Analyses Factorielles et Classification

EDEN

Denis Laloë
GABI - PSGen

27 octobre 2016

Introduction

Données

- *Le modèle doit suivre les données, non l'inverse, J P Benzécri*
- Observation *vs* Expérimentation
  - Donnée préexistante (Sciences sociales / Ecologie)
  - Pas de structure a priori : induction
  - Synthèse (vision holistique / corrélations partielles / causalité) :
    - Approche de Durkheim : Pour dégager des relations causales, une relation binaire ne suffit pas, il faut intégrer plusieurs variables dans l'analyse et considérer leurs relations
    - Approche de Benzécri : C'est de la synthèse, sans a priori, que les causes émergent.

Quelques références : Armatte, Benzecri, Bressoux, Rouanet et Leroux

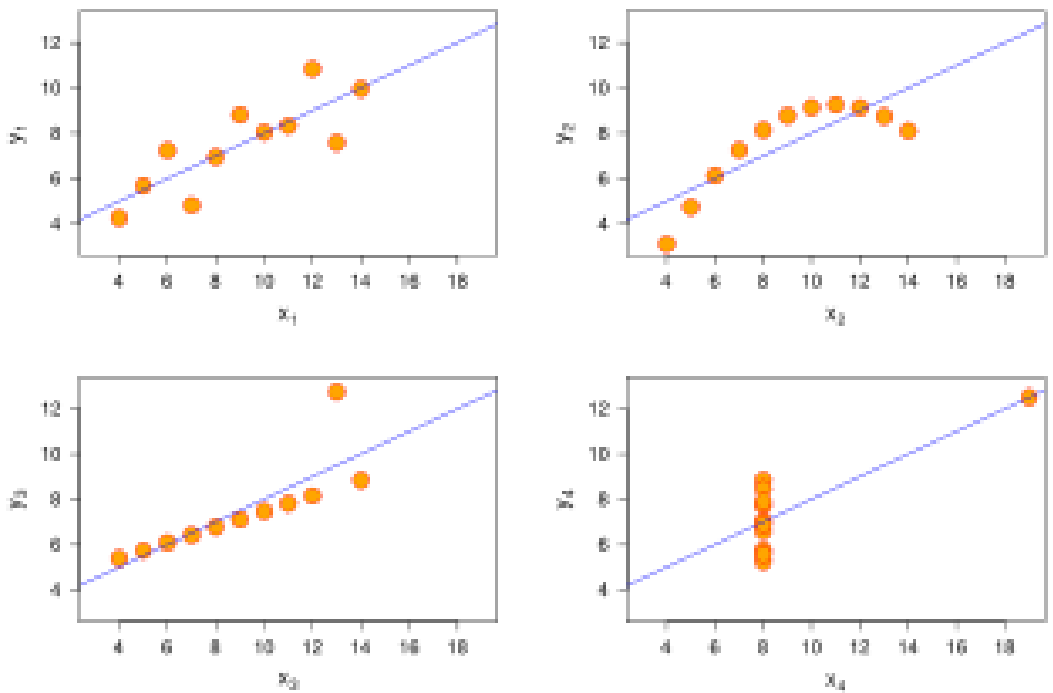
Intégration de données

Données : Synthèse, représentation, description,...

- Simplicité
- Clarté
  - *Un bon croquis vaut mieux qu'un long discours, N Bonaparte*
  - *An arrow is not a set of coordinates, J W Tukey*
  - *Graphs are essential to good statistical analysis, F J Anscombe*
- Efficience
- Adéquation / Fidélité / Justesse
  - *Si les sens sont faux, alors la raison l'est tout autant, Lucrèce*
- Lisibilité / Objet d'étude

Représentation des données

Approche géométrique : représentation de données sous forme de nuages de points (plutôt que des résumés quantitatifs) F J Anscombe, 1973



Représentation des données

Approche géométrique : représentation de données sous forme de nuages de points (plutôt que des résumés quantitatifs)

Cleveland and McGill, 1984

The real power of a Cartesian graph does not derive only from one's ability to perceive the x and y values separately but from one's ability to understand the relationship of x and y.

Lewandowski et Spence, 1989

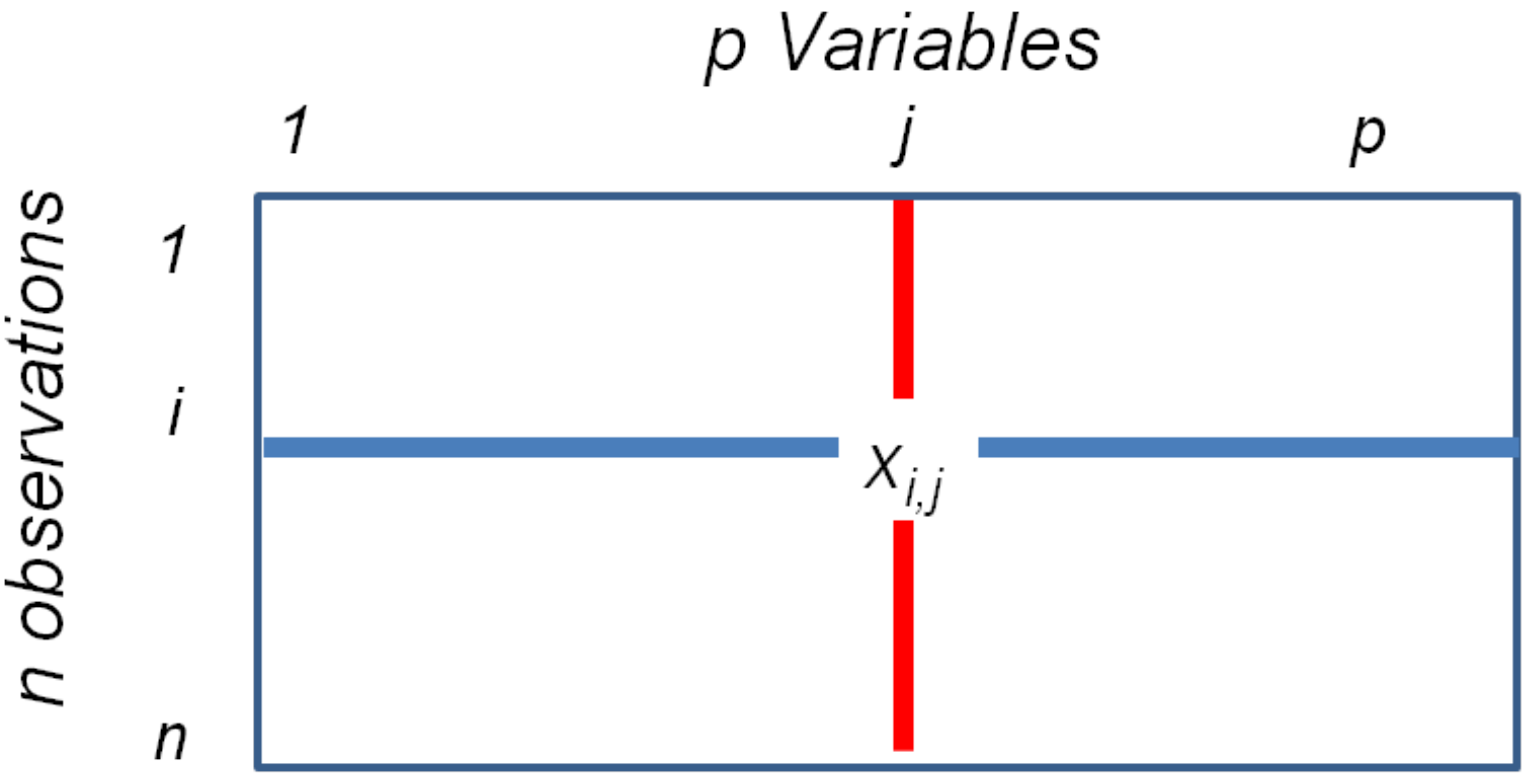
- Conservative judges of correlation, tending to estimate the squares of the correlation
- If outliers are present, they exhibit less bias in their estimates of correlation than do some robust numerical estimates

Analyse géométrique des données
Analyses factorielles

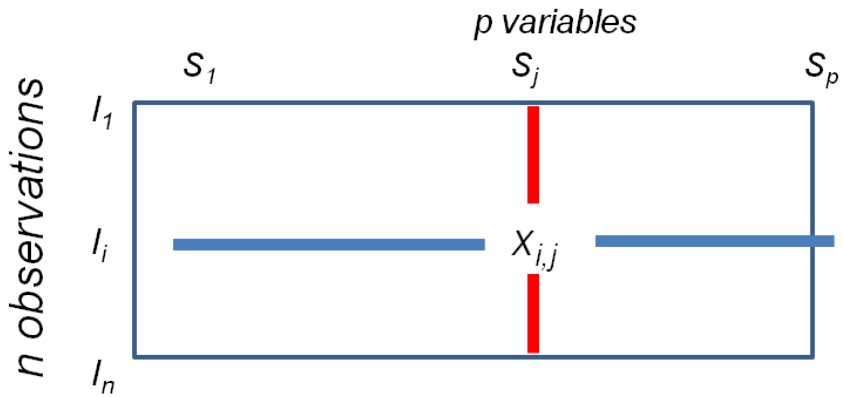
- Approche géométrique : représentation de données sous forme de nuages de points (plutôt que des résumés quantitatifs)
- Efficience : synthèse par optimisation d'un critère (Inertie,...)

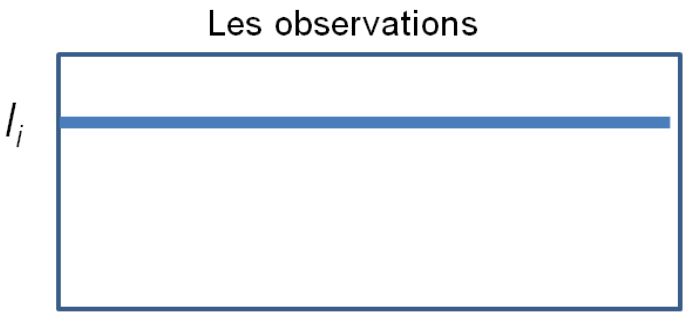
L'ACP

Un tableau de variables quantitatives

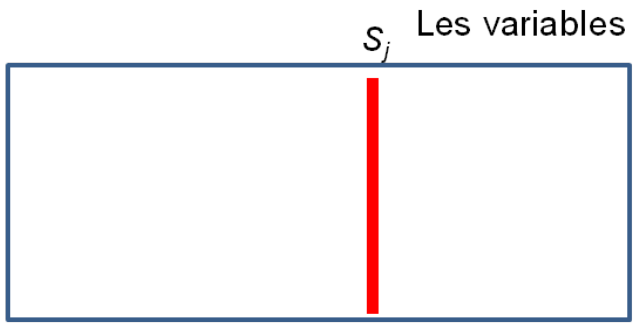


Un tableau, deux points de vue





La ième observation définie par p variables

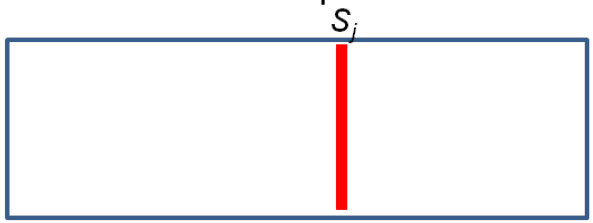


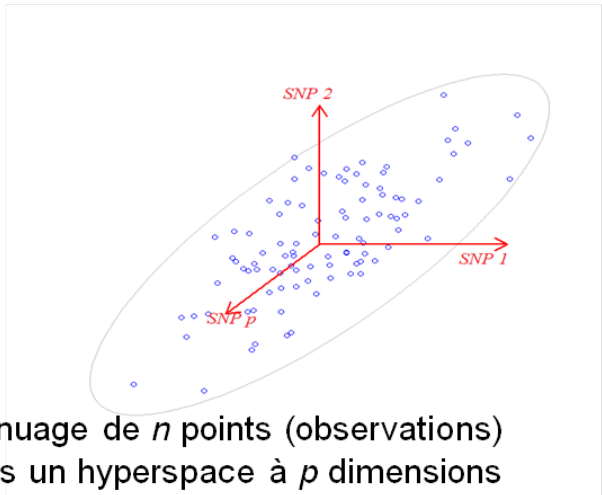
La jème variable définie par by n observations

