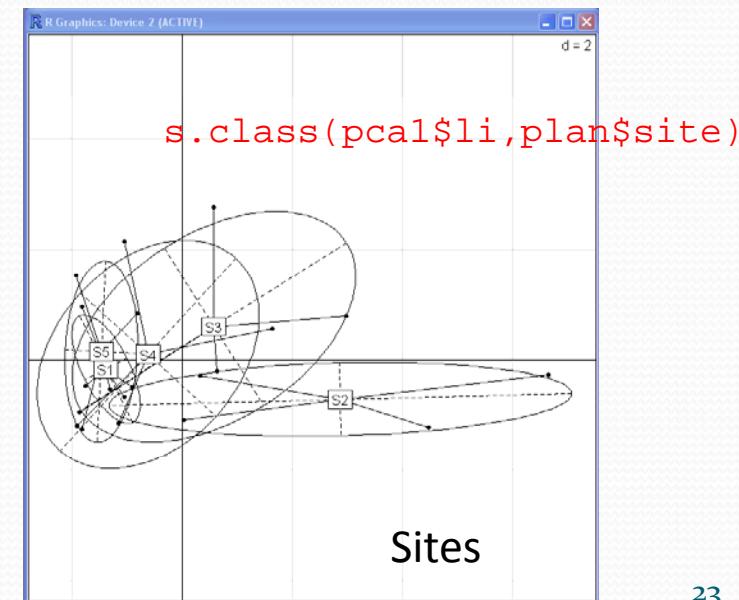
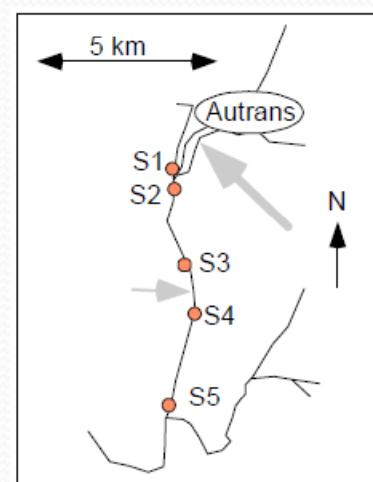


Interprétation de l'ACP : projeter la structure spatio temporelle des échantillons



`s.class(pca1$li, plan$season)`

- Forte dispersion en été et automne et pour sites 2, 3, 4
- Homogénéité en hiver et printemps et pour sites 1 et 5
- Gradient saisonnier sur l'axe 1
- La pollution organique se produit en été, surtout en amont de la station 2. Le site 1 n'est pas touché et au niveau du site 5 (aval), la pollution s'est diluée. En automne, S3 et S4 se restaurent mais pas S2.



Travaux pratiques : Le jeu de données

The screenshots illustrate the PPEAO (Peuplements de poissons et la Pêche artisanale des Ecosystèmes estuariens, lagunaires ou continentaux d'Afrique de l'Ouest) database interface. The left screenshot shows the main landing page with a brief description of the dataset and some sample images. The right screenshot shows a more detailed view where users can select specific data to extract, including environmental variables like sediment and vegetation measurements.

- Données extraites de ppeao.ird.fr
- Pêches expérimentales dans 4 systèmes estuariens d'Afrique de l'Ouest (Ebrié, Fatala, Gambie, Saloum), à 2 saisons (sèche / humide) => **256 coups de pêche**
- **Envir** = 14 variables : mesure de 5 variables environnementales lors des 256 coups de pêche (Salinité, Température, Transparence, Profondeur, Distance à l'embouchure) + divers identifiants (saison, système...)
- **Faune** = abondance de 111 espèces de poissons pour les 256 coups de pêche
- **Categ** = catégories écologiques et trophiques des 111 espèces

Exercice 1 : Analyse de l'environnement

- Importer le fichier **Envir.csv** sous R => dataframe
- Identifier les 5 variables environnementales parmi les 14 variables et en faire des boxplots par système et saison
- Faire l'ACP de ces 5 variables
- Décrire les valeurs propres (combien en garder, quel % ?)
- Représenter le cercle des corrélations des variables
- Représenter le nuage des individus
- Représenter le nuage des individus en les regroupant par saison, par système et par système-saison

Analyse en Composantes Principales avec la fonction **PCA()** de FactoMineR

PCA(df, scale.unit = TRUE, ncp = 5, ind.sup = NULL, quanti.sup = NULL, quali.sup = NULL, row.w = NULL, col.w = NULL, graph = TRUE, axes = c(1,2))

df	Un “data frame” à n lignes (unités statistiques) et p colonnes (variables numériques)
scale.unit	Un logique (TRUE/FALSE). Si TRUE (option par défaut), les données sont normées
ncp	Nombre d’axes conservés dans les résultats (par défaut 5)
ind.sup	Un vecteur indiquant les indices des individus supplémentaires
quanti.sup	Un vecteur indiquant les indices des variables quantitatives supplémentaires
quali.sup	Un vecteur indiquant les indices des variables qualitatives supplémentaires
row.w	Un vecteur optionnel des poids des individus actifs (par défaut, poids uniformes : 1/n)
col.w	Un vecteur optionnel des poids des variables actives (par défaut, poids unitaire : 1)
graph	Un logique, si TRUE (option par défaut), un graphique est affiché
axes	Un vecteur de longueur 2 indiquant les axes à afficher

ACP normée avec FactoMineR sur Méaudret environnement

```
# ACP avec FactoMineR sur meaudret$env
data(meaudret)

# Concaténer env (9 variables) et design (season, site)
meo <- cbind(meaudret$env, meaudret$design)

# ACP du tableau meo (col. 1 à 9), season (10) et site (11) en qualitatif
res.meo <- PCA(meo, quali.sup=c(10,11),graph=FALSE)
res.meo

**Results for the Principal Component Analysis (PCA)**
The analysis was performed on 20 individuals, described by 11 variables
*The results are available in the following objects:
  name          description
1  "$eig"        "eigenvalues"
2  "$var"         "results for the variables"
3  "$var$coord"   "coord. for the variables"
4  "$var$cor"     "correlations variables - dimensions"
5  "$var$cos2"    "cos2 for the variables"
6  "$var$contrib" "contributions of the variables"
7  "$ind"         "results for the individuals"
8  "$ind$coord"   "coord. for the individuals"
9  "$ind$cos2"    "cos2 for the individuals"
10 "$ind$contrib" "contributions of the individuals"
11 "$quali.sup"   "results for the supplementary categorical variables"
12 "$quali.sup$coord" "coord. for the supplementary categories"
13 "$quali.sup$v.test" "v-test of the supplementary categories"
14 "$call"        "summary statistics"
15 "$call$centre"  "mean of the variables"
16 "$call$ecart.type" "standard error of the variables"
17 "$call$row.w"   "weights for the individuals"
18 "$call$col.w"   "weights for the variables"
```

ACP normée avec FactoMineR sur Méaudret environnement

`summary(res.meo)`

Call:
`PCA(X = meo, quali.sup = c(10, 11), graph = FALSE)`

Eigenvalues

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5	Dim.6	Dim.7	Dim.8	Dim.9
Variance	5.175	1.320	1.093	0.732	0.490	0.110	0.053	0.020	0.006
% of var.	57.497	14.671	12.149	8.135	5.447	1.220	0.588	0.223	0.070
Cumulative % of var.	57.497	72.168	84.317	92.452	97.898	99.119	99.707	99.930	100.000

Individuals (the 10 first)

	Dist	Dim.1	ctr	cos2	Dim.2	ctr	cos2	Dim.3	ctr	cos2
sp_1	2.446	-1.756	2.981	0.516	-0.464	0.814	0.036	-1.129	5.831	0.213
sp_2	1.376	0.039	0.002	0.001	-1.074	4.369	0.609	-0.604	1.667	0.192
sp_3	2.237	-1.860	3.342	0.691	-0.936	3.314	0.175	-0.612	1.713	0.075
sp_4	2.495	-1.900	3.487	0.580	-1.187	5.331	0.226	-0.357	0.583	0.020
sp_5	2.533	-1.808	3.158	0.510	-1.241	5.834	0.240	-0.374	0.639	0.022
su_1	2.076	-1.036	1.036	0.249	-0.661	1.654	0.101	-1.369	8.570	0.435
su_2	4.861	4.469	19.300	0.845	-1.222	5.658	0.063	-0.917	3.843	0.036
su_3	3.633	2.979	8.573	0.672	0.802	2.435	0.049	-1.195	6.529	0.108
su_4	2.503	1.635	2.584	0.427	0.571	1.235	0.052	-1.269	7.359	0.257
su_5	2.214	-0.812	0.637	0.134	0.849	2.732	0.147	-1.114	5.674	0.253

Variables

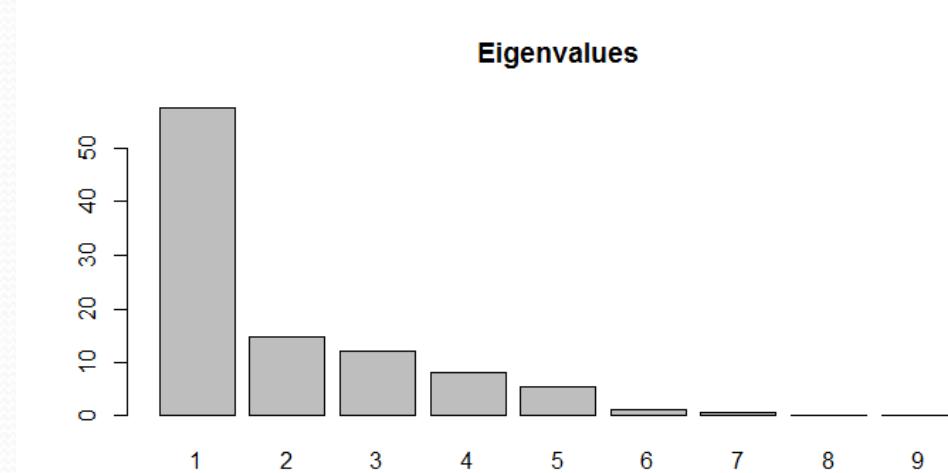
	Dim.1	ctr	cos2	Dim.2	ctr	cos2	Dim.3	ctr	cos2
Temp	0.105	0.215	0.011	-0.321	7.799	0.103	-0.837	64.147	0.701
Flow	-0.273	1.438	0.074	-0.474	17.039	0.225	0.581	30.891	0.338
pH	-0.793	12.167	0.630	-0.040	0.121	0.002	0.076	0.528	0.006
Cond	0.910	15.989	0.827	0.187	2.637	0.035	0.194	3.445	0.038
Bdo5	0.965	18.003	0.932	-0.108	0.878	0.012	0.069	0.435	0.005
Oxyd	0.921	16.388	0.848	-0.223	3.763	0.050	0.053	0.253	0.003
Ammo	0.985	18.734	0.969	-0.060	0.270	0.004	-0.029	0.079	0.001
Nitr	-0.190	0.697	0.036	0.906	62.213	0.821	-0.023	0.046	0.001
Phos	0.920	16.370	0.847	0.264	5.279	0.070	0.044	0.175	0.002

Supplementary categories

	Dist	Dim.1	cos2	v.test	Dim.2	cos2	v.test	Dim.3	cos2	v.test
spring	1.949	-1.457	0.559	-1.612	-0.980	0.253	-2.147	-0.615	0.100	-1.481
summer	1.979	1.447	0.535	1.601	0.068	0.001	0.148	-1.173	0.351	-2.822
autumn	1.737	0.490	0.080	0.542	1.434	0.682	3.141	0.427	0.060	1.028
winter	1.634	-0.480	0.086	-0.531	-0.522	0.102	-1.143	1.361	0.693	3.275
S1	1.934	-1.374	0.505	-1.317	-0.160	0.007	-0.304	-0.518	0.072	-1.080
S2	3.009	2.868	0.909	2.748	-0.707	0.055	-1.341	0.194	0.004	0.403
S3	0.970	0.582	0.360	0.558	0.608	0.393	1.153	0.055	0.003	0.115
S4	0.959	-0.615	0.411	-0.589	0.103	0.012	0.196	0.162	0.028	0.337
S5	1.571	-1.461	0.865	-1.400	0.156	0.010	0.296	0.107	0.005	0.224

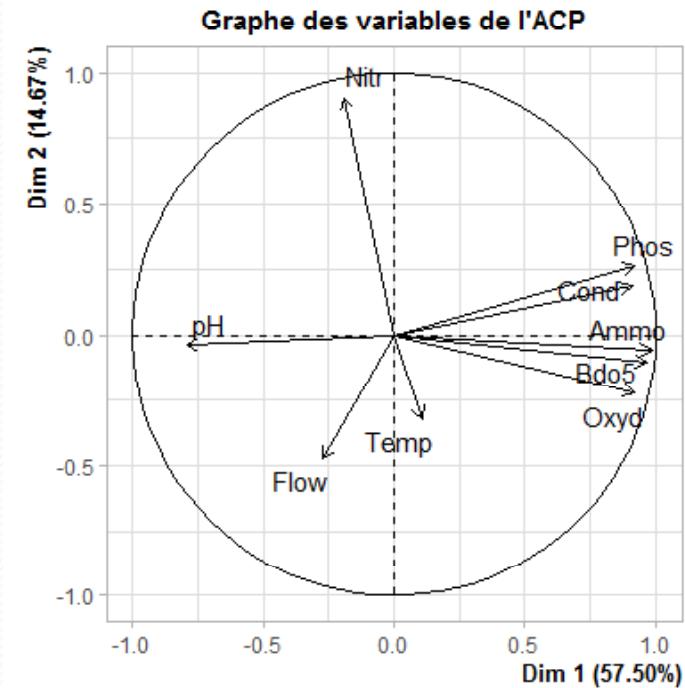
ACP normée avec FactoMineR sur Méaudret environnement

```
# Diagramme des valeurs propres  
barplot(res.meo$eig[,2],main="Eigenvalues",names.arg=1:nrow(res.meo$eig))
```



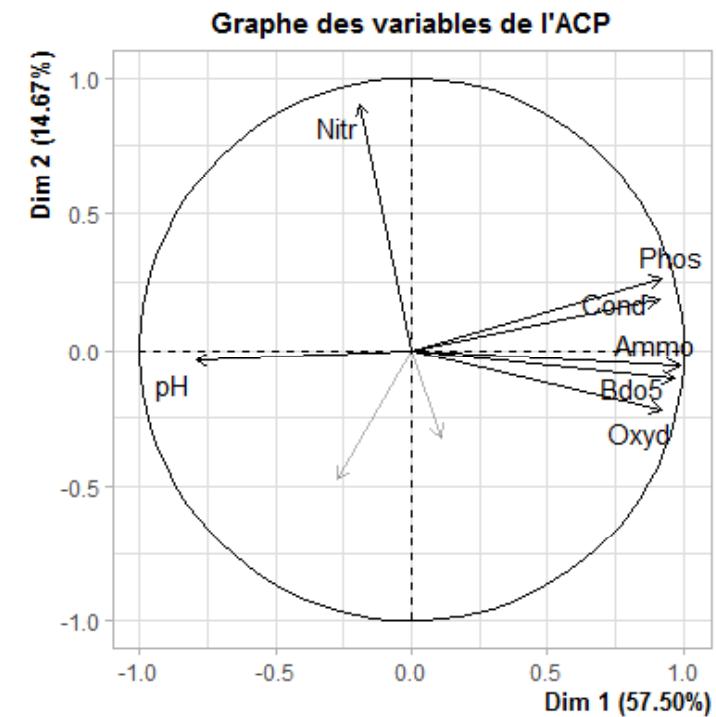
ACP normée avec FactoMineR sur Méaudret environnement

```
# Nuage des variables ( cercle des corrélations )
plot.PCA(res.meo,
          choix='var',
          title="Graphe des variables de l'ACP" )
```



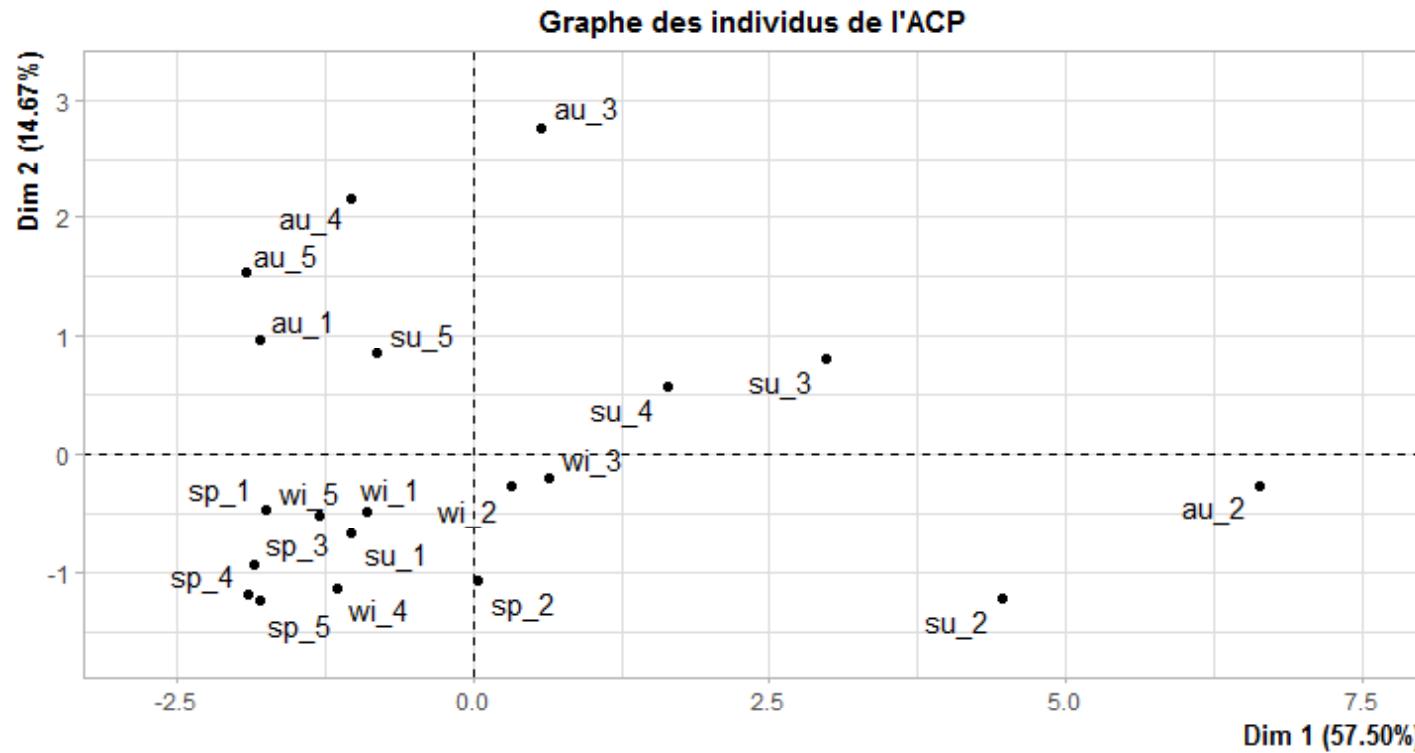
ACP normée avec FactoMineR sur Méaudret environnement

```
# On ne garde que les variables dont le cos2 (qualité de la
représentation) est supérieur à 0.5
plot.PCA(res.meo,
          choix='var',
          select='cos2 0.5',
          title="Graphe des variables de l'ACP")
```



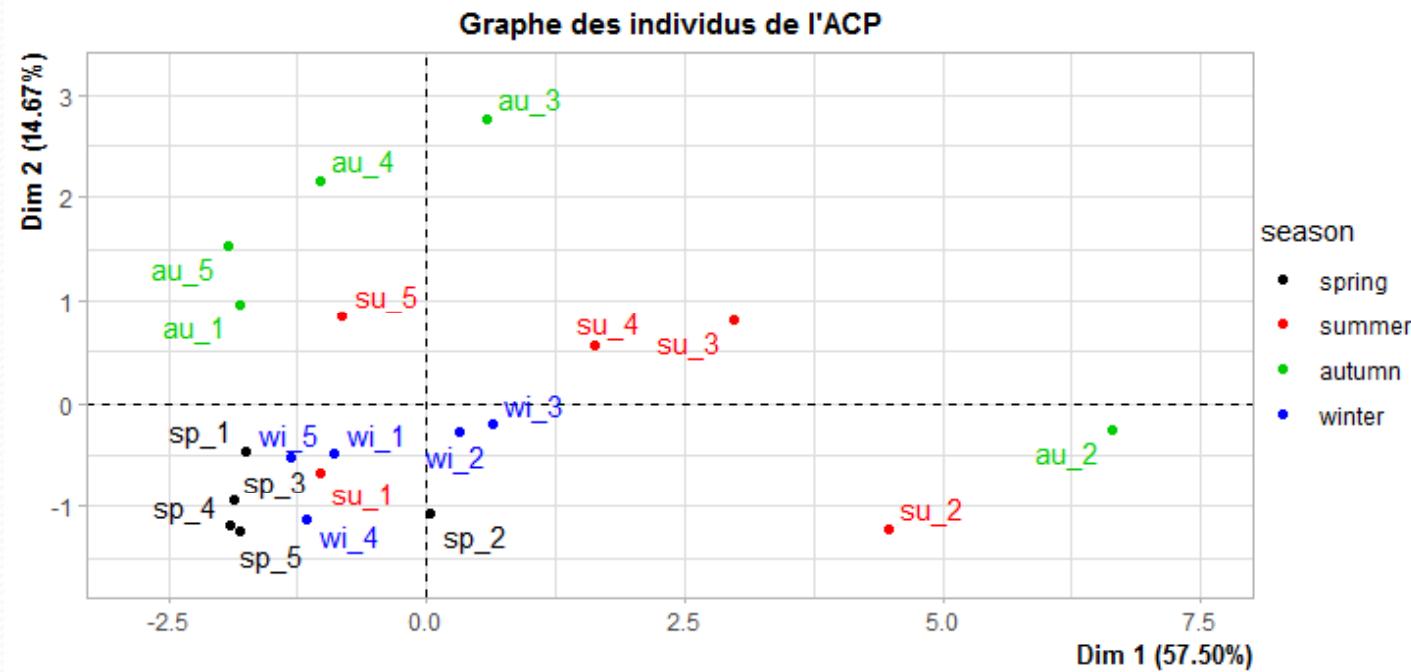
ACP normée avec FactoMineR sur Méaudret environnement

```
# Nuage des individus
plot.PCA(res.meo, choix='ind', invisible=c('quali', 'ind.sup'),
          title="Graphe des individus de l'ACP", label = c('ind'))
```



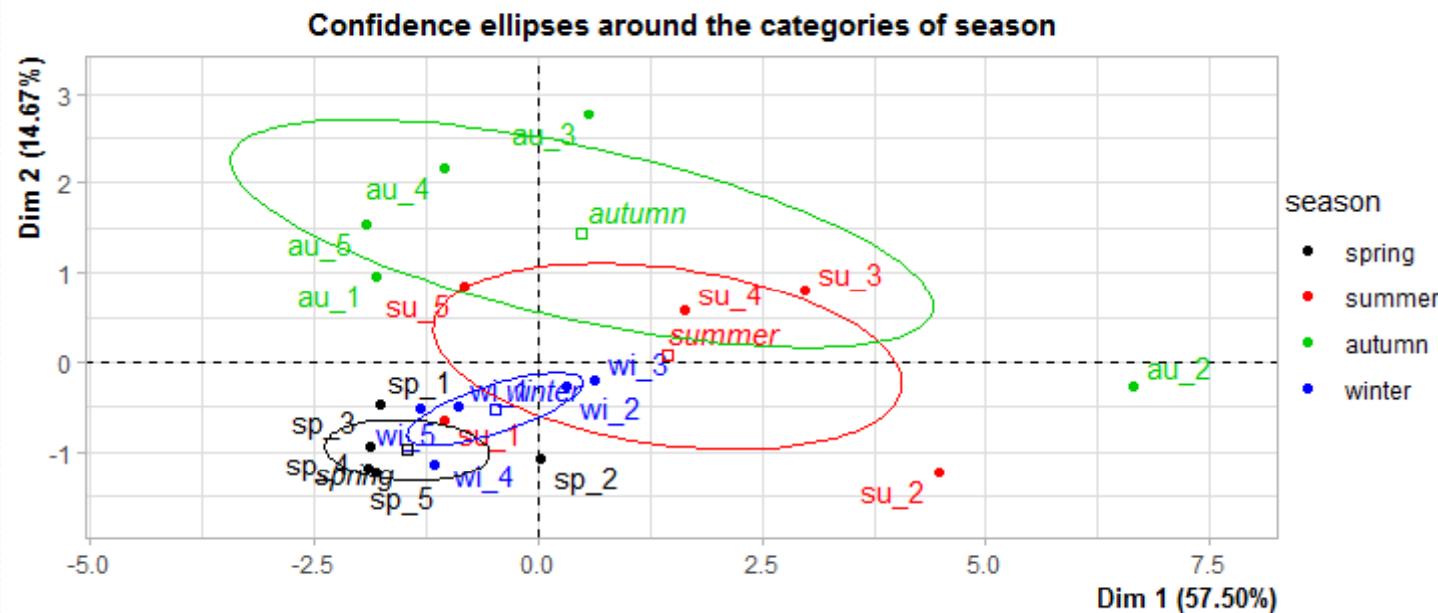
ACP normée avec FactoMineR sur Méaudret environnement

```
# Nuage des individus avec des couleurs différentes selon la saison (colonne 10)
plot.PCA(res.meo, choix='ind', invisible=c('quali', 'ind.sup'),
          title="Graphe des individus de l'ACP", label = c('ind'), habillage=10)
```



ACP normée avec FactoMineR sur Méaudret environnement

```
# Tracer des ellipses correspondant aux saisons  
plotellipses(res.meo, 10, level=0.95)
```



ACP avec données manquantes avec FactoMineR::missMDA

- La librairie **missMDA** permet d'estimer les valeurs manquantes dans un tableau de données pour réaliser une ACP, en utilisant un algorithme itératif :

```
require(missMDA)
data(orange)
summary(orange) # valeurs manquantes

res.pca <- PCA(orange) # Les valeurs manquantes sont remplacées par la
# moyenne de la variable
plot(res.pca, choix = "ind")

# Estimation des valeurs manquantes avec la fonction imputePCA
nb <- estim_ncpPCA(orange, ncp.min=0, ncp.max=5, method=cv="Kfold", nbsim=50)
imputed <- imputePCA(orange, ncp=nb$ncp)
imputed

# Refaire l'ACP sur les données complètes
res.pca <- PCA(imputed$completeObs)
plot(res.pca, choix = "ind")
```

Les bonus de FactoMineR: l'utilitaire FactoInvestigate et l'interface Factoshiny

- L'utilitaire **FactoInvestigate** rédige automatiquement en format PDF, Word ou html, la description (en français ou en anglais) d'une analyse factorielle de type ACP, AFC ou ACM réalisée par FactoMineR

```
require(FactoInvestigate)  
Investigate(res.meo)  
?Investigate
```

- **Factoshiny** est une interface graphique pour FactoMineR qui permet de réaliser via une application à menus les analyses factorielles classiques. Shiny est un outil de RStudio qui permet de réaliser une application WEB interactive utilisant R

```
require(Factoshiny)  
res <- Factoshiny(meo)  
# NB : Utiliser le bouton sous forme de panneau STOP de la console pour re  
prendre la main dans Rstudio
```

Cf présentation de François Husson le 1/04/2019 pour le réseau MATE-SHS du CNRS:

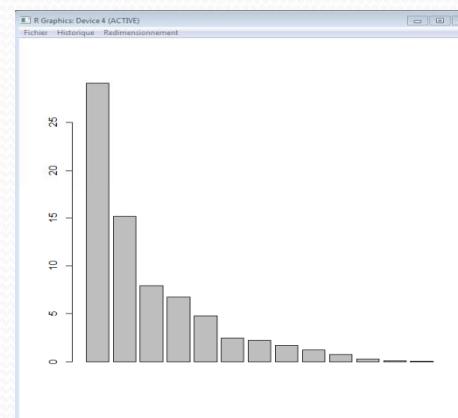
<https://mate-shs.cnrs.fr/actions/tutomate/tuto21-husson-factoshiny/>

Etude 2 : ACP centrée sur le tableau

Méaudret faune avec ade4

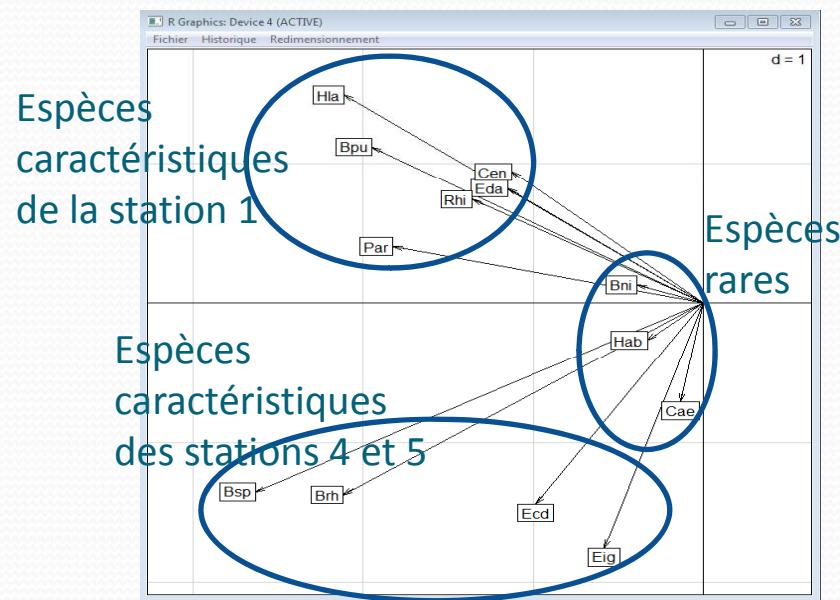
	Eda	Bsp	Brh	Bni	Bpu	Cen	Ecd	Rhi	Hla	Hab	Par	Cae	Eig
sp_1	4	7	10	9	0	0	0	5	9	0	4	0	0
sp_2	0	0	8	0	0	0	0	0	4	0	0	0	0
sp_3	0	5	5	0	0	0	0	2	5	0	0	0	0
sp_4	0	3	6	0	0	0	0	3	6	0	0	0	0
sp_5	0	5	6	0	0	0	5	0	4	0	0	0	4
su_1	6	7	10	0	10	0	0	2	7	0	0	0	2
su_2	0	0	9	0	0	0	0	0	0	0	0	0	0
su_3	0	6	8	0	0	2	0	0	0	0	0	0	0
su_4	0	7	11	0	0	2	0	0	2	0	0	5	5
su_5	0	6	9	2	3	0	4	0	0	0	0	2	7
au_1	4	5	8	0	9	6	0	5	9	0	7	0	0
au_2	0	0	1	0	0	0	0	0	0	0	0	0	0
au_3	0	9	10	0	0	0	0	0	4	0	3	0	0
au_4	0	10	13	0	0	3	0	5	5	1	4	2	4
au_5	2	10	12	0	4	0	8	4	4	2	5	1	6
wi_1	3	6	7	0	6	7	0	4	8	0	4	0	0
wi_2	0	3	6	0	0	5	0	4	3	0	1	0	0
wi_3	0	0	3	0	0	1	0	1	0	0	0	0	0
wi_4	0	6	10	0	0	5	1	3	5	0	2	0	0
wi_5	1	9	11	0	3	6	8	3	5	2	5	0	0

```
fau <- meaudret$spe  
dim(fau)  
[1] 20 13  
pca2 <- dudi.pca(fau , center=T, scale=F)  
Select the number of axes: 2
```

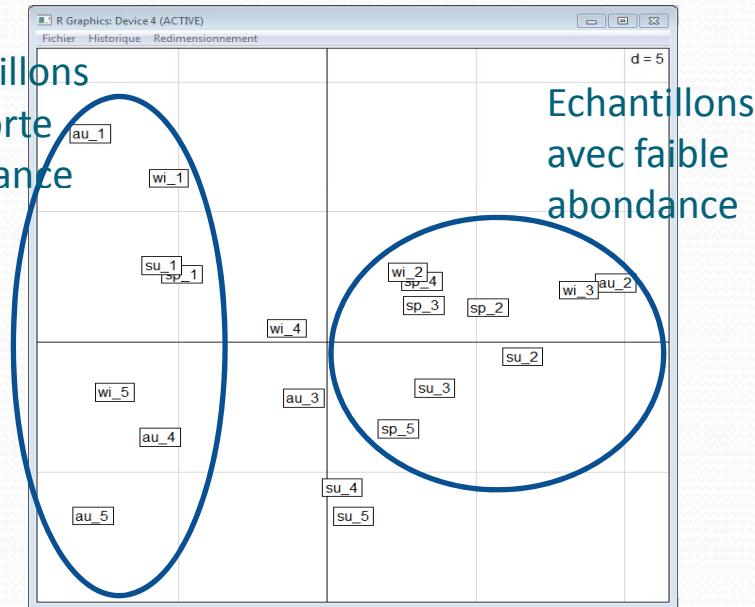


ACP centrée : interprétation

`s.arrow(pca2$co)`

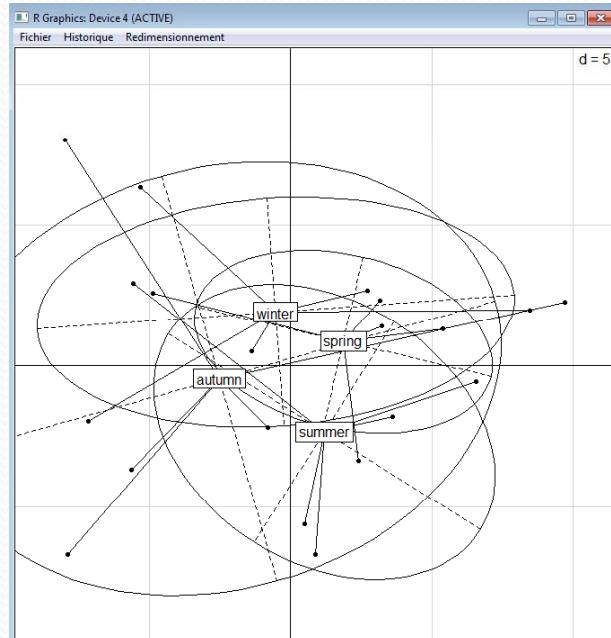


`s.label(pca2$li)`



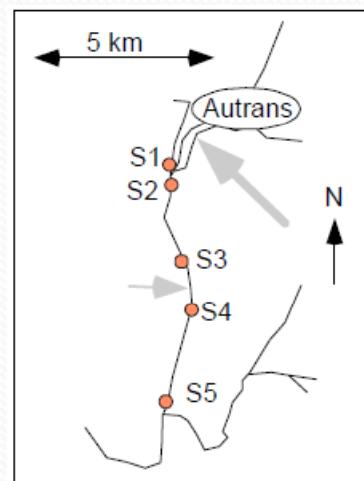
- Les variables (espèces) ne s'inscrivent pas dans un cercle de rayon 1 car centrage uniquement des données => utilisation de `s.label` ou de `s.arrow()` au lieu de `s.corcircle()`
- Espèces rares proches de l'origine des axes (Bni, Hab, Cae) car elles contribuent peu.
- Axe 1 : toutes les espèces sont du même côté : effet « taille » opposant les échantillons avec forte abondance globale (à gauche) aux échantillons avec faible abondance (à droite).
- Axe 2 : distingue 2 groupes parmi les espèces les plus abondantes (flèches longues)

ACP centrée : aide des variables illustratives pour l'interprétation

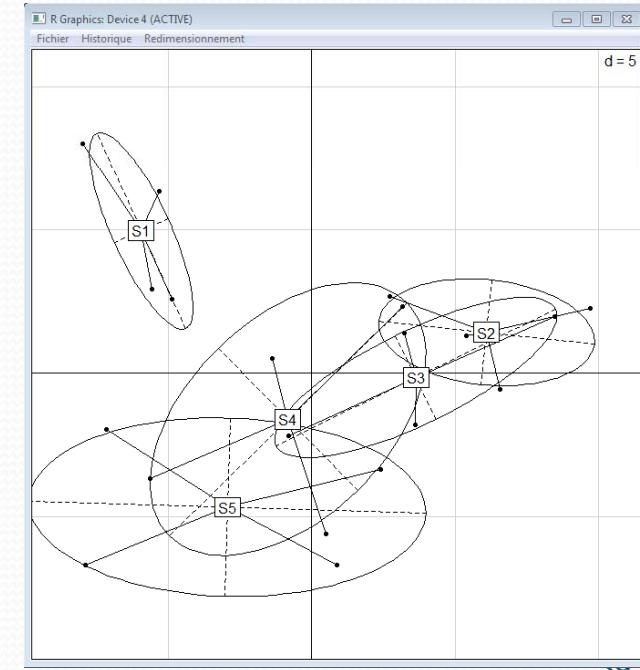


`s.class(pca2$li, plan$season)`

- La communauté suit un gradient amont-aval dans les sites 2 à 5 (selon diagonale des axes 1 et 2).
- Le site 1 a un peuplement particulier et très stable.
- Forte variabilité saisonnière surtout pour S4 et S5



`s.class(pca2$li, plan$site)`



Plan

- **1. Introduction**
 - Contexte et objectifs de l'analyse multivariée
 - Ordination ou classification ?
 - Ordination basée sur les distances ou sur les valeurs propres (analyses factorielles)
 - Les packages R spécialisés : ade4 et FactoMineR
- **2. Analyses factorielles à un tableau de données, avec les packages ade4 et FactoMineR**
 - 2.1 Analyse en Composantes Principales (ACP) : plusieurs variables quantitatives
 - **2.2 Analyse Factorielle des Correspondances (AFC) : deux variables qualitatives**
 - 2.3 Analyse des Correspondances Multiples (ACM) : plusieurs variables qualitatives
- **3. Classification Automatique**
 - 3.1 Classification Ascendante Hiérarchique (CAH) avec la fonction hclust
 - 3.2 Partitionnement avec la fonction kmeans
- **4. Méthodes de couplage de tableaux avec le package ade4**
 - 4.1 Analyses Inter et Intra-classes
 - 4.2 Analyse Discriminante
 - 4.3 Analyses sur variables Instrumentales (ACPVI ou RDA, AFCVI ou CCA)
 - 4.4 Analyse de Co-inertie
- **5. Aperçu des méthodes d'analyse multi-tableaux**
 - 5.1 Analyse Triadique Partielle
 - 5.2 STATIS
 - 5.3 Analyse Factorielle Multiple

2.2 Analyse factorielle des correspondances simples (AFC)

- Application : tableau croisant deux variables **qualitatives** (table de contingence). Dans chaque case (i,j) on a le nombre d'observations appartenant à la fois à la modalité i de la variable ligne et à la modalité j de la variable colonne.
- Ce type de tableau fait habituellement l'objet d'un **test de chi-2** pour tester l'indépendance entre les deux variables.
- L'AFC est une méthode permettant de définir pour un tableau croisé un **scoring** sur les colonnes tel que les scores moyens des lignes (obtenus en utilisant les fréquences du tableau des profils) soient les plus discriminants possible, au sens de la variance de ces scores moyens.
- **p-1 axes** sont calculés, où p est la plus petite dimension du tableau
- L'AFC est **symétrique** : contrairement à l'ACP, si on transpose le tableau on obtient les mêmes résultats

La fonction **dudi.coa()** d'ade4

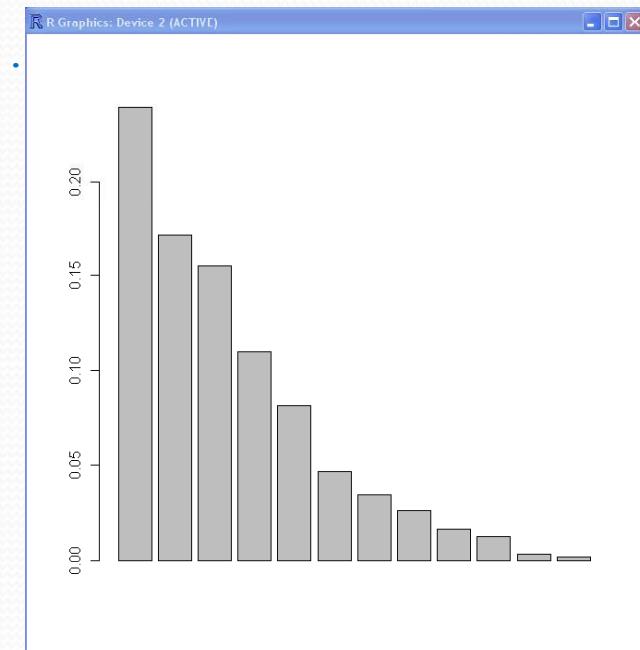
dudi.coa(df, scannf = TRUE, nf = 2)

df	un “data frame” à n lignes (unités statistiques) et p colonnes (variables numériques)
scannf	une valeur logique (TRUE/FALSE) indiquant si le diagramme des valeurs propres doit être affiché
nf	si scannf =FALSE, un entier indiquant le nombre d’axes à conserver

Cette fonction **dudi.coa** réalise une AFC simple. Elle n'a pas d'autre option que le nom du tableau à analyser (df), et les options classiques permettant d'afficher ou non les valeurs propres pour choisir le nombre d'axes à conserver.

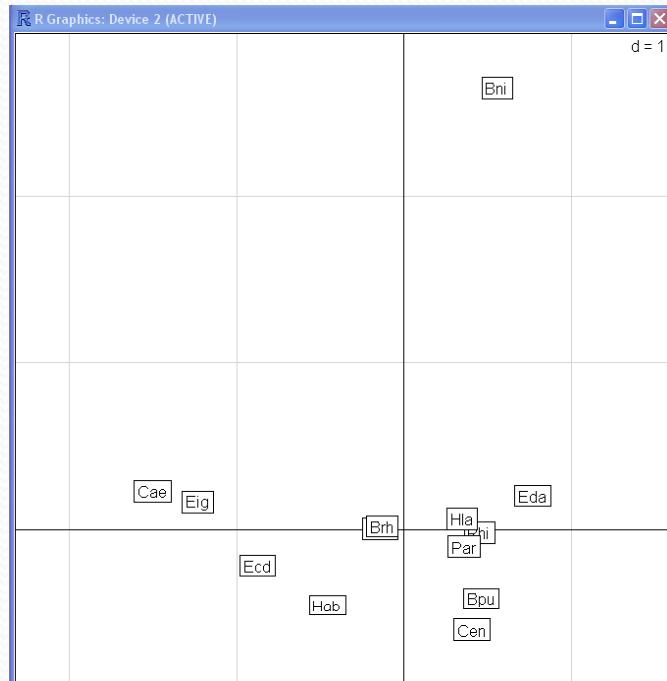
Etude 3 : AFC simple de Méaudret faune

```
coal <- dudi.coa(fau) # Même tableau que pour l'ACP centrée  
Select the number of axes: 5  
Duality diagramm  
class: coa dudi  
$call: dudi.coa(df = fau)  
  
$nf: 5 axis-components saved  
$rank: 12  
eigen values: 0.239 0.1712 0.1547 0.1102 0.0814 ...  
  vector length mode content  
1 $cw     13    numeric column weights  
2 $lw     20    numeric row weights  
3 $eig    12    numeric eigen values  
  
  data.frame nrow ncol content  
1 $stab      20   13  modified array  
2 $li        20    5  row coordinates  
3 $l1        20    5  row normed scores  
4 $co        13    5  column coordinates  
5 $c1        13    5  column normed scores  
other elements: N
```

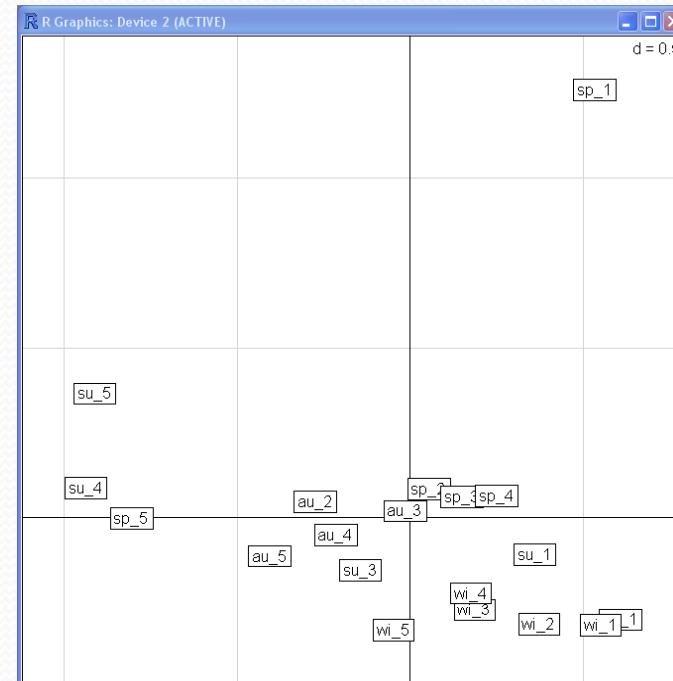


AFC simple de Méaudret faune

`s.label(coal$co)`



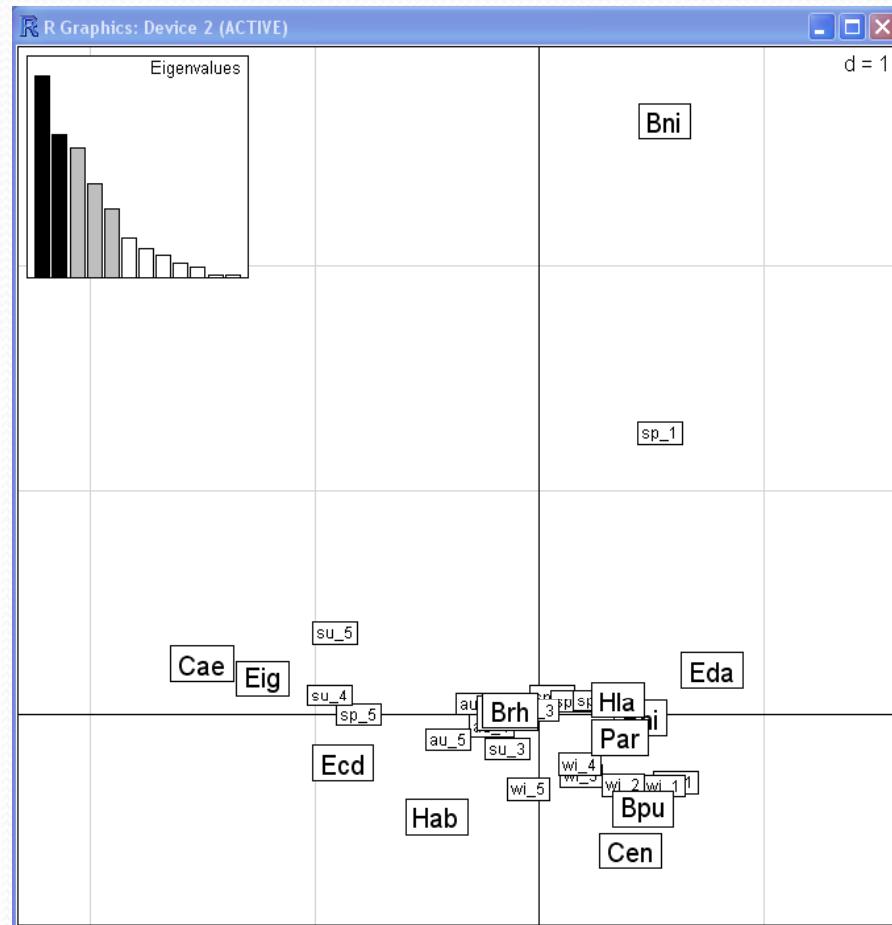
`s.label(coal$li)`



AFC, représentation optimale : le biplot

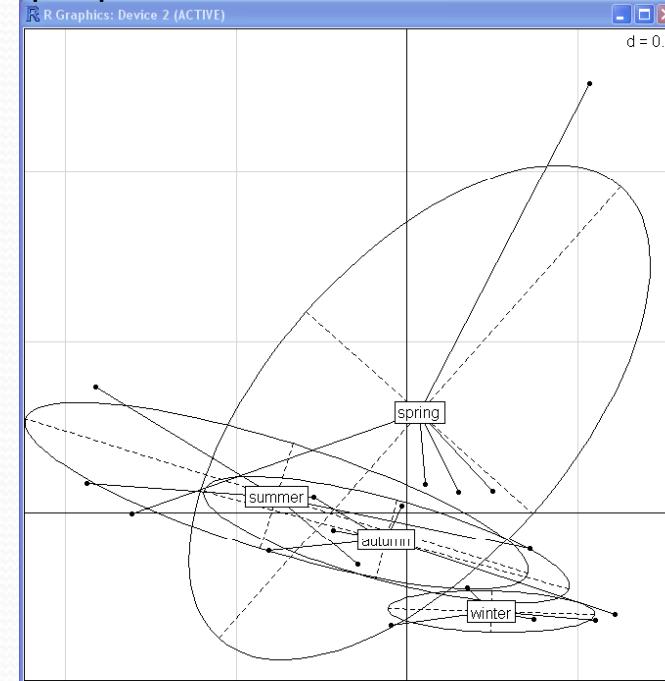
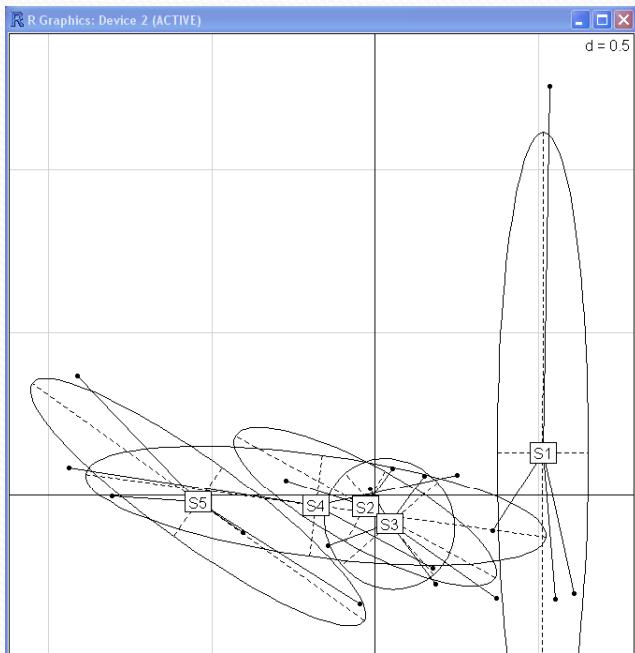
Avec ade4, quelle que soit l'analyse de type dudi, la fonction scatter() permet d'obtenir la représentation optimale : ici **scatter(coa1)** donne un biplot = représentation simultanée des lignes et des colonnes sur le même graphe, avec le diagramme des valeurs propres dans un coin.

	Eda	Bsp	Brh	Bni	Bpu	Cen	Ecd	Rhi	Hla	Hab	Par	Cae	Eig
sp_1	4	7	10	9	0	0	0	5	9	0	4	0	0
sp_2	0	0	8	0	0	0	0	0	4	0	0	0	0
sp_3	0	5	5	0	0	0	0	2	5	0	0	0	0
sp_4	0	3	6	0	0	0	0	3	6	0	0	0	0
sp_5	0	5	6	0	0	0	5	0	4	0	0	0	4
su_1	6	7	10	0	10	0	0	2	7	0	0	0	2
su_2	0	0	9	0	0	0	0	0	0	0	0	0	0
su_3	0	6	8	0	0	2	0	0	0	0	0	0	0
su_4	0	7	11	0	0	2	0	0	2	0	0	5	5
su_5	0	6	9	2	3	0	4	0	0	0	0	2	7
au_1	4	5	8	0	9	6	0	5	9	0	7	0	0
au_2	0	0	1	0	0	0	0	0	0	0	0	0	0
au_3	0	9	10	0	0	0	0	0	4	0	3	0	0
au_4	0	10	13	0	0	3	0	5	5	1	4	2	4
au_5	2	10	12	0	4	0	8	4	4	2	5	1	6
wi_1	3	6	7	0	6	7	0	4	8	0	4	0	0
wi_2	0	3	6	0	0	5	0	4	3	0	1	0	0
wi_3	0	0	3	0	0	1	0	1	0	0	0	0	0
wi_4	0	6	10	0	0	5	1	3	5	0	2	0	0
wi_5	1	9	11	0	3	6	8	3	5	2	5	0	0



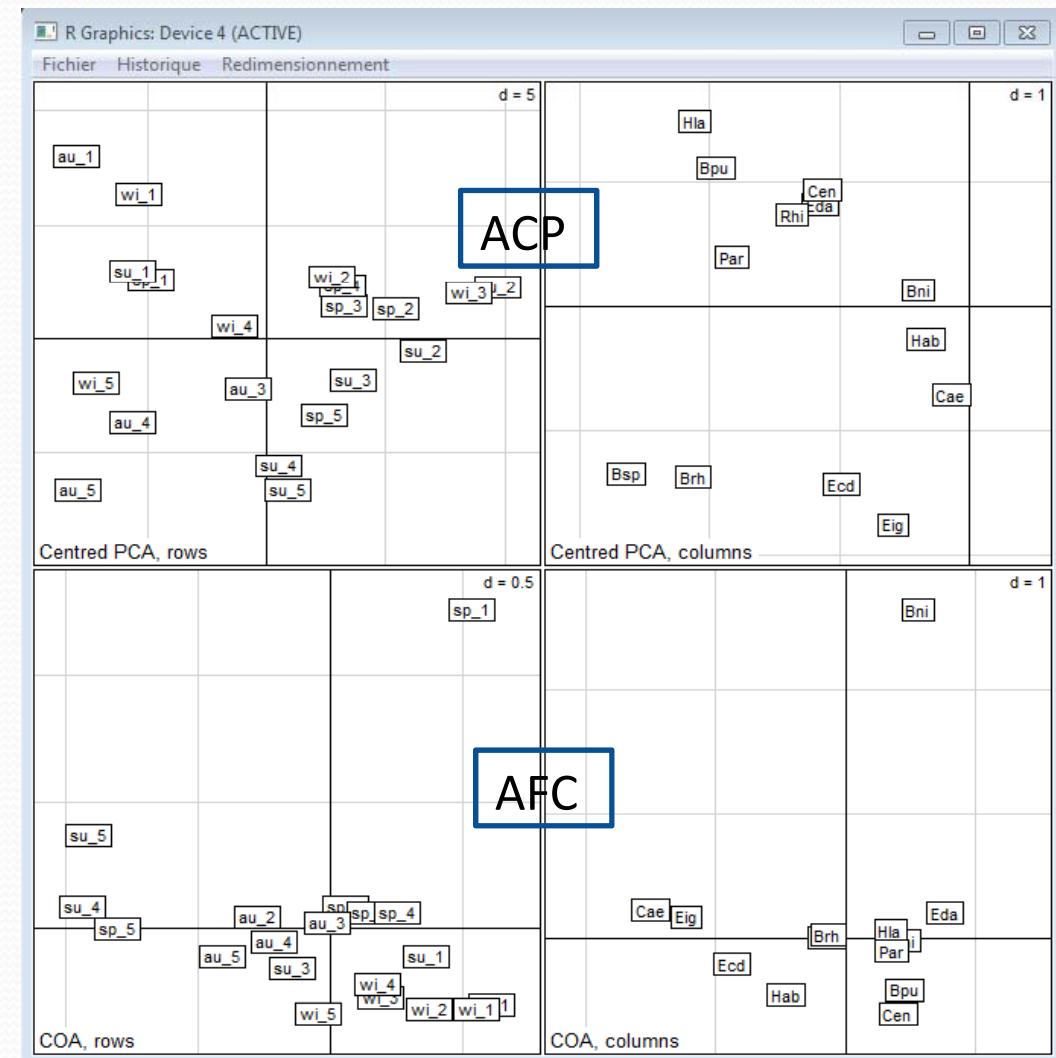
AFC : aide des variables illustratives pour l'interprétation

- La communauté d'éphéméroptères suit un gradient amont-aval (selon axe1) et un gradient saisonnier (selon axe2).
- Analyse perturbée par l'abondance exceptionnelle d'une espèce rare dans un échantillon (Bni pour sp_1)
- Pas de trace des épisodes de pollution qui agissent de manière globale sur l'abondance sans modifier la composition des peuplements



AFC vs. ACP centrée sur méaudret faune

	Eda	Bsp	Brh	Bni	Bpu	Cen	Ecd	Rhi	Hla	Hab	Par	Cae	Eig
sp_1	4	7	10	9	0	0	0	0	5	9	0	4	0
sp_2	0	0	8	0	0	0	0	0	4	0	0	0	0
sp_3	0	5	5	0	0	0	0	2	5	0	0	0	0
sp_4	0	3	6	0	0	0	0	3	6	0	0	0	0
sp_5	0	5	6	0	0	0	0	5	0	4	0	0	4
su_1	6	7	10	0	10	0	0	2	7	0	0	0	2
su_2	0	0	9	0	0	0	0	0	0	0	0	0	0
su_3	0	6	8	0	0	2	0	0	0	0	0	0	0
su_4	0	7	11	0	0	2	0	0	2	0	0	5	5
su_5	0	6	9	2	3	0	4	0	0	0	0	2	7
au_1	4	5	8	0	9	6	0	5	9	0	7	0	0
au_2	0	0	1	0	0	0	0	0	0	0	0	0	0
au_3	0	9	10	0	0	0	0	0	4	0	3	0	0
au_4	0	10	13	0	0	3	0	5	5	1	4	2	4
au_5	2	10	12	0	4	0	8	4	4	2	5	1	6
wi_1	3	6	7	0	6	7	0	4	8	0	4	0	0
wi_2	0	3	6	0	0	5	0	4	3	0	1	0	0
wi_3	0	0	3	0	0	1	0	1	0	0	0	0	0
wi_4	0	6	10	0	0	5	1	3	5	0	2	0	0
wi_5	1	9	11	0	3	6	8	3	5	2	5	0	0



Travaux pratiques : Le jeu de données

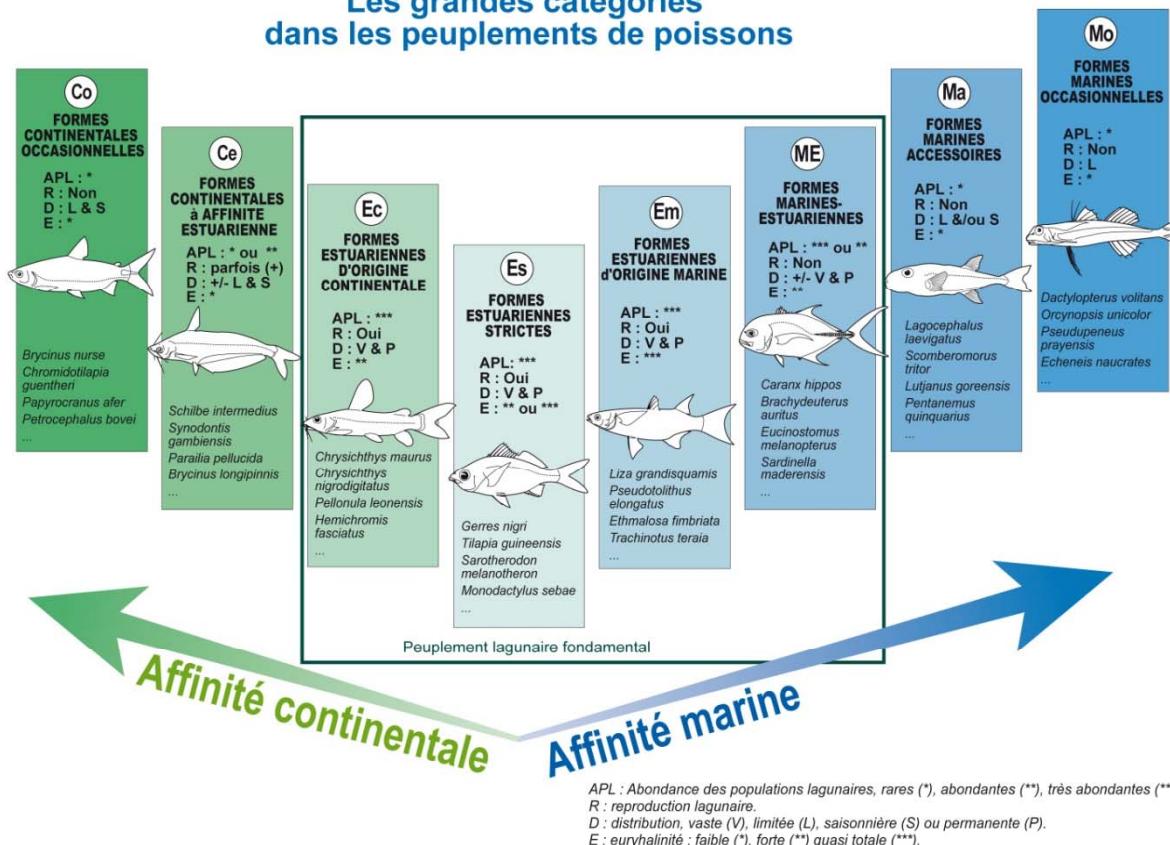
The left screenshot shows the PPEAO homepage with a banner about fish populations and artisanal fishing in West African estuarine, lacustrine, or continental ecosystems. It features several small images of water bodies and boats.

The right screenshot shows a 'consulter' (consult) page for experimental fishing data extraction. It includes a sidebar for selecting sectors (Peuplement - Environnement - NE/Pt - Biologie - Trophique), a list of variables for export (Station_id, Station, Site, Distance_embouchure, Debris_id, Position_station_id, Sediment_id, Vegetation_id, Station_latitude, Station_longitude, Station_memo, Debris, Position_station), and a summary section indicating 180 fishing coups.

- Données extraites de ppeao.ird.fr
- Pêches expérimentales dans 4 systèmes estuariens d'Afrique de l'Ouest (Ebrié, Fatala, Gambie, Saloum), à 2 saisons (sèche / humide) => **256 coups de pêche**
- **Envir** = 14 variables : mesure de 5 variables environnementales lors des 256 coups de pêche (Salinité, Température, Transparence, Profondeur, Distance à l'embouchure) + divers identifiants (saison, système...)
- **Faune** = abondance de 111 espèces de poissons pour les 256 coups de pêche
- **Categ** = catégories écologiques et trophiques des 111 espèces

Catégories écologiques

Les grandes catégories dans les peuplements de poissons



Exercice 2 : Analyse du tableau Faune

- Importer le fichier **Faune.csv** sous R => dataframe **Faune**
- Importer le tableau **Categ.csv** sous R => dataframe **Categ**
- Faire l'AFC de **Faune**
 - Décrire les valeurs propres (combien en garder, quel % ?)
 - Représenter le nuage des espèces regroupées par catégories écologiques et trophiques en utilisant **Categ**
 - Représenter le nuage des individus regroupés par saison, par système et par système-saison (dans **Envir**)
 - Quelles conclusions en tirez-vous ?

Exercice 2 : Analyse du tableau Faune - suite

- Transformer le data frame **Faune** en $\log(x+1)$ => **Faulog**
- Faire l'AFC de **Faulog**
 - Décrire les valeurs propres (combien en garder, quel % ?)
 - Représenter le nuage des espèces regroupées par catégories écologiques et trophiques
 - Représenter le nuage des individus regroupés par saison, par système et par système-saison
- Faire l'ACP centrée de **Faulog** avec le même protocole de description
- Comparer AFC et ACP centrée sur ce jeu de données faunistique

Plan

- **1. Introduction**
 - Contexte et objectifs de l'analyse multivariée
 - Ordination ou classification ?
 - Ordination basée sur les distances ou sur les valeurs propres (analyses factorielles)
 - Les packages R spécialisés : ade4 et FactoMineR
- **2. Analyses factorielles à un tableau de données, avec les packages ade4 et FactoMineR**
 - 2.1 Analyse en Composantes Principales (ACP) : plusieurs variables quantitatives
 - 2.2 Analyse Factorielle des Correspondances (AFC) : deux variables qualitatives
 - **2.3 Analyse des Correspondances Multiples (ACM) : plusieurs variables qualitatives**
- **3. Classification Automatique**
 - 3.1 Classification Ascendante Hiérarchique (CAH) avec la fonction hclust
 - 3.2 Partitionnement avec la fonction kmeans
- **4. Méthodes de couplage de tableaux avec le package ade4**
 - 4.1 Analyses Inter et Intra-classes
 - 4.2 Analyse Discriminante
 - 4.3 Analyses sur variables Instrumentales (ACPVI ou RDA, AFCVI ou CCA)
 - 4.4 Analyse de Co-inertie
- **5. Aperçu des méthodes d'analyse multi-tableaux**
 - 5.1 Analyse Triadique Partielle
 - 5.2 STATIS
 - 5.3 Analyse Factorielle Multiple

2.3 Analyse factorielle des correspondances multiples (ACM)

- L'Analyse des Correspondances Multiples (ACM) est la généralisation de l'Analyse des Correspondances Simples à plus de deux variables qualitatives.
- Cette méthode permet d'analyser un tableau rectangulaire où les colonnes sont des **variables qualitatives** et les lignes des observations ou des individus (ex : réponses à un questionnaire à choix multiples).
- La table est automatiquement transformée en **tableau disjonctif complet**. Un tableau disjonctif complet a autant de colonnes que le nombre total de modalités de l'ensemble des variables, chaque colonne/modalité valant 1 ou 0 selon la réalisation ou non de cette modalité pour l'individu/l'observation correspondant.
- Le **nombre d'axes** est obtenu en sommant pour chaque variable le nombre de modalités moins 1.
- La fonction **dudi.acm()** de la librairie **ade4** et la fonction **MCA()** de la librairie **FactoMineR** permettent de traiter ce type de données. Les variables qualitatives doivent être codées comme des variables de type **factor**.

La fonction `dudi.acm()` d'ade4

`dudi.acm(df, row.w = rep(1,nrow(df)), scannf = TRUE, nf = 2)`

<code>df</code>	un “data frame” à n lignes (unités statistiques) et p colonnes (variables qualitatives)
<code>row.w</code>	optionnel : un vecteur de longueur égale au nombre de lignes de df permettant de fournir des pondérations particulières pour les individus . Par défaut, chaque ligne a un poids uniforme.
<code>scannf</code>	une valeur logique (TRUE/FALSE) indiquant si le diagramme des valeurs propres doit être affiché
<code>nf</code>	si scannf =FALSE, un entier indiquant le nombre d’axes à conserver

Cette fonction réalise une **Analyse factorielle des Correspondances Multiples (ACM)**. En plus du nom du tableau à analyser (`df`) et des options classiques permettant d'afficher ou non les valeurs propres pour choisir le nombre d'axes à conserver, elle permet d'indiquer si besoin le nom d'un vecteur de poids des lignes (par exemple si chaque ligne représente un groupe d'individus).

ACM sur données ours (ade4)

Données d'une étude de Georges Erome (1989)

La réintroduction de l'ours dans les Alpes est-elle possible ?

38 zones de l'Inventaire National Forestier (en lignes) et 10 variables qualitatives (en colonnes). Les 8 premières sont codées de 1 à 3 (1=défavorable à l'implantation de l'ours, 2 moyennement favorable, 3 très favorable).

```
> str(ours)
'data.frame': 38 obs. of 10 variables:
 $ altit : Factor w/ 3 levels "1","2","3": 2 1 3 3 3 3 2 1 2 2 ...
 $ deniv : Factor w/ 3 levels "1","2","3": 3 2 3 3 3 3 2 1 3 2 ...
 $ cloiso: Factor w/ 3 levels "1","2","3": 3 1 3 3 1 3 3 2 1 3 ...
 $ domain: Factor w/ 3 levels "1","2","3": 2 2 2 1 2 1 2 2 2 1 ...
 $ boise : Factor w/ 3 levels "1","2","3": 2 1 2 3 2 3 2 1 3 3 ...
 $ hetra : Factor w/ 3 levels "1","2","3": 3 1 2 3 3 3 1 1 2 3 ...
 $ favor : Factor w/ 3 levels "1","2","3": 3 2 3 3 2 3 2 3 3 2 ...
 $ inexp : Factor w/ 3 levels "1","2","3": 2 2 3 2 3 3 3 2 3 3 ...
 $ citat : Factor w/ 4 levels "1","2","3","4": 1 2 2 3 1 3 1 2 4 1 ...
 $ depart: Factor w/ 7 levels "AHP","AM","D",...: 5 5 5 5 7 7 7 7 7 7 ...
> summary(ours)
altit  deniv  cloiso  domain  boise  hetra  favor  inexp  citat  depart 
1: 8   1:13   1:12   1: 9   1:10   1:19   1:15   1:20   1:22   AHP:5 
2:17   2:14   2: 4   2:13   2:15   2: 5   2:12   2:10   2: 7   AM :4  
3:13   3:11   3:22   3:16   3:13   3:14   3:11   3: 8   3: 4   D  :5  
                           4: 5   HP :8 
                           HS :4 
                           I  :5 
                           S  :7 
```

>



altit La variable donne, pour le district, l'importance de la tranche altitudinale habitée par l'ours (800-2000 m) sous forme d'un facteur à 3 modalités :

1. moins 50% de la surface se situe entre 800 et 2000 m
2. entre 50 et 70% de la surface se situe entre 800 et 2000 m
3. plus de 70% de la surface se situe entre 800 et 2000 m

deniv La variable donne, pour le district, l'importance de la dénivellation moyenne par carrés de 50 km² sous forme d'un facteur à 3 modalités :

1. moins de 700 m
2. entre 700 et 900 m
3. plus de 900 m

cloiso La variable décrit le cloisonnement du massif sous forme d'un facteur à 3 modalités :

1. une grande vallée ou une crête isole au moins un quart du massif
2. une grande vallée ou une crête isole moins d'un quart du massif
3. le massif ne présente pas de coupure

domain La variable donne l'importance du domaine forestier en contact avec le massif sous forme d'un facteur à 3 modalités :

1. moins de 400 km²
2. entre 400 et 1000 km²
3. plus de 1000 km²

boise La variable donne le taux (pourcentage de surface du district) de boisement du district sous forme d'un facteur à 3 modalités :

1. moins de 30%
2. entre 30 et 50%
3. plus de 50%

hetra La variable indique l'importance (pourcentage de surface du district) des hêtraies et forêts mixtes dans le district sous forme d'un facteur à 3 modalités :

1. moins de 5%
2. entre 5 et 10%
3. plus de 10%

favor La variable renseigne l'importance (pourcentage de surface du district) des forêts favorables, hêtraies, forêts mixtes, sapinières et pessières sous forme d'un facteur à 3 modalités :

1. moins de 5%
2. entre 5 et 10%
3. plus de 10% du massif

inexp La variable renseigne l'importance (pourcentage de surface du district) des forêts inexploitées sous forme d'un facteur à 3 modalités :

1. moins de 4%
2. entre 4 et 8%
3. plus de 8% du massif

Ces variables ont la même logique et peuvent être considérées comme des facteurs ou comme des indices quantitatifs, les notes 1,2 et 3 codant dans l'ordre une situation de plus en plus favorable à l'Ours brun (grands espaces forestiers, connexes, d'accès difficile).