Deux représentations géométriques

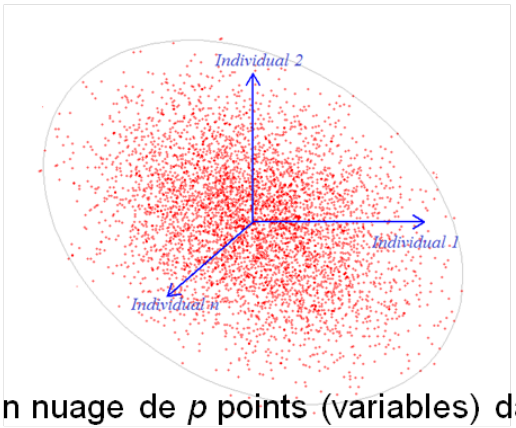
La ième observation définie par p variables

La ième variable definie par n observations





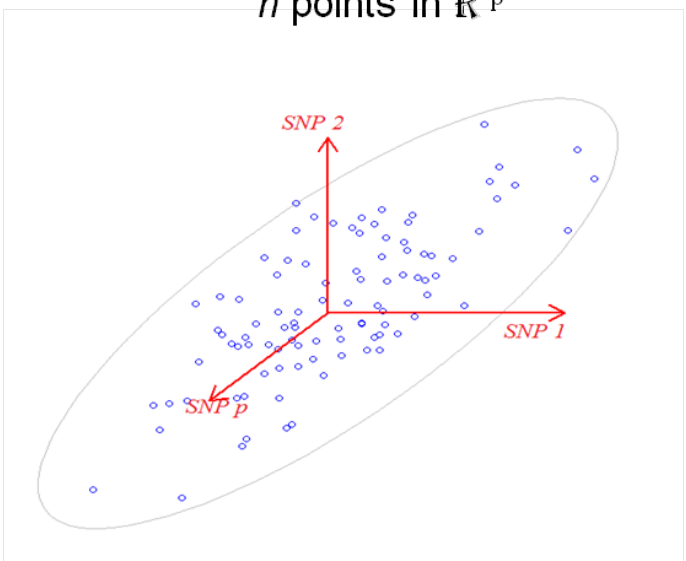
Un nuage de n points (observations) dans un hyperspace à p dimensions (variables)



Un nuage de p points (variables) dans un hyperspace à n dimensions (observations)

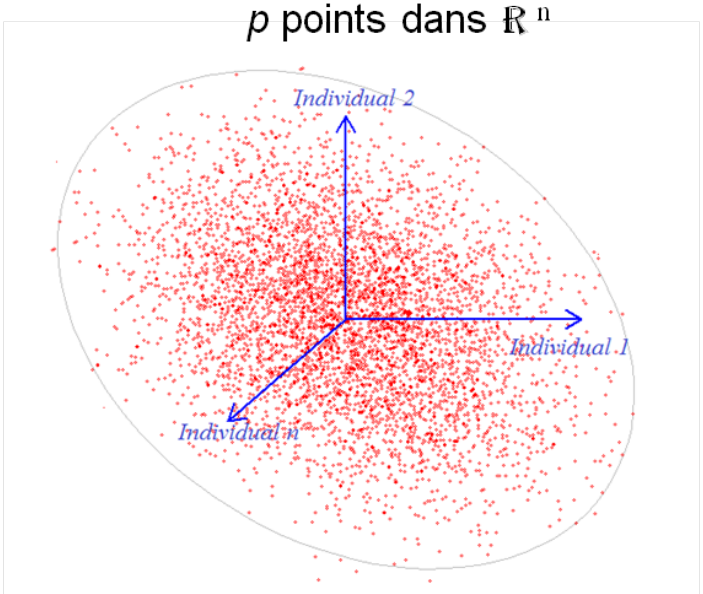
Deux interprétations

HyperSpace des observations
n points in $\mathbb{R}^p$



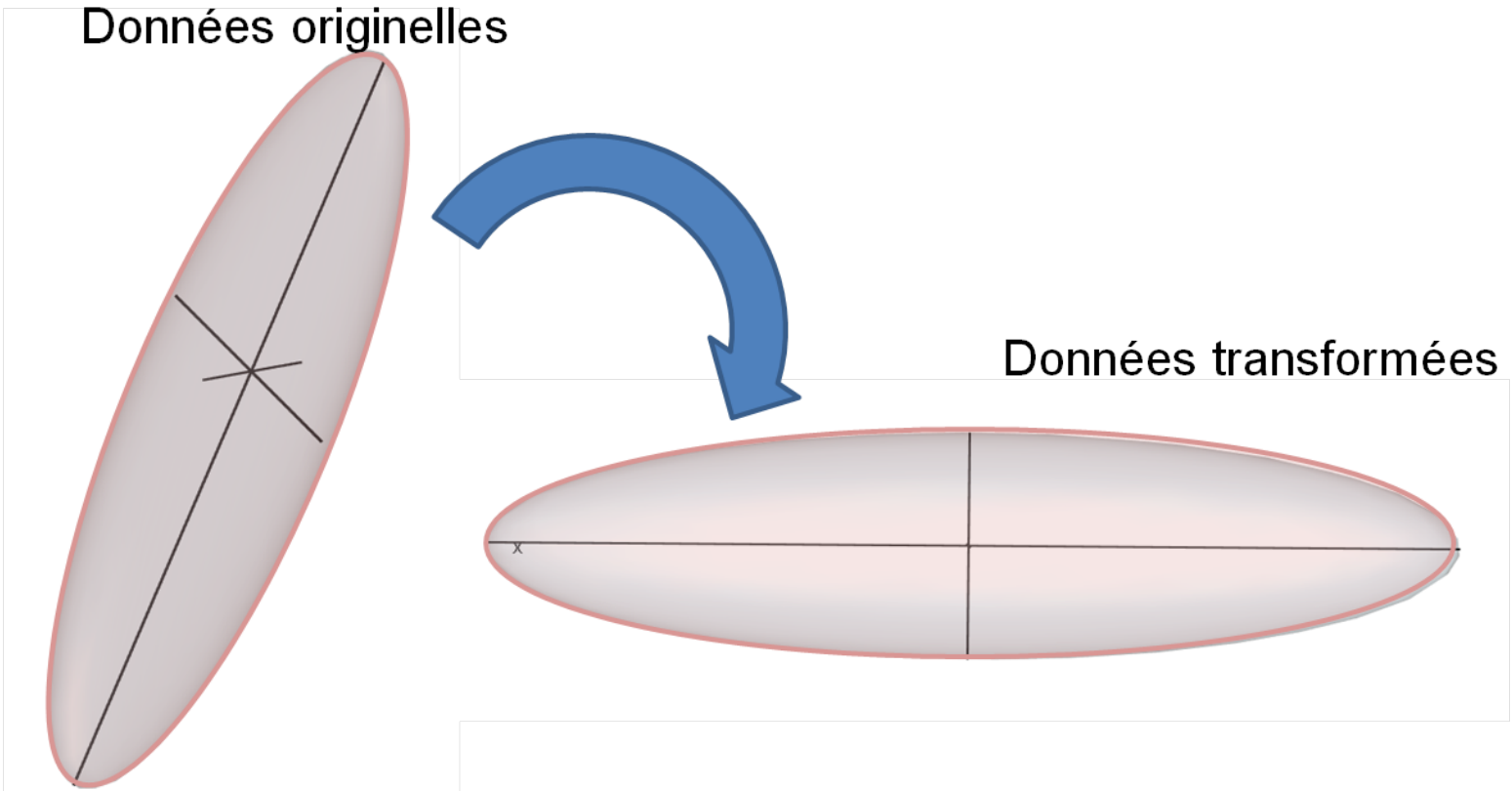
Similarités / différences entre observations

HyperSpace des variables
p points dans $\mathbb{R}^n$



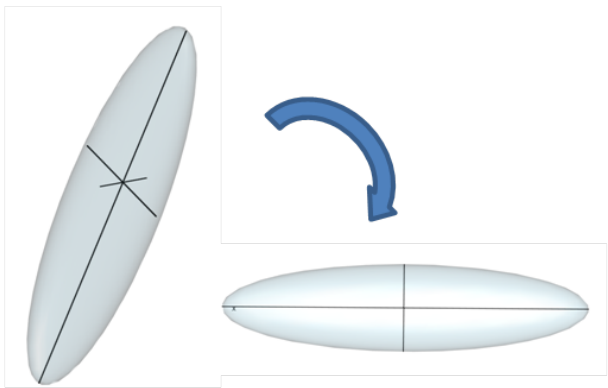
Relations entre variables

Transformation du nuage



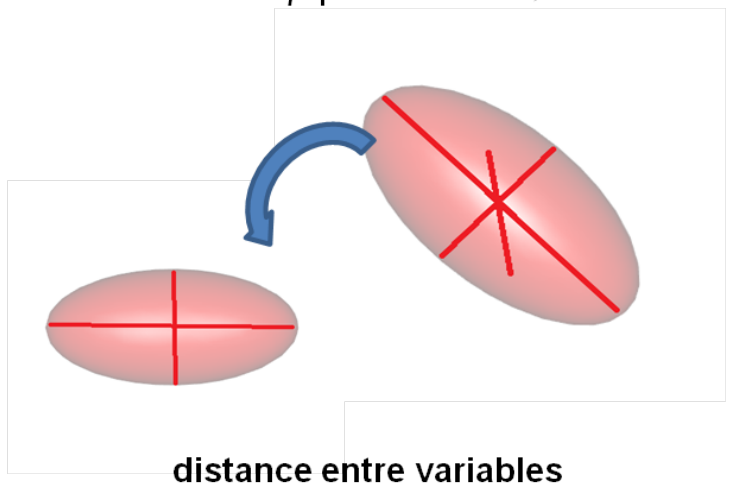
Le même mécanisme

Hyperspace des observations
n points dans $\mathbb{R}^p$



distance entre observations

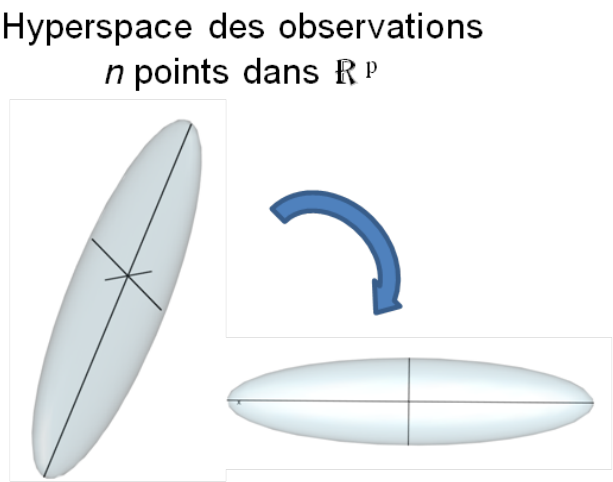
Hyperspace des variables
p points dans $\mathbb{R}^n$



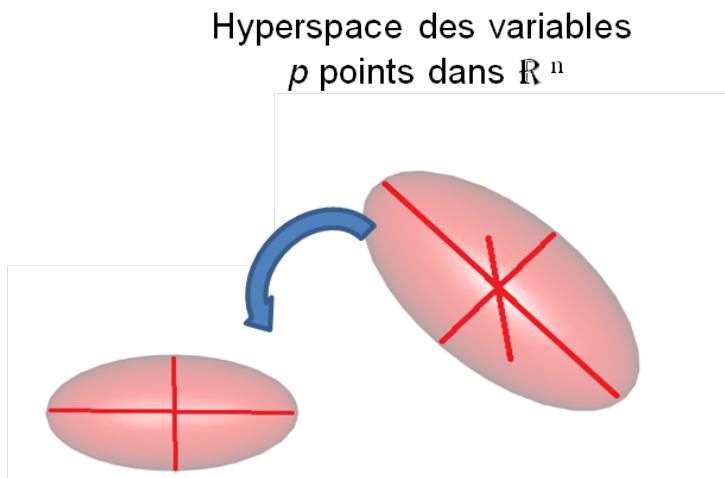
distance entre variables

Le schéma de dualité

- Décomposition canonique - valeurs et vecteurs propres
- Décomposition en valeurs singulières



distance entre observations



distance entre variables

Le schéma de dualité
Observations / Variables

Maximisation of the correlation between variables and components

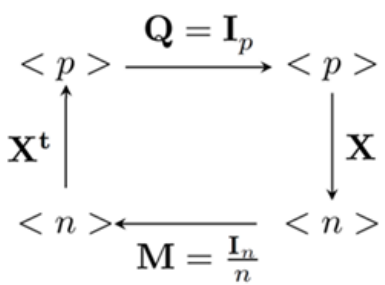
$$V = X'X/n$$

$$VA = A\Delta$$
$$A'A = I$$

Principal axes

Variable scores

$$C = X'B$$



Diagonalisation

$X'X$ XX'

mêmes valeurs propres non nulles

$\lambda 1 > \lambda 2 > \ldots > \lambda r > 0$

$\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_r)$

Transition formulae

$$XA\Delta^{-0,5} = B$$
$$X'B\Delta^{-0,5} = A$$

Singular value decomposition

$$X = B\Lambda^{0.5}A'$$

Best approximation (rank l)

Eckart and Young

$$\hat{\mathbf{X}}_l = \sum_{i=1,l} \sqrt{\lambda_i}\, \mathbf{b_i a_i'}$$

Maximisation of the dispersion of individuals

Observations

$$W = XX'/n$$

$$WB = B\Delta$$
$$B'B = I$$

Principal components

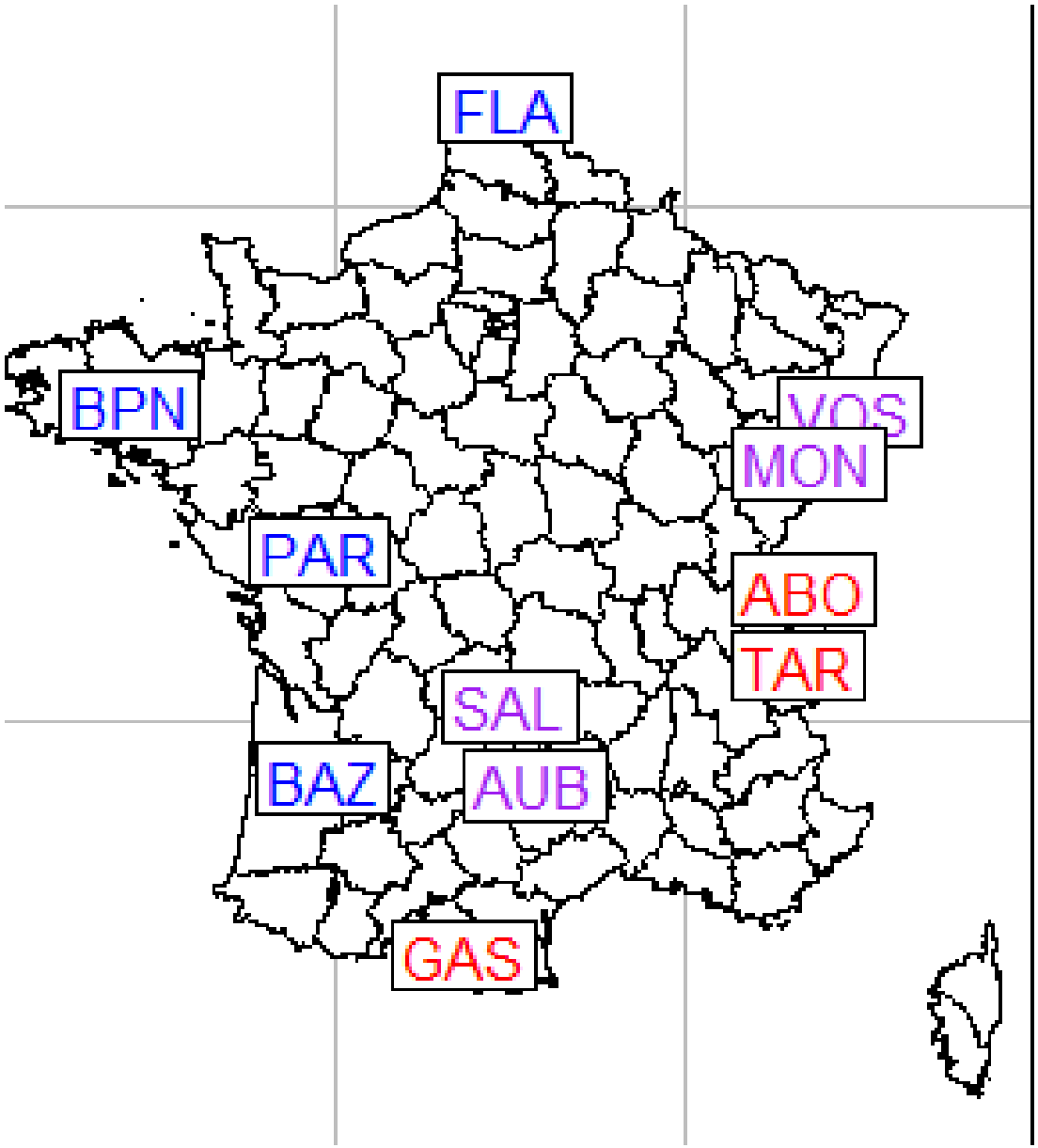
Observation scores

$$L = XA$$

Déclinaison du schéma de dualité

Analyse	Nature des données	Tableau de données	Pondération lignes	Pondérations colonnes
ACP	Quantitatif	$\frac{x_{ij} - x_{.j}}{s_{.j}}$	$\frac{\mathbf{I}}{n}$	$\mathbf{I}$
AFC	Tableau de contingence	$\frac{f_{ij} - f_{i.} f_{.j}}{f_{i.} f_{.j}}$	$f_{i.}$	$f_{.j}$
ACM	Qualitatif (tableau disjonctif)	$\frac{\delta_{ikj} - f_{kj}}{f_{kj}}$	$\frac{\mathbf{I}}{n}$	$diag(f_i)$/nfacteurs

Climat, géographie et races bovines



Données climatiques

climond.org

1	temp_moy	Annual mean temperature
2	temp_rang0	Mean diurnal temperature range
3	temp_iso	Isothermality
4	temp_sai	Temperature seasonality
5	temp_max0	Max temperature of warmest week
6	temp_min0	Min temperature of coldest week
7	temp_rang1	Temperature annual range
8	pluie_moy	Annual precipitation
9	pluie_max0	Precipitation of wettest week
10	pluie_min0	Precipitation of driest week
11	pluie_sais	Precipitation seasonality
12	pluie_max1	Precipitation of wettest quarter
13	pluie_min1	Precipitation of driest quarter
14	rad_moy	Annual mean radiation
15	rad_max0	Highest weekly radiation
16	rad_min0	Lowest weekly radiation
17	rad_sais	Radiation seasonality
18	moist_moy	Annual mean moisture index
19	moist_max	Highest weekly moisture index
20	moist_min	Lowest weekly moisture index
21	moist_sais	Moisture index seasonality

Climat, géographie et races bovines

	Longitude	Latitude	Bio01	Bio02	Bio03	Bio04	...	Race
ABO	6.72	46.28	6.83	8.04	0.32	0.02	...	ABO
AUB	2.85	44.37	10.11	9.65	0.37	0.02	...	AUB
BPN	-2.98	48.07	10.77	6.56	0.37	0.01	...	BPN
GAS	1.83	42.72	7.66	9.43	0.37	0.02	...	GAS
SAL	2.50	45.14	9.40	9.65	0.38	0.02	...	SAL
TAR	6.65	45.54	2.83	6.48	0.29	0.02	...	TAR

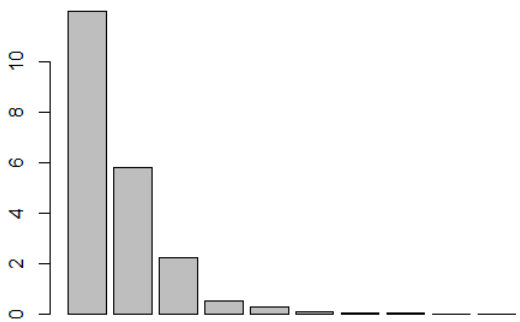
ACP
Les packages R

- *ade4*: dudi.pca
- fichier.pca← dudi.pca(df = fichier1[, -c(1, 2, 24)],scannf=F,nf=3)
- *FactoMineR*: PCA
- fichier.PCA ← PCA(fichier1,ncp=3,quanti.sup=1:2,quali.sup=24)
- *vegan*: rda
- *base* : princomp,prcomp

ACP
Choix de la dimension

- ACP sur variables normées : valeur propre supérieure à 1
- éboulis des valeurs propres
- interprétabilité des axes
- tests à partir de distributions
- ...

```
barplot(fic.pca$eig)
barplot(fic.PCA$eig[,1])
```



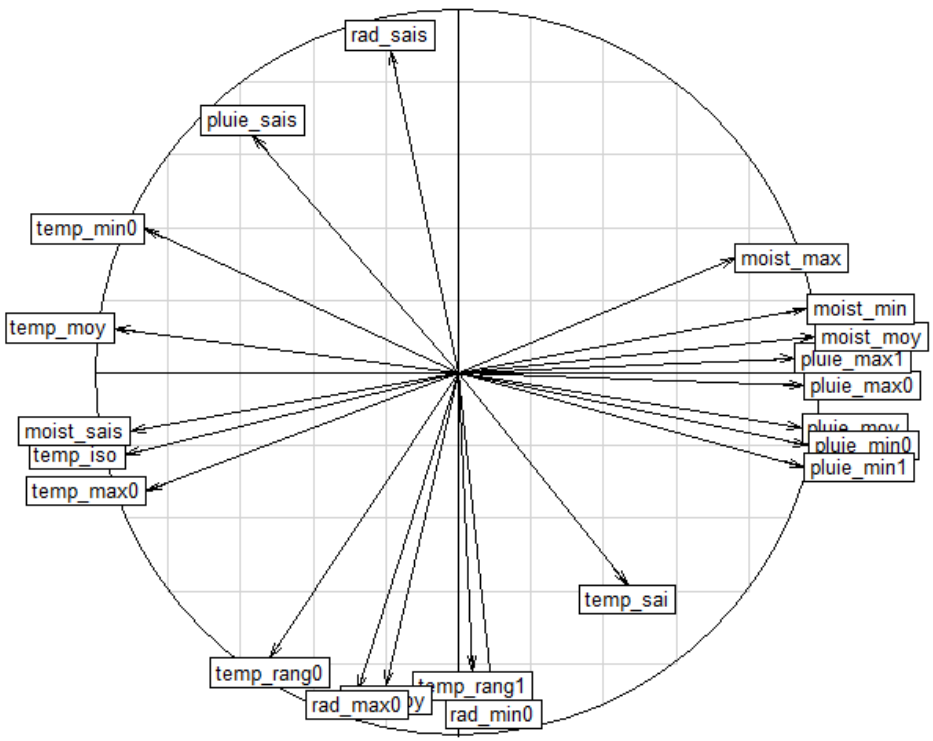
Solving the number of factors is easy, I do it everyday before breakfast. But knowing the right solution is harder. Henry Kaiser

ACP

Variables : Cercle des corrélations

s.corcircle(fic.acp$co,...)

plot(fichier.PCA,choix="var")

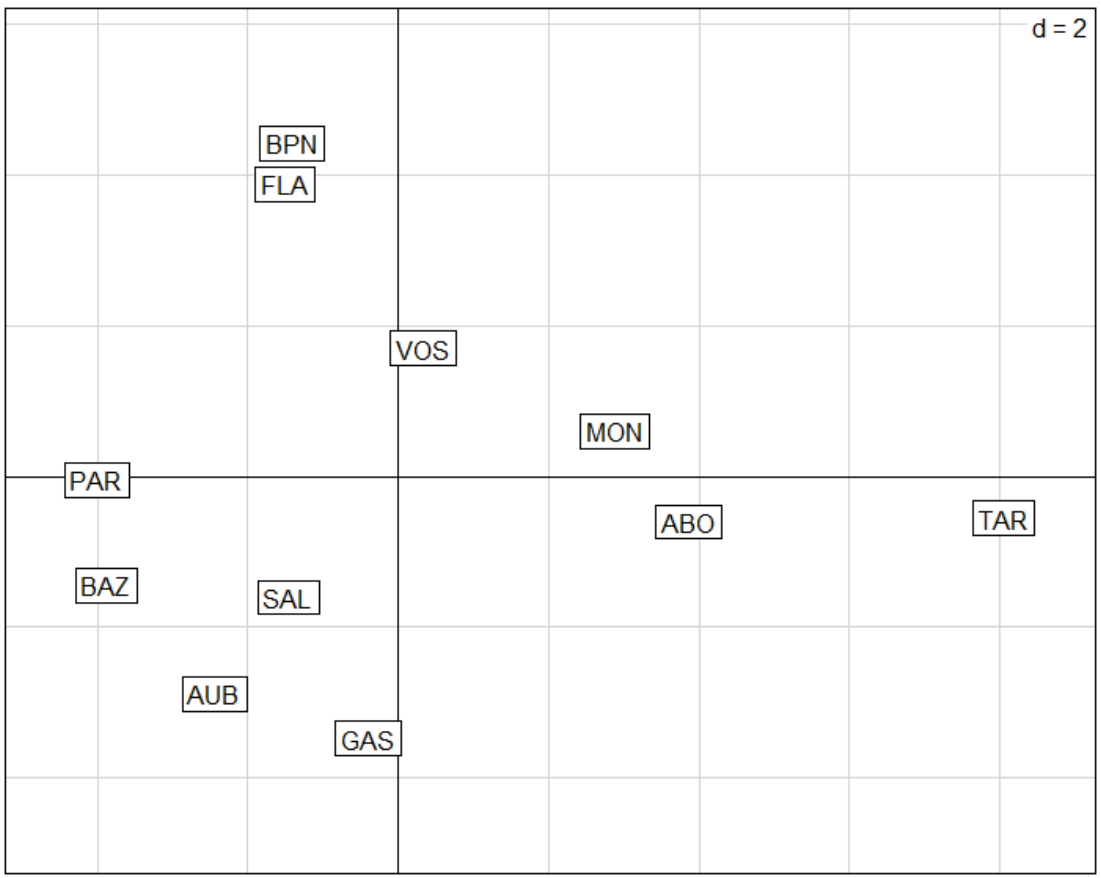


Axes 1 & 2

ACP

Individus : Cartes factorielles (scatterplots)