Le tableau contient en outre deux facteurs d'une autre signification :

citat La variable donne l'information sur la date de disparition de l'Ours sous la forme d'un facteur à 4 modalités :

1. aucune citation de l'espèce depuis 1840
2. 1 à 3 citations avant 1900 et aucune après
3. 4 citations et plus avant 1900 et aucune après
4. 1 citation ou plus entre 1900 et 1940

depart La variable indique le département qui contient le district sous la forme d'un facteur à 7 modalités :

AHP Alpes-de-Haute-Provence

AM Alpes-Maritimes

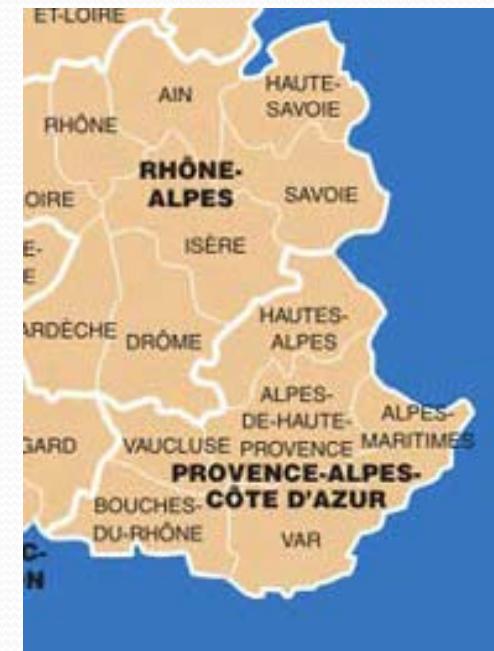
D Drôme

HP Hautes-Alpes

HS Haute-Savoie

I Isère

S Savoie



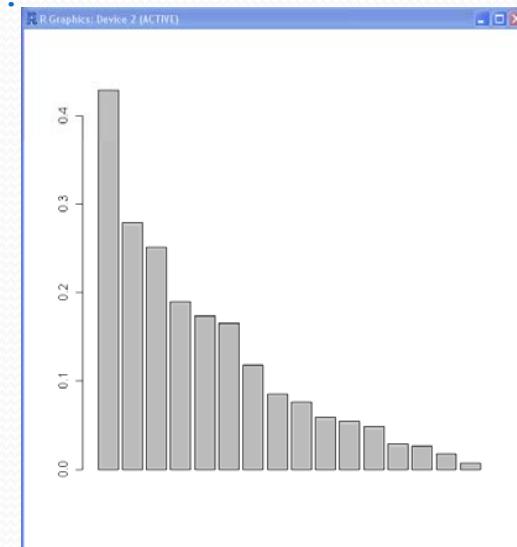
ACM sur données ours

```
acm <- dudi.acm(ours[, 1:8])
acm
Duality diagramm
class: acm dudi
$call: dudi.acm(df = ours[, 1:8], scannf = FALSE)

$nf: 2 axis-components saved
$rank: 16 ←
eigen values: 0.4283 0.2783 0.2501 0.1888 0.1733 ...
  vector length mode    content
1 $cw     24   numeric column weights
2 $lw     38   numeric row weights
3 $eig    16   numeric eigen values

  data.frame nrow ncol content
1 $stab      38   24 modified array
2 $li        38    2  row coordinates
3 $l1        38    2  row normed scores
4 $co        24    2  column coordinates
5 $c1        24    2  column normed scores
other elements: cr
```

16 valeurs propres car 8 variables à chacune 3 modalités = $8 \times (3-1) = 16$



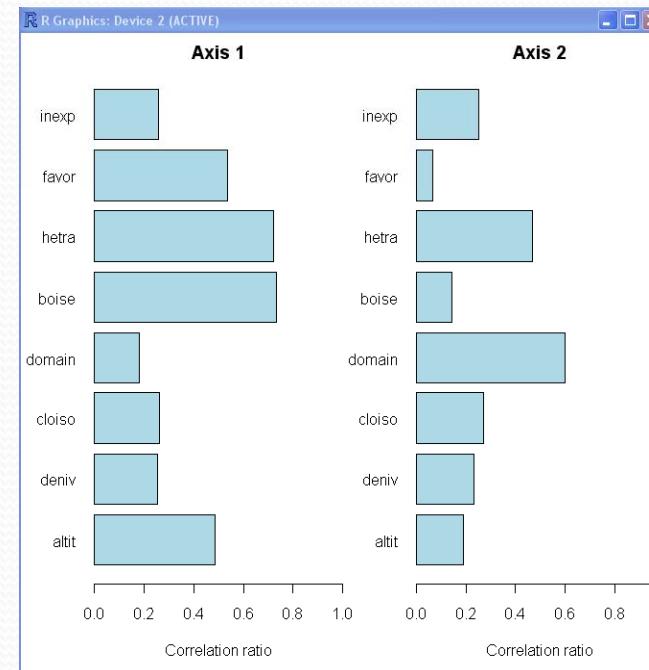
ACM sur données ours

`acm$cr` : rapports de corrélation entre les variables et les axes

`acm$cr`

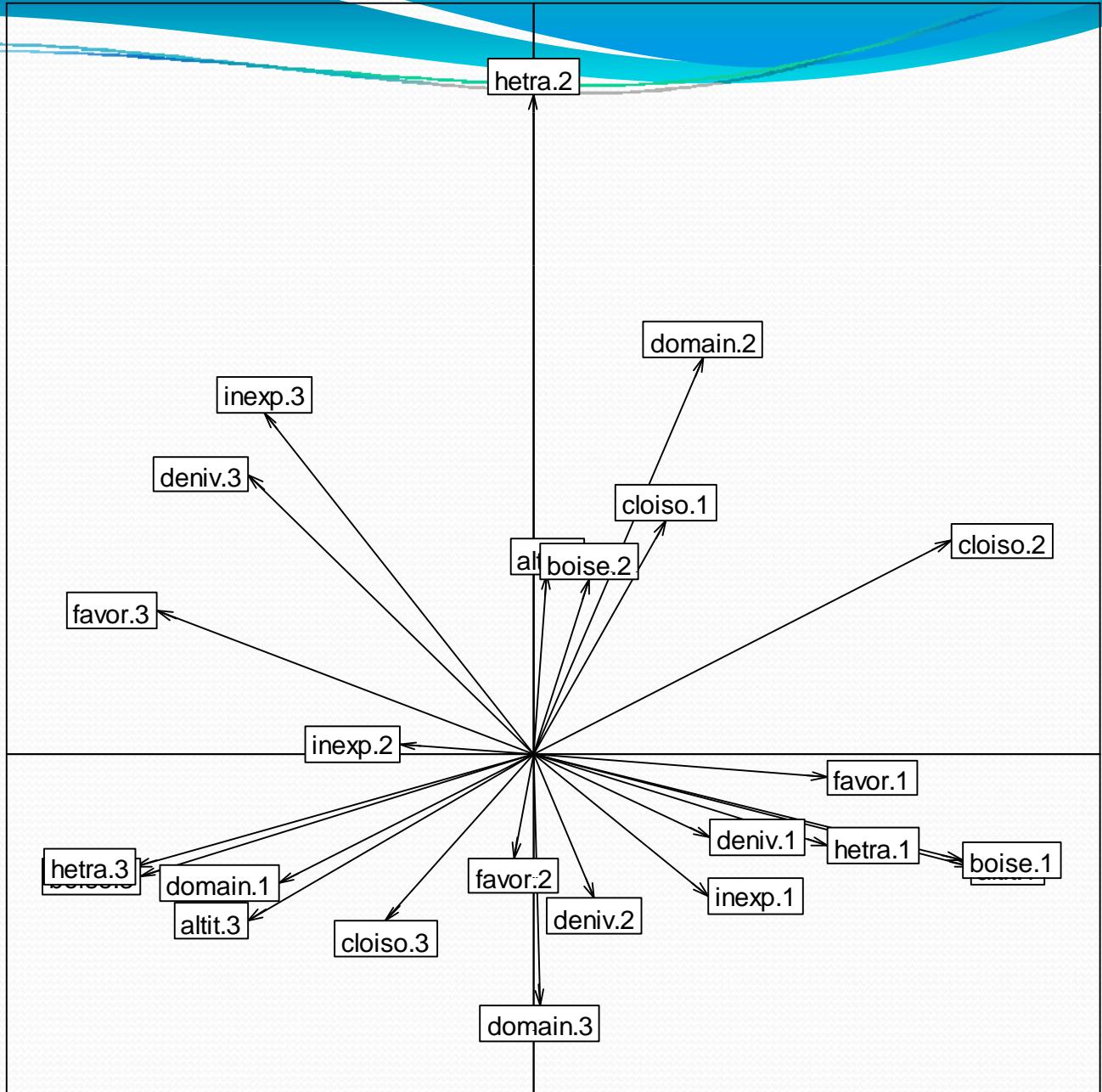
	RS1	RS2
altit	0.4867906	0.1893131
deniv	0.2528392	0.2325247
cloiso	0.2598395	0.2724303
domain	0.1796741	0.6002266
boise	0.7317972	0.1429825
hetra	0.7233694	0.4699515
favor	0.5352790	0.0682467
inexp	0.2568897	0.2505504

- Axe 1 : boise, hetra, favor et altit
- Axe 2: domain et hetra



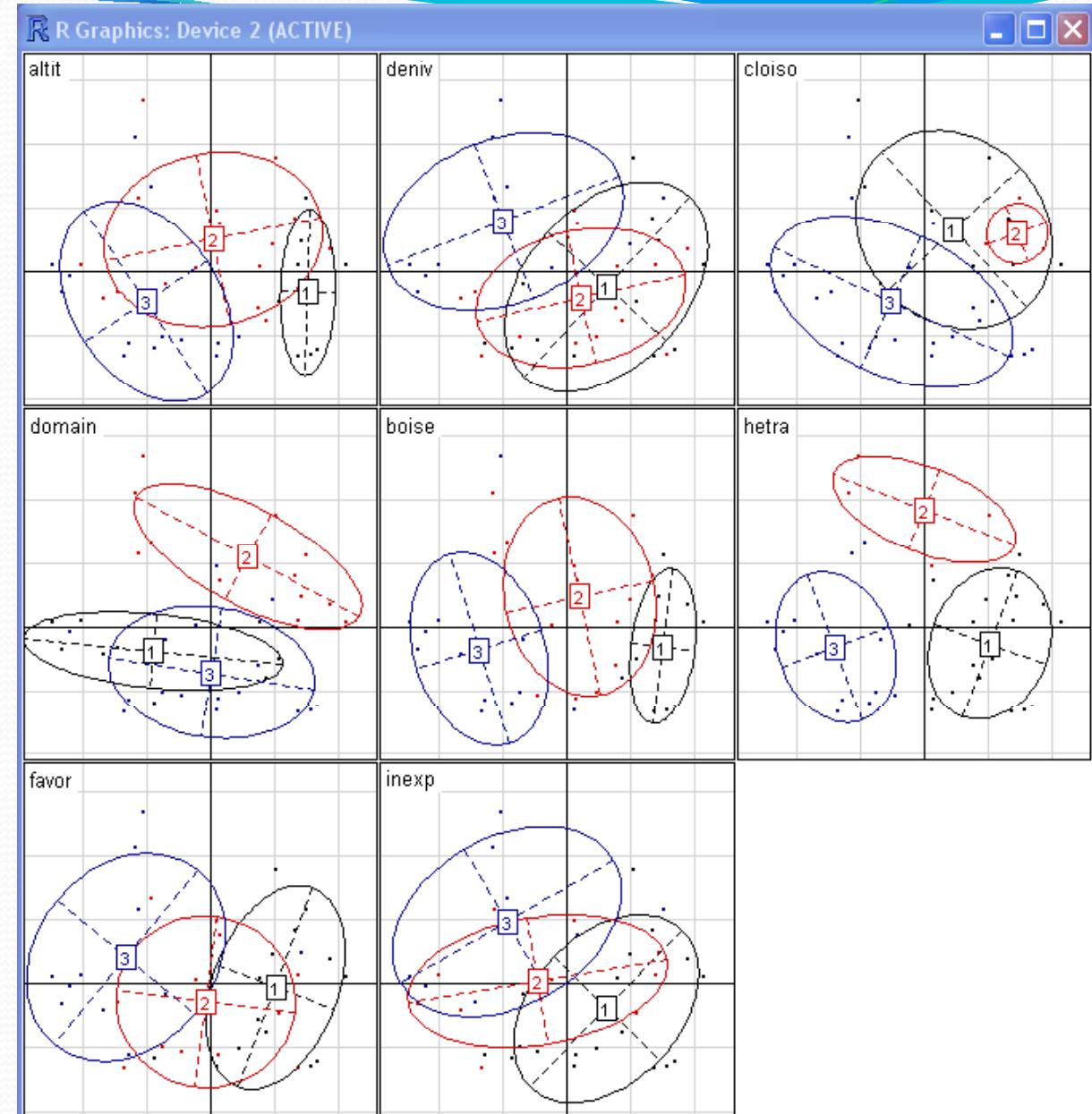
ACM ours

s.arrow(acm\$co)

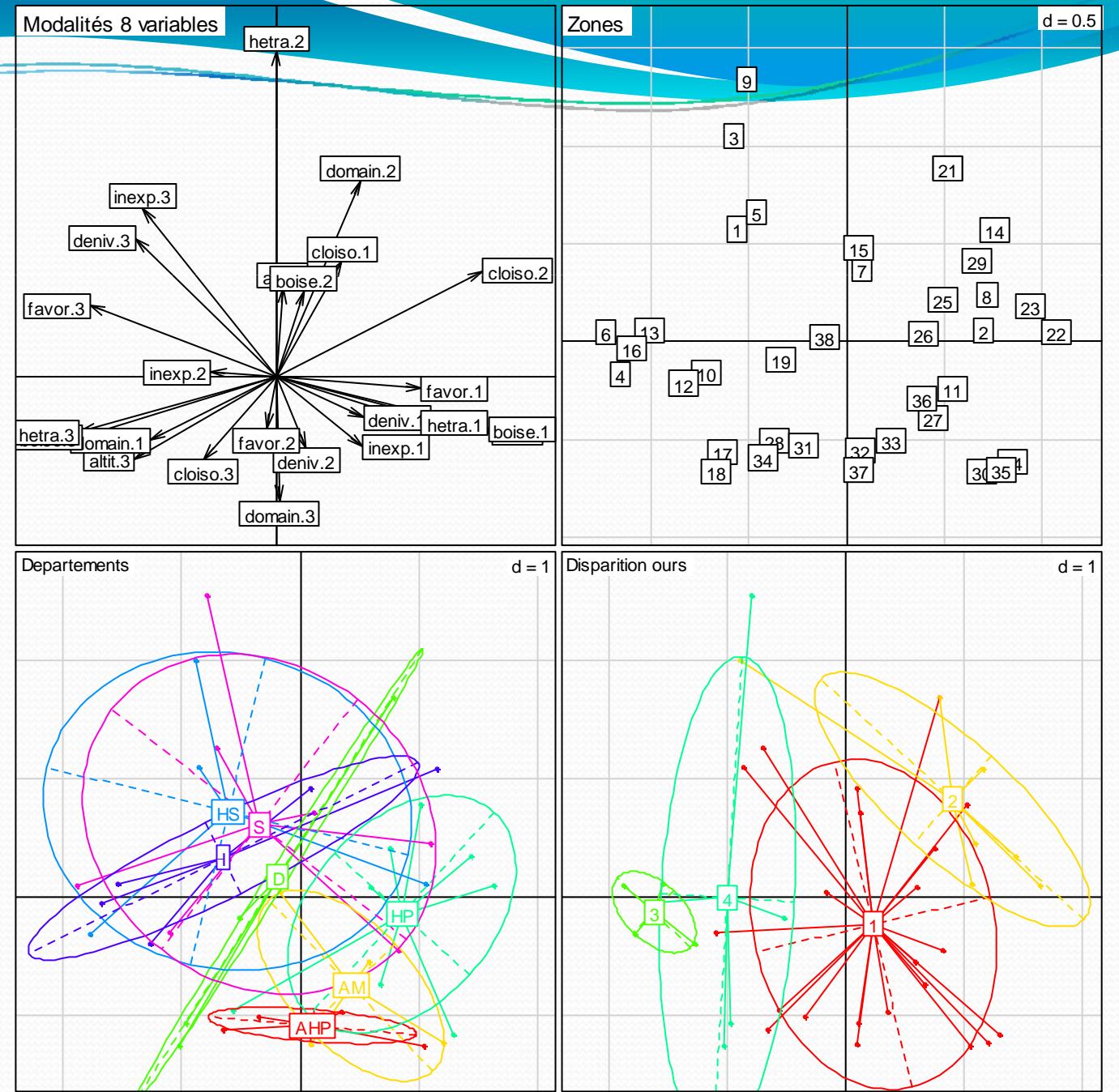


ACM ours

```
scatter(acm, col =  
       rep(c("black",  
             "red3",  
             "darkblue"), 2))
```



ACM ours



Analyse Factorielle des Correspondances Multiples avec la fonction **MCA()** de FactoMineR

MCA(df, ncp = 5, ind.sup = NULL, quanti.sup = NULL, quali.sup = NULL, excl=NULL, graph = TRUE, level.ventil = 0, axes = c(1,2), row.w = NULL, method="Indicator", na.method="NA", tab.disj=NULL)

df	Un “data frame” à n lignes (unités statistiques) et p colonnes (variables qualitatives)
ncp	Nombre d’axes conservés dans les résultats (par défaut 5)
ind.sup	Un vecteur indiquant les indices des individus supplémentaires
quanti.sup	Un vecteur indiquant les indices des variables quantitatives supplémentaires
quali.sup	Un vecteur indiquant les indices des variables qualitatives supplémentaires
excl	Un vecteur indiquant les catégories à exclure
graph	Un logique, si TRUE (option par défaut), un graphique est affiché
level.ventil	Une proportion correspondant au niveau en dessous duquel la catégorie est ventilée (0 par défaut)
axes	Un vecteur de longueur 2 indiquant les axes à afficher
row.w	Un vecteur optionnel des poids des individus actifs (par défaut, poids uniformes : 1/n)
method	Nom de la méthode utilisée: par défaut “Indicator” = AFC sur tableau individus x modalités, ou “Burt” = AFC du tableau de Burt (tableau croisé des modalités)
na.method	Nom de la méthode appliquée pour gérer les valeurs manquantes, “NA” par défaut ou “Average”
tab.disj	df optionnel correspondant au tableau disjonctif utilisé pour l’analyse

ACM sur Ours avec la fonction **MCA()** de FactoMineR

```
res.mca <- MCA(ours,quali.sup=9:10)
res.mca
**Results of the Multiple Correspondence Analysis (MCA)**
The analysis was performed on 38 individuals, described by 10 variables
*The results are available in the following objects:
```

	name	description
1	"\$eig"	"eigenvalues"
2	"\$var"	"results for the variables"
3	"\$var\$coord"	"coord. of the categories"
4	"\$var\$cos2"	"cos2 for the categories"
5	"\$var\$contrib"	"contributions of the categories"
6	"\$var\$v.test"	"v-test for the categories"
7	"\$ind"	"results for the individuals"
8	"\$ind\$coord"	"coord. for the individuals"
9	"\$ind\$cos2"	"cos2 for the individuals"
10	"\$ind\$contrib"	"contributions of the individuals"
11	"\$quali.sup"	"results for the supplementary categorical variables"
12	"\$quali.sup\$coord"	"coord. for the supplementary categories"
13	"\$quali.sup\$cos2"	"cos2 for the supplementary categories"
14	"\$quali.sup\$v.test"	"v-test for the supplementary categories"
15	"\$call"	"intermediate results"
16	"\$call\$marge.col"	"weights of columns"
17	"\$call\$marge.li"	"weights of rows"

ACM sur Ours avec la fonction **MCA()** de FactoMineR

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

```
$`Dim 1`  
$quali  
      R2      p.value  
boise 0.7317972 9.956504e-11  
hetra 0.7233694 1.711015e-10  
citat 0.5864353 1.123836e-06  
favor 0.5352790 1.499217e-06  
altilt 0.4867906 8.514438e-06  
cloiso 0.2598395 5.166578e-03  
inexp 0.2568897 5.539008e-03  
deniv 0.2528392 6.091806e-03  
domain 0.1796741 3.124371e-02
```

```
$category
```

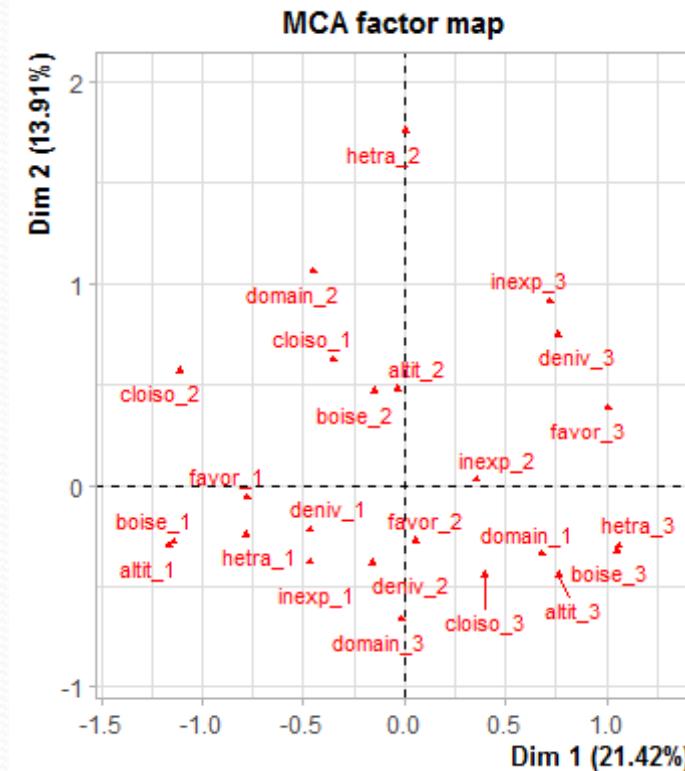
```
      Estimate      p.value  
hetra=hetra_3 0.6342321 6.235888e-10  
boise=boise_3 0.7409956 3.451193e-08  
favor=favor_3 0.5978541 1.367785e-05  
citat=citat_3 0.8159148 3.042320e-04  
altilt=altilt_3 0.5956423 3.361059e-04  
deniv=deniv_3 0.4698980 1.985691e-03  
cloiso=cloiso_3 0.4919272 3.422737e-03  
citat=citat_4 0.4099779 1.616510e-02  
domain=domain_1 0.3988456 1.893311e-02  
inexp=inexp_3 0.3382306 2.149642e-02  
domain=domain_2 -0.3423348 4.576240e-02  
deniv=deniv_1 -0.3356020 3.755617e-02  
cloiso=cloiso_2 0.4943687 1.811479e 02  
citat=citat_2 -0.8309049 7.400855e-03  
depart=HP -0.5989381 5.445917e-03  
inexp=inexp_1 -0.4381363 1.734563e-03  
altilt=altilt_1 -0.6678671 6.306611e-05  
favor=favor_1 -0.5723378 2.158902e-05  
boise=boise_1 -0.6962592 2.215711e-06  
hetra=hetra_1 -0.5743240 5.917410e-09
```

```
attr(", "class")  
[1] "condes" "list"
```

Description des dimensions:

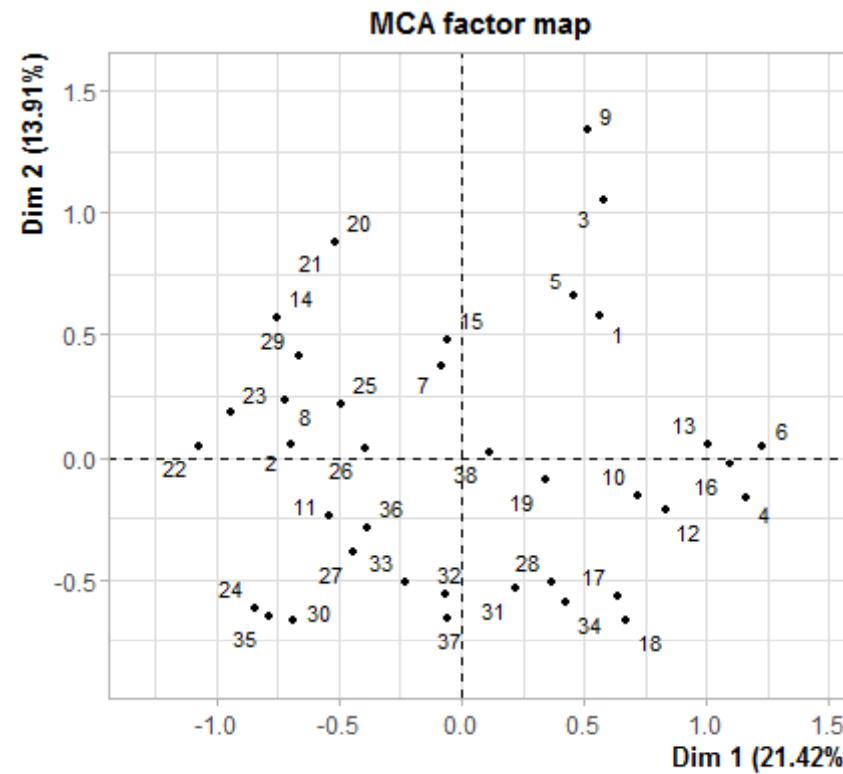
Liste des **variables** (\$quali) et des **modalités** (\$category) qui caractérisent le mieux chaque dimension de l'analyse. (Ici axe 1 seulement)

```
plot.MCA(res.mca,  
         invisible=c("ind", "quali.sup", "quanti.sup"),  
         cex=0.7)
```



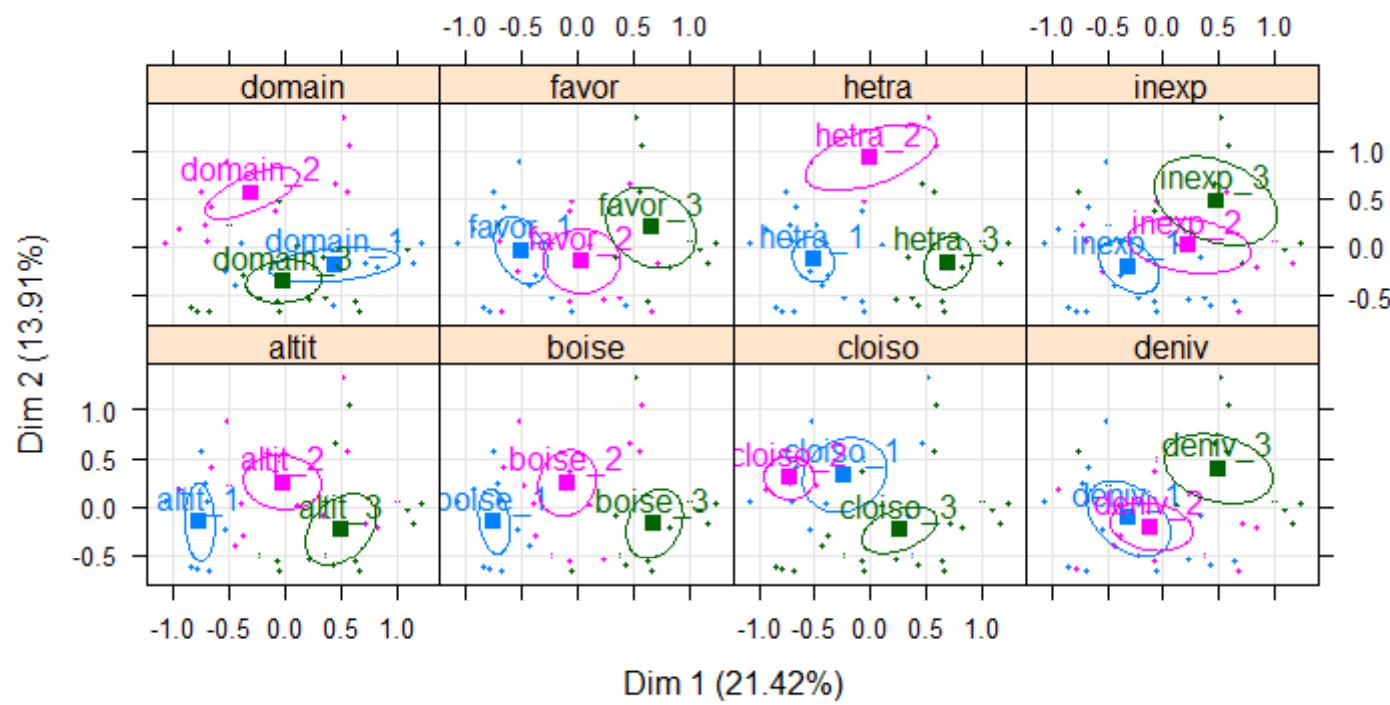
ACM sur Ours avec la fonction **MCA()** de FactoMineR

```
# Les individus  
plot.MCA(res.mca,invisible=c("var","quali.sup","quanti.sup"),cex=0.7)
```



ACM sur Ours avec la fonction **MCA()** de FactoMineR

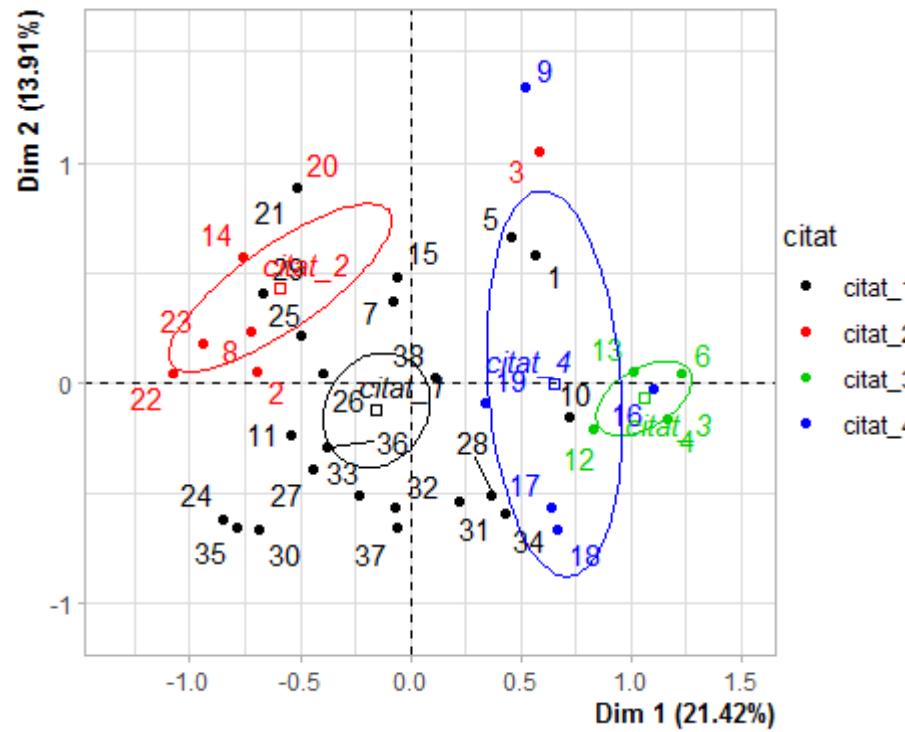
```
# Les individus avec des ellipses par modalité  
plotellipses(res.mca, keepvar=1:8)
```



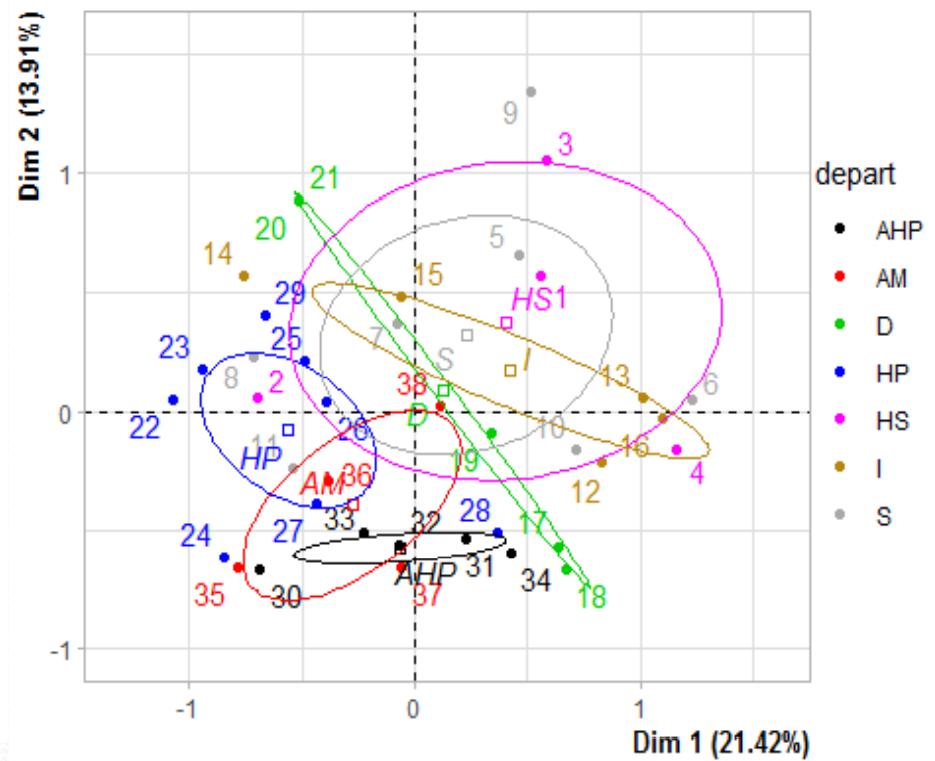
ACM sur Ours avec la fonction MCA() de FactoMineR

```
# Les individus avec des ellipses par variables illustratives  
plotellipses(res.mca,keepvar="citat")  
plotellipses(res.mca,keepvar="depart")
```

Confidence ellipses around the categories of citat



Confidence ellipses around the categories of depart



Plan

- **1. Introduction**
 - Contexte et objectifs de l'analyse multivariée
 - Ordination ou classification ?
 - Ordination basée sur les distances ou sur les valeurs propres (analyses factorielles)
 - Les packages R spécialisés : ade4 et FactoMineR
- **2. Analyses factorielles à un tableau de données, avec les packages ade4 et FactoMineR**
 - 2.1 Analyse en Composantes Principales (ACP) : plusieurs variables quantitatives
 - 2.2 Analyse Factorielle des Correspondances (AFC) : deux variables qualitatives
 - 2.3 Analyse des Correspondances Multiples (ACM) : plusieurs variables qualitatives
- **3. Classification Automatique**
 - 3.1 Classification Ascendante Hiérarchique (CAH) avec la fonction hclust
 - 3.2 Partitionnement avec la fonction kmeans
- **4. Méthodes de couplage de tableaux avec le package ade4**
 - 4.1 Analyses Inter et Intra-classes
 - 4.2 Analyse Discriminante
 - 4.3 Analyses sur variables Instrumentales (ACPVI ou RDA, AFCVI ou CCA)
 - 4.4 Analyse de Co-inertie
- **5. Aperçu des méthodes d'analyse multi-tableaux**
 - 5.1 Analyse Triadique Partielle
 - 5.2 STATIS
 - 5.3 Analyse Factorielle Multiple

3. Classification Automatique

- L'objectif principal des méthodes de Classification Automatique (ou **clustering**) est de répartir les individus statistiques en **groupes** ou **clusters**, c'est-à-dire d'établir une **partition de l'ensemble des individus**.
- Différentes contraintes sont imposées : chaque groupe doit être le plus **homogène** possible et les groupes doivent être les plus **différents** possibles entre eux.
- On peut rechercher une **hiérarchie** des groupes : **Classification Hiérarchique**, qui s'appuie sur des **distances** entre individus. => **fonction hclust()**
- On peut rechercher seulement une **partition** des individus (cas des très grands jeux de données où le calcul de la matrice de distance demande trop de mémoire) => **fonction kmeans()**
- Ces fonctions de Classification Automatique sont dans la **librairie stats de base de R**

3.1 Classification Ascendante Hiérarchique (CAH)

Etape 1: Calcul de distances avec `dist()`

`dist(x, method = "euclidean", diag = FALSE, upper = FALSE, p = 2)`

x	Matrice numérique ou data frame ou objet de type dist
method	La distance à calculer. Parmi "euclidean", "maximum", "manhattan", "canberra", "binary" or "minkowski".
diag	Valeur logique indiquant si la diagonale de la matrice de distance doit être affichée par print.dist
upper	Valeur logique indiquant si le triangle supérieur de la matrice de distance doit être affiché par print.dist
p	La puissance dans le cas de la distance de Minkowski.

La **distance euclidienne** est la plus courante des distances pour des données quantitatives. D'autres distances sont disponibles dans d'autres fonctions :

- Distance de **Jaccard** (données en 0/1) et bien d'autres dans `dist.binary()` de ade4
- Distance de **Mahalanobis** (données morphométriques) dans `dist.quant()` de ade4
- Distances **génétiques** dans `dist.genpop()` de la librairie adegenet...

3.1 Classification Ascendante Hiérarchique (CAH)

Etape 2 : Classification Hiérarchique avec **hclust()**

hclust(d, method = "complete")

d	Un objet généré par la fonction dist.
method	La méthode d'agrégation à utiliser. Parmi "ward.D", "ward.D2", "single", "complete", "average" (= UPGMA), "mcquitty" (= WPGMA), "median" (= WPGMC) ou "centroid" (= UPGMC).

La fonction **hclust()** réalise une **Classification Ascendante Hiérarchique** à partir d'une matrice de distances entre les n individus à classifier.

- A l'étape 1, chaque individu constitue son propre groupe.
- L'algorithme d'agrégation recherche les deux individus les plus similaires au sens de la méthode choisie
- A l'étape 2 on a donc n-1 groupes.
- L'algorithme est répété itérativement, regroupant à chaque étape les deux individus ou groupes les plus similaires, jusqu'à obtenir un seul groupe de tous les individus.

3.1 Classification Ascendante Hiérarchique (CAH)

Etape 2 : Classification Hiérarchique avec **hclust()**

Les méthodes d'agrégation pour agréger les groupes:

"**single**" = lien simple / saut minimum /plus proche voisin (distance entre les deux individus les plus proches)

"**complete**" = [défaut] lien complet / diamètre (distance entre les deux individus les plus éloignés)

"**average**" = UPGMA (Unweighted Pair Group Method of Agregation) ou lien moyen = moyenne des distances arithmétiques entre tous les couples d'individus

"**mcquitty**" = WPGMA (Weighted Pair Group Method of Agregation) , lien moyen avec pondérations

"**centroid**" = UPGMC , Méthode du centroïde (non pondérée)

"**median**" = WPGMC , idem avec pondérations

"**ward.D**" = méthode de variance minimum de Ward : clusters compact et sphériques

"**ward.D2**" = idem, distances au carré (vrai critère de Ward, 1963)

3.1 Classification Ascendante Hiérarchique (CAH)

Etape 3 : le dendrogramme

`plot(h, hang=-1)`

h	Un objet généré par la fonction hclust.
hang	hang=-1 permet d'avoir tous les labels sur la même ligne

La fonction `plot()` appliquée à un objet généré par la fonction `hclust()` permet de tracer le **dendrogramme** dans une fenêtre graphique.

`cutree(h, k=n)`

h	Un objet généré par la fonction hclust.
k	Nombre de groupes choisi

La fonction `cutree()` permet, au vu du dendrogramme de choisir le nombre de groupes et de **créer un vecteur associant chaque individu à un groupe** (NB : ce vecteur n'est pas un facteur).

3.1 Classification Ascendante Hiérarchique

Exemple : les iris de Fisher

```
data(iris)
str(iris)
'data.frame': 150 obs. of 5 variables:
$ Sepal.Length: num 5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
$ Sepal.Width : num 3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
$ Petal.Length: num 1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
$ Petal.Width : num 0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
$ Species    : Factor w/ 3 levels "setosa","versicolor",...: 1 1 1 1 1 1 1 1 1 1 ...
a <- iris[,1:4]
iris.dist <- dist(a, method="euclidean")
iris.dist
      1         2         3         4         5         6         7
2  0.5385165
3  0.5099020  0.3000000
4  0.6480741  0.3316625  0.2449490
5  0.1414214  0.6082763  0.5099020  0.6480741
6  0.6164414  1.0908712  1.0862780  1.1661904  0.6164414
7  0.5196152  0.5099020  0.2645751  0.3316625  0.4582576  0.9949874 ...
```

3.1 Classification Ascendante Hiérarchique

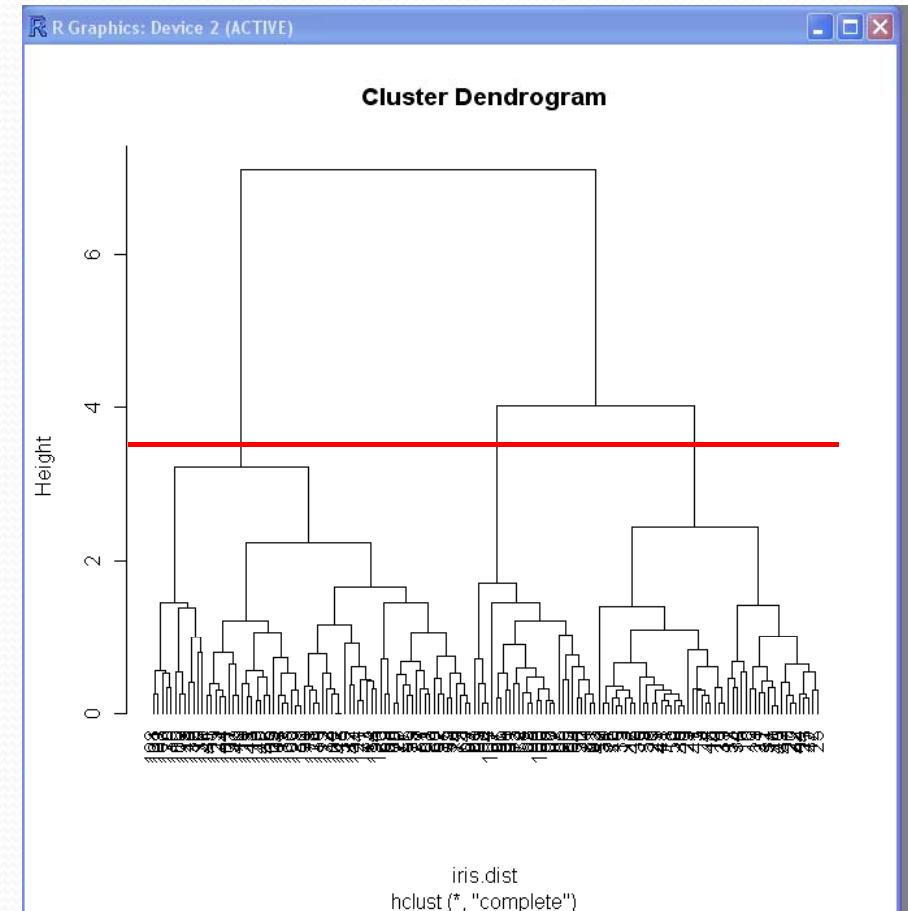
Exemple : les iris de Fisher

```
# Calcul de la hiérarchie
iris.hclust.lc <- hclust(iris.dist) # default method : complete linkage
iris.hclust.lc
Call:
hclust(d = iris.dist)

Cluster method : complete
Distance       : euclidean
Number of objects: 150

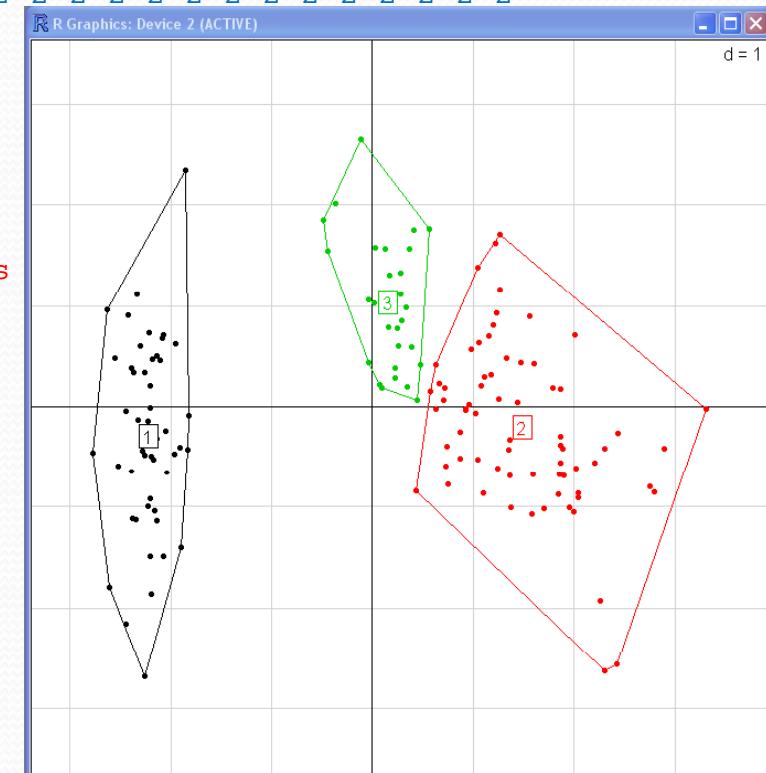
# Affichage du dendrogramme
plot(iris.hclust.lc, hang=-1)

# Coupe du dendrogramme (Faire un choix)
iris.k.lc <- cutree(iris.hclust.lc, k=3)
```



3.1 Classification Ascendante Hiérarchique

Exemple : les iris de Fisher



3.2 Recherche d'une partition avec `kmeans()`

```
kmeans(x, centers, iter.max = 10, nstart = 1,  
        algorithm = c("Hartigan-Wong", "Lloyd", "Forgy", "MacQueen"))
```

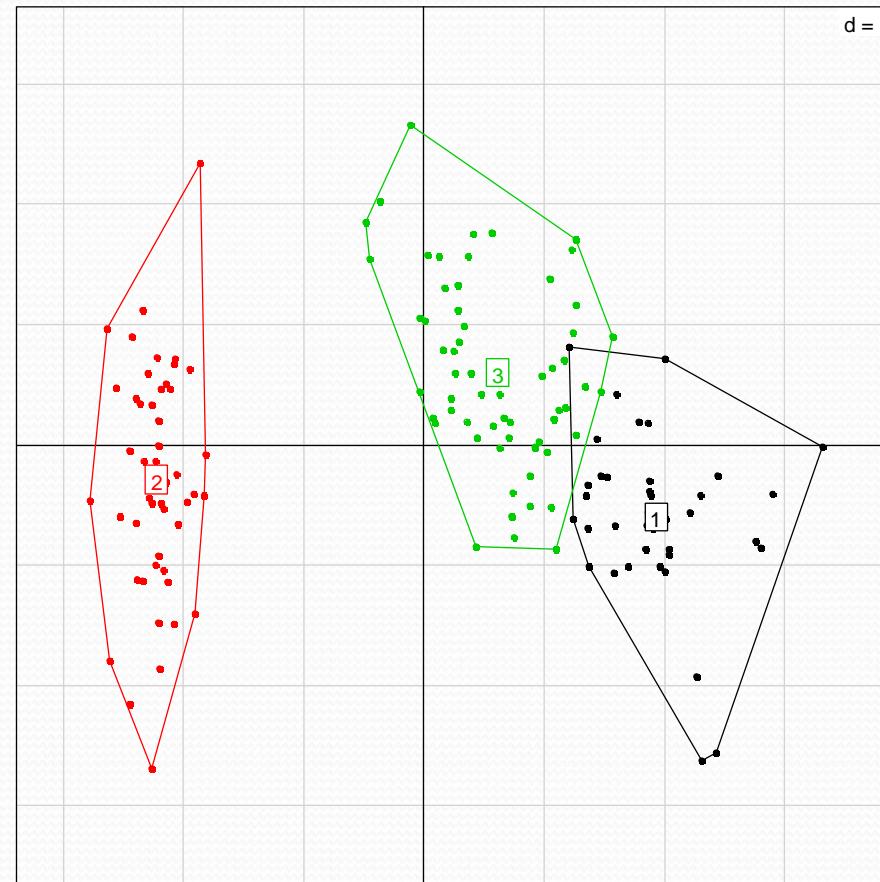
x	Matrice de données numériques ou data frame
centers	k le nombre de groupes souhaités, ou un ensemble de centres de groupes distincts. Si centers est un nombre, un ensemble aléatoire de lignes est choisi comme centres des clusters initiaux.
iter.max	Le nombre maximal d'itérations autorisé
nstart	Si centers est un nombre, combien d'ensembles aléatoires initiaux ?
algorithm	Algorithme parmi "Hartigan-Wong", "Lloyd", "Forgy", "MacQueen" "Hartigan-Wong" est l'algorithme par défaut

La fonction `kmeans()` recherche une partition des individus de x minimisant la somme des carrés des distances entre les points et le centre du groupe qui leur est attribué. Il faut fixer a priori le nombre de groupes souhaités (avec l'option `centers`).

3.2 Recherche d'une partition avec `kmeans()`

3.2 Recherche d'une partition avec kmeans()

```
s.chull(iris.pca$li, as.factor(cl.iris$cluster), optchull =1, cpoint=1,col=c(1,2,3))
```



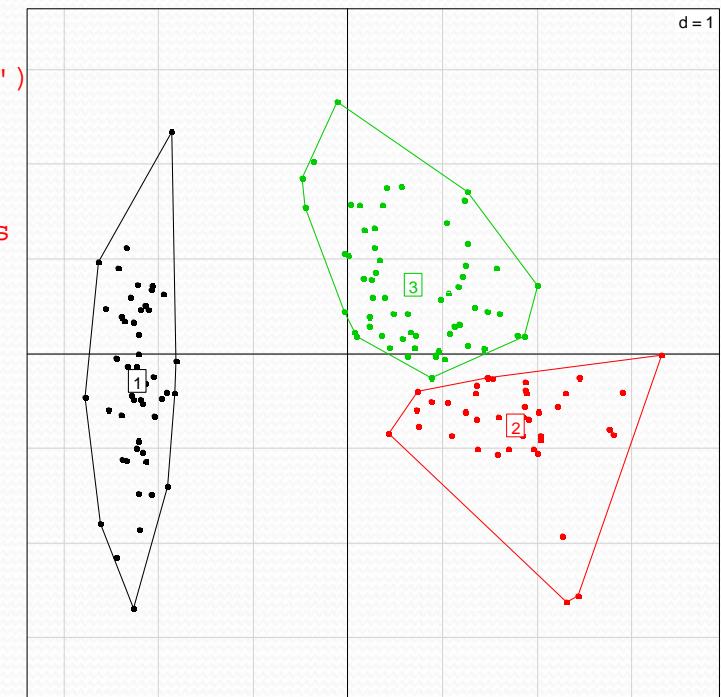
Clustering sur les coordonnées des individus après une analyse factorielle (ACP, AFC, ACM...)

```
# ACP (ou autre analyse factorielle de type dudi)
iris.pca <- dudi.pca(a, scannf=F, nf=2)

# Calcul des distances à partir des coordonnées des individus dans l'ACP
iris.dist <- dist(iris.pca$li, method="euclidean")

# Classification sur cette matrice de distances
iris.hclust.acp <- hclust(iris.dist, method="ward.D2")
plot(iris.hclust.acp, hang=-1)
iris.k.acp <- cutree(iris.hclust.acp, k=3)

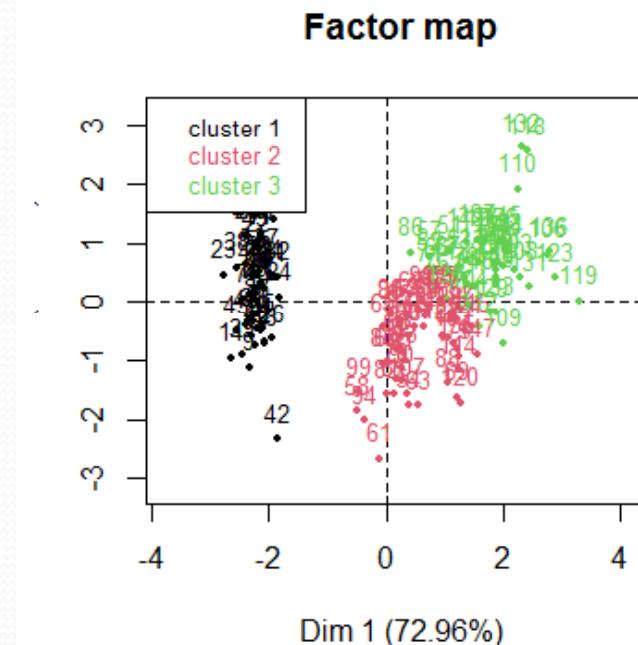
# Projection de ces groupes sur l'ACP du tableau iris
s.chull(iris.pca$li,
as.factor(iris.k.acp),
optchull =1,
cpoint=1,
col=c(1,2,3))
```



La fonction HCPC de FactoMineR : Classification Hiérarchique sur Composantes Principales

```
res.pca <- PCA(iris[1:4], graph=F)
HCPC(res.pca)
# nb.clust pour choix du nombre de cluster,
# nb.clust=0: cliquer sur le dendrogramme pour choisir le niveau de coupure
# nb.clust=-1 pour choix automatique

HCPC(res,
      nb.clust=0,
      consol=TRUE,
      iter.max=10,
      min=3,
      max=NULL,
      metric="euclidean",
      method="ward",
      order=TRUE,
      graph.scale="inertia",
      nb.par=5,
      graph=TRUE,
      proba=0.05,
      cluster.CA="rows",
      kk=Inf,
      description=TRUE, ...)
```



Travaux pratiques : Le jeu de données

accueil consulter gérer s'informer liens contacter IRD déconnecter monique simier

PPEAO

Système d'informations sur les Peuplements de poissons et la Pêche artisanale des Ecosystèmes estuariens, lagunaires ou continentaux d'Afrique de l'Ouest

La base de données PPEAO archive des informations sur les poissons vivant dans différents écosystèmes aquatiques tant continentaux que lagunaires, estuariens ou côtiers de l'Afrique de l'Ouest. Les données collectées concernent aussi bien l'écologie des espèces que leur exploitation par la pêche artisanale.

Ces informations sont le résultat de programmes de recherche menés sur ces écosystèmes à partir de 1978.

Cette base de données a été conçue et réalisée par l'Unité de Recherches RAP (Réponses adaptatives des populations et peuplements de poissons aux pressions de l'environnement) de l'IRD (Institut de Recherche pour le Développement).

version 2.0b | copyright © 2008-2011 IRD, tous droits réservés | code & design originalisé par studio8

Terminé

accueil consulter gérer s'informer liens contacter IRD déconnecter monique simier

consulter des données : extraction des pêches expérimentales

choix des filières

aide >>
[afficher/modifier ma sélection]

Choix de la filière : Peuplement - Environnement - NE/Pt - Biologie - Trophique

critères généraux | cat. écologiques | cat. trophiques | sélection de variables optionnelles | espèces

tout

Pays
Système
Secteur
Station
Campagne
Coup de pêche
Environnement
Fraction
Espèce

liste des colonnes exportées pour Station

Vous pouvez les sélectionner en les cochant quand elles ne sont pas grises

<input checked="" type="checkbox"/> Station_id	<input type="checkbox"/> Sediment_id
<input checked="" type="checkbox"/> Station	<input type="checkbox"/> Vegetation_id
<input type="checkbox"/> Site	<input type="checkbox"/> Station_latitude
<input type="checkbox"/> Distance_embouchure	<input type="checkbox"/> Station_longitude
<input type="checkbox"/> Debris_id	<input type="checkbox"/> Station_memo
<input type="checkbox"/> Position_station_id	<input type="checkbox"/> Debris
	<input type="checkbox"/> Position_station

Afficher le résultat
Exporter en fichier

votre sélection correspond à :

- 180 coup(s) de pêche

- Données extraites de ppeao.ird.fr
- Pêches expérimentales dans 4 systèmes estuariens d'Afrique de l'Ouest (Ebrié, Fatala, Gambie, Saloum), à 2 saisons (sèche / humide) => **256 coups de pêche**
- **Envir** = 14 variables : mesure de 5 variables environnementales lors des 256 coups de pêche (Salinité, Température, Transparence, Profondeur, Distance à l'embouchure) + divers identifiants (saison, système...)
- **Faune** = abondance de 111 espèces de poissons pour les 256 coups de pêche
- **Categ** = catégories écologiques et trophiques des 111 espèces

Exercice 3 : Classification des coups de pêche après l'ACP de l'environnement

- Faire une classification ascendante hiérarchique des coups de pêche en utilisant pour le calcul de la matrice de distance les coordonnées des coups de pêche dans l'ACP de Envir.
- Représenter les groupes obtenus sur le plan 1-2 de l'ACP de Envir
- Comparer le résultat obtenu avec les trois approches étudiées (hclust / kmeans / HCPC)

Bibliographie

- **ade4** : pbil.univ-lyon1.fr/ADE-4/
 - Multivariate Analysis of Ecological Data with ade4. 2018. J. Thioulouse, S. Dray, A.-B. Dufour, A. Siberchicot, T. Jombart et S. Pavoine. <https://link.springer.com/book/10.1007/978-1-4939-8850-1>
- **FactoMineR** : factominer.free.fr
 - Analyse de données avec R. 2016. F. Husson, S. Lê et J. Pagès. Presses Universitaires de Rennes.

Bibliographie - analyses à 1 tableau

PDF files

- Champely S. (2005) **Introduction à l'analyse multivariée (factorielle).pdf**
- Cours ade4, Novembre 2010, Université de Lyon :
 - Partie2-1_Initiation_ACP.pdf
 - Partie2-2_ACP.pdf
 - Partie2-3_Initiation_AFC.pdf
 - Partie2-4_Initiation_ACM.pdf
 - Partie2_5_Ordination_Tableaux_Ecologiques.pdf
- **Introduction Classification Hierarchique Chessel_etal(fr).pdf** : Fiche de Biostatistique Stage 7, par D. Chessel, J. Thioulouse et A.B. Dufour.
- **tdr69_classification.pdf** par A.B. Dufour et S. Dray

Plan

- **1. Introduction**
 - Contexte et objectifs de l'analyse multivariée
 - Ordination ou classification ?
 - Ordination basée sur les distances ou sur les valeurs propres (analyses factorielles)
 - Les packages R spécialisés : ade4 et FactoMineR
- **2. Analyses factorielles à un tableau de données, avec les packages ade4 et FactoMineR**
 - 2.1 Analyse en Composantes Principales (ACP) : plusieurs variables quantitatives
 - 2.2 Analyse Factorielle des Correspondances (AFC) : deux variables qualitatives
 - 2.3 Analyse des Correspondances Multiples (ACM) : plusieurs variables qualitatives
- **3. Classification Automatique**
 - 3.1 Classification Ascendante Hiérarchique (CAH) avec la fonction hclust
 - 3.2 Partitionnement avec la fonction kmeans
- **4. Méthodes de couplage de tableaux avec le package ade4**
 - 4.1 Analyses Inter et Intra-classes
 - 4.2 Analyse Discriminante
 - 4.3 Analyses sur variables Instrumentales (ACPVI ou RDA, AFCVI ou CCA)
 - 4.4 Analyse de Co-inertie
- **5. Aperçu des méthodes d'analyse multi-tableaux**
 - 5.1 Analyse Triadique Partielle
 - 5.2 STATIS
 - 5.3 Analyse Factorielle Multiple

4. Méthodes de couplage de tableaux

Nous revenons maintenant sur les analyses factorielles (ACP, AFC, ACM...) et nous allons voir les méthodes permettant d'analyser simultanément deux tableaux portant sur les mêmes individus (lignes)

Tableau Y	Tableau X	Objectif	Méthode	
1 tableau à expliquer (n lignes, p colonnes)	1 vecteur = classes (n lignes)	Prendre en compte ou éliminer la part de variabilité due à la classe	Analyse Inter-classes et Intra- classes	bca, wca (ade4)
			Analyse discriminante linéaire	lda (MASS, R de base) discrimin (ade4)
1 tableau à expliquer (n lignes, p colonnes) Ex : abondance faune	1 tableau explicatif (n lignes, q colonnes) Ex: variables de milieu	Prendre en compte ou éliminer la part de variabilité due au tableau explicatif ($X \Leftrightarrow Y$)	Analyse sur variables instrumentales	pcaiv (ade4) cca, rda (vegan)
1 tableau à expliquer (n lignes, p colonnes) Ex : abondance faune	1 tableau explicatif (n lignes, q colonnes) Ex: variables de milieu	Rechercher la structure des échantillons commune aux deux tableaux ($X \Leftrightarrow Y$)	Analyse de coinertie	coi (ade4)

4.1. Analyses Inter et Intra-classes

- Les **analyses inter et intra-classes** peuvent être vues comme une extension à un schéma de dualité (ACP, AFC, ACM...) de l'analyse de variance à un facteur : on décompose la variabilité d'une variable (ou ici d'un tableau) Y en une part expliquée par une variable qualitative X (**variance inter**) et une part non expliquée par X (**variance intra**).
- La fonction **bca()** de la librairie ade4 réalise une analyse **inter-classes** ou inter-groupes, en anglais **between-class**, permettant de se focaliser sur la part prise en compte par une variable explicative unique de type facteur (X) dans une analyse simple de type dudi sur un tableau Y.
- La fonction **wca()** de la librairie ade4 réalise l'analyse intra-classes (**within-class**), c'est-à-dire l'analyse du tableau Y débarrassé de l'effet de X. Sa synthaxe est similaire à celle de bca()

4.1. Analyses Inter-classes (Between-Class Analyses)

```
library(ade4)
bca(x, fac, scannf = TRUE, nf = 2)
```

x	objet de classe dudi issu de l'analyse factorielle du tableau Y (n,p) avec ade4 : dudi.pca, dudi.coa, dudi.mca...
fac	objet (vecteur) de classe factor de longueur n, partitionnant les individus de Y
scannf	une valeur logique (TRUE/FALSE) indiquant si le diagramme des valeurs propres doit être affiché
nf	si scannf =FALSE, un entier indiquant le nombre d'axes à conserver (2 par défaut)

La **bca()** est elle-même une analyse d'inertie portant sur le tableau des moyennes des variables par classe, les individus du tableau d'origine étant ensuite projetés sur les axes en individus supplémentaires.

4.1. Analyses Inter classes

Exemple : Méaudret milieu

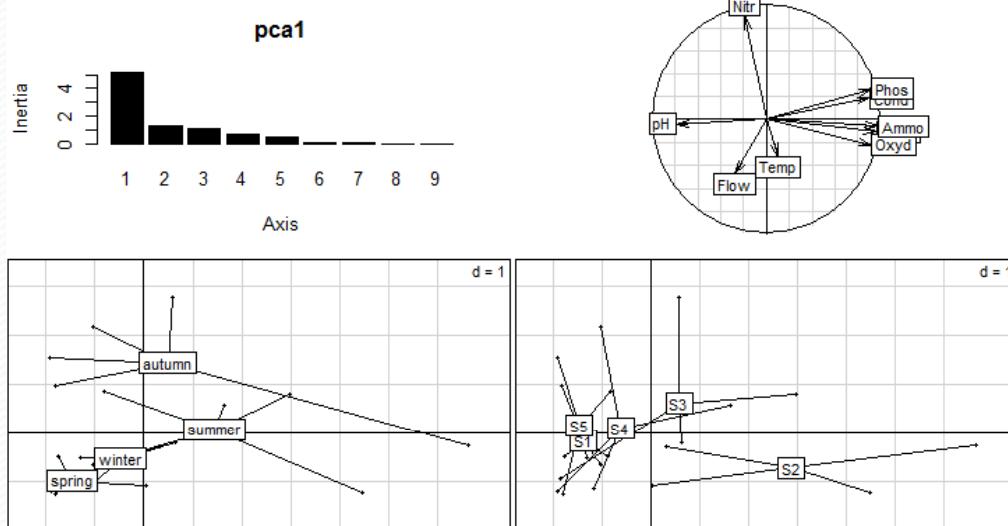
Objectif : quantifier les effets dus au site ou à la saison dans l'ACP du tableau Méaudret milieu.

- Première étape : effectuer l'ACP du tableau mil de meaudret:

```
library(ade4)
data(meaudret)
mil <- meaudret$env
pca1 <- dudi.pca(mil, scannf=F, nf=5)
```

- Suite à cette analyse, on procède à titre exploratoire à la projection sur les axes des points moyens par site et par saison : mélange des deux effets

```
screeplot(pca1)
s.corcircle(pca1$co)
plan <- meaudret$design
s.class(pca1$li,
         plan$season,
         cellipse = 0)
s.class(pca1$li,
         plan$site,
         cellipse = 0)
```



4.1. Analyses Inter classes

Exemple : Méaudret milieu, effet site

Analyse inter-sites : se focaliser sur l'effet **site**, le quantifier (ratio) et tester sa significativité (par permutations)

```
bcal <- bca(pcal, plan$site, scannf=F, nf=2)
```

```
$nf (axis saved) : 2
```

```
$rank: 4
```

```
$ratio: 0.3805115
```

```
eigen values: 2.681 0.6208 0.1132 0.009503
```

```
vector length mode content
1 $eig    4     numeric eigen values
2 $lw     5     numeric group weights
3 $cw     9     numeric col weights
```

```
data.frame nrow ncol content
1 $stab     5    9     array class-variables
2 $li       5    2     class coordinates
3 $l1       5    2     class normed scores
4 $co       9    2     column coordinates
5 $c1       9    2     column normed scores
6 $ls      20    2     row coordinates
7 $as       5    2     inertia axis onto between axis
```

bca1\$ratio (entre 0 et 1) est le rapport entre inertie totale de l'ACP intra / inertie de l'ACP simple = % pris en compte par effet site.

```
sum(bcal$eig)/sum(pca1$eig)
```

```
0.3805115
```

Ratio = 0.38 = 38%

Test de permutations

```
rtl <- rtest(bcal, nrep=1000)
rtl
Monte-Carlo test
Call: rtest.between(xtest = bcal, nrep = 1000)
```

```
Observation: 0.3805115
```

```
Based on 1000 replicates
```

```
Simulated p-value: 0.02597403
```

```
Alternative hypothesis: greater
```

```
Std.Obs Expectation Variance
2.522453851 0.208637853 0.004642718
```

Principe du test : permuter aléatoirement les individus par rapport à leur classe (ici le site) et refaire l'ACP inter-sites.

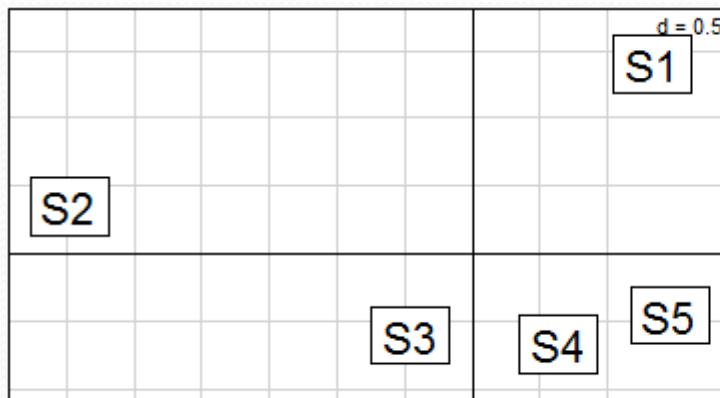
On répète cette opération 1000 fois.

A chaque fois, le ratio est calculé. On peut ensuite comparer le ratio observé (0.38) à la distribution des ratios simulés et calculer la p-value

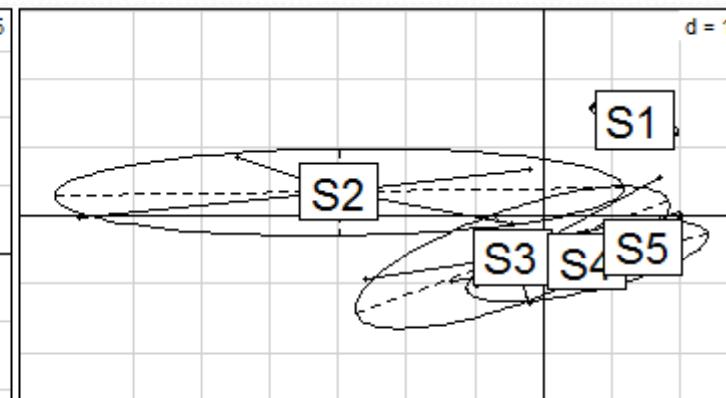
4.1. Analyses Inter classes

Exemple : Méaudret milieu, effet site

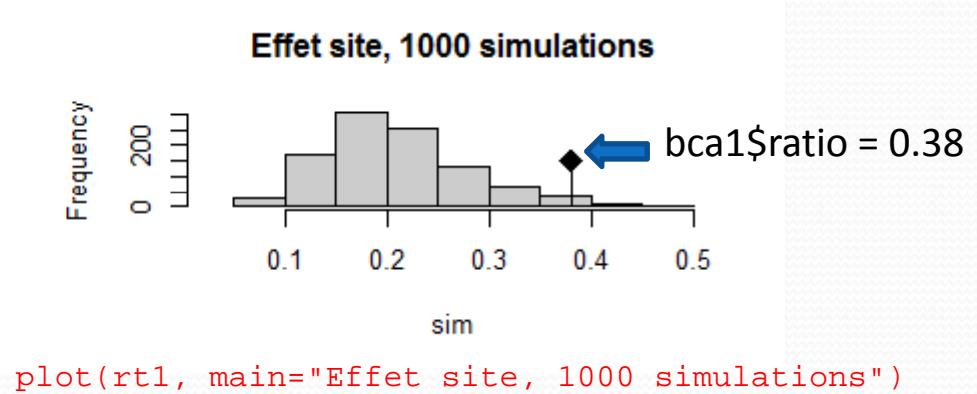
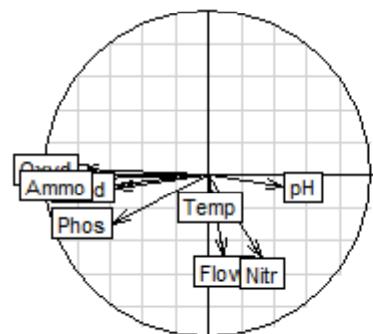
```
s.label(bcal$li, clab=2)
```



```
s.class(bcal$ls, plan$site, clab=2)
```



```
s.corcircle(bcal$co)
```



4.1. Analyses Inter classes

Exemple : Méaudret milieu, effet saison

Analyse inter-saisons : se focaliser sur l'effet **saison**, le quantifier (ratio) et tester sa significativité (par permutations)

```
bca2 <- bca(pcal,plan$saison, scannf=F, nf=2)
$nf (axis saved) : 2
$rank: 3
$ratio: 0.3722686

eigen values: 1.707 1.078 0.5652

  vector length mode      content
1 $eig    3     numeric eigen values
2 $lw     4     numeric group weights
3 $cw     9     numeric col weights

  data.frame nrow ncol content
1 $stab     4    9     array class-variables
2 $li       4    2     class coordinates
3 $l1       4    2     class normed scores
4 $co       9    2     column coordinates
5 $c1       9    2     column normed scores
6 $ls      20    2     row coordinates
7 $as       5    2     inertia axis onto between axis
```

bca1\$ratio est le rapport entre inertie totale de l'ACP intra / inertie de l'ACP simple = % pris en compte par effet site.

```
sum(bcal$eig)/sum(pcal$eig)
0.3722686
```

Ratio = 0.37 = 37%

Test de permutations

```
rt2 <- rtest(bca2, nrepel=1000)
rt2
Monte-Carlo test
Call: rtest.between(xtest = bca2, nrepel = 1000)
```

Observation: 0.3722686

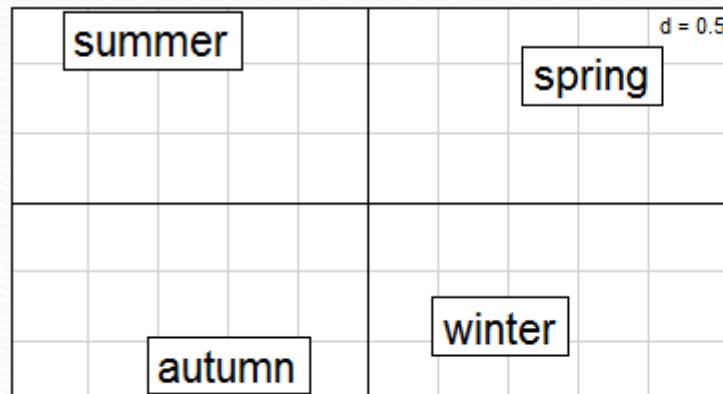
```
Based on 1000 replicates
Simulated p-value: 0.004995005
Alternative hypothesis: greater
```

```
Std.Obs  Expectation   Variance
3.552057650 0.154996737 0.003741511
```

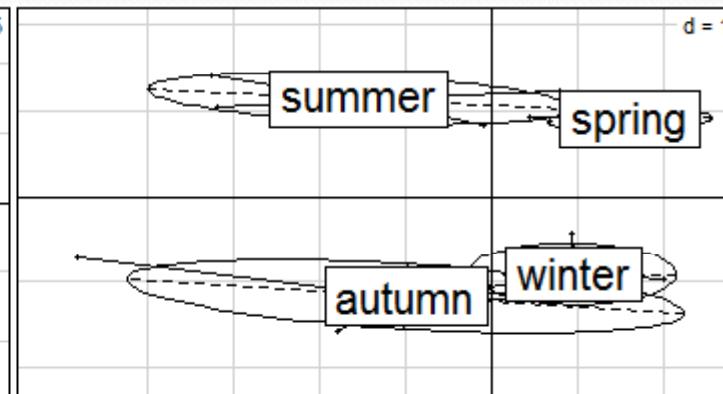
4.1. Analyses Inter classes

Exemple : Méaudret milieu, effet saison

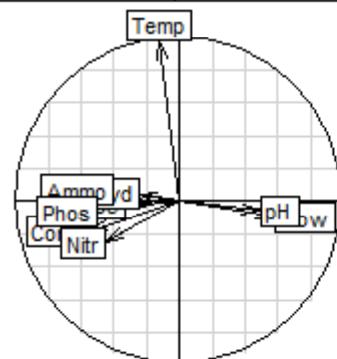
```
s.label(bca2$li, clab=2)
```



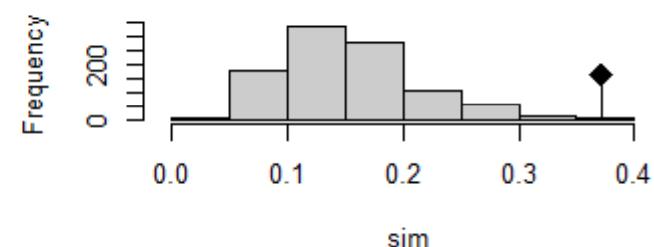
```
s.class(bca2$ls, plan$saison, clab=2)
```



```
s.corcircle(bca2$co)
```



Effet saison, 1000 simulations



```
plot(rt2, main="Effet saison, 1000 simulations")
```

4.1. Analyses Intra-classes (Within-Class Analyses)

```
library(ade4)  
wca(x, fac, scannf = TRUE, nf = 2)
```

x	objet de classe dudi issu de l'analyse factorielle du tableau Y (n,p) avec ade4 : dudi.pca, dudi.coa, dudi.mca...
fac	objet (vecteur) de classe factor de longueur n, partitionnant les individus de Y
scannf	une valeur logique (TRUE/FALSE) indiquant si le diagramme des valeurs propres doit être affiché
nf	si scannf =FALSE, un entier indiquant le nombre d'axes à conserver (2 par défaut)

La **wca()** réalise l'analyse intra-classes (**within-class**), c'est-à-dire l'analyse du tableau Y débarrassé de l'effet considéré. Sa syntaxe est similaire à celle de bca().