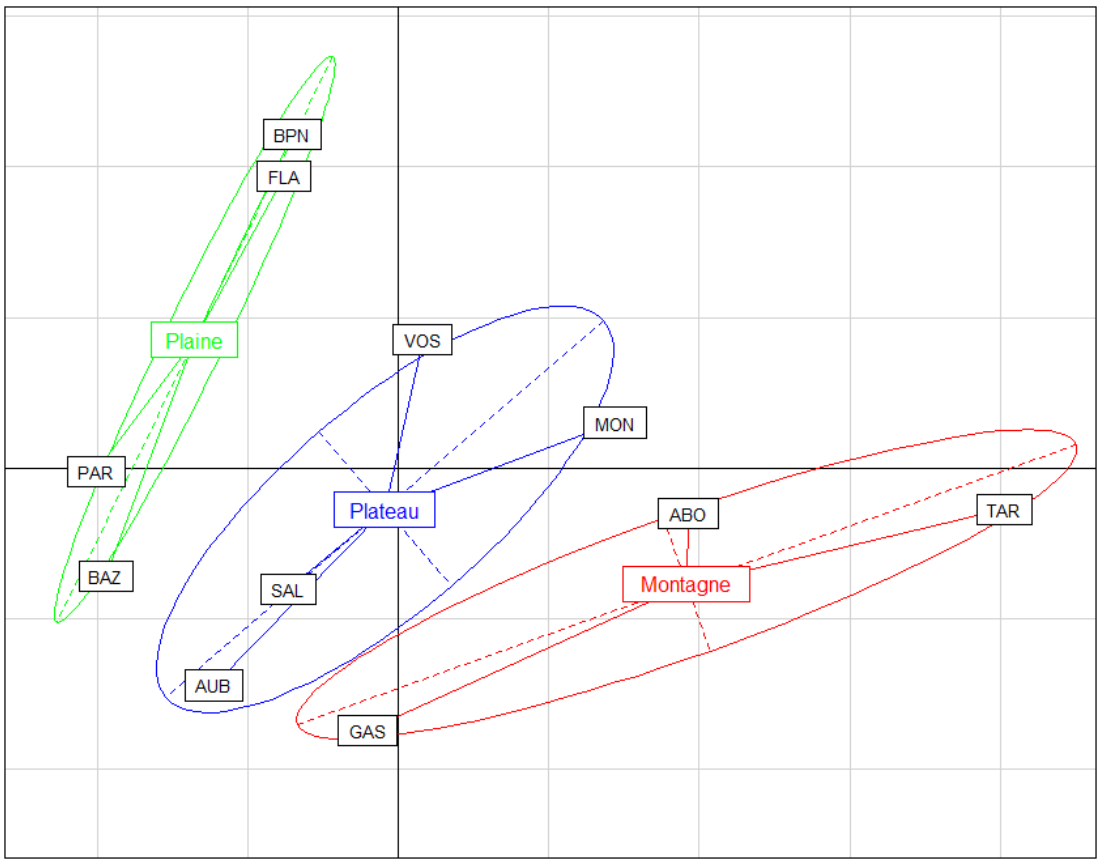
```
s.label(fichier.pca$li)
plot(fichier.PCA,choix="ind")
```



ACP
Individus/Interprétation

```
s.class(fichier.pca$li,fac=fichier$Zone,col=rainbow(3)[1:3])
s.label(fichier.pca$li,label=fichier$Race,clab=.8,add.plot=T)

plotellipses(fichier.PCA)
```

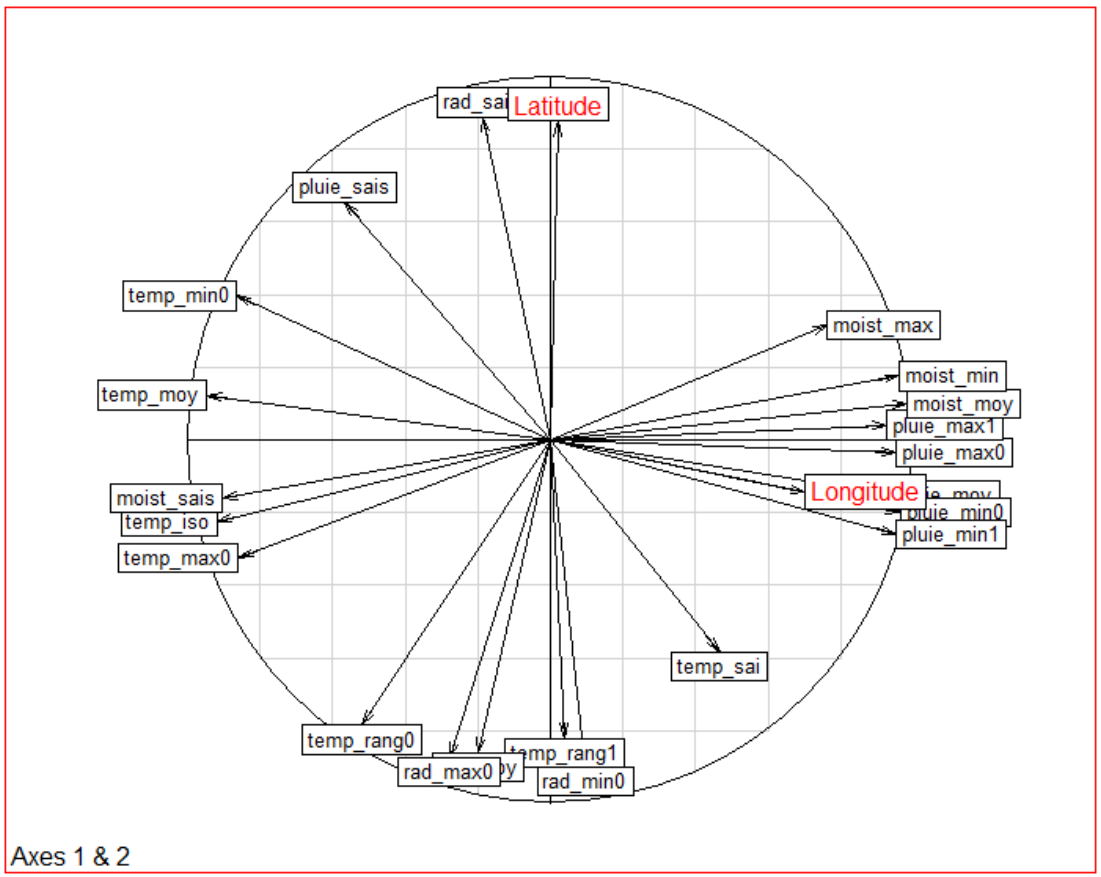


ACP

Variables/Interprétation avec les variables supplémentaires

geog¡-scale(fichier[,c(1,2)])
geog.supcol¡-supcol(fichier.pca,geog)$cosup
par(col="black") s.corcircle(fichier.pca$co,lab=climlab$V2,clab=.8,sub="Axes 1 et 2")

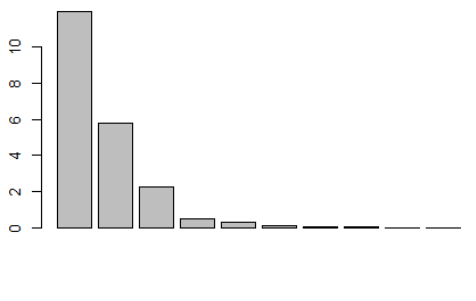
par(col="red") s.arrow(geog.supcol,add.plot=T)

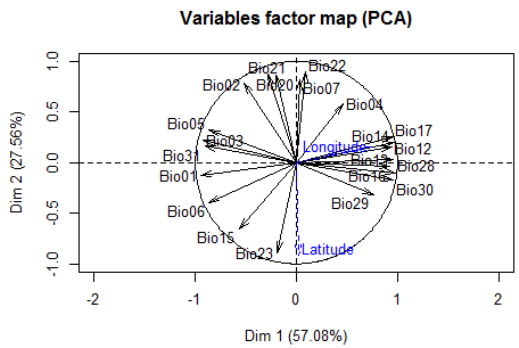


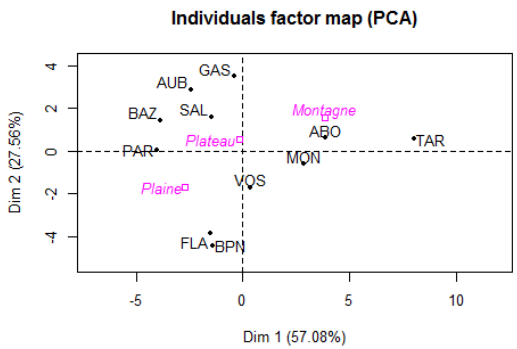
ACP
FactoMineR

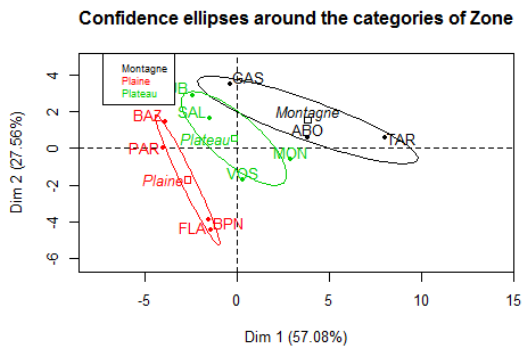
```
fichier.PCA ← =PCA(fichier1,ncp=3,quanti.sup=1:2,quali.sup=24)
par(mfrow=c(2,2))
barplot(fichier.PCA$eig[,1])
plot(fichier.PCA,choix="var")
plot(fichier.PCA,choix="ind")

plotellipses(fichier.PCA)
```