4.1. Analyses Intra classes

Exemple : Méaudret milieu, éliminer l'effet site

Analyse intra-sites : Eliminer l'effet site

```
wcal <- wca(pcal,plan$site, scannf=F, nf=2)
wcal
Within analysis
call: wca.dudi(x = pcal, fac = plan$site, scannf = F, nf = 2)
class: within dudi

$nf (axis saved) : 2
$rank: 9
$ratio: 0.6194885 ← Ratio = 0.62

eigen values: 2.703 1.146 0.9934 0.4422 0.1846 ...

      vector length mode      content
1 $eig     9    numeric eigen values
2 $lw     20    numeric row weights
3 $cw     9    numeric col weights
4 $stabw  5    numeric class weights
5 $fac    20    numeric factor for grouping

      data.frame nrow ncol
1 $stab     20   9
2 $li       20   2
3 $l1       20   2
4 $co       9   2
5 $c1       9   2
6 $ls       20   2
7 $as       5   2
      content
1 array class-variables
2 row coordinates
3 row normed scores
4 column coordinates
5 column normed scores
6 supplementary row coordinates
7 inertia axis onto within axis
```

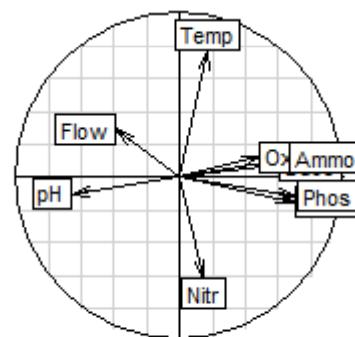
Contrairement à bca qui s'applique au tableau des moyennes par classe, on analyse ici un tableau wca1\$tab de **même dimension** que le tableau d'origine mil où les données sont centrées par classe

Variance totale = variance inter-site + variance intra-site
(0.38 + 0.62) car plan d'échantillonnage équilibré

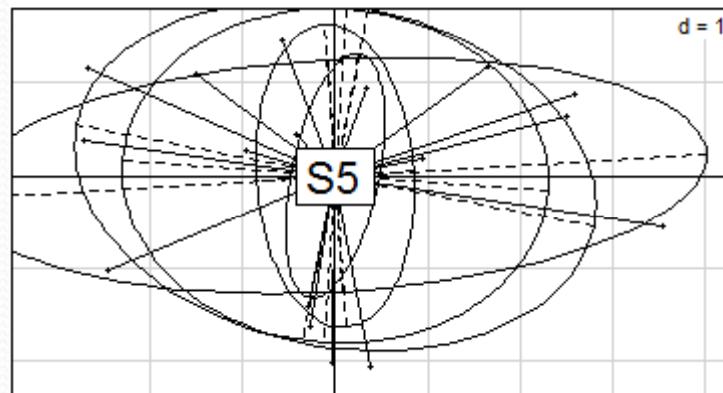
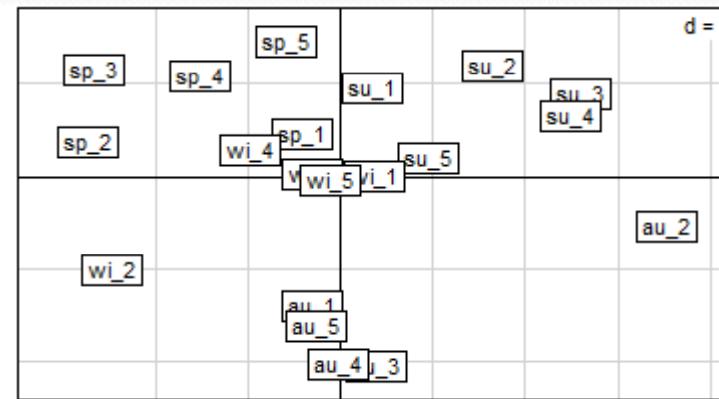
4.1. Analyses Intra classes

Exemple : Méaudret milieu, éliminer l'effet site

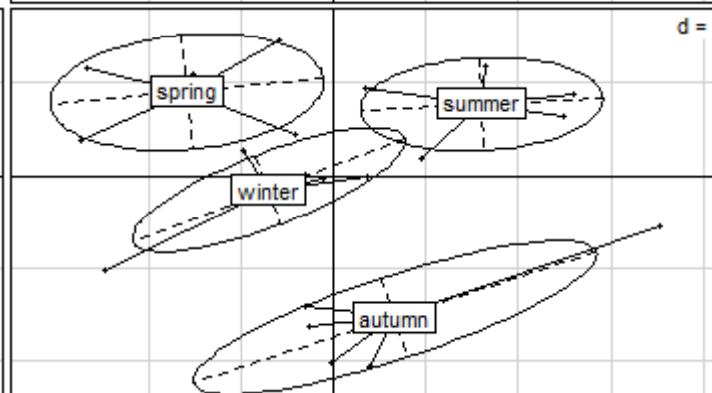
`s.corcircle(wcal$co)`



`s.label(wcal$li)`



`s.class(wcal$l, plan$site, clab=2)`



`s.class(wcal$ls, plan$season)`

4.1. Analyses Intra classes

Exemple : Méaudret milieu, éliminer l'effet saison

Analyse intra-saisons : Eliminer l'effet saison

```
wca2 <- wca(pcal,plan$saison, scannf=F, nf=2)
wca2
Within analysis
call: wca.dudi(x = pcal, fac = plan$season, scannf = F, nf = 2)
class: within dudi

$nf (axis saved) : 2
$rank: 9
$ratio: 0.6277314 ← Ratio = 0.63

eigen values: 4.158 0.7531 0.4054 0.228 0.05361 ...

      vector length mode     content
1 $eig    9     numeric eigen values
2 $lw     20    numeric row weights
3 $cw     9     numeric col weights
4 $stabw  4     numeric class weights
5 $fac    20    numeric factor for grouping

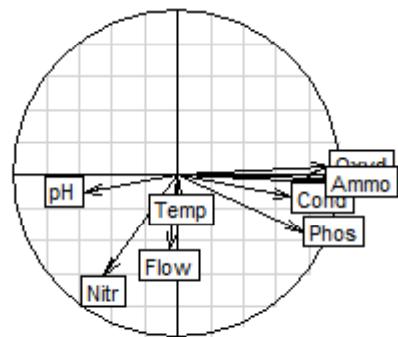
      data.frame nrow ncol content
1 $tab      20   9   array class-variables
2 $li       20   2   row coordinates
3 $l1       20   2   row normed scores
4 $co       9    2   column coordinates
5 $c1       9    2   column normed scores
6 $ls       20   2   supplementary row coordinates
7 $as       5    2   inertia axis onto within axis
```

Variance totale = variance inter-saison + variance intra-saison
(0.37+ 0.63) car plan d'échantillonnage équilibré

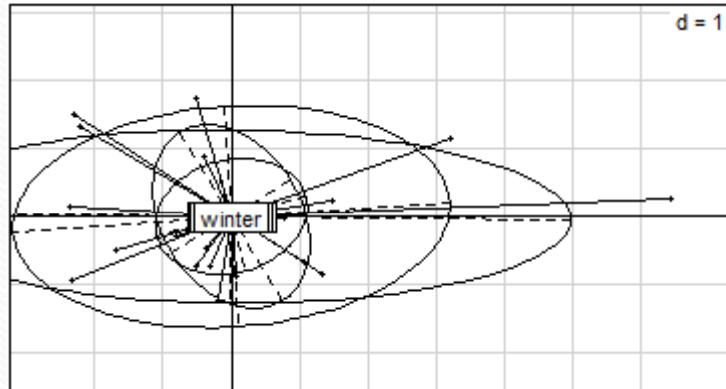
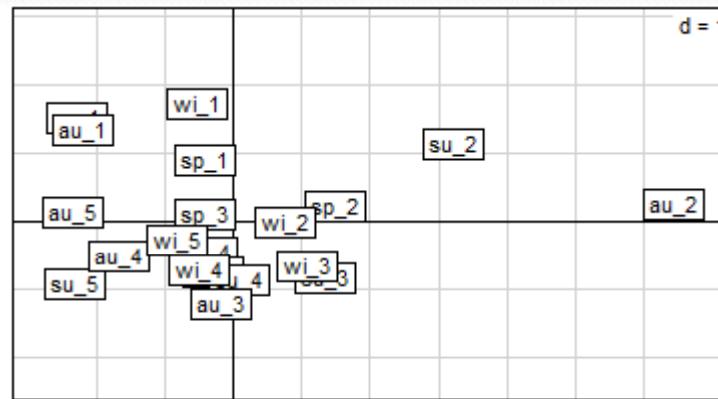
4.1. Analyses Intra classes

Exemple : Méaudret milieu, éliminer l'effet saison

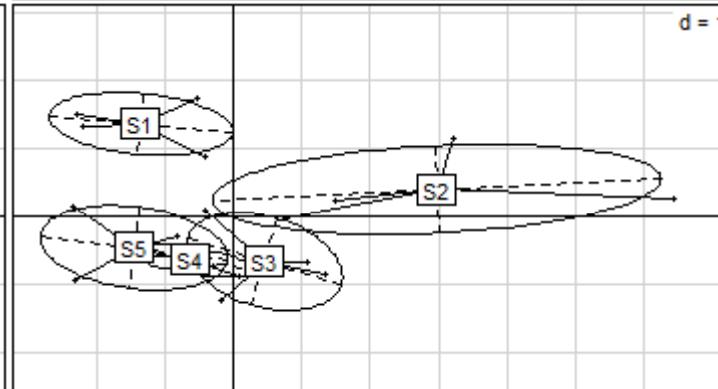
`s.corcircle(wca2$co)`



`s.label(wca2$li)`



`s.class(wca2$li, plan$saison)`



`s.class(wca2$li, plan$site)`

Travaux pratiques : Le jeu de données

PPEAO

Système d'informations sur les Peuplements de poissons et la Pêche artisanale des Ecosystèmes estuariens, lagunaires ou continentaux d'Afrique de l'Ouest

La base de données PPEAO archive des informations sur les poissons vivant dans différents écosystèmes aquatiques tant continentaux que lagunaires, estuariens ou côtiers de l'Afrique de l'Ouest. Les données collectées concernent aussi bien l'écologie des espèces que leur exploitation par la pêche artisanale.

Ces informations sont le résultat de programmes de recherche menés sur ces écosystèmes à partir de 1978.

Cette base de données a été conçue et réalisée par l'Unité de Recherches RAP (Réponses adaptatives des populations et peuplements de poissons aux pressions de l'environnement) de l'IRD (Institut de Recherche pour le Développement).

choix des filières

Choix de la filière : Peuplement - Environnement - NE/Pt - Biologie - Trophique

listé des colonnes exportées pour Station

<input checked="" type="checkbox"/> Station_id	<input type="checkbox"/> Sediment_id
<input checked="" type="checkbox"/> Station	<input type="checkbox"/> Vegetation_id
<input type="checkbox"/> Site	<input type="checkbox"/> Station_latitude
<input type="checkbox"/> Distance_embouchure	<input type="checkbox"/> Station_longitude
<input type="checkbox"/> Debris_id	<input type="checkbox"/> Station_memo
<input type="checkbox"/> Position_station_id	<input type="checkbox"/> Debris
	<input type="checkbox"/> Position_station

- Données extraites de **ppeao.ird.fr**
- Pêches dans 4 systèmes estuariens d'Afrique de l'Ouest (Ebrié, Fatala, Gambie, Saloum), à deux saisons (sèche / humide) => **256 coups de pêche**
- **Faune** = abondance de 111 espèces de poissons
- **Categ** = catégories écologiques et trophiques des espèces
- **Envir** = mesure de 5 variables environnementales : Salinité, Température, Transparence, Profondeur, Distance à l'embouchure + divers identifiants (saison, système...)

Exercice 4 : Analyses inter et intra Saison et Système

- Faire les analyses inter-classes et intra-classes pour l'ACP de **Envir** et l'AFC de **Faulog** en utilisant les facteurs système, saison et systsais
- Comparer le % de la variance totale de chaque analyse pris en compte par les facteurs :

Analyse	%Syst	%Saison	%SystSais
Between-ACP Envir	39.23	3.69	56.17
Within-ACP Envir	60.77	96.31	43.83
Between- AFC Faulog	11.51	1.04	15.39
Within- AFC Faulog	88.49	96.96	84.61

Plan

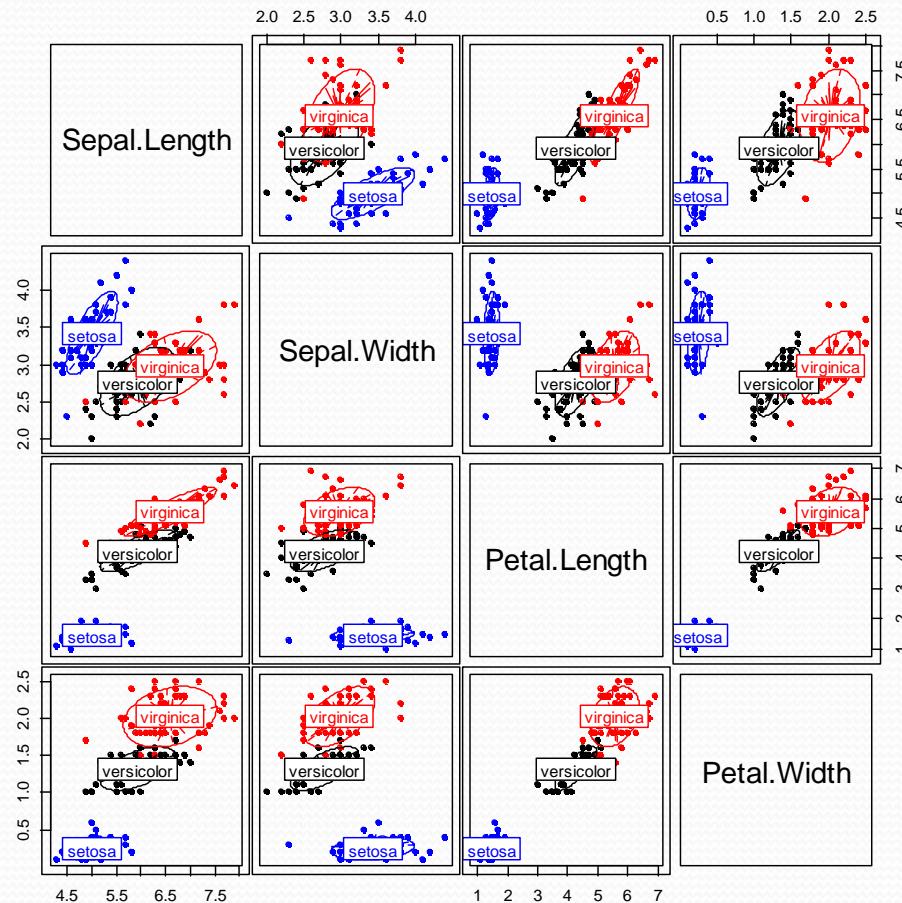
- **1. Introduction**
 - Contexte et objectifs de l'analyse multivariée
 - Ordination ou classification ?
 - Ordination basée sur les distances ou sur les valeurs propres (analyses factorielles)
 - Les packages R spécialisés : ade4 et FactoMineR
- **2. Analyses factorielles à un tableau de données, avec les packages ade4 et FactoMineR**
 - 2.1 Analyse en Composantes Principales (ACP) : plusieurs variables quantitatives
 - 2.2 Analyse Factorielle des Correspondances (AFC) : deux variables qualitatives
 - 2.3 Analyse des Correspondances Multiples (ACM) : plusieurs variables qualitatives
- **3. Classification Automatique**
 - 3.1 Classification Ascendante Hiérarchique (CAH) avec la fonction hclust
 - 3.2 Partitionnement avec la fonction kmeans
- **4. Méthodes de couplage de tableaux avec le package ade4**
 - 4.1 Analyses Inter et Intra-classes
 - **4.2 Analyse Discriminante**
 - 4.3 Analyses sur variables Instrumentales (ACPVI ou RDA, AFCVI ou CCA)
 - 4.4 Analyse de Co-inertie
- **5. Aperçu des méthodes d'analyse multi-tableaux**
 - 5.1 Analyse Triadique Partielle
 - 5.2 STATIS
 - 5.3 Analyse Factorielle Multiple

4.2. Analyse Discriminante Linéaire

- L'**analyse factorielle discriminante ou analyse discriminante linéaire** cherche à discriminer des groupes connus *a priori* au sein d'un ensemble d'observations, à partir d'un ensemble de variables prédictives (mesures).
- L'analyse discriminante est utilisée dans de nombreux domaines, par exemple en biologie, lorsque l'on veut affecter un objet à sa famille d'appartenance à partir de ses caractéristiques physiques. Les iris de Sir Ronald Fisher — qui est à l'origine de cette méthode — en est un exemple, il s'agit de reconnaître la variété d'iris (setosa, virginica, et versicolor) à partir de la longueur/largeur de ses pétales et sépales.
- L'idée de l'analyse discriminante est de chercher une **variable discriminante synthétique** qui est une combinaison linéaire des variables d'origine, permettant de discriminer au mieux les groupes.
- On parle de **discrimination descriptive** quand la question est : « qu'est-ce qui sépare les groupes ? », et de **discrimination prédictive** quand la question est : « à quel groupe est-ce que je peux affecter un nouvel individu dont je connais les mesures, mais pas le groupe, et avec quelle erreur ? »
- La fonction **discrimin()** de la librairie ade4 réalise une analyse discriminante linéaire dans un but plutôt descriptif.
- La fonction **lda()** de la librairie MASS (dans R de base) réalise une analyse discriminante linéaire focalisée sur la **prédiction** de la classe à laquelle appartient un individu nouveau.

4.2. Analyse Discriminante du tableau iris (150,4)

```
data(iris)
# Nuages de points de toutes les variables
# numériques 2 à 2
# Avec des étoiles par Species
panl <- function(x, y, ...) {
  xy <- cbind.data.frame(x, y)
  s.class(xy,
    iris$Species,
    include.origin = F,
    add.plot = T,
    clab = 1.5,
    col = c("blue", "black", "red"),
    cpoint = 2,
    cstar = 0.5)
}
pairs(iris[, 1:4], panel = panl)
```



4.2. Analyse Discriminante avec discrimin (ade4)

```
library(ade4)
discrimin(dudi, fac, scannf = TRUE, nf = 2)
```

dudi	objet de classe dudi issu de l'analyse factorielle du tableau Y (n,p) avec ade4 : dudi.pca, dudi.coa, dudi.mca...
fac	objet (vecteur) de classe factor de longueur n, partitionnant les individus de Y
scannf	une valeur logique (TRUE/FALSE) indiquant si le diagramme des valeurs propres doit être affiché
nf	si scannf =FALSE, un entier indiquant le nombre d'axes à conserver (2 par défaut)

4.2. Analyse Discriminante du tableau iris

```
dis1 <- discrimin(dudi.pca(iris[, 1:4], scannf = F), iris$Species, scannf = F)
dis1
Discriminant analysis
call: discrimin(dudi = dudi.pca(iris[, 1:4], scannf = F), fac = iris$Species,
scannf = F)
class: discrimin

$nf (axis saved) : 2

eigen values: 0.9699 0.222

  data.frame nrow ncol content
1 $fa       4     2   loadings / canonical weights
2 $li      150    2   canonical scores
3 $va       4     2   cos(variables, canonical scores)
4 $cp       4     2   cos(components, canonical scores)
5 $gc       3     2   class scores
```

discrimin() fournit une combinaison linéaire des variables normalisées, de variance totale=1, qui maximise la variance inter-classe. Les coefficients sont dans la composante `dis1$fa` (fa pour facteur, dans le vocabulaire du schéma de dualité).

```
dis1$fa
      DS1        DS2
Sepal.Length 0.1200150  0.01772302
Sepal.Width  0.1168775  0.83778380
Petal.Length -0.6790443 -1.46087856
Petal.Width  -0.3743571  1.92176982
```

4.2. Analyse Discriminante du tableau iris

Eléments générés par discrimin:

- Eigenvalues (`dis1$eig`) : valeurs propres.
- Canonical weights (`dis1$fa`) : poids canoniques ou loadings (coefficients des combinaisons linéaires de variance unité et de variance inter maximales). Les variables utilisées sont les colonnes normalisées de l'analyse en composantes principales préalable.
- Scores and classes (`dis1$li` avec `dis1$gc`) : variables canoniques ou scores (combinaisons linéaires de variance unité et de variance inter maximales) et groupes ou classes - ellipses - qui donnent le mode de discrimination opérée.
- Cos(variates, canonical weights) (`dis1$va`): corrélations entre variables canoniques et les variables de départ. Si pas cohérent avec Canonical weights, c'est l'indice d'une instabilité numérique qui remet en cause l'analyse.
- Cos(components, canonical variates) (`dis1$cp`) : corrélations entre les variables canoniques et les composantes principales de l'analyse de départ. On peut ainsi savoir si la discrimination se fait dans la partie interprétable de l'analyse préliminaire (sinon il faut être méfiant, des variables discriminantes pouvant être non interprétables).
- Class scores (`dis1$gc`): les moyennes des variables canoniques par classe.

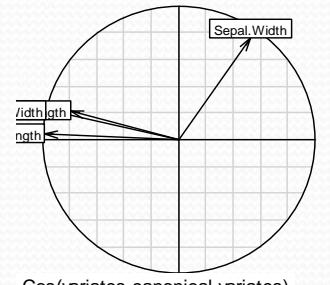
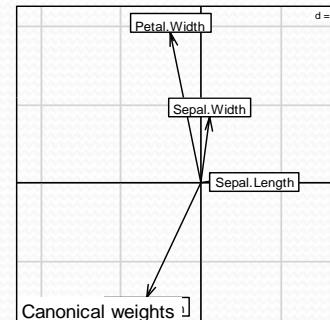
4.2. Analyse Discriminante du tableau iris

`plot(dis1)`

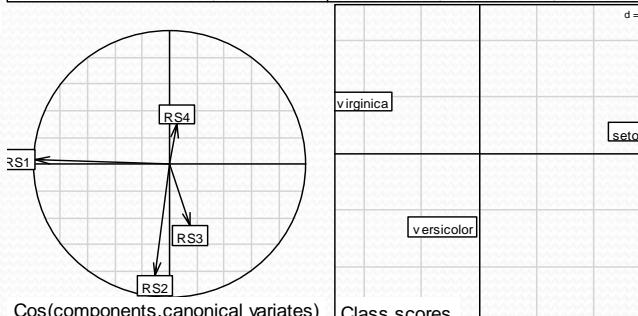
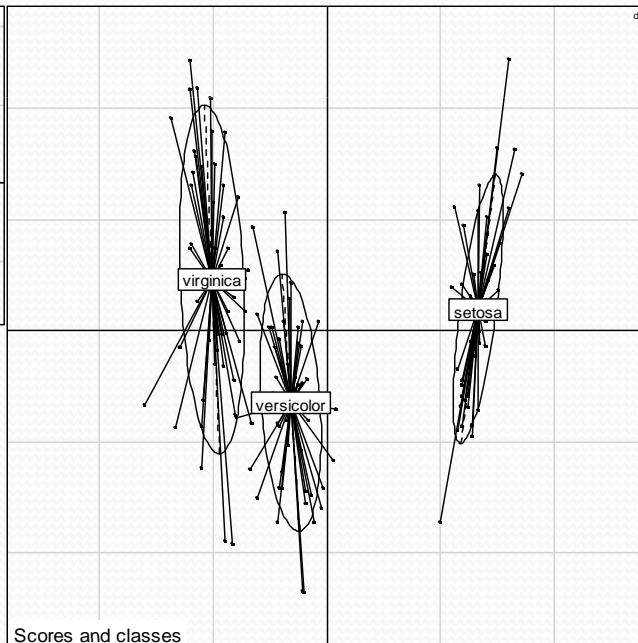
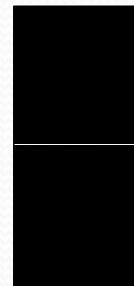
Canonical weights = loadings
`s.arrow(dis1$fa)`

Cos(variates,canonical variates)
`s.corcircle(dis1$va)`

Eigenvalues diagram
`barplot(dis1$eig)`



Eigenvalues



Cos(variates,canonical variates)
`s.corcircle(dis1$va)`

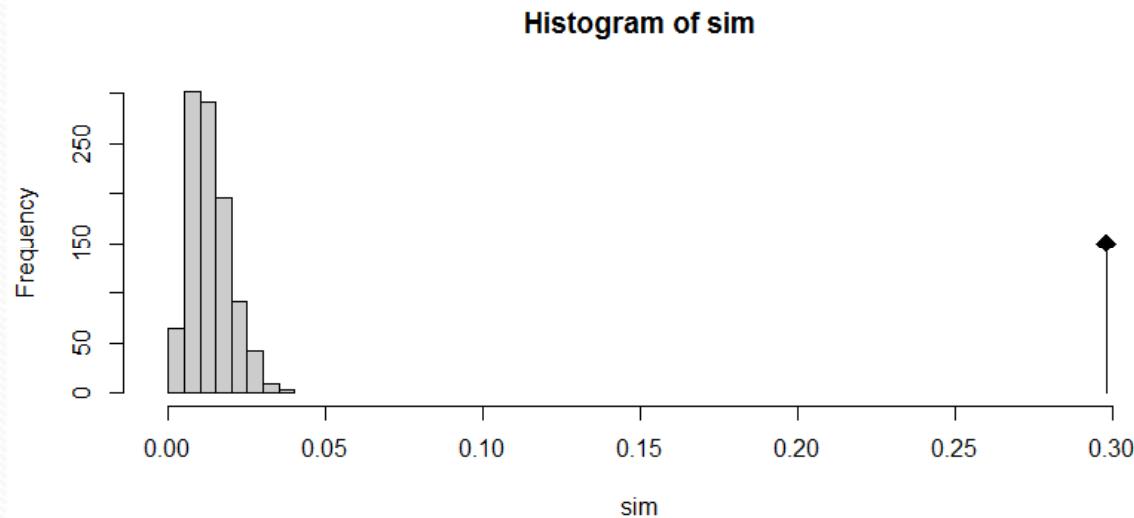
Scores and classes
`s.class(dis1$li,
iris$Species)`

Class scores
`s.label(dis1$gc)`

4.2. Analyse Discriminante du tableau iris

A la fonction discrimin() est associé un test non paramétrique de significativité **randtest()** basé sur le même principe de permutations aléatoires que celui de l'analyse inter-classes. La statistique observée est le critère de Pillai (somme des valeurs propres de l'analyse discriminante) divisé par le rang de l'analyse de départ :

```
# Valeur observée  
sum(dis1$eig)/4  
[1] 0.2979747  
  
plot(randtest(dis1))
```



4.1. Analyse Discriminante avec lda (MASS)

```
library(MASS)  
?lda  
lda(x, grouping,...)
```

x	objet de classe data frame ou matrice contenant les variables explicatives (Y)
grouping	objet (vecteur) de classe factor de longueur n, partitionnant les individus de Y

4.2. Analyse Discriminante du tableau iris

```
ldal <- lda(iris[, 1:4], iris$Species)
ldal
Call:
lda(iris[, 1:4], iris$Species)

Prior probabilities of groups:
  setosa versicolor virginica
0.3333333 0.3333333 0.3333333

Group means:
  Sepal.Length Sepal.Width Petal.Length Petal.Width
setosa        5.006      3.428      1.462      0.246
versicolor    5.936      2.770      4.260      1.326
virginica     6.588      2.974      5.552      2.026

Coefficients of linear discriminants:
            LD1         LD2
Sepal.Length 0.8293776 0.02410215
Sepal.Width   1.5344731 2.16452123
Petal.Length -2.2012117 -0.93192121
Petal.Width   -2.8104603 2.83918785

Proportion of trace:
LD1       LD2
0.9912 0.0088
```

lدا() fournit une combinaison linéaire des variables de départ, avec les coefficients qui sont dans la colonne LD1

```
ldal$scaling
          LD1         LD2
Sepal.Length 0.8293776 0.02410215
Sepal.Width   1.5344731 2.16452123
Petal.Length -2.2012117 -0.93192121
Petal.Width   -2.8104603 2.83918785
```

On peut calculer cette combinaison w1 :

```
w1 <- as.vector(as.matrix(iris[, 1:4]) %*% ldal$scaling[, 1]) # %*% = produit matriciel
head(w1)
[1] 5.956693 5.023581 5.384722 4.708094 6.027203 5.596840
tail(w1)
[1] -8.952466 -7.750110 -7.284671 -7.072847 -7.991252 -6.788261
```

Plan

- **1. Introduction**
 - Contexte et objectifs de l'analyse multivariée
 - Ordination ou classification ?
 - Ordination basée sur les distances ou sur les valeurs propres (analyses factorielles)
 - Les packages R spécialisés : ade4 et FactoMineR
- **2. Analyses factorielles à un tableau de données, avec les packages ade4 et FactoMineR**
 - 2.1 Analyse en Composantes Principales (ACP) : plusieurs variables quantitatives
 - 2.2 Analyse Factorielle des Correspondances (AFC) : deux variables qualitatives
 - 2.3 Analyse des Correspondances Multiples (ACM) : plusieurs variables qualitatives
- **3. Classification Automatique**
 - 3.1 Classification Ascendante Hiérarchique (CAH) avec la fonction hclust
 - 3.2 Partitionnement avec la fonction kmeans
- **4. Méthodes de couplage de tableaux avec le package ade4**
 - 4.1 Analyses Inter et Intra-classes
 - 4.2 Analyse Discriminante
 - **4.3 Analyses sur variables Instrumentales (ACPVI ou RDA, AFCVI ou CCA)**
 - 4.4 Analyse de Co-inertie
- **5. Aperçu des méthodes d'analyse multi-tableaux**
 - 5.1 Analyse Triadique Partielle
 - 5.2 STATIS
 - 5.3 Analyse Factorielle Multiple

4.3. Analyse sur variables instrumentales

- Les analyses sur variables instrumentales sont des **méthodes de couplage de tableaux dissymétriques**. Elles permettent de coupler
 - un tableau Y de variables à expliquer, préalablement soumis à une analyse de type dudi (ACP, AFC...),
 - un tableau X de variables explicatives.
- En écologie on se place dans le cas de l'analyse d'un tableau d'observations floristiques ou faunistiques que l'on cherche à projeter sur un tableau de variables environnementales mesurées sur les mêmes observations.
- Lorsque Y est analysé par ACP, on parle d'**ACPVI**, encore appelée **RDA** (Redundancy Analysis).
- Lorsque Y est analysé par une AFC, c'est une **AFCVI**, encore appelée **CCA** (Canonical Correspondence Analysis).
- Toutes ces analyses peuvent être réalisées avec la même fonction **pcaiv()** d'ade4.
- Remarque : les analyses inter et intra-classes (bca et wca) sont des cas particuliers d'analyses sur variables instrumentales, où le tableau X contient une seule variable qualitative. Dans ce cas, prédire Y par X revient à remplacer la valeur d'une variable pour un individu par la moyenne des individus de la même classe pour la même variable. L'analyse inter-classes est l'analyse de ce tableau de moyennes. Elle recherche des combinaisons des variables de Y maximisant la variance inter-classes. L'analyse intra-classes, qui étudie ce qui reste une fois enlevé l'effet de la variable qualitative est l'ACPVI orthogonale.

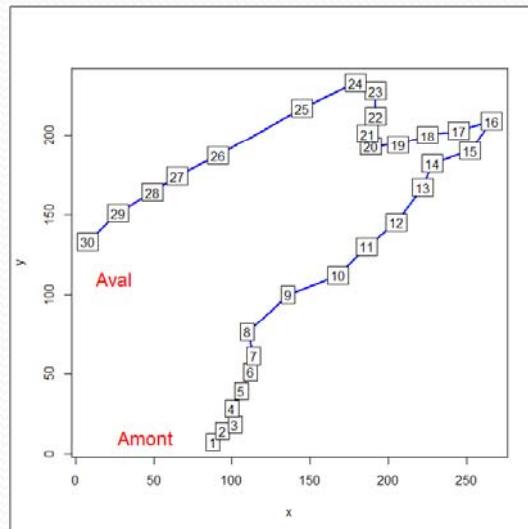
4.3. Analyse sur variables instrumentales avec pcaiv (ade4)

```
library(ade4)
pcaiv(dudi, df, scannf = TRUE, nf = 2)
```

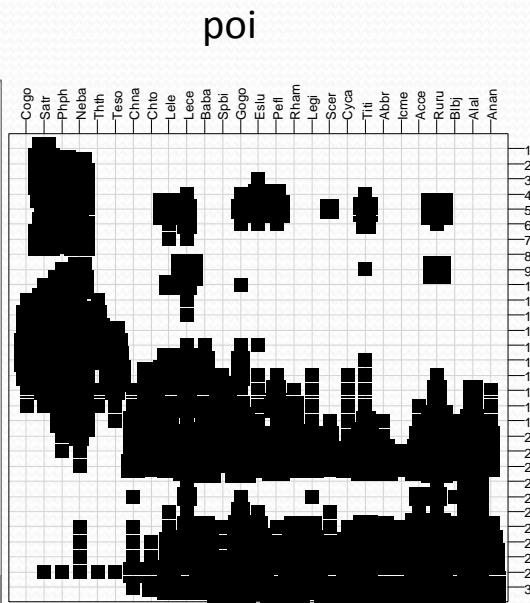
dudi	objet de classe dudi issu de l'analyse factorielle du tableau Y (n,p) avec ade4 : dudi.pca, dudi.coa, dudi.mca...
df	objet de classe data frame (n,q), portant sur les mêmes individus que Y et q variables explicatives
scannf	une valeur logique (TRUE/FALSE) indiquant si le diagramme des valeurs propres doit être affiché
nf	si scannf =FALSE, un entier indiquant le nombre d'axes à conserver (2 par défaut)

4.3. ACPVI avec pcaiv (ade4) sur données Doubs

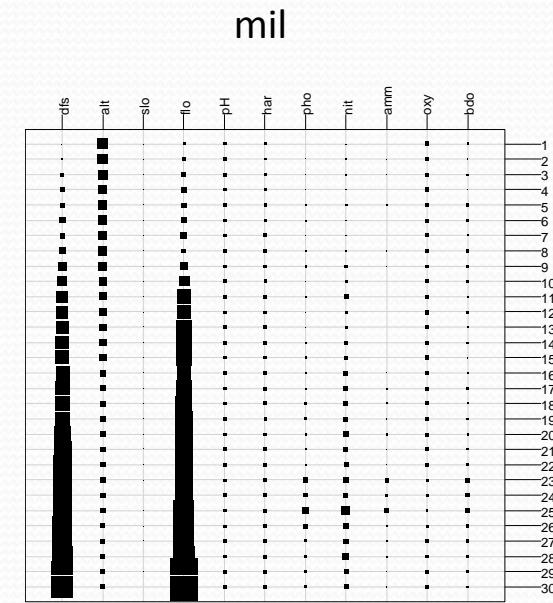
```
library(ade4)
data(doubs)
poi <- doubs$fish
mil <- doubs$env
```



Echantillonnage amont-aval



Abondances de 27 espèces de poissons



Mesures de 11 variables environnementales

Question : expliquer la structure faunistique à partir des variables environnementales ?

4.3. ACPVI avec pcaiv (ade4) sur données Doubs

```
pcafau <- dudi.pca(poi, scale = F, scannf = F, nf = 2)
pcaivdoub <- pcaiv(pcafau, mil, scannf = F, nf = 2)
pcaivdoub
Principal Component Analysis with Instrumental Variables
call: pcaiv(dudi = pcafau, df = mil, scannf = F, nf = 2)
class: pcaiv dudi

$rank (rank)      : 11
$nf (axis saved) : 2

eigen values: 38.42 5.954 2.416 1.339 0.7431 ...

vector length mode   content
$eig   11    numeric eigen values
$lw    30    numeric row weights (from dudi)
$cw    27    numeric col weights (from dudi)

data.frame nrow ncol content
$Y       30    27    Dependant variables
$X       30    11    Explanatory variables
$tab     30    27    modified array (projected variables)

data.frame nrow ncol content
$c1      27    2    PPA Pseudo Principal Axes
$as      2      2    Principal axis of dudi$tab on PAP
$ls      30    2    projection of lines of dudi$tab on PPA
$li      30    2    $ls predicted by X

data.frame nrow ncol content
$fa      12    2    Loadings (CPC as linear combinations of X
$ll      30    2    CPC Constraint Principal Components
$co      27    2    inner product CPC - Y
$cor     11    2    correlation CPC - X
```

4.3. ACPVI avec pcaiv (ade4) sur données Doubs

X: le tableau des variables explicatives

Y: le tableau des variables dépendantes

tab : le tableau des modèles linéaires des colonnes de Y par X (variables projetées) – mêmes dimensions que Y

eig : les valeurs propres, optimum du critère "somme des carrés des corrélations entre CPC et variables de Y".

cw : le poids des colonnes provenant du dudi (1 pour chacune des m colonnes de Y).

lw : le poids des lignes provenant du dudi (1/n pour chacune des n colonnes de Y).

Il existe deux possibilités pour interpréter une ACPVI :

Point de vue 1:

L'analyse recherche des coefficients ou loadings (**fa**) des variables de X. La combinaison linéaire obtenue est une composante principale ou composante explicative (**I1**). Cette composante explicative maximise la somme des carrés de corrélations (si Y est analysé par une ACP normée) ou de covariances (dans le cas d'une ACP centrée) avec les variables de milieu. Les colonnes de Y sont alors représentées par leurs corrélations ou covariances (**co**) avec la composante explicative. Les corrélations entre X et la composante explicative sont dans **cor**.

Point de vue 2:

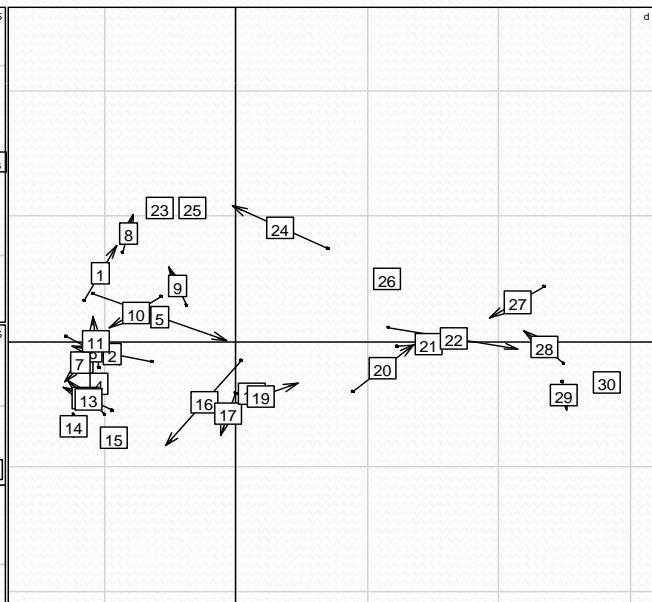
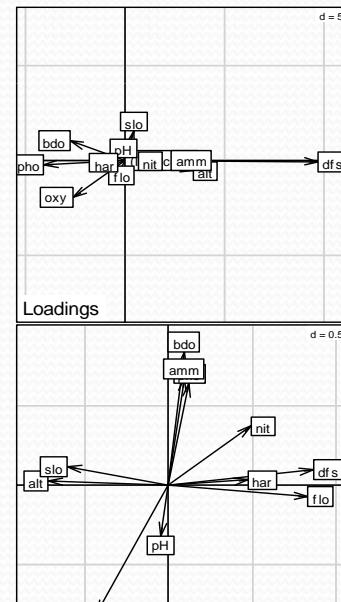
La deuxième interprétation de l'ACPVI consiste à calculer un pseudo axe principal ou PAP (**c1**). Les lignes de Y sont projetées sur les pseudo axes principaux. Ces projections ou coordonnées (**ls**) sont des combinaisons des variables de Y maximisant la variance expliquée par X. Les prédictions des coordonnées des projections des lignes de Y sur les PAP par régressions multiples sur X sont contenues dans **li**. Ces régressions définissent des carrés de corrélation multiple ou pourcentage de variance expliquée. On peut superposer ls (projections sur les PAP) et li (prédictions des positions).

4.3. ACPVI avec pcaiv (ade4) sur données Doubs

`plot(pcaivdoubs)`

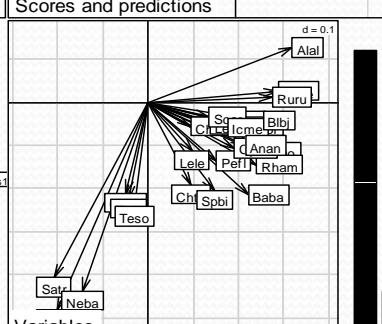
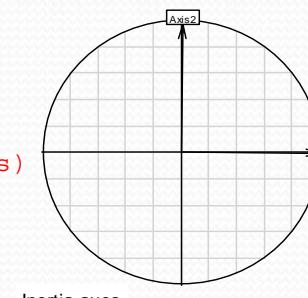
Loadings ou coefficients

`s.arrow(pcaivdoubs$fa)`



Correlations

`s.arrow(pcaivdoubs$cor)`



Eigenvalues

Inertia axes

`s.corcircle(pcaivdoubs$as)`

Scores and predictions

`s.match(pcaivdoubs$li, pcaivdoubs$ls)`

Variables

`s.arrow(pcaivdoubs$c1)`

Eigenvalues diagram

`barplot(pcaivdoubs$eig)`

4.3. ACPVI avec pcaiv (ade4) sur données Doubs

```
summary(pcaivdoub)
Principal component analysis with instrumental variables

Class: pcaiv dudi
Call: pcaiv(dudi = pcafau, df = mil, scannf = F, nf = 2)

Total inertia: 50.26

Eigenvalues:
    Ax1     Ax2     Ax3     Ax4     Ax5
38.4177  5.9540  2.4162  1.3387  0.7431

Projected inertia (%):
    Ax1     Ax2     Ax3     Ax4     Ax5
76.441  11.847  4.808   2.664   1.478

Cumulative projected inertia (%):
    Ax1     Ax1:2   Ax1:3   Ax1:4   Ax1:5
76.44   88.29   93.10   95.76   97.24

(Only 5 dimensions (out of 11) are shown)

Total unconstrained inertia (pcafau): 66.08

Inertia of pcafau explained by mil (%): 76.06

Decomposition per axis:
  iner inercum inerC inercumC ratio      R2 lambda
1 42.75     42.7 42.59      42.6 0.996 0.902 38.42
2  8.16     50.9  7.76      50.4 0.989 0.767  5.95
```

iner = valeurs propres de l'ACP simple

inercum = valeurs propres cumulées de l'ACP simple

inerC = somme (pondérée par les **lw**) des carrés des coordonnées des lignes de l'ACPVI (**ls**)

inercumC = inerC cumulés

ratio = rapport inerC / iner

R2 = carré de corrélation multiple

lambda = valeurs propres de l'ACPVI

L'analyse simple (ACP) de Y trouve des combinaisons des variables de Y de variance maximale (**iner** et **inercum** en cumulé).

Les valeurs propres de l'ACPVI (**lambda**) sont des variances expliquées. Elles correspondent au produit de la variance (**inerC**) par le carré de la corrélation multiple (**R2**). Exemple pour l'axe 1 : $38.42 = 42.6 * 0.902$

En maximisant un compromis (la variance expliquée), on rajoute une contrainte (prédiction par les variables de X) et la maximisation de la variance n'est donc plus optimale (elle l'est pour l'analyse simple). On mesure l'importance de cette contrainte par le **ratio** des variances des combinaisons des variables de Y des deux analyses

```
# inerC
sum(pcaivdoub$lw * pcaivdoub$ls[, 1]^2)
[1] 42.59456
```

```
# ratio inerC/iner
sum(pcaivdoub$lw * pcaivdoub$ls[, 1]^2)/pcafau$eig[1]
[1] 0.9964509
```

4.3. AFCVI avec pcaiv (ade4) sur données Doubs

```

coafau <- dudi.coa(poi, scannf = F, nf = 2)
coadoubs <- pcaiv(coafau, mil, scannf = F, nf = 2)
coadoubs
plot(ccadoubs)
summary(ccadoubs)
Canonical correspondence analysis

Class: caiv pcaiv dudi
Call: pcaiv(dudi = coafau, df = mil, scannf = F, nf = 2)

Total inertia: 0.8369

Eigenvalues:
    Ax1     Ax2     Ax3     Ax4     Ax5
0.53452 0.12184 0.06870 0.04917 0.02709

Projected inertia (%):
    Ax1     Ax2     Ax3     Ax4     Ax5
63.872 14.559  8.210  5.875  3.237

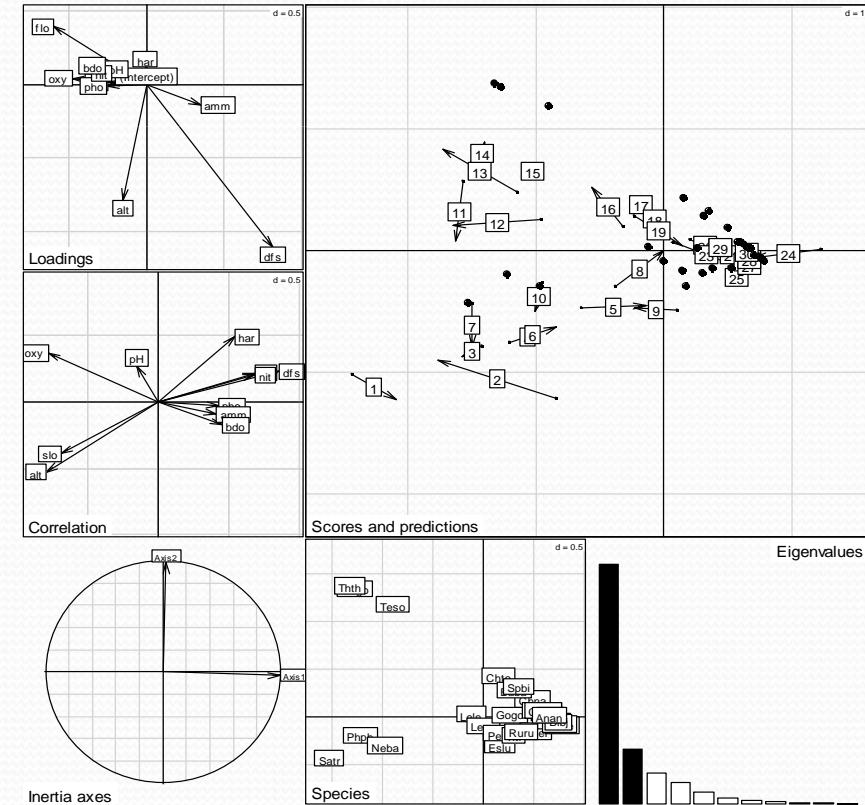
Cumulative projected inertia (%):
    Ax1   Ax1:2   Ax1:3   Ax1:4   Ax1:5
63.87   78.43   86.64   92.52   95.75

(Only 5 dimensions (out of 11) are shown)

Total unconstrained inertia (coafau): 1.167

Inertia of coafau explained by mil (%): 71.7

```



Dans l'**ACFVI**, Analyse Factorielle des Correspondances sur Variables Instrumentales (ou CCA, Canonical Correspondence Analysis), l'analyse appliquée au tableau Y est une **AFC**. On procède alors au couplage de cette AFC avec le tableau de variables instrumentales. A cette différence près, le déroulement de l'analyse est identique à celui de l'ACPVI et son dépouillement également.

4.3. AFCVI avec pcaiv (ade4) sur données Doubs

X: le tableau des variables explicatives

Y: le tableau des variables dépendantes

tab : le tableau des modèles linéaires des colonnes de Y par X (variables projetées) – mêmes dimensions que Y

eig : les valeurs propres, optimum du critère "somme des carrés des corrélations entre CPC et variables de Y".

cw : le poids des colonnes provenant du dudi (1 pour chacune des m colonnes de Y).

lw : le poids des lignes provenant du dudi (1/n pour chacune des n colonnes de Y).

Point de vue 1:

L'analyse recherche des coefficients ou loadings (**fa**) des variables de X. La combinaison linéaire obtenue est une composante principale ou composante explicative (**l1**). C'est un score des relevés de variance unité, combinaison linéaire des variables de milieu. Les espèces (**co**) sont positionnées à la moyenne des relevés. L'analyse maximise la variance des moyennes conditionnelles par un double centrage. Cette vision est parfaitement adaptée à la vision de la **niche écologique** et des **gradients environnementaux** sur lesquels se séparent les niches des espèces. On réalise une **Analyse Canonique des Correspondances (CCA)** selon **Ter Braak, 1986**.

Point de vue 2:

Il existe un deuxième point de vue qui consiste à calculer un pseudo axe principal (**c1**). Les lignes de Y (sites) sont projetées sur les pseudo-axes principaux et positionnés à la moyenne des espèces qu'ils contiennent (**ls**). Les prédictions de ces projections par X sont contenues dans **li**. Ce deuxième point de vue est celui de **l'AFCVI de Lebreton et al., 1991**.

Exercice 4 : AFCVI

Faune - Environnement

- Faire l'AFCVI entre le l'AFC de Faulog et les 5 variables environnementales

Plan

- **1. Introduction**
 - Contexte et objectifs de l'analyse multivariée
 - Ordination ou classification ?
 - Ordination basée sur les distances ou sur les valeurs propres (analyses factorielles)
 - Les packages R spécialisés : ade4 et FactoMineR
- **2. Analyses factorielles à un tableau de données, avec les packages ade4 et FactoMineR**
 - 2.1 Analyse en Composantes Principales (ACP) : plusieurs variables quantitatives
 - 2.2 Analyse Factorielle des Correspondances (AFC) : deux variables qualitatives
 - 2.3 Analyse des Correspondances Multiples (ACM) : plusieurs variables qualitatives
- **3. Classification Automatique**
 - 3.1 Classification Ascendante Hiérarchique (CAH) avec la fonction hclust
 - 3.2 Partitionnement avec la fonction kmeans
- **4. Méthodes de couplage de tableaux avec le package ade4**
 - 4.1 Analyses Inter et Intra-classes
 - 4.2 Analyse Discriminante
 - 4.3 Analyses sur variables Instrumentales (ACPVI ou RDA, AFCVI ou CCA)
 - **4.4 Analyse de Co-inertie**
- **5. Aperçu des méthodes d'analyse multi-tableaux**
 - 5.1 Analyse Triadique Partielle
 - 5.2 STATIS
 - 5.3 Analyse Factorielle Multiple

4.4. Analyse de coinertie

- L'analyse de **coinertie** (ou costructure) propose une approche symétrique du couplage de tableaux. Chacun des tableaux (Y et X) fait l'objet d'une analyse factorielle préalable de type dudi. L'analyse de coinertie est une approche unifiée qui regroupe entre autres méthodes :
 - **l'analyse inter-batterie** de Tucker, 1958 (couplage ACP-ACP),
 - **l'analyse canonique sur variables explicatives** de Cazes, 1980 (couplage ACM-ACM),
 - **l'analyse des correspondances d'un tableau de profils écologiques** (Romane, 1972).
- L'analyse de coinertie a été proposée par **Dolédec et Chessel (1994)** et par **Dray, Chessel et Thioulouse (2003)**. Elle est disponible uniquement dans la librairie ade4 par la fonction **coinertia()**.

Dolédec, S. and Chessel, D. (1994) Co-inertia analysis: an alternative method for studying species-environment relationships. *Freshwater Biology*, **31**, 277–294.

Dray, S., Chessel, D. and J. Thioulouse (2003) Co-inertia analysis and the linking of the ecological data tables. *Ecology*, **84**, 11, 3078–3089.
- L'analyse de coinertie présente deux avantages :
 - elle permet le couplage de n'importe quel type d'analyses factorielles (par exemple le cas où les variables de X sont qualitatives).
 - elle n'est pas sensible à la proportion entre nombre de lignes et nombre de variables du tableau X, contrairement aux analyses sur variables instrumentales où un trop petit nombre d'observations par rapport au nombre de variables de X peut conduire à un résultat non fiable, en raison de la surparamétrisation du modèle (comme en régression linéaire multiple).

4.3. Analyse de coinertie (ade4)

```
library(ade4)
coinertia(dudiX, dudiY, scannf = TRUE, nf = 2)
```

dudiX	objet de classe dudi issu de l'analyse factorielle du tableau X(n,q) avec ade4 : dudi.pca, dudi.coa, dudi.mca...
dudiY	objet de classe dudi issu de l'analyse factorielle du tableau Y(n,p) avec ade4 : dudi.pca, dudi.coa, dudi.mca...
scannf	une valeur logique (TRUE/FALSE) indiquant si le diagramme des valeurs propres doit être affiché
nf	si scannf =FALSE, un entier indiquant le nombre d'axes à conserver (2 par défaut)

4.4. Analyse de coinertie

- Dans la pratique, on commence par effectuer une analyse simple de type dudi (ACP, AFC, ACM) sur chaque tableau de données à coupler (Y et X), puis on couple les deux analyses en les appelant par **coinertia(dudiX, dudiY)**.
- Une contrainte à respecter est que les deux analyses doivent avoir la **même pondération sur les lignes**. Cela n'est pas un problème en ACP ou en ACM, mais dans le cas du couplage entre une AFC et une ACP, il faut imposer à l'ACP la pondération des lignes de l'AFC.
- Le coefficient '**RV**' (Escoufier, 1973) est un coefficient de **corrélation vectorielle** entre les deux tableaux, extension du coefficient de corrélation entre deux variables. Il permet de juger de la ressemblance entre les deux tableaux.
- Le tableau croisé '**tab**' donne les covariances entre les variables de Y et celles de X .
- C'est ce tableau qui est diagonalisé pour en extraire des axes, sur lesquels on pourra projeter les variables de X (**co**) et les variables de Y (**li**).
- Enfin on pourra projeter en éléments supplémentaires sur ces axes les lignes du tableau X (**IX**) et du tableau Y (**IY**) et visualiser graphiquement la ressemblance entre les deux structures.

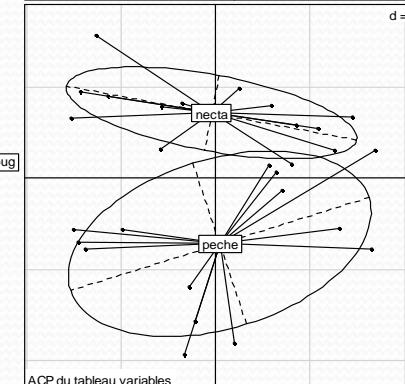
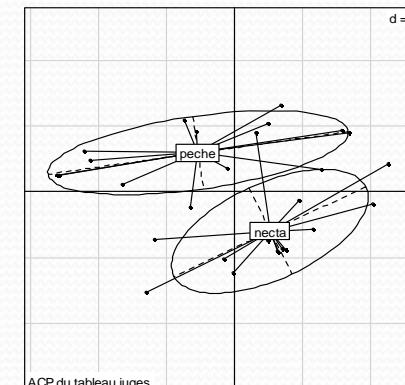
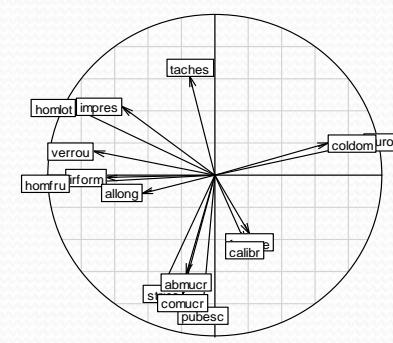
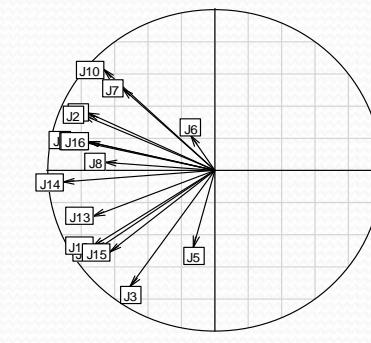
4.4. Analyse de coinertie - Couplage ACP-ACP sur fruits

- On prendra pour exemple le jeu de données **fruits** d'ade4, portant sur 28 lots de fruits (pêches et nectarines) jugés de deux manières différentes :
 - Le classement par ordre de préférence sans ex aequo par 16 juges (**fruits\$jug**)
 - 15 variables quantitatives décrivant chaque lot (**fruits\$var**)
 - fruits\$typ** est un vecteur de longueur 28 identifiant chaque lot
- Sur chacun des tableaux **jug** et **var** on effectue d'abord une ACP normée que l'on dépouille

```

data(fruits)
?fruits
pcajug <- dudi.pca(fruits$jug, scann = F)
pcavar <- dudi.pca(fruits$var, scann = F)

par(mfrow = c(2,2))
s.corcircle(pcajug$co,
            sub="ACP du tableau juges")
s.class(pcajug$li, fac = fruits$type,
         sub="ACP du tableau juges")
s.corcircle(pcavar$co,
            sub="ACP du tableau variables")
s.class(pcavar$li, fac = fruits$type,
         sub="ACP du tableau variables")
par(mfrow = c(1,1))
    
```



4.4. Analyse de coinertie - Couplage ACP-ACP sur fruits

- On couple ensuite ces deux ACP dans une analyse de coinertie

```
coi.fruits <- coinertia(pcajug, pcavar, scan = FALSE)
coi.fruits
$rank (rank)      : 15
$nf (axis saved) : 2
$RV (RV coeff)   : 0.4927474

eigenvalues: 15.13 5.704 2.728 0.8568 0.5648 ...

  vector length mode     content
1 $eig    15    numeric Eigenvalues
2 $lw     15    numeric Row weights (for pcavar cols)
3 $cw     16    numeric Col weights (for pcajug cols)

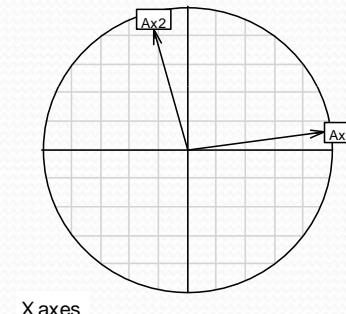
  data.frame nrow ncol content
1 $tab      15   16  Crossed Table (CT): cols(pcavar) x cols(pcajug)
2 $li       15   2   CT row scores (cols of pcavar)
3 $l1       15   2   Principal components (loadings for pcavar cols)
4 $co       16   2   CT col scores (cols of pcajug)
5 $c1       16   2   Principal axes (loadings for pcajug)
6 $lx       28   2   Row scores (rows of pcajug cols)
7 $mX      28   2   Normed row scores (rows of pcajug)
8 $ly       28   2   Row scores (rows of pcavar)
9 $mY      28   2   Normed row scores (rows of pcavar)
10 $aX      2     2   Corr pcajug axes / coinertia axes
11 $aY      2     2   Corr pcavar axes / coinertia axes

CT rows = cols of pcavar (15) / CT cols = cols of pcajug (16)
```

4.4. Analyse de coinertie - Couplage ACP-ACP sur fruits

`plot(coi.fruits)`

Projections des axes de l'ACP de X (jug)
sur les axes de coinertie
`s.corcircle(coi.fruits$aX)`



Projections des axes de l'ACP de Y
(var) sur les axes de coinertie
`s.corcircle(coi.fruits$aY)`

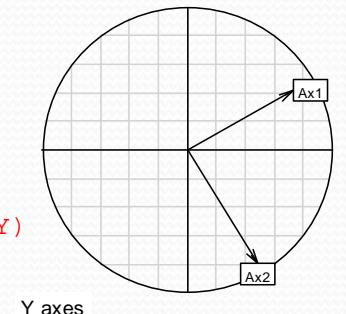
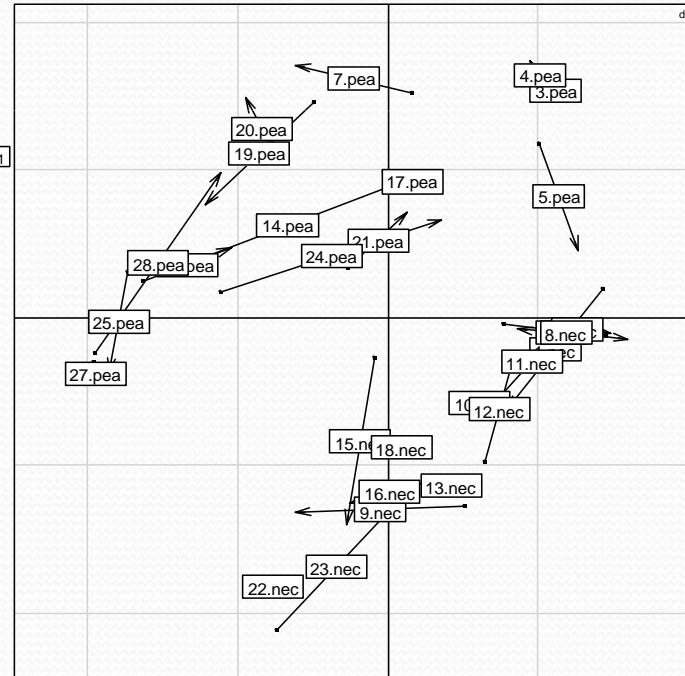
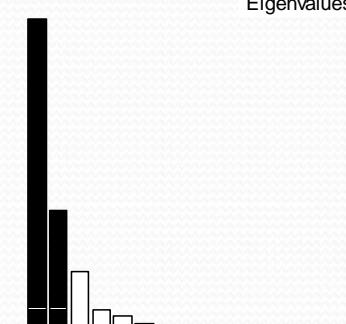
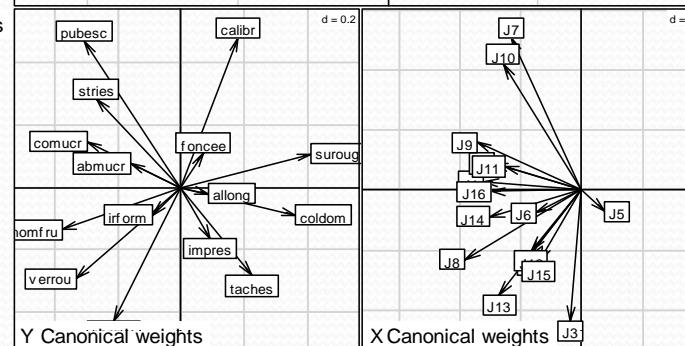


Diagramme des valeurs propres
`barplot(coi.fruits$eig)`



Projection des lignes des
tableaux X et Y
`s.match(coi.fruits$mX,
coi.fruits$mY)`



4.4. Analyse de coinertie - Couplage ACP-ACP sur fruits

```
summary(coi.fruits)
Total inertia: 25.35

Eigenvalues:
    Ax1      Ax2      Ax3      Ax4      Ax5
15.1338  5.7037  2.7282  0.8568  0.5648

Projected inertia (%):
    Ax1      Ax2      Ax3      Ax4      Ax5
59.690  22.496  10.761   3.379   2.228

Cumulative projected inertia (%):
    Ax1     Ax1:2     Ax1:3     Ax1:4     Ax1:5
59.69    82.19    92.95    96.33    98.55

(Only 5 dimensions (out of 15) are shown)
```

```
Eigenvalues decomposition:
    eig      covar      sdX      sdY      corr
1 15.133835 3.890223 2.607581 1.864335 0.8002263
2  5.703734 2.388249 1.550666 1.776134 0.8671329
```

```
Inertia & coinertia X (pcajug):
    inertia      max      ratio
1  6.799477 7.318882 0.9290322
12 9.204041 9.930650 0.9268317
```

```
Inertia & coinertia Y (pcavar):
    inertia      max      ratio
1  3.475745 4.391663 0.7914416
12 6.630397 7.620306 0.8700960
```

```
RV:
0.4927474
```

- La première valeur propre de la coinertie vaut **15.134**, pour une covariance de **3.89**, qui est le produit d'une corrélation de **0.8002** par les deux écarts types dans sdX et sdY , respectivement **2.607** et **1.864**. La corrélation (entre 0 et 1) exprime la ressemblance entre les deux ACP de départ.
- Les tableaux ‘Inertia & coinertia X’ et ‘Inertia & coinertia Y’ permettent de donner le **ratio d'inertie projetée sur le premier axe et sur le plan 1-2 pour X et pour Y**.

- Dans le cas du couplage de deux ACP normées, on peut retrouver le coefficient RV par :
`sum(cor(pcajug$tab, pcavar$tab)^2)/sqrt(sum(cor(pcajug$tab, pcajug$tab)^2) * sum(cor(pcavar$tab, pcavar$tab)^2))`
[1] 0.4927474

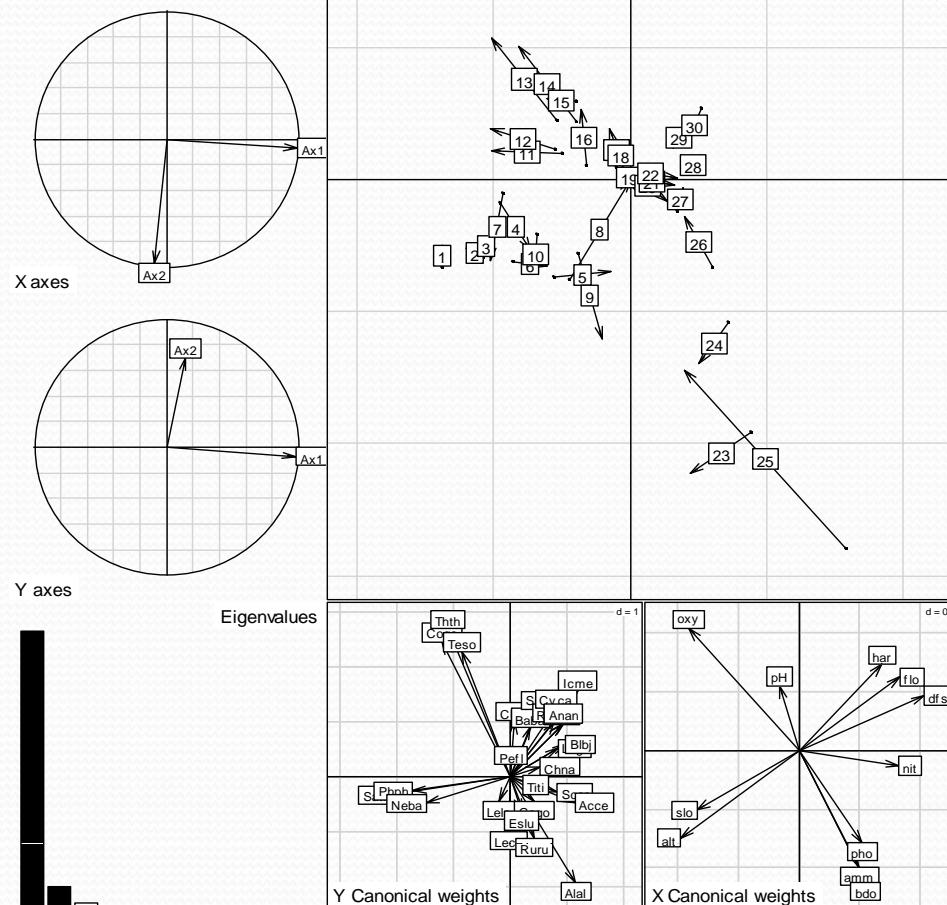
4.4. Analyse de coinertie - Couplage AFC-ACP sur Doubs

On revient sur le jeu de données sur le doubs déjà analysé en AFCVI.

Le couplage AFC-ACP nécessite d'imposer une pondération des lignes dans l'ACP à partir du vecteur de pondérations de l'AFC.

L'analyse est dépouillée comme précédemment

```
coafau <- dudi.coa(poi, scannf = F, nf = 2)
pcamil <- dudi.pca(mil, row.w = coafau$lw,
                     scannf = F, nf = 2)
coidoubs <- coinertia(pcamil, coafau,
                      scannf = F, nf = 2)
plot(coidoubs)
```



4.4. Analyse de coinertie - Couplage AFC-ACP sur Doubs

```
summary(coidoubs)
Total inertia: 2.59

Eigenvalues:
    Ax1      Ax2      Ax3      Ax4      Ax5
2.34163 0.17496 0.03947 0.01908 0.00658

Projected inertia (%):
    Ax1      Ax2      Ax3      Ax4      Ax5
90.4004 6.7545 1.5238 0.7367 0.2540

Cumulative projected inertia (%):
    Ax1    Ax1:2    Ax1:3    Ax1:4    Ax1:5
90.40    97.15    98.68    99.42    99.67

(Only 5 dimensions (out of 11) are shown)
```

```
Eigenvalues decomposition:
  eig   covar     sdX     sdY     corr
1 2.3416297 1.5302384 2.366115 0.7591393 0.8519259
2 0.1749618 0.4182844 1.533416 0.3336151 0.8176473
```

```
Inertia & coinertia X (pcamil):
  inertia     max     ratio
1 5.598498 5.727595 0.9774606
12 7.949863 8.153563 0.9750170
```

```
Inertia & coinertia Y (coafau):
  inertia     max     ratio
1 0.5762925 0.6009926 0.9589011
12 0.6875916 0.7453635 0.9224915
```

```
RV:
0.636319
```

- La première valeur propre de la coinertie vaut **2.34**, pour une covariance de **1.53**, qui est le produit d'une corrélation de **0.852** par les deux écarts types dans sdX et sdY , respectivement **2.366** et **0.759**. La corrélation (entre 0 et 1) exprime la ressemblance entre les deux ACP de départ.
- Les tableaux ‘Inertia & coinertia X’ et ‘Inertia & coinertia Y’ permettent de donner le ratio d’inertie projetée sur le premier axe et sur le plan 1-2 pour X et pour Y.