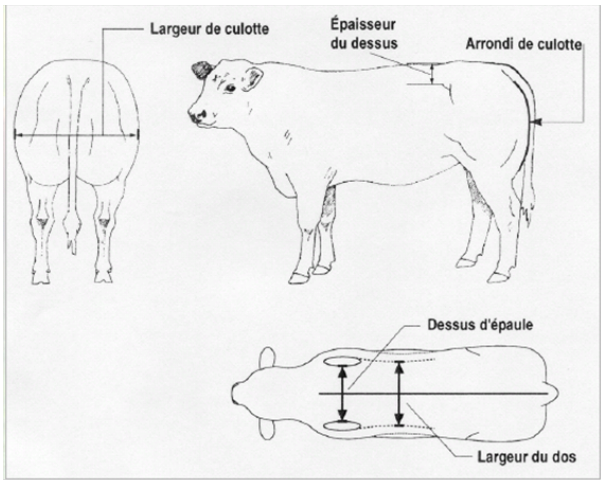


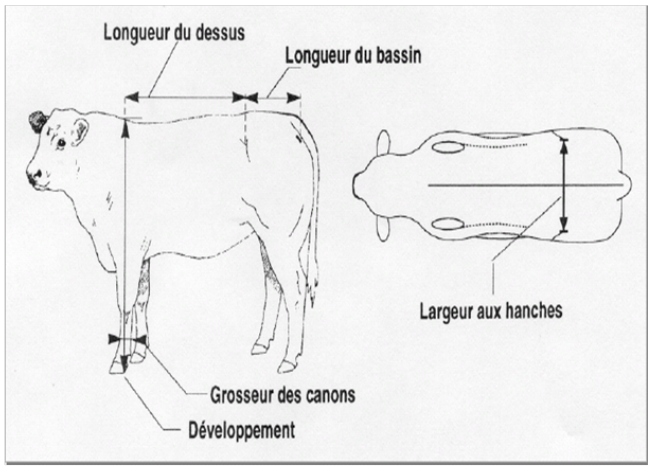


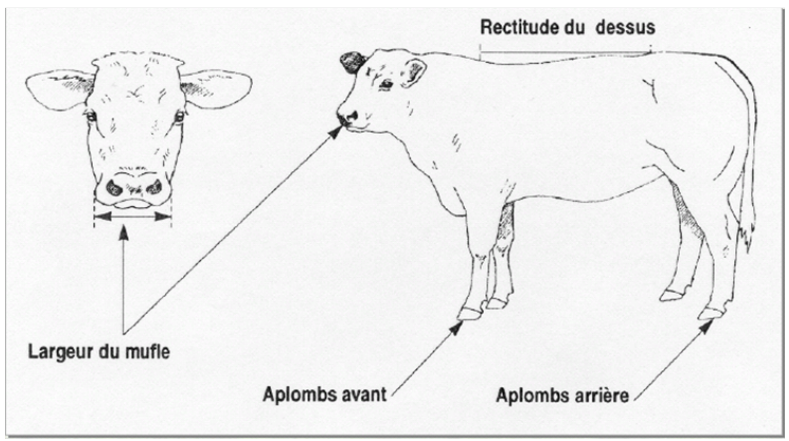




ACP entre classes
Pointage des bovins allaitants







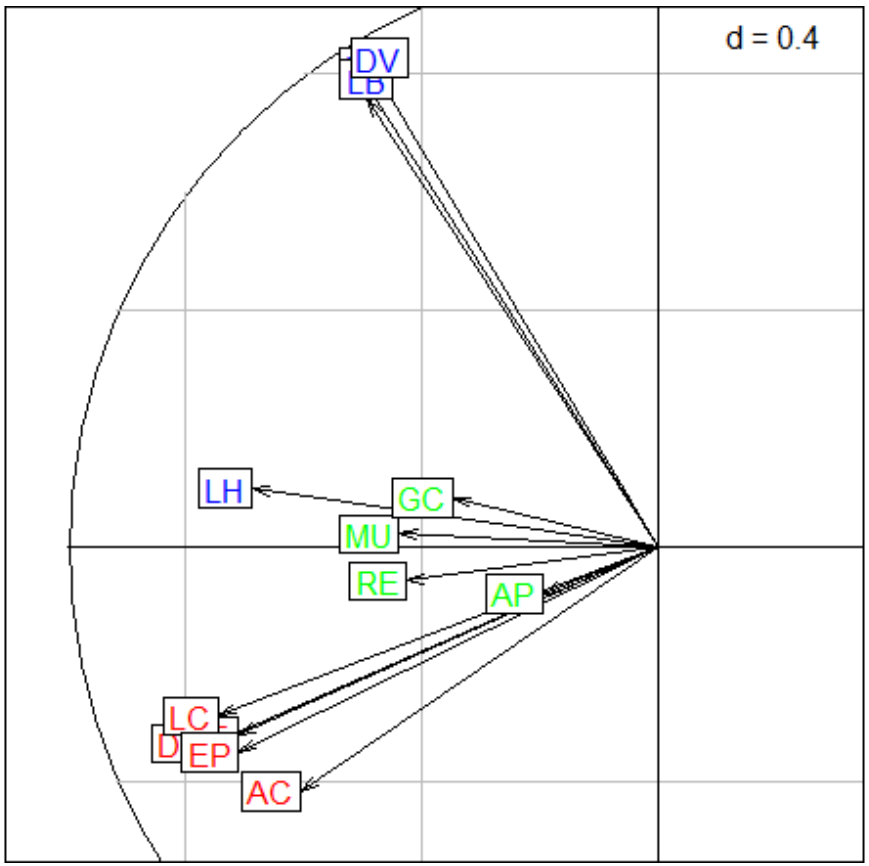
Pointage des bovins allaitants

Le tableau des corrélations

	DEPA	LADO	ARCU	LACU	EPDE	GRCA	LODE	LBA	LAH	DEV	TETE	APV	APR	REDE
DEPA	1	0,89	0,73	0,78	0,78	0,19	0,22	0,21	0,54	0,1	0,23	0,13	0,17	0,23
LADO	0,89	1	0,74	0,78	0,79	0,19	0,21	0,2	0,53	0,09	0,22	0,16	0,19	0,33
ARCU	0,73	0,74	1	0,81	0,8	0,06	0,09	0,07	0,43	-0,04	0,13	0,05	0,14	0,23
LACU	0,78	0,78	0,81	1	0,8	0,17	0,21	0,21	0,6	0,08	0,2	0,12	0,17	0,24
EPDE	0,78	0,79	0,8	0,8	1	0,15	0,2	0,19	0,52	0,08	0,21	0,13	0,17	0,29
GRCA	0,19	0,19	0,06	0,17	0,15	1	0,46	0,5	0,46	0,48	0,42	0,28	0,29	0,07
LODE	0,22	0,21	0,09	0,21	0,2	0,46	1	0,84	0,52	0,82	0,35	0,27	0,27	0,17
LOBA	0,21	0,2	0,07	0,21	0,19	0,5	0,84	1	0,54	0,8	0,37	0,28	0,28	0,16
LAH1	0,54	0,53	0,43	0,6	0,52	0,46	0,52	0,54	1	0,45	0,37	0,27	0,28	0,19
DEVE	0,1	0,09	-0,04	0,08	0,08	0,48	0,82	0,8	0,45	1	0,34	0,28	0,26	0,16
TETE	0,23	0,22	0,13	0,2	0,21	0,42	0,35	0,37	0,37	0,34	1	0,3	0,29	0,12
APAV	0,13	0,16	0,05	0,12	0,13	0,28	0,27	0,28	0,27	0,28	0,3	1	0,5	0,29
APAR	0,17	0,19	0,14	0,17	0,17	0,29	0,27	0,28	0,28	0,26	0,29	0,5	1	0,24
REDE	0,23	0,33	0,23	0,24	0,29	0,07	0,17	0,16	0,19	0,16	0,12	0,29	0,24	1

Pointage des bovins allaitants

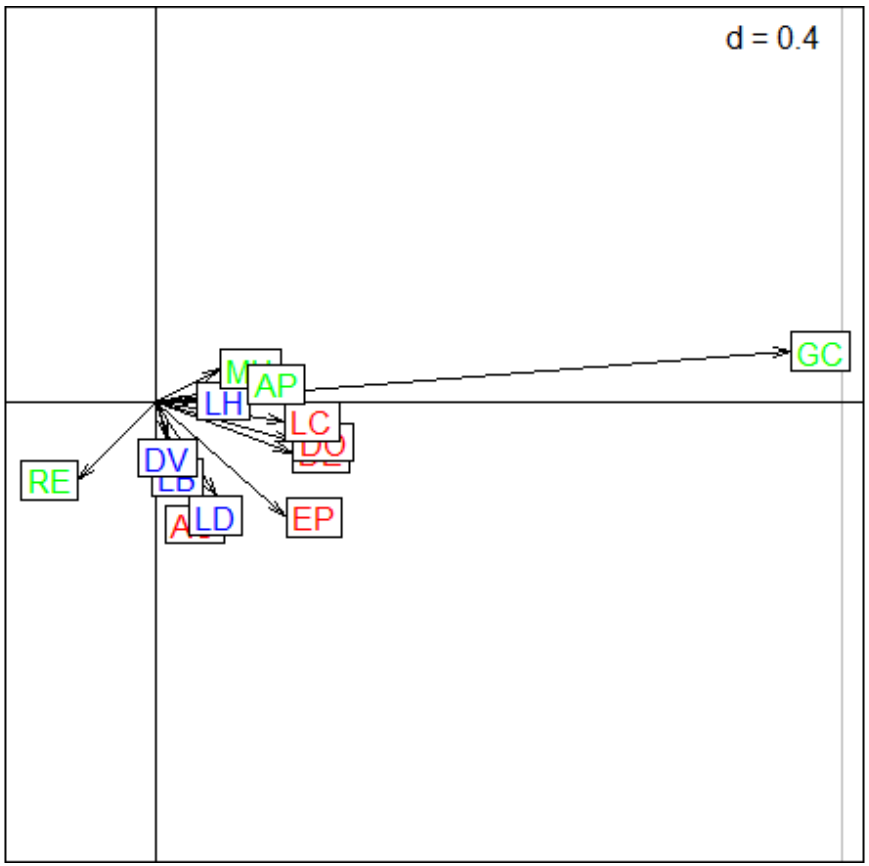
Le cercle des corrélations



ACP entre campagnes

L'ACP est réalisée sur le tableau des moyennes de variables par campagne

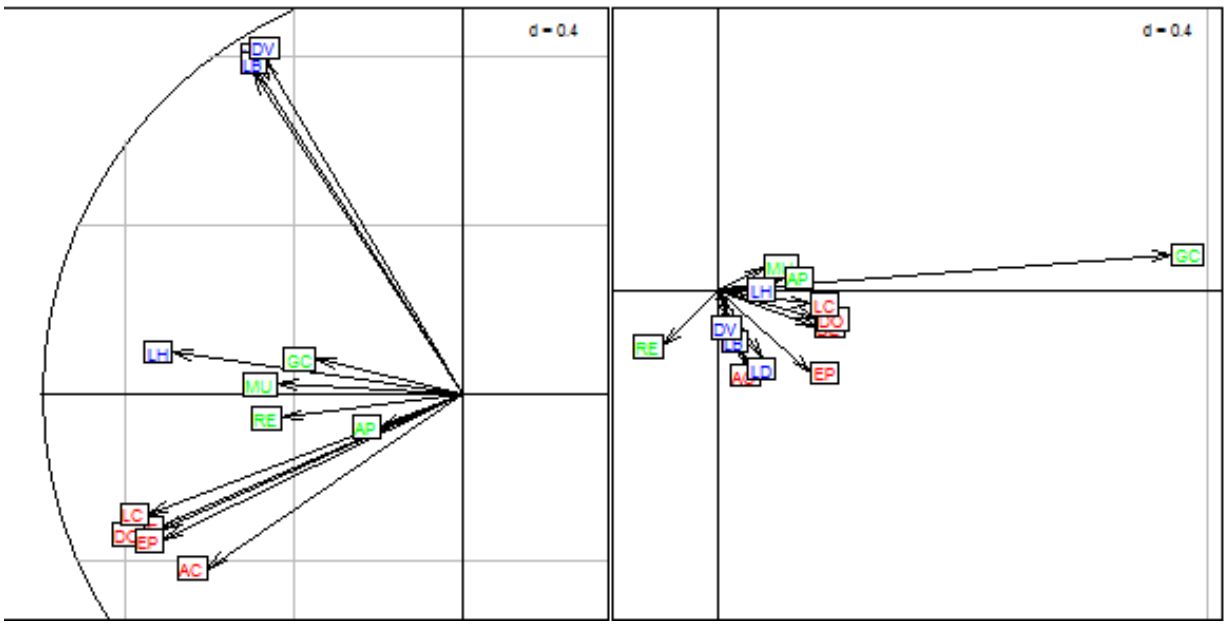
Le cercle des corrélations



ACP entre campagnes

L'ACP est réalisée sur le tableau des moyennes de variables par campagne

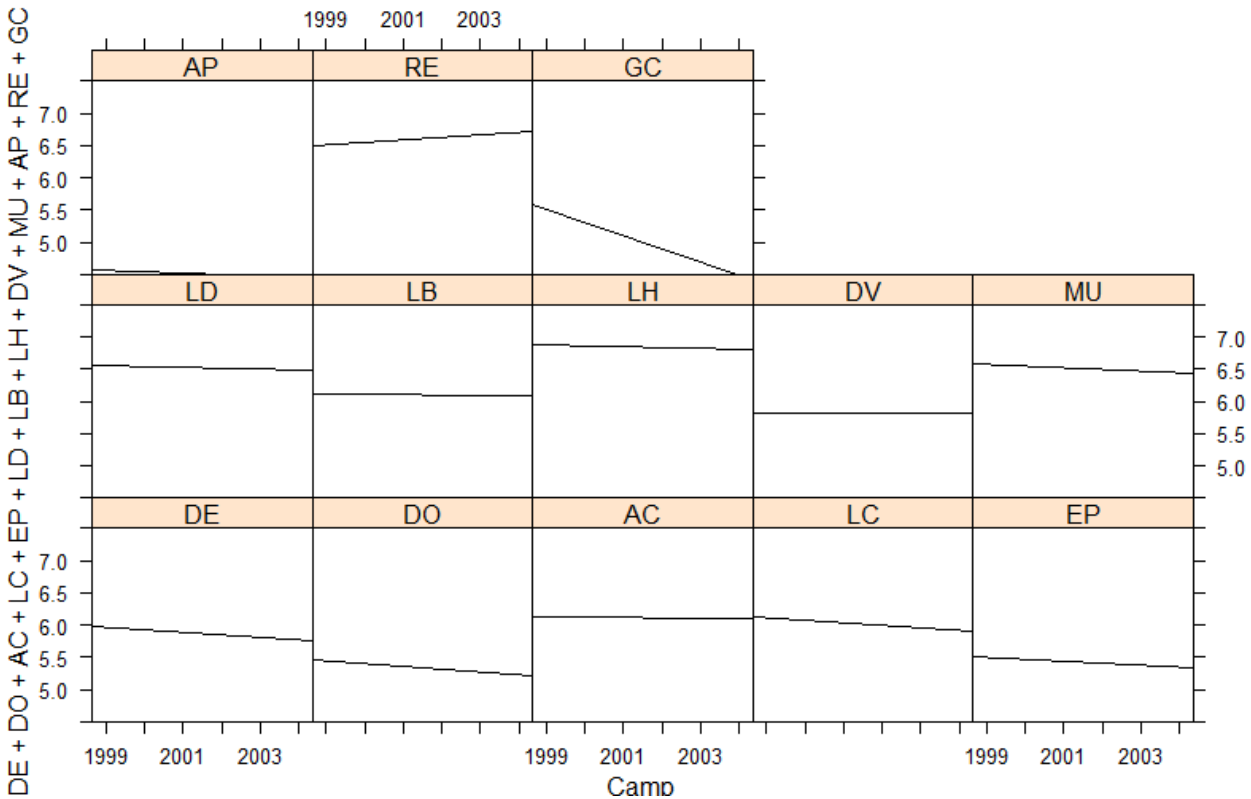
Les deux cercles des corrélations



Grosseur des canons

Package *lattice*

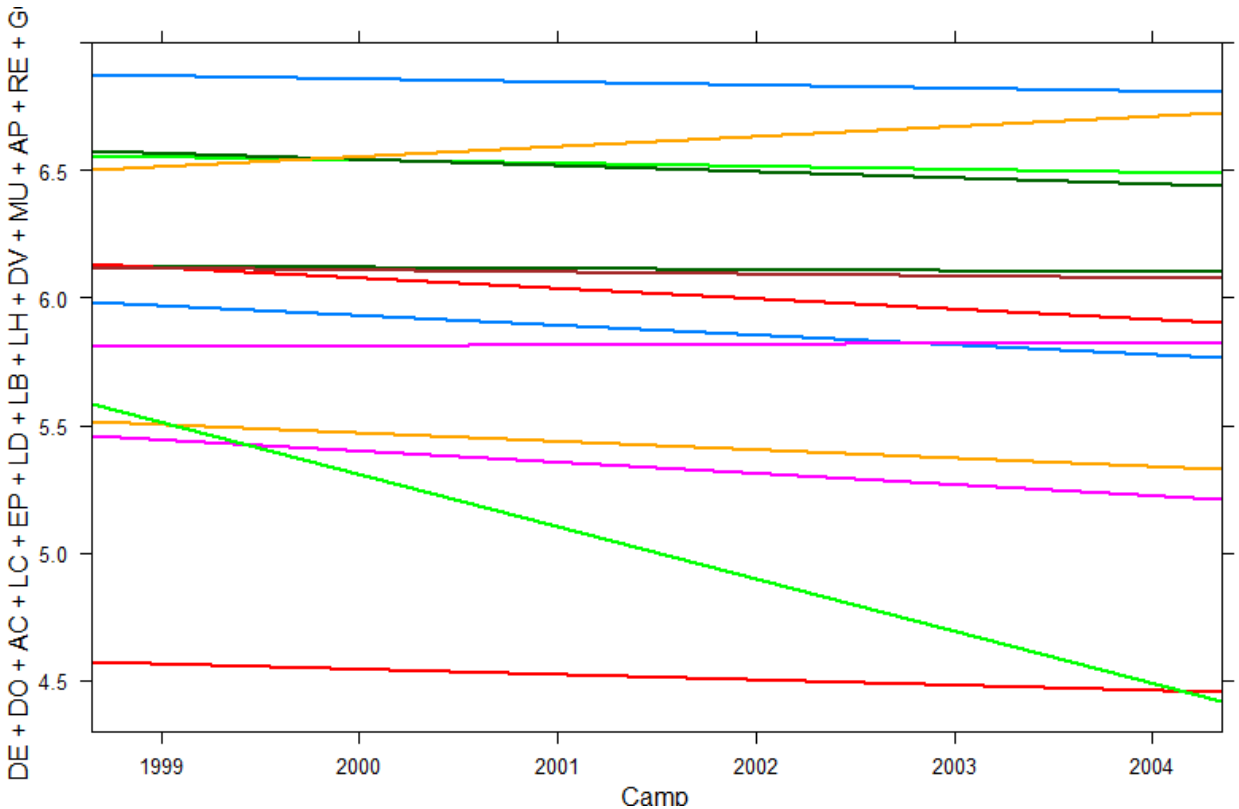
Evolution du GC par campagne



Grosseur des canons

Evolution du GC par campagne

Package *lattice*

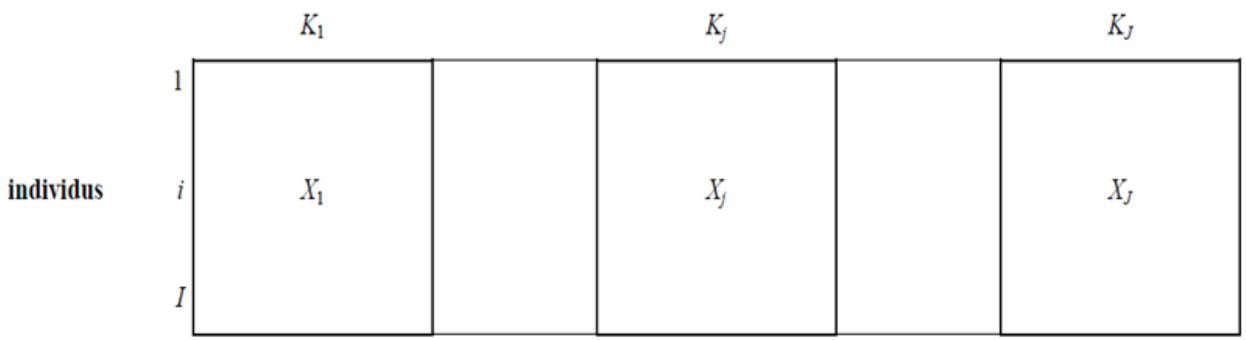


L'intégration de données

Appréhender un phénomène selon différents points de vue

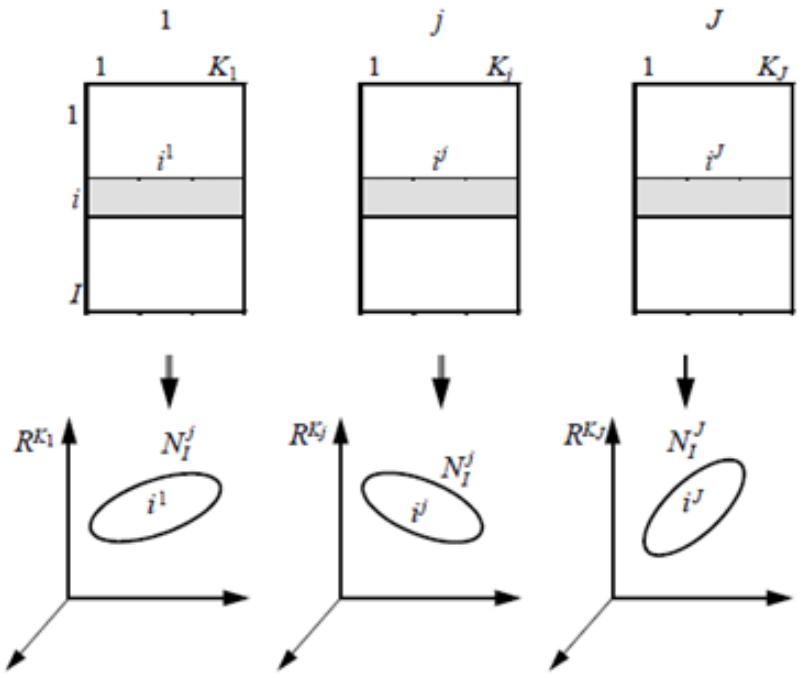
L'Analyse Factorielle Multiple (AFM, Escofier et Pagès, 1997)

- L'AFM traite simultanément des tableaux croisant les mêmes individus
- Représentation superposée des ACP partielles
- Pondération équilibrée des différents groupes de variables



L'Analyse Factorielle Multiple

Représentation superposée des ACP partielles



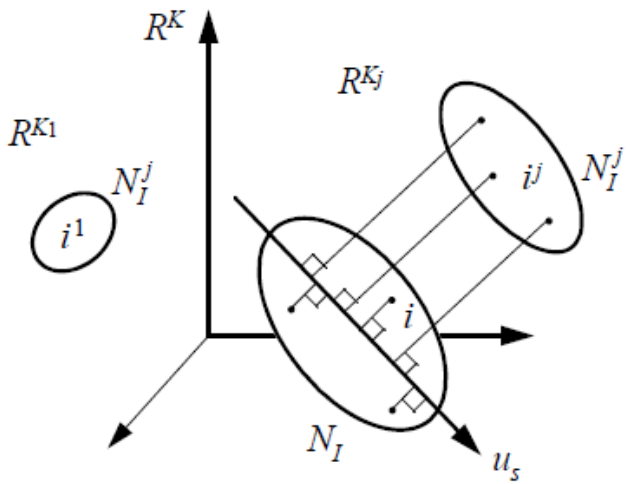
N_j : nuage partiel (des individus vus par le groupe j)

http://factominer.free/docs/AFM.pdf

L'Analyse Factorielle Multiple

Représentation superposée des ACP partielles

Principe



Les nuages partiels sont projetés sur les axes principaux d'inertie du nuage moyen

http://factominer.free/docs/AFM.pdf

L'Analyse Factorielle Multiple

(Gelein et Sautory, 2007)

TOTALE		INTER		INTRA
Inertie totale de l'union des nuages partiels	=	Inertie du nuage compromis des individus	+	Inertie des individus partiels autour de leur compromis

Objectifs:

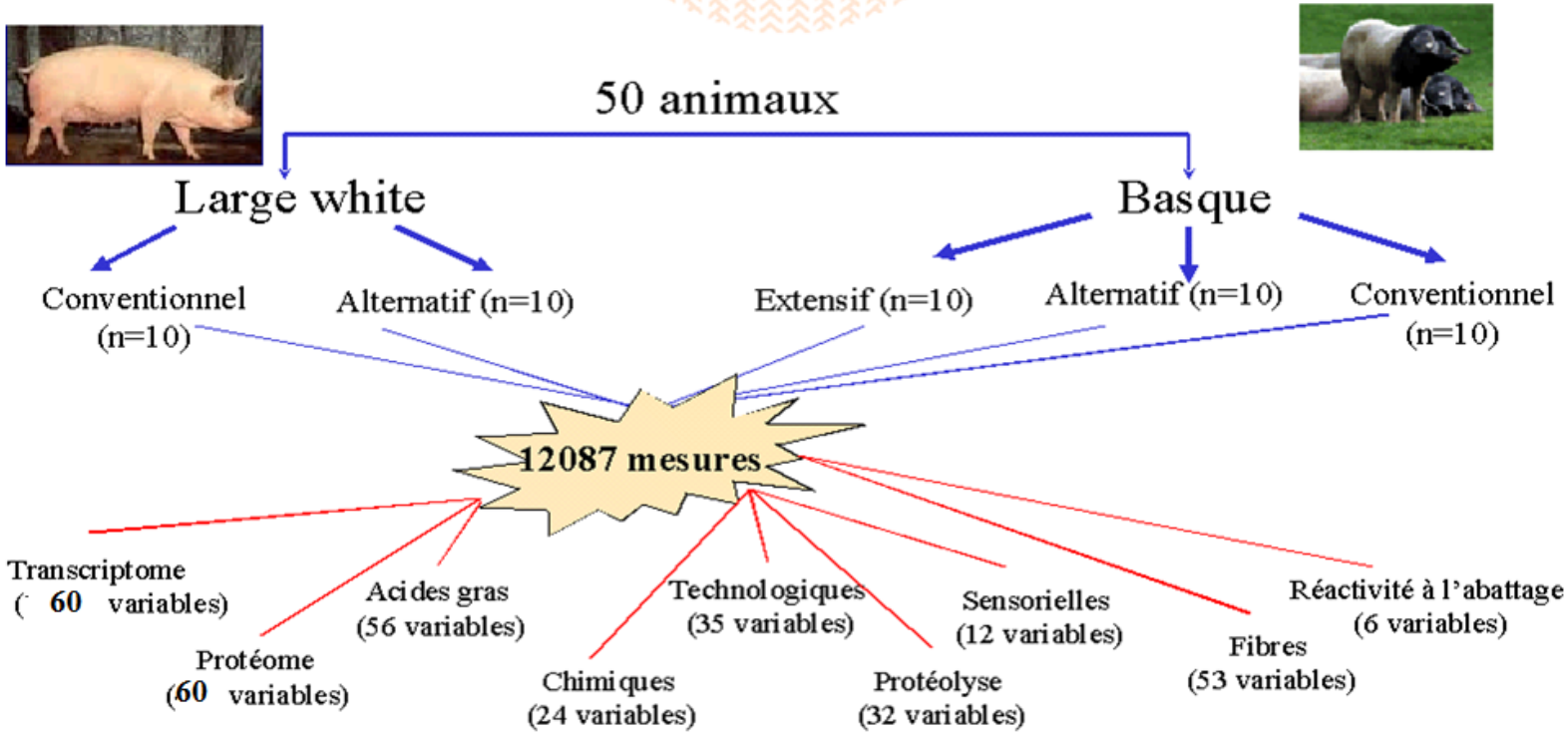
ACP : maximiser Totale = Inter + Intra

AFM: maximiser Inter - Intra

Pondérations : $\left\{ \frac{1}{\lambda_1^{[1]}}, \frac{1}{\lambda_1^{[2]}}, \ldots, \frac{1}{\lambda_1^{[j]}}, \ldots, \frac{1}{\lambda_1^{[J]}} \right\}$

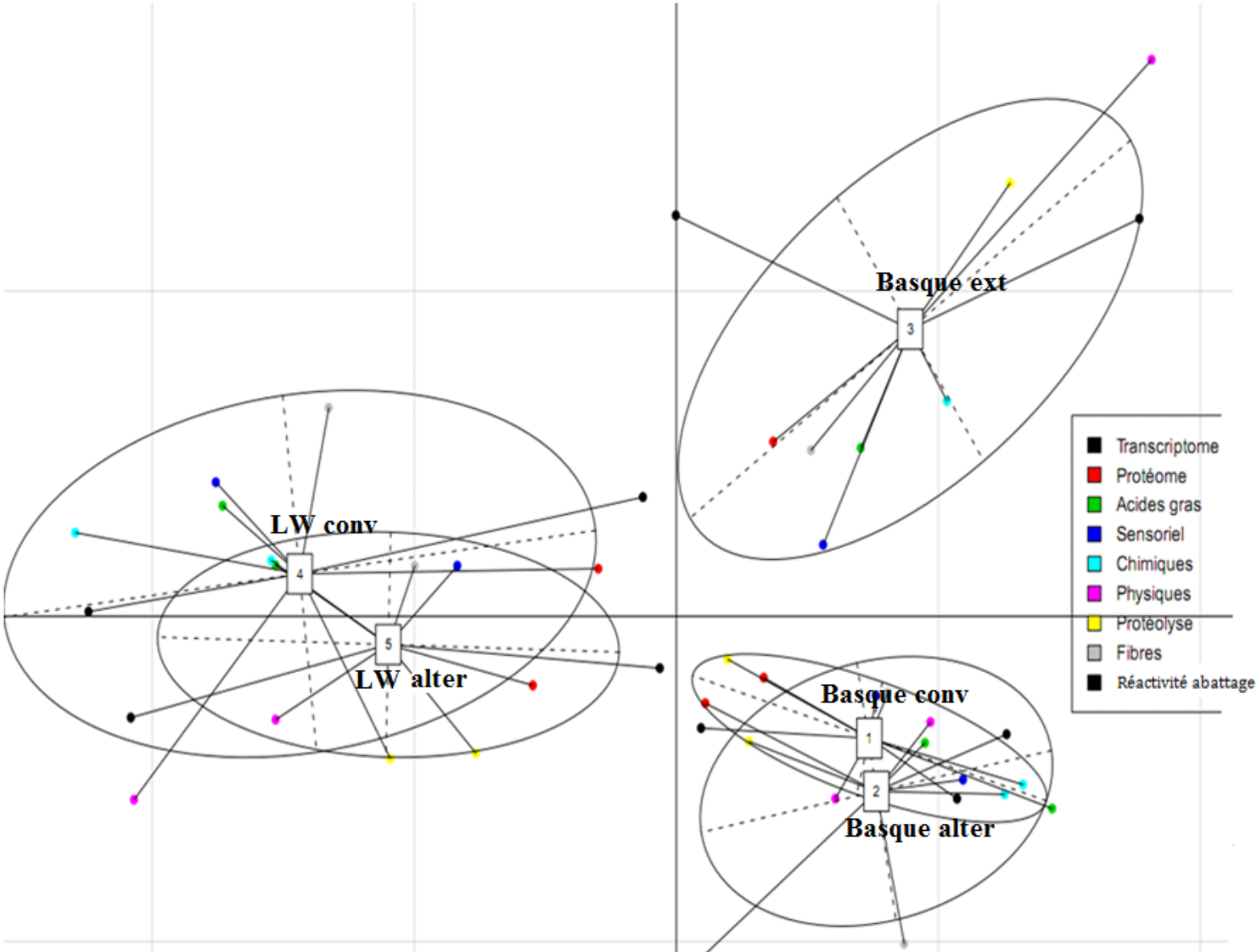
Une application sur le porc

Thèse de B Salmi



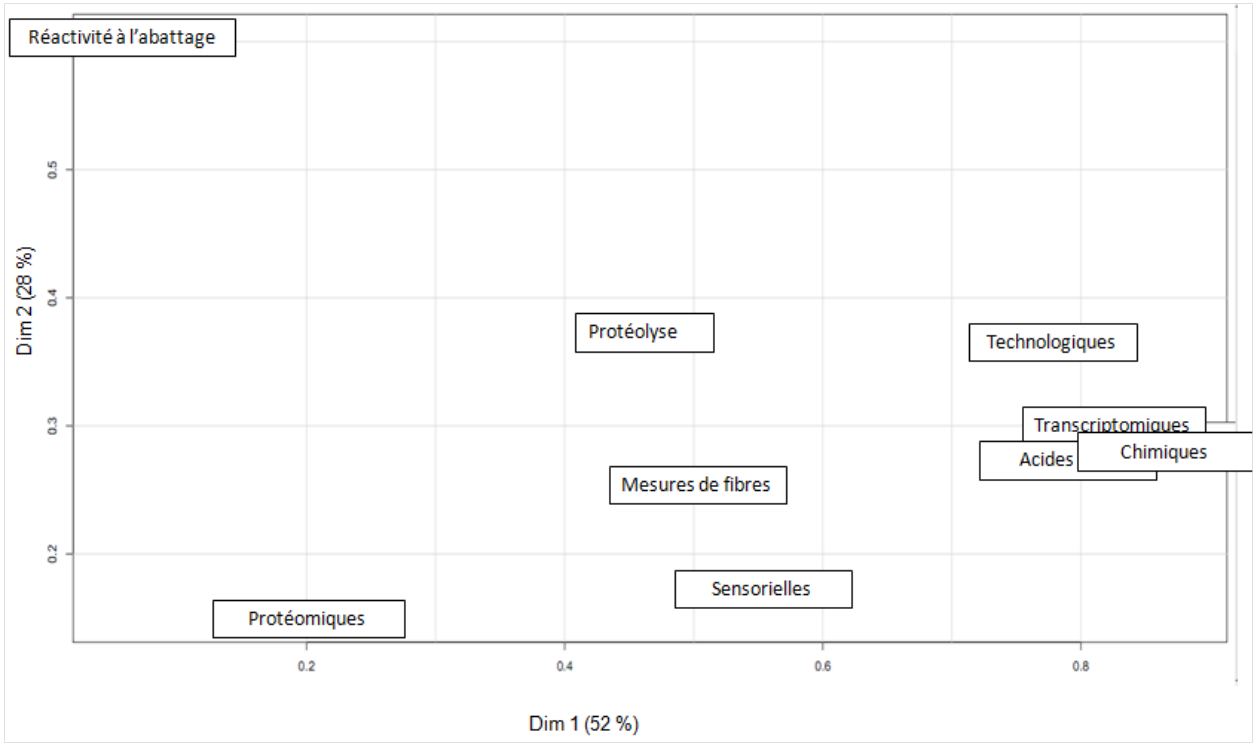
AFM entre races*systèmes d'élevage

Consensus et nuages partiels



AFM entre races*systèmes d'élevage

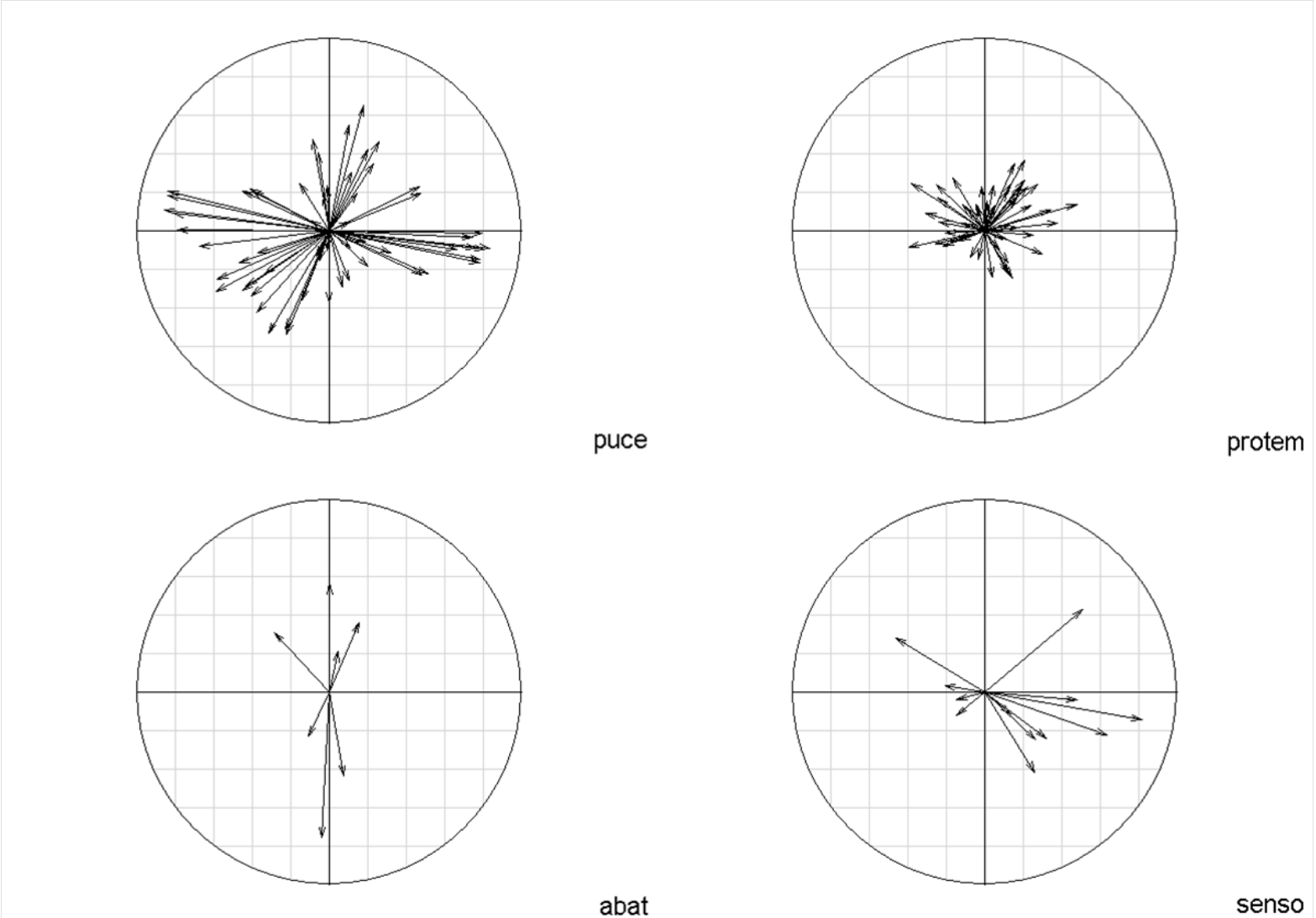
Lien des groupes de variables et des composantes