Exercice 4 : Coinertie Faune - Environnement

- Faire l'analyse de coinertie entre l'AFC de Faulog et l'ACP des 5 variables environnementales

Bibliographie - analyses à 2 tableaux

Fichiers PDF

- **tdr621_ACP_Inter_Intra.pdf** (*A.B. Dufour*)
- **tdr63_Analyses_Discriminantes.pdf** (*D. Chessel, A.B. Dufour & J. Thioulouse*)
- **tdr65_VI.pdf** (*A.B. Dufour, D. Chessel & J. Thioulouse*)
- **tdr64_coinertie.pdf** (*D. Chessel A.B. Dufour & S. Dray*)
- **stage5_couplage_tableaux.pdf** (*D. Chessel, A.B. Dufour & J. Thioulouse*)

Plan

- **1. Introduction**
 - Contexte et objectifs de l'analyse multivariée
 - Ordination ou classification ?
 - Ordination basée sur les distances ou sur les valeurs propres (analyses factorielles)
 - Les packages R spécialisés : ade4 et FactoMineR
- **2. Analyses factorielles à un tableau de données, avec les packages ade4 et FactoMineR**
 - 2.1 Analyse en Composantes Principales (ACP) : plusieurs variables quantitatives
 - 2.2 Analyse Factorielle des Correspondances (AFC) : deux variables qualitatives
 - 2.3 Analyse des Correspondances Multiples (ACM) : plusieurs variables qualitatives
- **3. Classification Automatique**
 - 3.1 Classification Ascendante Hiérarchique (CAH) avec la fonction hclust
 - 3.2 Partitionnement avec la fonction kmeans
- **4. Méthodes de couplage de tableaux avec le package ade4**
 - 4.1 Analyses Inter et Intra-classes
 - 4.2 Analyse Discriminante
 - 4.3 Analyses sur variables Instrumentales (ACPVI ou RDA, AFCVI ou CCA)
 - 4.4 Analyse de Co-inertie
- **5. Aperçu des méthodes d'analyse multi-tableaux**
 - 5.1 Analyse Triadique Partielle
 - 5.2 STATIS
 - 5.3 Analyse Factorielle Multiple

5. Analyses Multi-tableaux

- Les analyses multi-tableaux s'appliquent quand on considère **plusieurs (K) tableaux de données ayant au moins une dimension commune**:
- Si cette dimension commune est les **individus** (lignes), on peut avoir K groupes de variables (colonnes) mesurées sur les mêmes individus. Exemple: plusieurs questionnaires sur différents sujets soumis aux mêmes individus ; analyse sensorielle avec plusieurs groupes de variables...
- Si cette dimension commune est les **variables** (colonnes), on a alors plusieurs groupes d'individus sur lesquels on a réalisé les mêmes mesures. Exemple : mesures physico-chimiques le long du gradient amont aval dans différentes rivières ; mesures prises sur les élèves de différentes écoles.
- On peut aussi avoir **les mêmes individus et les mêmes variables** pour chacun des k tableaux, on se trouve alors dans le cas d'un **cube de données**. Exemple: mesures physico-chimiques réalisées sur les mêmes sites à des saisons différentes (cas de meaudret\$env)
- On peut dans tous ces cas regrouper les K tableaux en un seul et en faire une analyse simple (ACP, AFC, ACM), mais les méthodes d'analyse multi-tableaux permettent d'aller plus loin en mettant en évidence **les ressemblances / dissemblances entre les structures des individus / variables générées par les différents groupes de variables / individus**.
- Elles permettent de répondre à des questions telles que :
 - Tous les tableaux sont-ils structurés de la même façon ?
 - La typologie des individus et/ou celle des variables est-elle la même d'un tableau à l'autre ?

Analyses multi-tableaux: mise en œuvre avec R

Méthode	Variables	ade4	FactoMineR
Analyse Triadique Partielle	Cube de données (mêmes individus et mêmes variables) où chaque tableau relève d'une ACP	pta	-
AFC de Foucart	Cube de données (mêmes individus et mêmes variables) où chaque tableau relève d'une AFC	foucart	-
STATIS sur les opérateurs	Plusieurs groupes de variables sur les mêmes individus (STATIS sur les WD) ou Mêmes variables pour plusieurs groupes d'individus (STATIS sur les VQ)	statis	-
Analyse Factorielle Multiple	Plusieurs groupes de variables (quantitatives ou qualitatives) sur les mêmes individus	mfa	MFA
Analyse de coinertie multiple	Plusieurs groupes de variables (quantitatives ou qualitatives) sur les mêmes individus	mcoa	-

Plan

- **1. Introduction**
 - Contexte et objectifs de l'analyse multivariée
 - Ordination ou classification ?
 - Ordination basée sur les distances ou sur les valeurs propres (analyses factorielles)
 - Les packages R spécialisés : ade4 et FactoMineR
- **2. Analyses factorielles à un tableau de données, avec les packages ade4 et FactoMineR**
 - 2.1 Analyse en Composantes Principales (ACP) : plusieurs variables quantitatives
 - 2.2 Analyse Factorielle des Correspondances (AFC) : deux variables qualitatives
 - 2.3 Analyse des Correspondances Multiples (ACM) : plusieurs variables qualitatives
- **3. Classification Automatique**
 - 3.1 Classification Ascendante Hiérarchique (CAH) avec la fonction hclust
 - 3.2 Partitionnement avec la fonction kmeans
- **4. Méthodes de couplage de tableaux avec le package ade4**
 - 4.1 Analyses Inter et Intra-classes
 - 4.2 Analyse Discriminante
 - 4.3 Analyses sur variables Instrumentales (ACPVI ou RDA, AFCVI ou CCA)
 - 4.4 Analyse de Co-inertie
- **5. Aperçu des méthodes d'analyse multi-tableaux**
 - 5.1 Analyse Triadique Partielle
 - 5.2 STATIS
 - 5.3 Analyse Factorielle Multiple

5.1. Analyse Triadique Partielle

- **Application** : K-tableaux avec **mêmes individus et mêmes variables** = cas d'un **cube de données**
- **Exemple** : meaudret\$env (vu pour l'ACP et l'ACP inter/intra-classes) : 9 mesures physico-chimiques réalisées sur les 5 mêmes sites aux 4 saisons
- => 4 tableaux (saisons) avec les 5 sites en lignes et les 9 variables en colonnes
- **Mise en œuvre** : fonction **pta()** de la librairie ade4
- **Principe de la méthode:**
 - 1. **Interstructure** = ressemblances entre les 4 tableaux
 - 2. **Compromis** = tableau moyen (sites x variables) pondéré selon l'interstructure => Analyse (ACP)
 - 3. **Infrastructure ou Trajectoires** = projection des éléments des tableaux de départ sur les axes du compromis qui sert de référentiel commun pour comparer la structure des tableaux (ici saisons)
- **Références** : Thioulouse et Chessel, 1987 ; Blanc et al., 1998

5.1. Préparation du K-tableaux

```
data(meaudret)
# Analyse intra-saisons
wit1 <- withinpca(df = meaudret$env, fac = meaudret$design$season, scaling = "partial", scannf = FALSE)
df : a data frame with quantitative variables
fac      : a factor partitioning the rows of df in classes
scaling   : a string of characters as a scaling option :
  if "partial", the sub-table corresponding to each class is centred and normed.
  if "total", the sub-table corresponding to each class is centred and the total table is then normed.
scannf    : a logical value indicating whether the eigenvalues bar plot should be displayed
nf : if scannf FALSE, an integer indicating the number of kept axes
```

```
wit1
Within analysis
$nf (axis saved) : 2
$rank: 9
$ratio: 0.7287203
eigen values: 5.095 2.3 0.7259 0.4445 0.2553 ...
  vector length mode content
1 $eig   9     numeric eigen values
2 $lw    20    numeric row weigths
3 $cw    9     numeric col weigths
4 $stabw 4     numeric class weigths
5 $fac   20    numeric factor for grouping
```

```
  data.frame nrow ncol content
1 $stab     20   9   array class-variables
2 $li       20   2   row coordinates
3 $l1       20   2   row normed scores
4 $co       9    2   column coordinates
5 $cl       9    2   column normed scores
6 $ls       20   2   supplementary row coordinates
7 $as       4    2   inertia axis onto within axis
```

Etape préliminaire : Analyse intra-saisons de meaudret\$env avec **withinpca()**, qui permet deux options :
-partial=center et réduire chaque tableau
-total=centerer chaque tableau et réduire globalement

5.1. Préparation du K-tableaux

```
# Transformation de l'intra-saisons en objet "k-tableaux"
# càd un tableau par saison avec les variables en lignes et les sites en colonnes
# Utilisation de la fonction ktab.within
ktal <- ktab.within(dudiwit = wit1, colnames = rep(c("S1", "S2", "S3", "S4", "S5"), 4))

dudiwit      : an objet of class within
rownames : the row names of the K-tables (otherwise the row names of dudiwit$tab)
colnames : the column names of the K-tables (otherwise the column names of dudiwit$tab)
tabnames : the names of the arrays of the K-tables (otherwise the levels of the factor which defines
           the within-classes)

ktal
class: ktab
tab number: 4
  data.frame nrow ncol
1 spring      9   5
2 summer      9   5
3 autumn      9   5
4 winter      9   5
  vector length mode    content
5 $lw     9     numeric row weights
6 $cw    20     numeric column weights
7 $blo    4     numeric column numbers
8 $stabw  4     numeric array weights
  data.frame nrow ncol content
9 $TL      36    2   Factors Table number Line number
10 $TC     20    2   Factors Table number Col number
11 $T4     16    2   Factors Table number 1234
```

Etape préliminaire, suite : Transformer l'intra-saison en k-tableau avec la fonction **ktab.within**

!!! Variables en lignes et sites en colonnes

```
12 $call: ktab.within(dudiwit = wit1,
colnames = rep(c("S1", "S2",
"S3", "S4", "S5"), 4))
names :
spring : S1 S2 S3 S4 S5
summer : S1 S2 S3 S4 S5
autumn : S1 S2 S3 S4 S5
winter : S1 S2 S3 S4 S5

Col weights :
spring : 0.2 0.2 0.2 0.2 0.2
summer : 0.2 0.2 0.2 0.2 0.2
autumn : 0.2 0.2 0.2 0.2 0.2
winter : 0.2 0.2 0.2 0.2 0.2

Row weights :
1 1 1 1 1 1 1 1 1
```

5.1. Le K-tableaux

```
# Transposition du k-tableaux pour avoir les variables en colonnes et les sites en lignes
kta2 <- t(ktal)
kta2
class: ktab

tab number: 4
  data.frame nrow ncol
1 spring      5    9
2 summer     5    9
3 autumn     5    9
4 winter     5    9          => Variables en colonnes
                                et sites en lignes : OK

  vector length mode   content
5 $lw      5     numeric row weights
6 $cw     36     numeric column weights
7 $blo     4     numeric column numbers
8 $stabw   4     numeric array weights

  data.frame nrow ncol content
9 $TL       20    2 Factors Table number Line number
10 $TC      36    2 Factors Table number Col number
11 $T4      16    2 Factors Table number 1234

12 $call: t.ktab(x = ktal)

names :
spring : Temp Flow pH Cond Bdo5 Oxyd
          Ammo Nitr Phos
summer : Temp Flow pH Cond Bdo5 Oxyd
          Ammo Nitr Phos
autumn : Temp Flow pH Cond Bdo5 Oxyd
          Ammo Nitr Phos
winter : Temp Flow pH Cond Bdo5 Oxyd
          Ammo Nitr Phos

Col weights :
spring : 1 1 1 1 1 1 1 1 1 1
summer : 1 1 1 1 1 1 1 1 1 1
autumn : 1 1 1 1 1 1 1 1 1 1
winter : 1 1 1 1 1 1 1 1 1 1

Row weights :
0.2 0.2 0.2 0.2 0.2
```

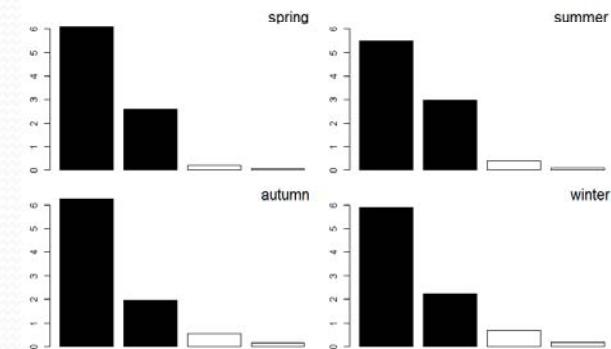
5.1. Analyses séparées

```
# Optionnel (juste pour voir) : Analyses séparées de chaque tableau
sep1 <- sepan(kta2)

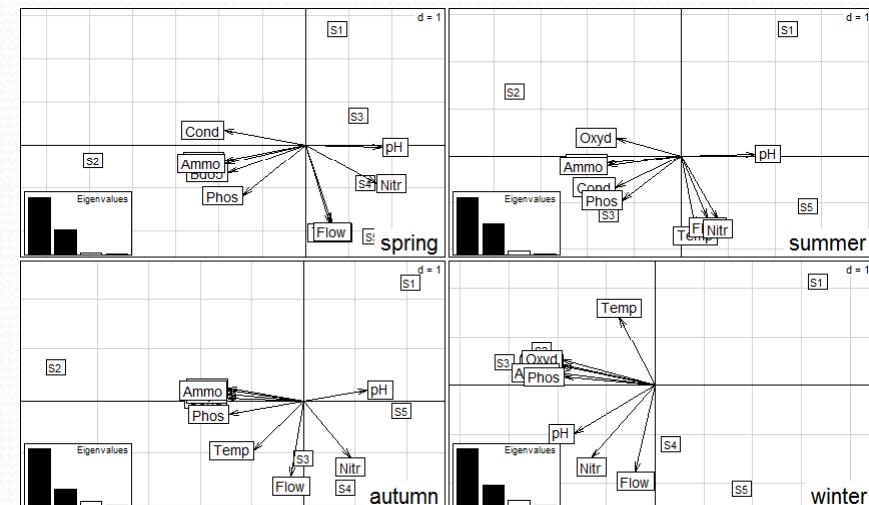
sep1
class: sepan list
$call: sepan(X = kta2)
  vector length mode content
1 $tab.names 4 character tab names
2 $blo      4 numeric column number
3 $rank     4 numeric tab rank
4 $Eig      16 numeric All the eigen values
  data.frame nrow ncol content
1 $Li        20  2 row coordinates
2 $L1        20  2 row normed scores
3 $Co        36  2 column coordinates
4 $C1        36  2 column normed coordinates
5 $TL        20  2 factors for Li L1
6 $TC        36  2 factors for Co C1
summary(sep1)
Separate Analyses of a 'ktab' object
  names nrow ncol rank lambda1 lambda2 lambda3 lambda4
1 spring  5    4   4   6.1    2.603   0.212   0.085
2 summer  5    4   4   5.505   2.983   0.402   0.11
3 autumn  5    4   4   6.328   1.969   0.552   0.152
4 winter  5    4   4   5.887   2.237   0.695   0.181
```

sepan() pour "separate analyses"

plot(sep1)



kplot(sep1)



5.1. Analyse Triadique Partielle – Interstructure

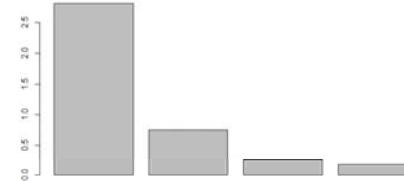
```
# Réalisation de l'Analyse triadique partielle
pta1 <- pta(kta2, scann = FALSE)
pta1
Partial Triadic Analysis
class:pta dudi
table number: 4
row number: 5    column number: 9

***** Interstructure *****

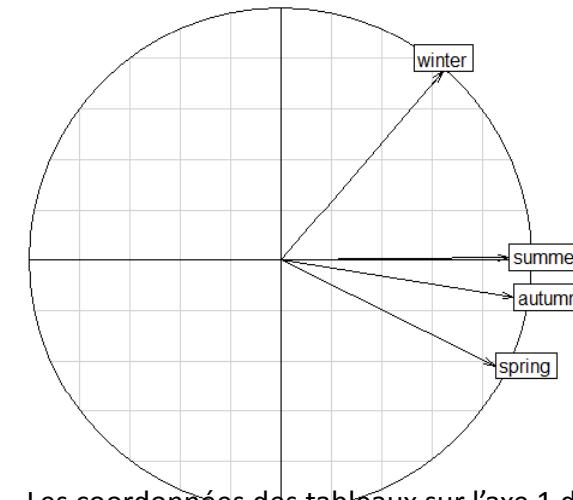
eigen values: 2.812 0.7541 0.2537 0.18
$RV      matrix      4      4      RV coefficients
$RV.eig   vector      4      eigenvalues
$RV.coo   data.frame  4      4      array scores
$stab.names   vector      4      array names
(...)
# 1) Interstructure ($RV)
# La matrice RV donne les corrélations (vectorielles) entre tableaux
pta1$RV
spring     summer    autumn   winter
spring 1.0000000 0.6934558 0.7886185 0.2834592
summer  0.6934558 1.0000000 0.7671756 0.5340456
autumn   0.7886185 0.7671756 1.0000000 0.4794976
winter   0.2834592 0.5340456 0.4794976 1.0000000
```

INTERSTRUCTURE

```
barplot(pta1$RV.eig)
```



```
s.corcircle(pta1$RV.coo)
```



Les coordonnées des tableaux sur l'axe 1 de l'interstructure seront utilisées pour pondérer les tableaux dans le calcul du tableau moyen compromis => favorise les tableaux qui se ressemblent

5.1. Analyse Triadique Partielle – Compromis

```
# 2) Compromis = tableau "moyen" sites X especes  
**** Compromise ****
```

```
eigen values: 17.2 7.298 0.6099 0.2008
```

```
$nf: 2 axis-components saved
```

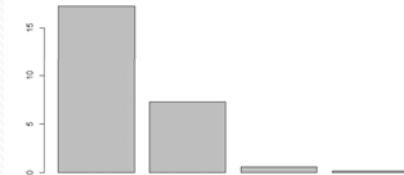
```
$rank: 4
```

```
vector length mode content  
$tabw 4 numeric array weights  
$cw 9 numeric column weights  
$lw 5 numeric row weights  
$eig 4 numeric eigen values  
$cos2 4 numeric cosine^2 between  
compromise and arrays
```

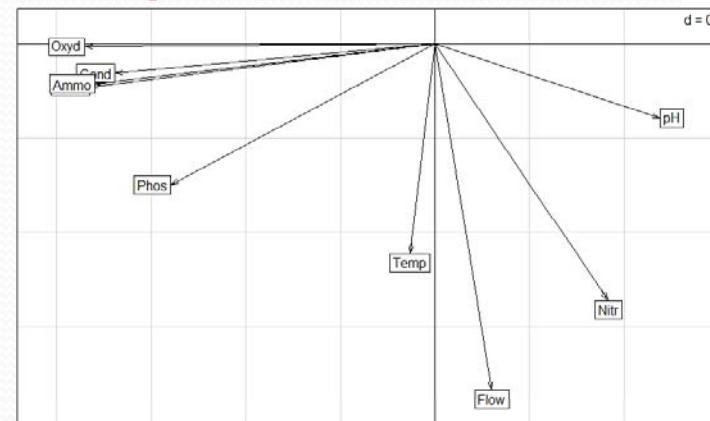
```
data.frame nrow ncol content  
$tab 5 9 modified array  
$li 5 2 row coordinates  
$l1 5 2 row normed scores  
$co 9 2 column coordinates  
$cl 9 2 column normed scores
```

COMPROMIS

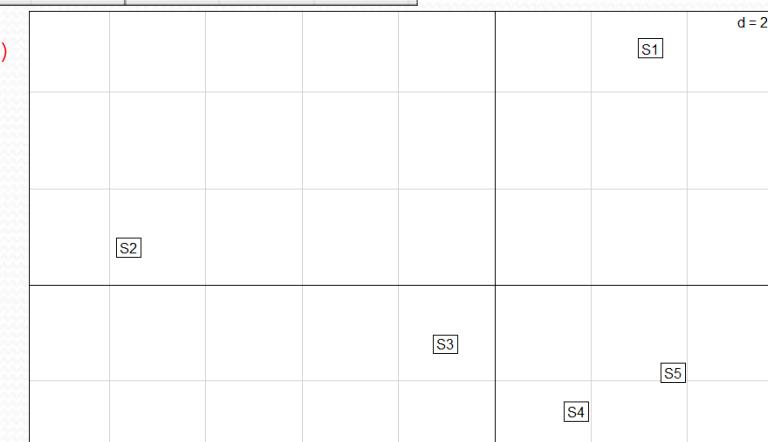
```
barplot(ptal$eig)
```



```
s.arrow(ptal$co)
```



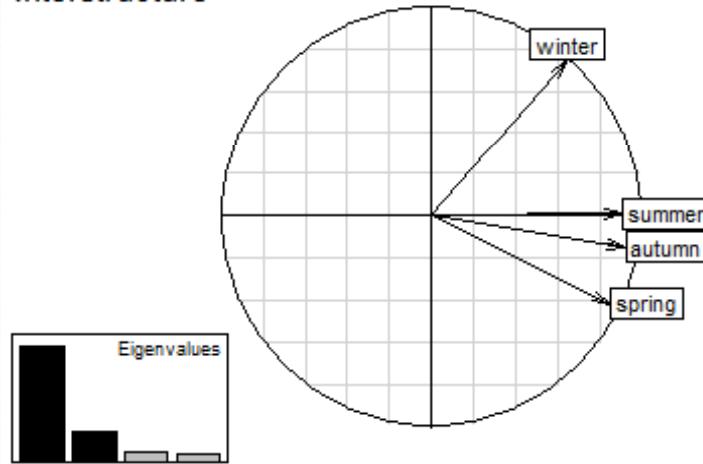
```
s.label(ptal$li)
```



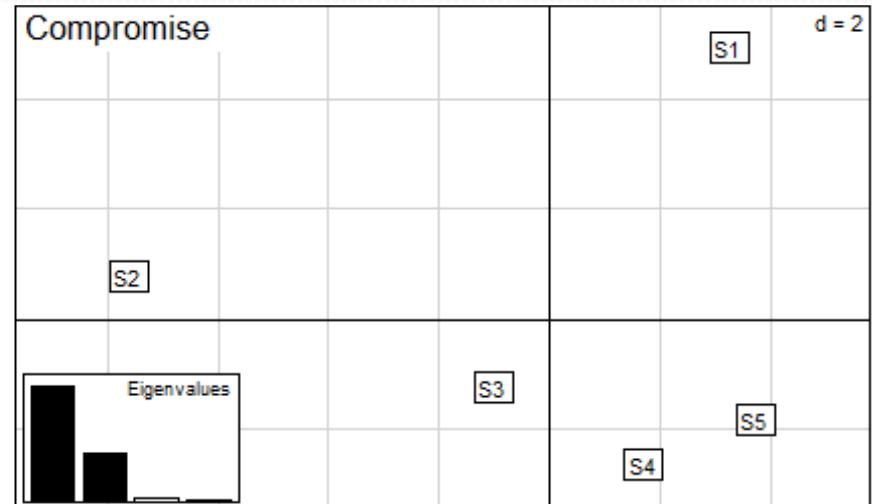
5.1. Analyse Triadique Partielle – résumé

`plot(pta1)`

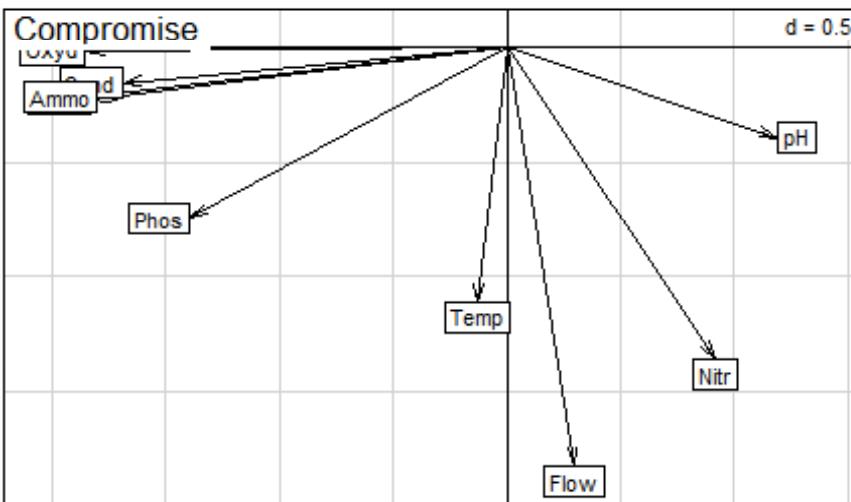
Interstructure



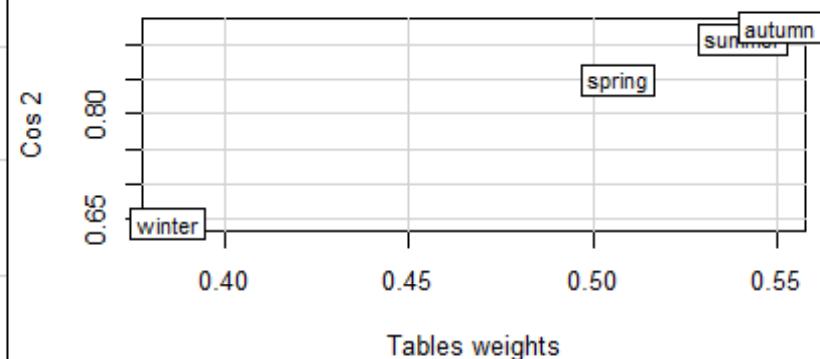
Compromise



Compromise



Typological value



5.1. Analyse Triadique Partielle – Inrastructure

```
# 3) Inrastructure ou Trajectoires :  
# projections des éléments des tableaux de départ sur les axes du compromis  
# qui sert de référentiel commun pour comparer la structure des tableaux (ici saisons)
```

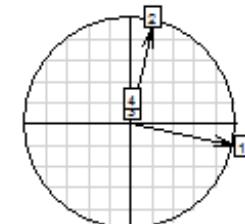
**** Inrastructure ****

```
data.frame nrow ncol content  
$Tli      20   2   row coordinates (each table)  
$Tco      36   2   col coordinates (each table)  
$Tcomp    16   2   principal components (each table)  
$Tax      16   2   principal axis (each table)  
$TL       20   2   factors for Tli  
$TC       36   2   factors for Tco  
$T4       16   2   factors for Tax Tcomp
```

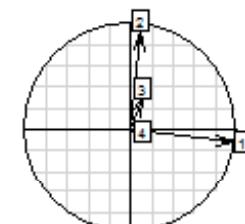
INRASTRUCTURE

5.1. Analyse Triadique Partielle – Infrastructure

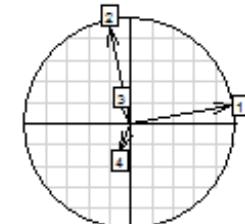
`kplot(pta1)`



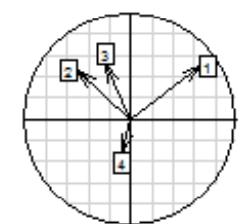
spring



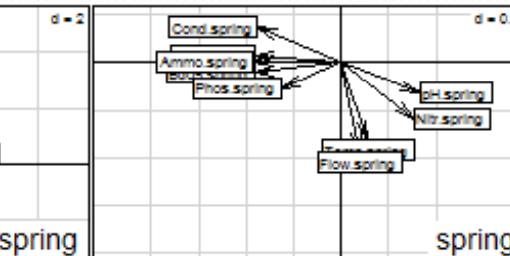
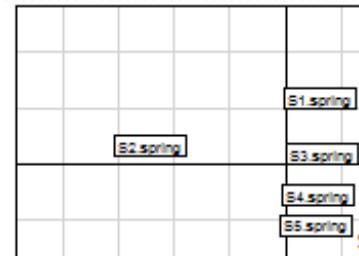
summer



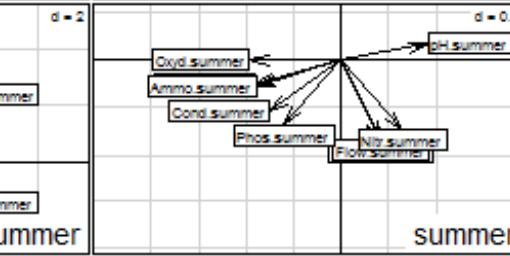
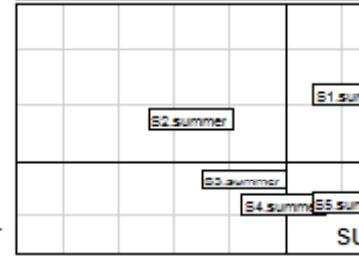
autumn



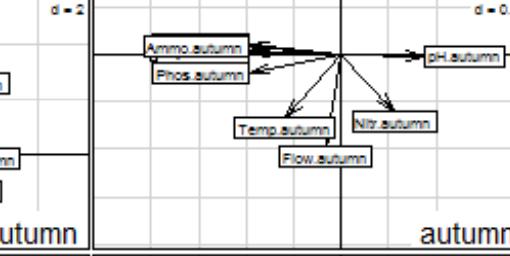
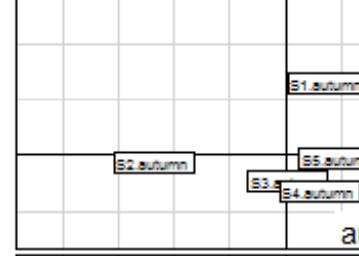
winter



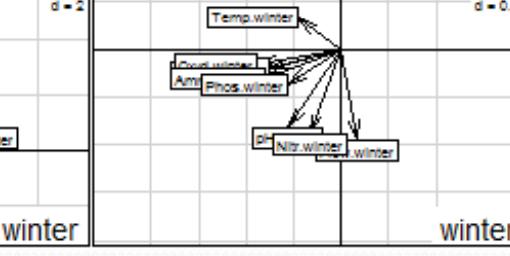
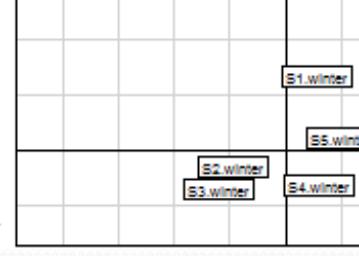
spring



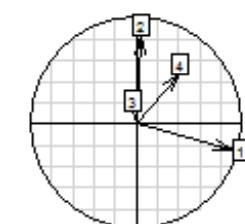
summer



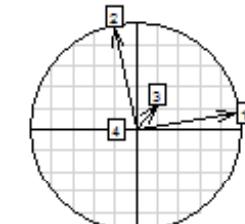
autumn



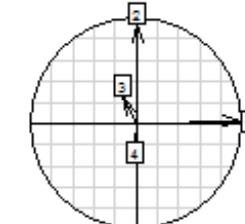
winter



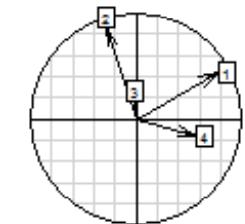
spring



summer



autumn



winter

Plan

- **1. Introduction**
 - Contexte et objectifs de l'analyse multivariée
 - Ordination ou classification ?
 - Ordination basée sur les distances ou sur les valeurs propres (analyses factorielles)
 - Les packages R spécialisés : ade4 et FactoMineR
- **2. Analyses factorielles à un tableau de données, avec les packages ade4 et FactoMineR**
 - 2.1 Analyse en Composantes Principales (ACP) : plusieurs variables quantitatives
 - 2.2 Analyse Factorielle des Correspondances (AFC) : deux variables qualitatives
 - 2.3 Analyse des Correspondances Multiples (ACM) : plusieurs variables qualitatives
- **3. Classification Automatique**
 - 3.1 Classification Ascendante Hiérarchique (CAH) avec la fonction hclust
 - 3.2 Partitionnement avec la fonction kmeans
- **4. Méthodes de couplage de tableaux avec le package ade4**
 - 4.1 Analyses Inter et Intra-classes
 - 4.2 Analyse Discriminante
 - 4.3 Analyses sur variables Instrumentales (ACPVI ou RDA, AFCVI ou CCA)
 - 4.4 Analyse de Co-inertie
- **5. Aperçu des méthodes d'analyse multi-tableaux**
 - 5.1 Analyse Triadique Partielle
 - **5.2 STATIS**
 - 5.3 Analyse Factorielle Multiple

5.2. Analyse STATIS

- **Application** : Quand les tableaux n'ont qu'une seule dimension en commun : K-tableaux avec **mêmes individus OU mêmes variables**
- **Exemple** : friday87 : 91 espèces de macro-invertébrés réparties en 10 groupes taxonomiques et 1 tableau de variables environnementales => 11 tableaux avec les mêmes lignes : 16 mares (en lignes) et les espèces et les variables en colonnes soit 102 colonnes au total
- **Mise en œuvre** : fonction **statis()** de la librairie ade4
- **Principe de la méthode**: identique à la PTA
 - 1. **Interstructure** = ressemblances entre les 4 tableaux
 - 2. **Compromis** = tableau moyen (sites x variables) pondéré selon l'interstructure => Analyse (ACP)
 - 3. **Infrastructure ou Trajectoires** = projection des éléments des tableaux de départ sur les axes du compromis qui sert de référentiel commun pour comparer la structure des tableaux (ici saisons)
- Nécessite une étape préliminaire par rapport à la PTA pour comparer les tableaux qui n'ont pas les mêmes colonnes : à partir de chaque tableau, calcul d'un **opérateur** = matrice de dimension identique (ici les 16 mares)
- **Références** : Lavit, 1988 ; Lavit et al., 1994

5.2. Création du K-tableaux – Ex: friday87

```
data(friday87)
?friday87
names(friday87)
[1] "fau"       "mil"        "fau.blo"    "tab.names"
dim(friday87$fau) # 16 échantillons et 91 taxons
[1] 16 91
dim(friday87$mil) # 16 échantillons et 11 variables environnementales
[1] 16 11
friday87$fau.blo # les 91 taxons se répartissent en 10 blocs = groupes d'espèces
      Hemiptera   Odonata Trichoptera Ephemeroptera Coleoptera   Diptera
      11           7         13          4          13          22
      Hydracarina Malacostraca Mollusca   Oligochaeta
      4            3          8          6
# On crée un dataframe (ici w1) qui concatène
# le tableau faune à 91 colonnes et le tableau milieu à 11 colonnes
# En faisant un centrage sur la faune et un centrage-réduction sur le milieu
w1 <- cbind.data.frame(scale(friday87$fau,scale=F), scale(friday87$mil))

# Création du Ktableau, cette fois-ci à partir du data.frame w1 avec ktab.data.frame
ktal <- ktab.data.frame(w1,c(friday87$fau.blo,11), tabnames=c(friday87$tab.names, "Milieu"))

df : a data frame
blocks      : an integer vector for which the sum must be the number of variables of df.
               Its length is the number of arrays of the K-tables
rownames    : the row names of the K-tables (otherwise the row names of df)
colnames    : the column names of the K-tables (otherwise the column names of df)
tabnames    : the names of the arrays of the K-tables (otherwise "Anal", "Ana2", ... )
w.row       : a vector of the row weightings
w.col       : a vector of the column weightings
```

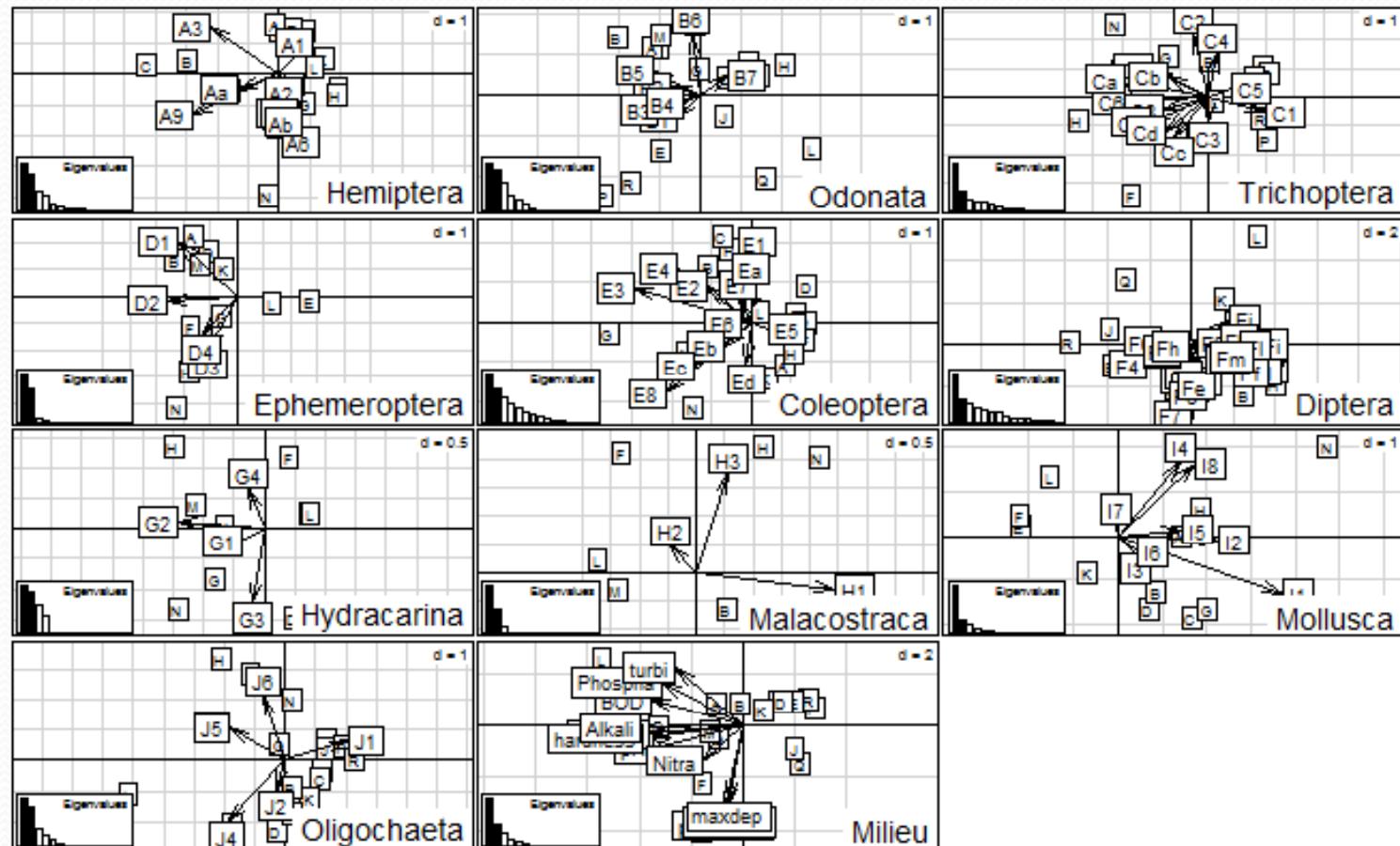
Utilisation de **scale()** pour centrer ou normer les tableaux et de **cbind.data.frame()** pour les mettre côté à côté

Créer le k-tableau avec **ktab.data.frame()**

5.2. Le K-tableaux

5.2. Analyses séparées des K tableaux

```
sep1 <- sepan(ktal)
kplot(sep1)
```



5.2. STATIS – Interstructure

```

statis1 <- statis(ktal, scannf=F)
statis1
STATIS Analysis
class:statis
table number: 11
row number: 16    total column number: 102

***** Interstructure *****

eigen values: 5.542 1.219 0.9506 0.7153 0.5889 ...
$RV      matrix      11      11    RV coefficients
$RV.eig   vector      11      eigenvalues
$RV.coo   data.frame  11       4    array scores
$tab.names  vector      11      array names
$RV.tabw   vector      11      array weights

RV coefficient
          Hemiptera Odonata Trichoptera Ephemeroptera Coleoptera Diptera Hydracarina M
Hemiptera 1.0000000
Odonata   0.4417286 1.0000000
Trichoptera 0.5271259 0.5094746 1.0000000
Ephemeroptera 0.4390920 0.5712601 0.5434608 1.0000000
Coleoptera  0.4983868 0.4061728 0.4632209 0.3053675 1.0000000
Diptera    0.4278053 0.6396437 0.6144842 0.6238601 0.4509614 1.0000000
Hydracarina 0.5021736 0.3068471 0.4327090 0.3162107 0.4726492 0.3384773 1.0000000
Malacostraca 0.3474792 0.4323226 0.5100464 0.5964685 0.3101339 0.4944160 0.4177109
Mollusca   0.4066397 0.4914714 0.4243347 0.6104123 0.4962414 0.5281906 0.5943189
Oligochaeta 0.3888223 0.4104643 0.4417298 0.4074446 0.2539902 0.5137700 0.2839199
Milieu     0.3405098 0.4283602 0.4938589 0.4162155 0.2905822 0.6614336 0.3366027
...

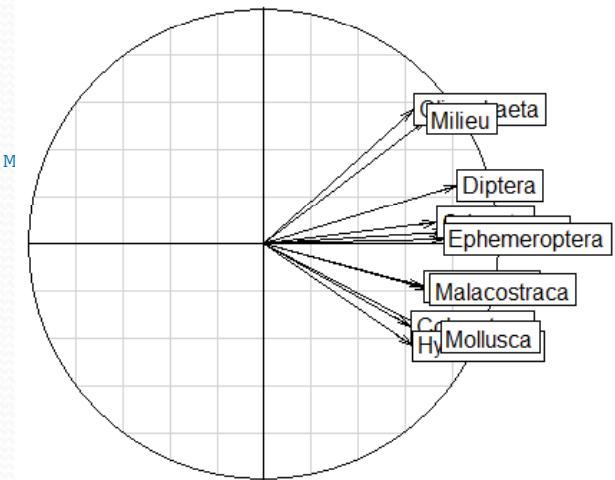
```

INTERSTRUCTURE

```
barplot(statis1$RV.eig)
```



```
s.corcircle(statis1$RV.coo)
```

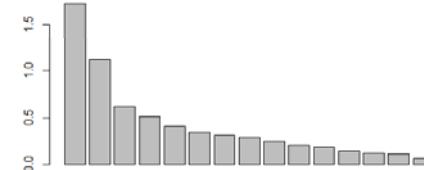


Les coordonnées des tableaux sur l'axe 1 de l'interstructure seront utilisées pour pondérer les tableaux dans le calcul du tableau moyen compromis => favorise les tableaux qui se ressemblent

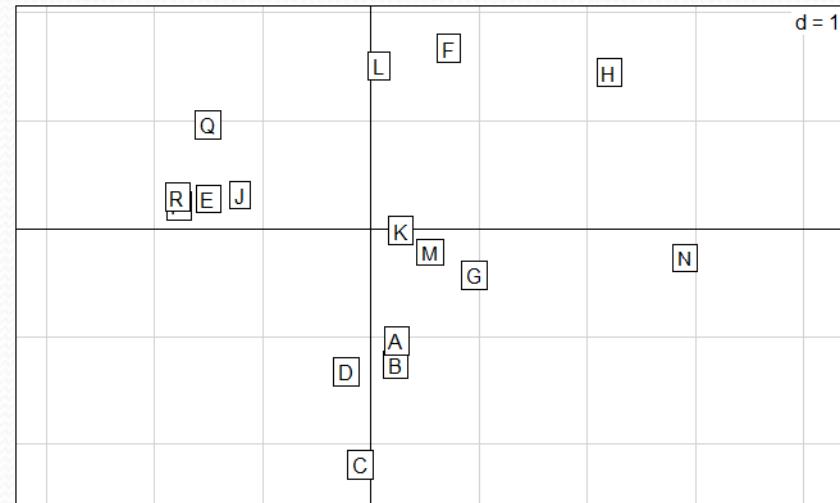
5.2. STATIS – Compromis

```
# 2) Compromis = tableau "moyen" sites X especes  
***** Compromise *****  
  
eigen values: 1.725 1.117 0.6198 0.5146 0.4093 ...  
  
$nf: 3 axis-components saved  
$rank: 15  
  
data.frame nrow ncol content  
$C.li      16   3    row coordinates  
$C.Co     102   3    column coordinates  
$C.T4      44   3    principal vectors (each table)  
$TL       176   2    factors (not used)  
$TC       102   2    factors for Co  
$T4       44   2    factors for T4
```

```
barplot(statis1$C.eig)
```

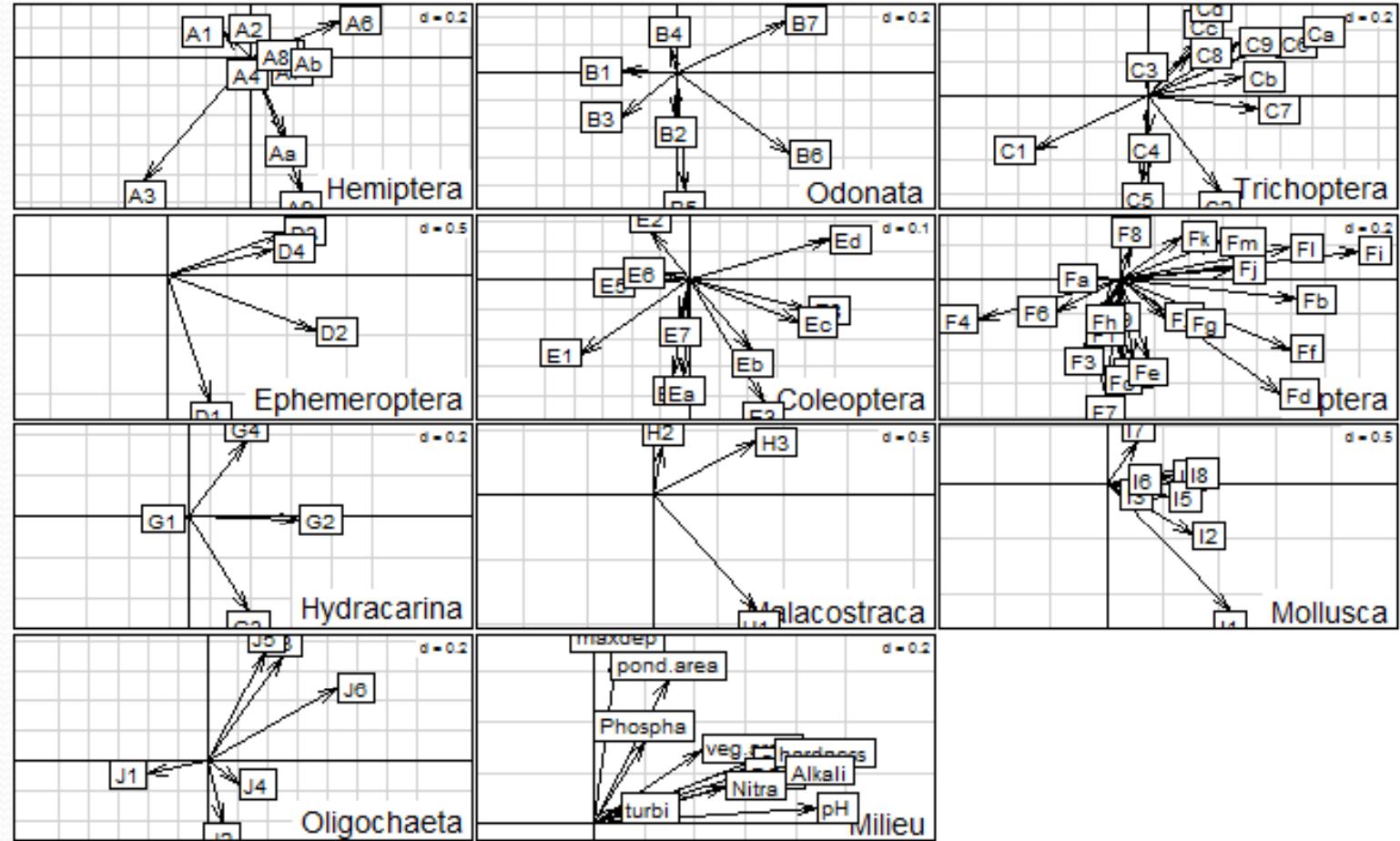


```
s.label(statis1$C.li)
```



5.2. STATIS – représentation des colonnes

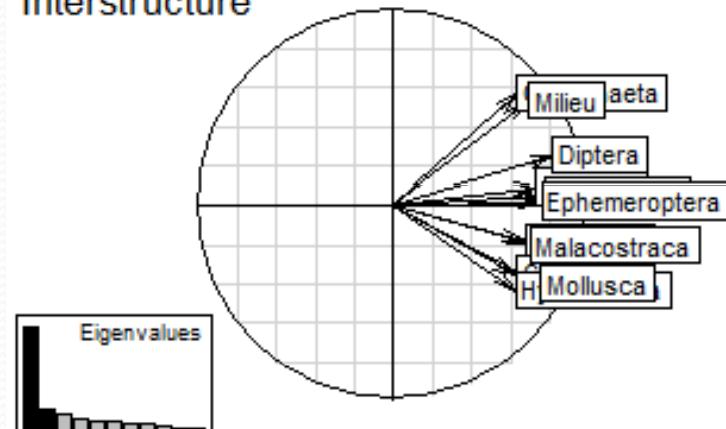
kplot(statis1)



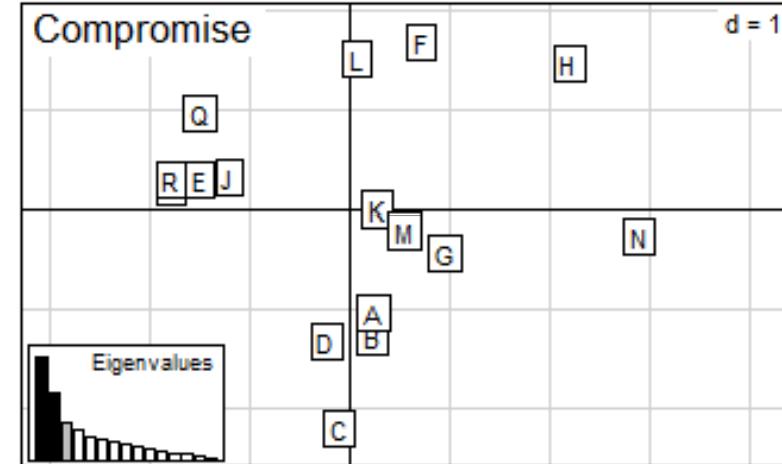
5.2. STATIS - résumé

plot(statist1)

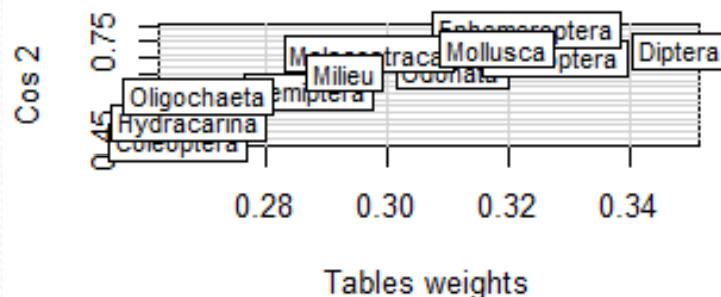
Interstructure



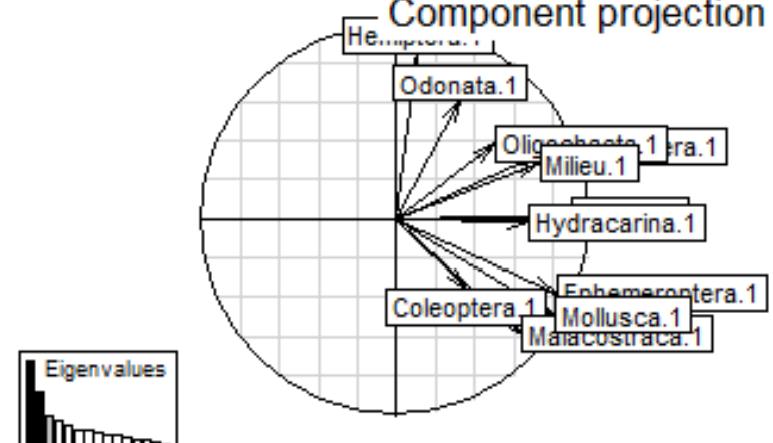
Compromise



Typological value



Component projection



Plan

- **1. Introduction**
 - Contexte et objectifs de l'analyse multivariée
 - Ordination ou classification ?
 - Ordination basée sur les distances ou sur les valeurs propres (analyses factorielles)
 - Les packages R spécialisés : ade4 et FactoMineR
- **2. Analyses factorielles à un tableau de données, avec les packages ade4 et FactoMineR**
 - 2.1 Analyse en Composantes Principales (ACP) : plusieurs variables quantitatives
 - 2.2 Analyse Factorielle des Correspondances (AFC) : deux variables qualitatives
 - 2.3 Analyse des Correspondances Multiples (ACM) : plusieurs variables qualitatives
- **3. Classification Automatique**
 - 3.1 Classification Ascendante Hiérarchique (CAH) avec la fonction hclust
 - 3.2 Partitionnement avec la fonction kmeans
- **4. Méthodes de couplage de tableaux avec le package ade4**
 - 4.1 Analyses Inter et Intra-classes
 - 4.2 Analyse Discriminante
 - 4.3 Analyses sur variables Instrumentales (ACPVI ou RDA, AFCVI ou CCA)
 - 4.4 Analyse de Co-inertie
- **5. Aperçu des méthodes d'analyse multi-tableaux**
 - 5.1 Analyse Triadique Partielle
 - 5.2 STATIS
 - **5.3 Analyse Factorielle Multiple**

5.3. Analyse Factorielle Multiple

- **Application** : Quand les tableaux n'ont qu'une seule dimension en commun : les **individus**
- **Exemple** : `friday87` : 91 espèces de macro-invertébrés réparties en 10 groupes taxonomiques et 1 tableau de variables environnementales => 11 tableaux avec les mêmes lignes : 16 mares (en lignes) et les espèces et les variables en colonnes soit 102 colonnes au total (même tableau que pour STATIS)
- **Mise en œuvre** : fonction **mfa()** de la librairie ade4 ou **MFA()** de FactoMineR
- **Principe de la méthode**: Les variables peuvent être quantitatives ou qualitatives. Si toutes quantitatives, AFM=variante de l'ACP en 3 étapes:
 - ACPs des tableaux séparés et normalisations
 - Réunion des tableaux normalisés en un seul tableau soumis à une ACP non normée
 - Chaque tableau est pondéré par un poids pour diminuer l'importance des grands tableaux et augmenter celle des petits. Par défaut, le poids d'un tableau est l'inverse de la première valeur propre de l'analyse du tableau
 - Projection de chacun des k tableaux sur l'analyse globale (=compromis ou consensus) mettant en évidence les ressemblances ou dissemblances des tableaux avec le compromis
- **Références** : Escofier et Pagès, 1994

5.3. AFM avec ade4

On utilise le même k-tableaux que pour STATIS :

```
# Réalisation de l'AFM
mfa1 <- mfa(ktal, scannf = FALSE)
mfa1
Multiple Factorial Analysis
list of class mfa list of class list
$call: mfa(X = ktal, scannf = FALSE)
$nf: 3 axis-components saved

  vector      length mode      content
1 $tab.names 11     character tab names
2 $blo       11     numeric  column number
3 $rank      1     numeric   tab rank
4 $eig       15     numeric  eigen values
5 $lw        16     numeric  row weights
6 $stabw     0      NULL    array weights

  data.frame nrow ncol content
1 $tab       16   102 modified array
2 $li        16     3  row coordinates
3 $l1        16     3  row normed scores
4 $co        102    3  column coordinates
5 $c1        102    3  column normed scores
6 $lisup     176    3  row coordinates from each table
7 $TL        176    2  factors for li l1
8 $TC        102    2  factors for co c1
9 $T4        44     2  factors for T4comp
10 $T4comp   44     3  component projection
11 $link      11     3  link array-total
other elements: NULL
```

```
mfa(X,
      option = c("lambda1", "inertia", "uniform", "internal"),
      scannf = TRUE, nf = 3)
```

X : objet de classe ktab

option : méthode de pondération des tableaux

'lambda1'= inverse de la première valeur propre [par défaut]

'inertia'= inverse de l'inertie total du tableau

'uniform'= poids uniforme

'internal'= poids inclus dans X\$tabw

mfa1\$li = coordonnées des lignes (16 mares)

mfa1\$lisup = coordonnées des lignes de chaque tableau (16 mares x 11 tableaux)

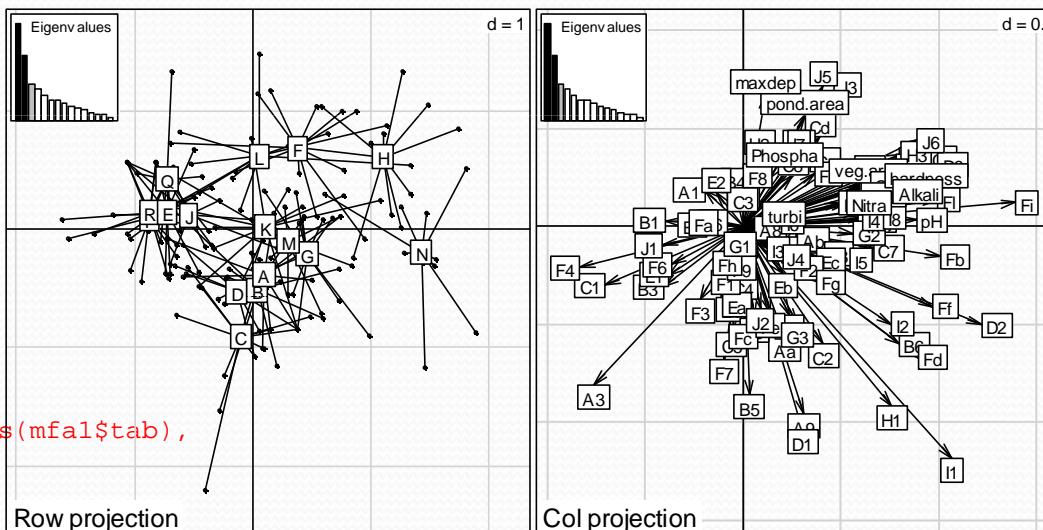
mfa1\$co = coordonnées des colonnes (91 taxons + 11 variables de milieu)

mfa1\$T4comp = projection des 4 premiers axes des ACP de chaque tableau sur les axes de l'analyse globale

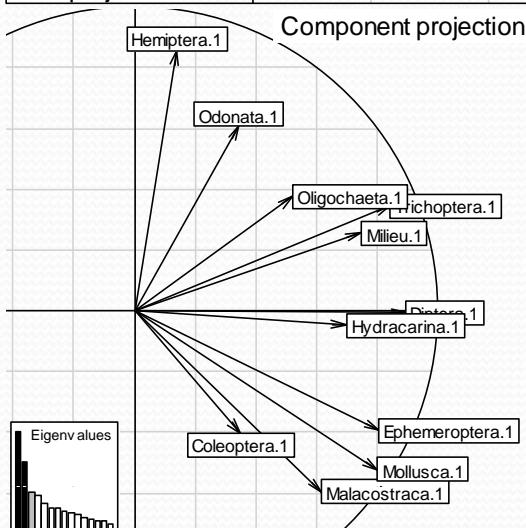
5.3. AFM avec ade4

```
plot(mfal)
```

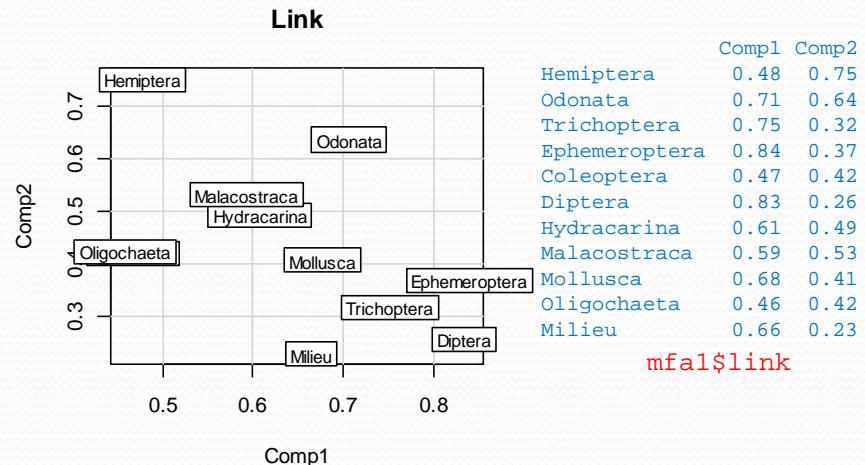
```
s.class(mfal$lisup,
mfal$TL[,2],
label=row.names(mfal$tab),
cellipse = 0)
```



```
s.arrow(mfal$co)
```



```
s.corcircle(mfal$T4comp[mfal$T4[,2]==1, ])
```



```
mfal$link
```

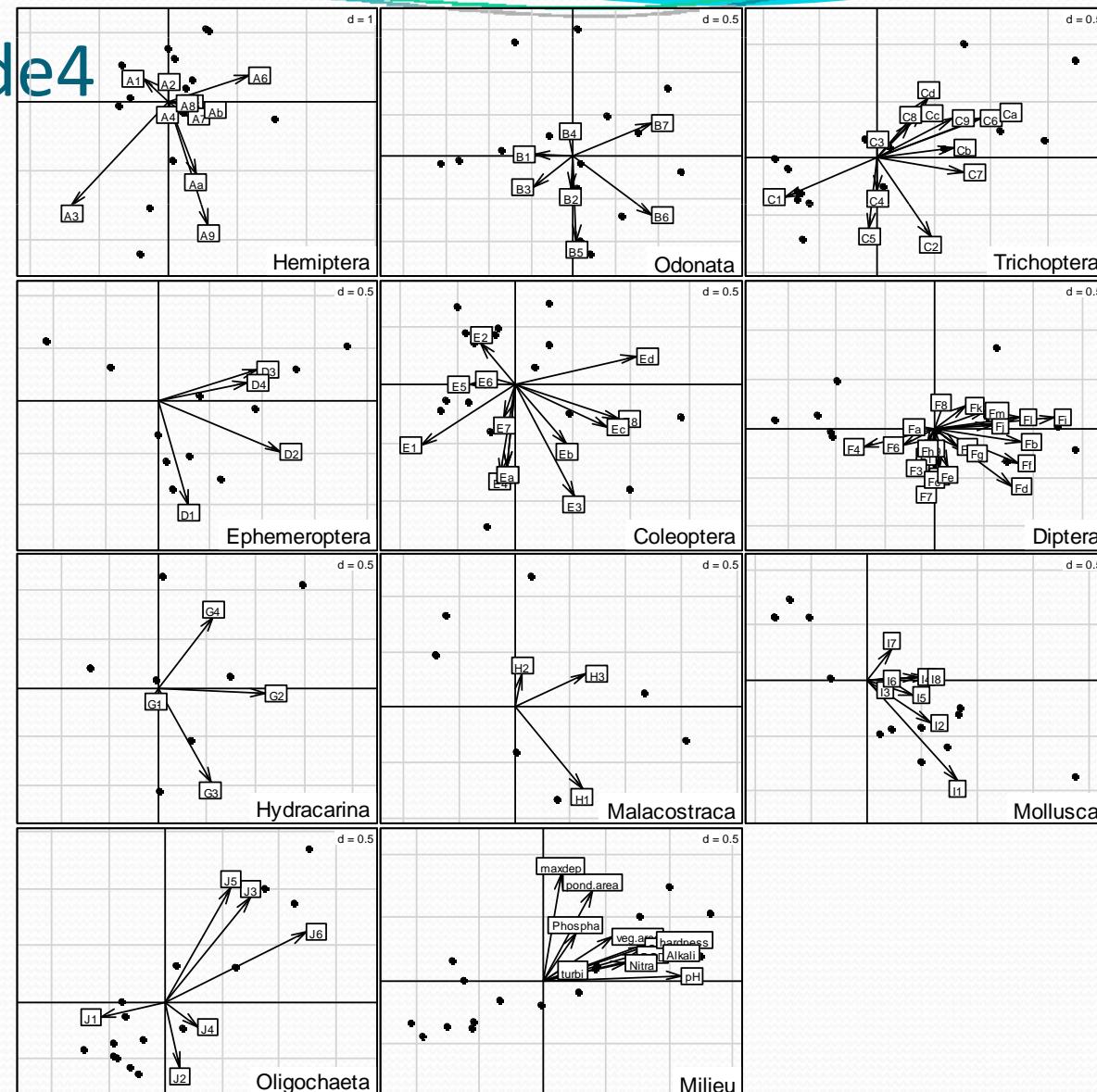
5.3. AFM avec ade4

Représentation :

- des lignes (marges) par des points
- des colonnes (taxons) par leur label

Multifenêtrage par tableau

`kplot(mfa1)`



5.3. AFM avec FactoMineR

```
MFA(base, group, type = rep("s", length(group)), excl = NULL, ind.sup = NULL, ncp = 5, name.group = NULL, num.group.sup = NULL, graph = TRUE, weight.col.mfa = NULL, row.w = NULL, axes = c(1,2), tab.comp=NULL))
```

base = tableau de données (ici friday87\$fau 16 x 91 + friday87\$mil)

group = vecteur avec le nombre de variables dans chaque sous-tableau

type = type de variable dans chaque sous-tableau

"c" = variable quantitative à centrer ;

"s" = variable quantitative à normer ;

"n" = variable qualitative ;

"f" = fréquences d'une table de contingence)

name.group= nom des sous-tableaux

```
fridaydf <- cbind(friday87$fau,friday87$mil)                      # Les 10 tableaux espèces et le tableau milieu
fridaygroup <-c(friday87$fau.blo,11)                                # nombre de colonnes de chaque tableau (11=milieu)
fridaytype <- c(rep("c", length(friday87$fau.blo)), "s") # centrer les tableaux espèces et normer milieu
fridaynames <- c(friday87$tab.names, "Milieu")                     # Noms des 11 tableaux

res.MFA = MFA(fridaydf,
              group=fridaygroup,
              type=fridaytype,
              name.group=fridaynames)
```

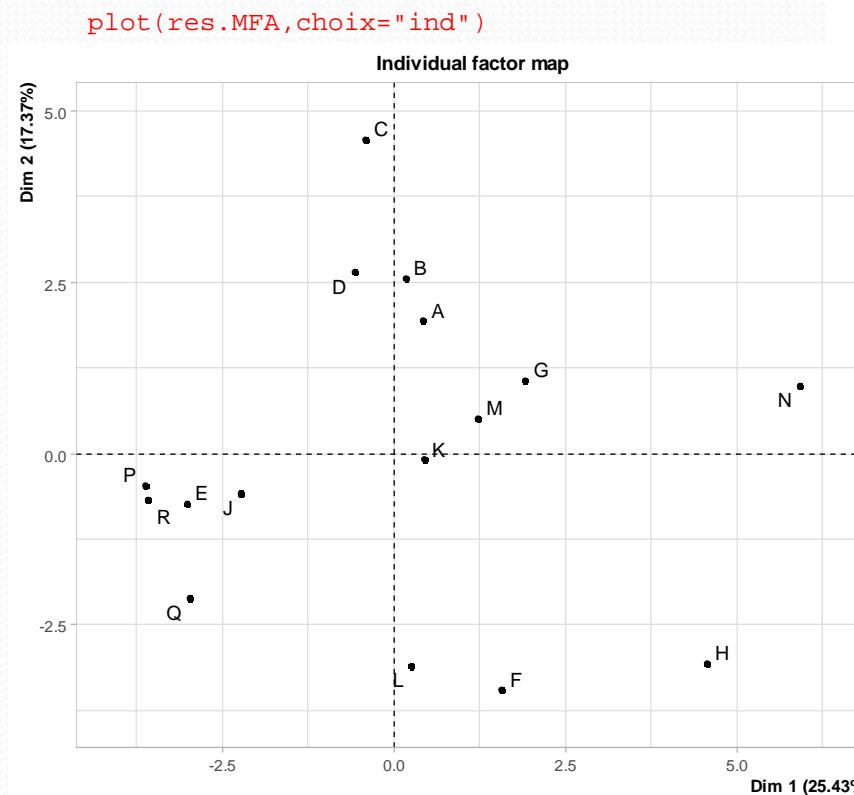
5.3. AFM avec FactoMineR

```
print(res.MFA)
**Results of the Multiple Factor Analysis (MFA)**
The analysis was performed on 16 individuals, described by 102 variables
*Results are available in the following objects :
```

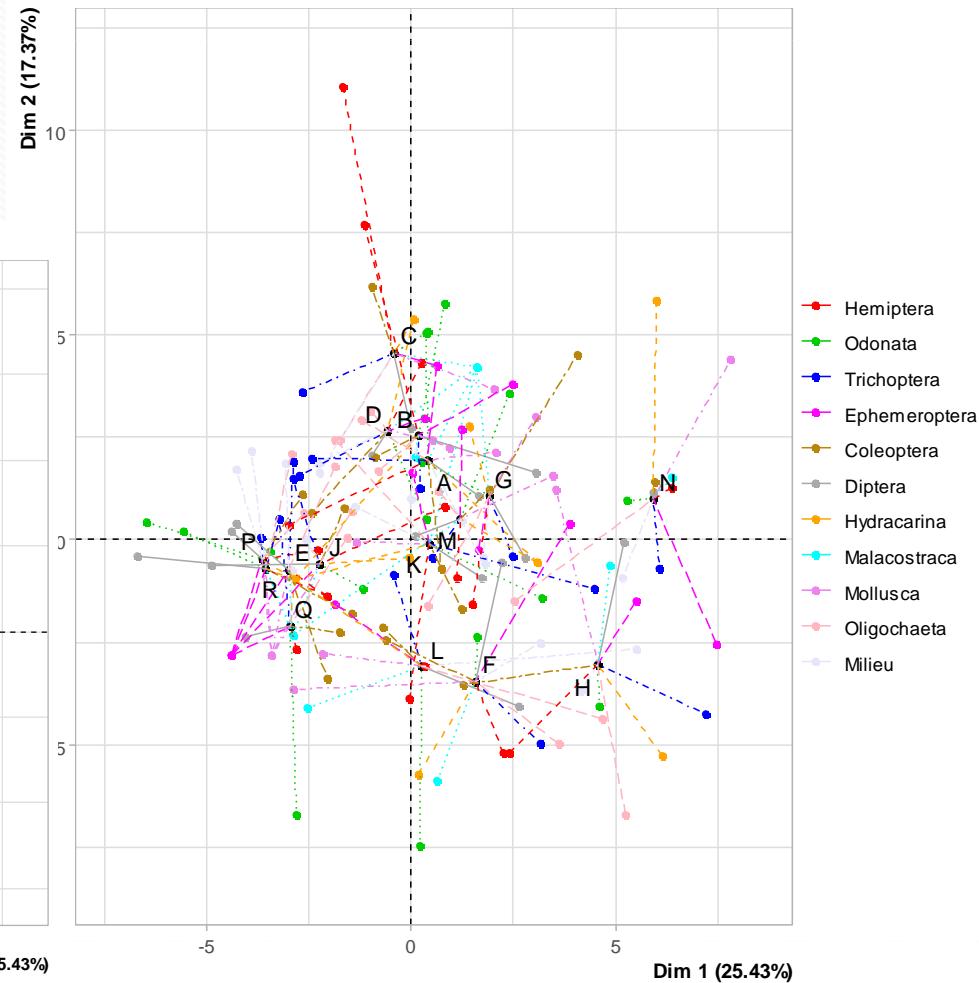
	name	description
1	"\$eig"	"eigenvalues"
2	"\$separate.analyses"	"separate analyses for each group of variables"
3	"\$group"	"results for all the groups"
4	"\$partial.axes"	"results for the partial axes"
5	"\$inertia.ratio"	"inertia ratio"
6	"\$ind"	"results for the individuals"
7	"\$quanti.var"	"results for the quantitative variables"
8	"\$summary.quanti"	"summary for the quantitative variables"
9	"\$global.pca"	"results for the global PCA")

5.3. AFM avec FactoMineR

Représentation des individus (mares)



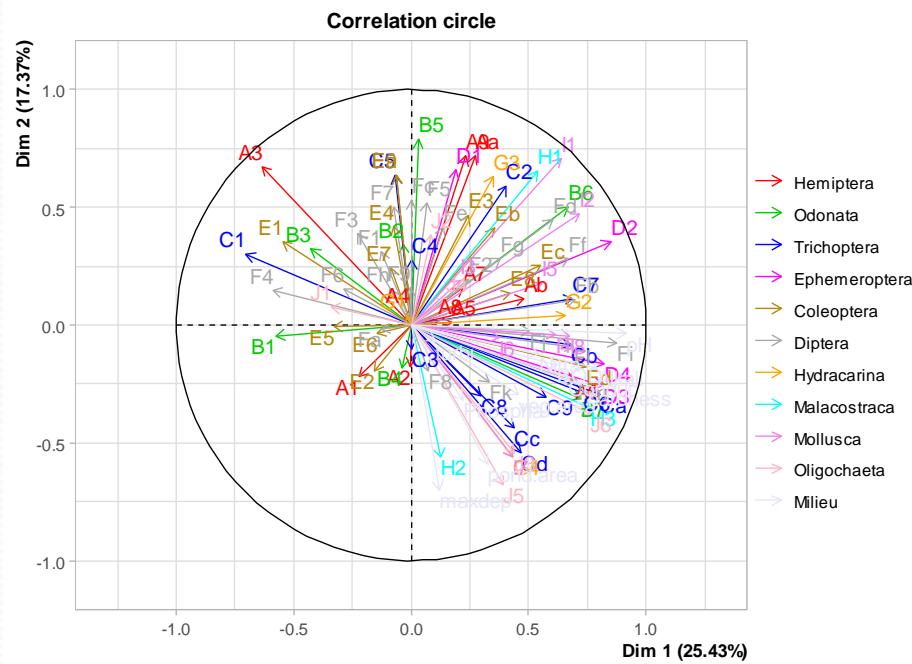
plot(res.MFA, choix="ind", partial="all")
Individual factor map



5.3. AFM avec FactoMineR

Représentation des colonnes (taxons et variables de milieu)

```
plot(res.MFA, choix="var" ,partial="all")
```



Bibliographie - analyses à K tableaux

Fichiers PDF

- **tdr68_KTab.pdf** - L'ordination simultanée de plusieurs tableaux