Lien du groupe k sur la composante i : inertie projetée du nuage k sur la composante i

AFM entre races*systèmes d'élevage

Retour aux variables



Classification Hiérarchique

- Classer / Partitionner des individus en classes
- Distances entre individus
- Homogénéité intra-groupes *vs* Hétérogénéité entre groupes
- Hiérarchisation dans un dendrogramme

Un grand nombre de distances
Distances écologiques

	1	0	Total
1	a	b	a+b
0	c	d	c+d
Total	a+c	c+d	n

- Indice de Jaccard $\frac{a}{a+b+c}$
- Indice de Sokal et Michener $\frac{a+d}{n}$
- ...

pbil.univ-lyon1.fr/R/stage/stage7.pdf

Un grand nombre de distances
Distances génétiques

A partir des fréquences alléliques

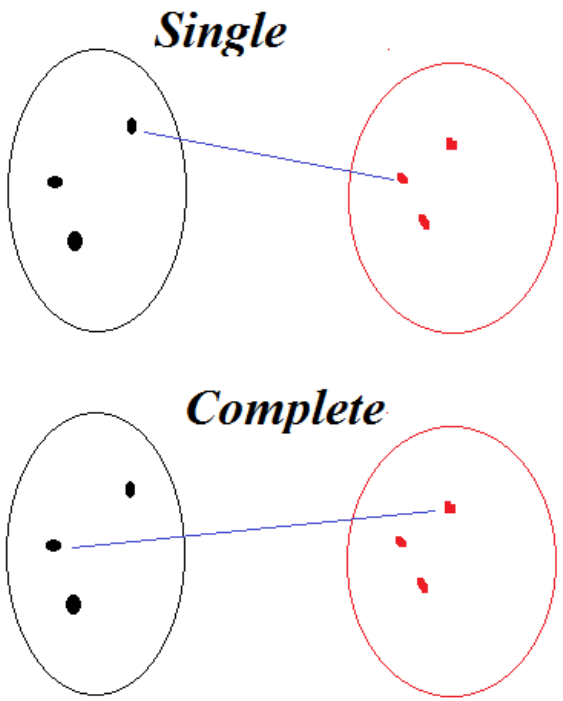
- distance de Nei
- distance de Reynolds
- distance de Rogers (euclidienne)
- ...

Distances dans le package *ade4*

A partir des fréquences alléliques

- Distances génétiques : dist.genet
- Distances écologiques : dist.binary
- Distances à partir d'analyses factorielles : dist.dudi
- ...

Un grand nombre de méthodes d'aggrégation



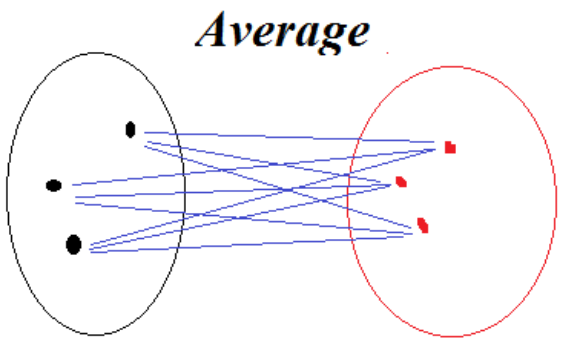
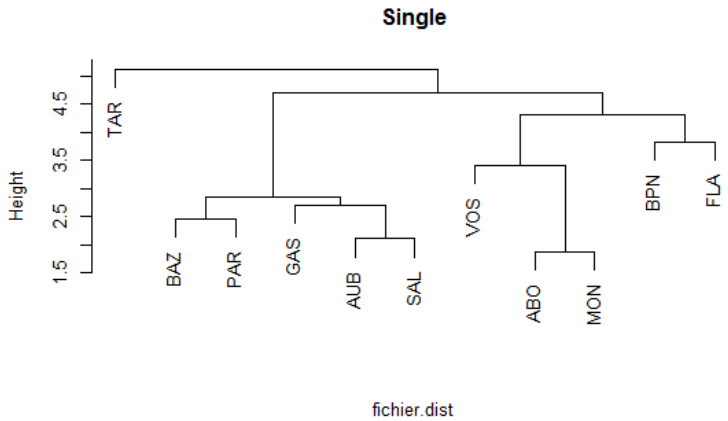
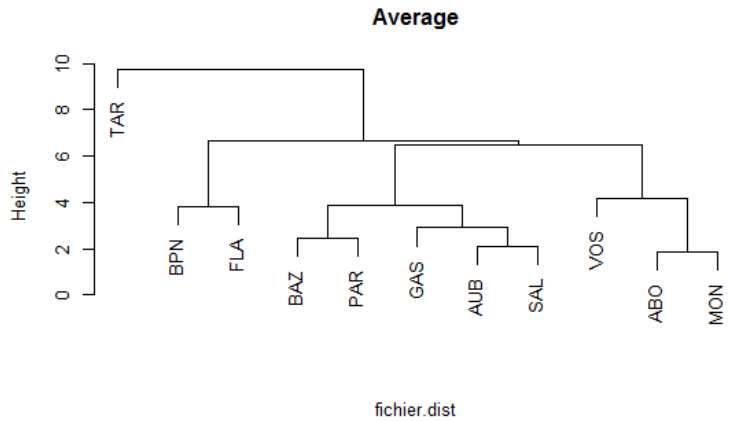


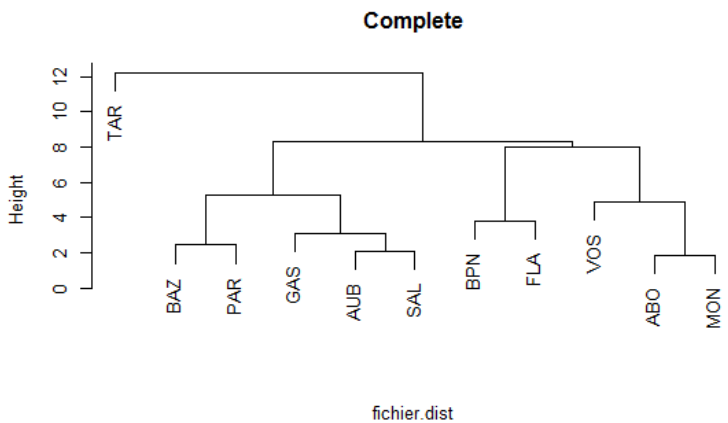
Schéma général d'une classification hiérarchique

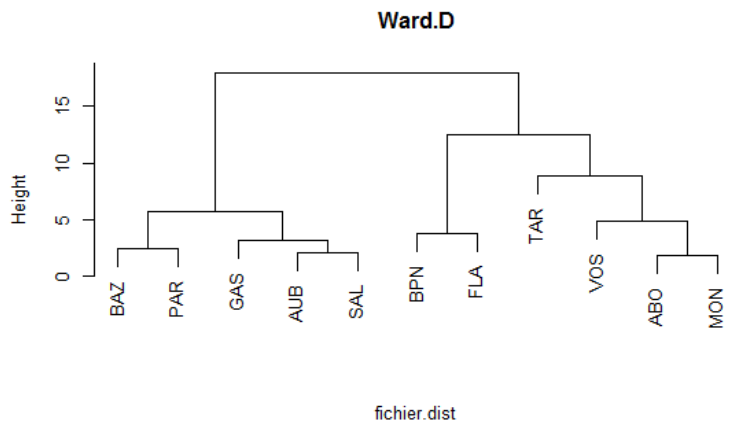
1. Identifier les individus / clusters les plus proches à partir d'une matrice de dissimilarités.
2. Les réunir dans un nouveau cluster. On passe de n à n-1 clusters.
3. Recalculer la matrice de dissimilarité.
4. Retourner à l'étape 1.

Exemple sur les données du climat





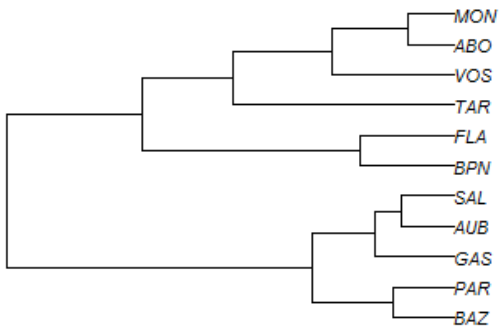




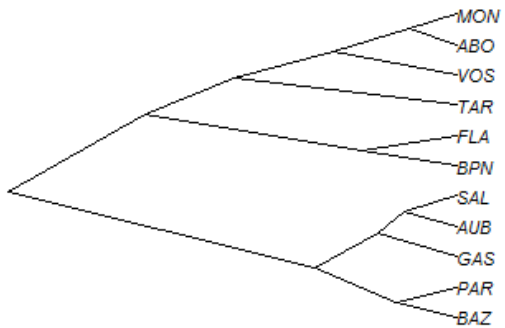
Exemple sur les données du climat
Un grand nombre de représentations

Package *ape*, objet *phylo*

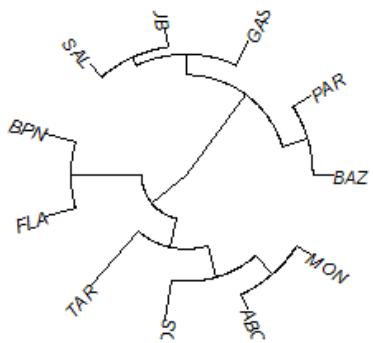
phylogramme



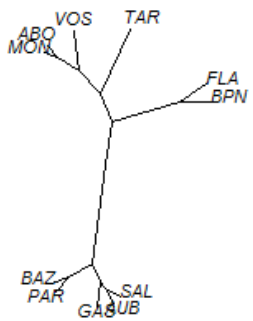
cladogramme



fan



unrooted



Partitionnement
Agrégation autour de centres mobiles (kmeans)

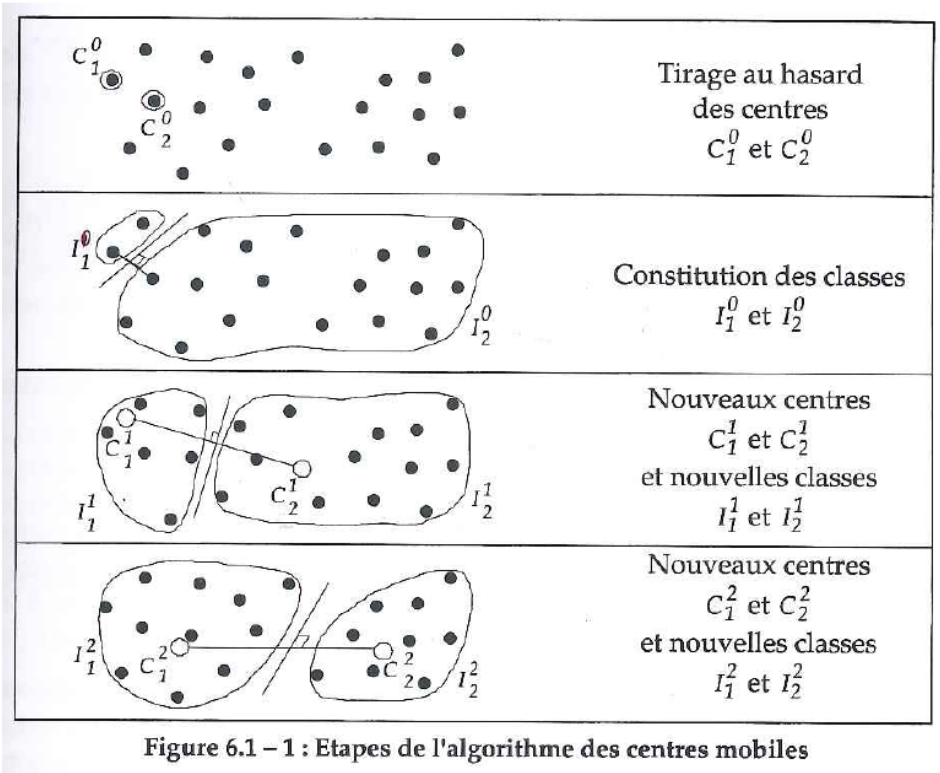
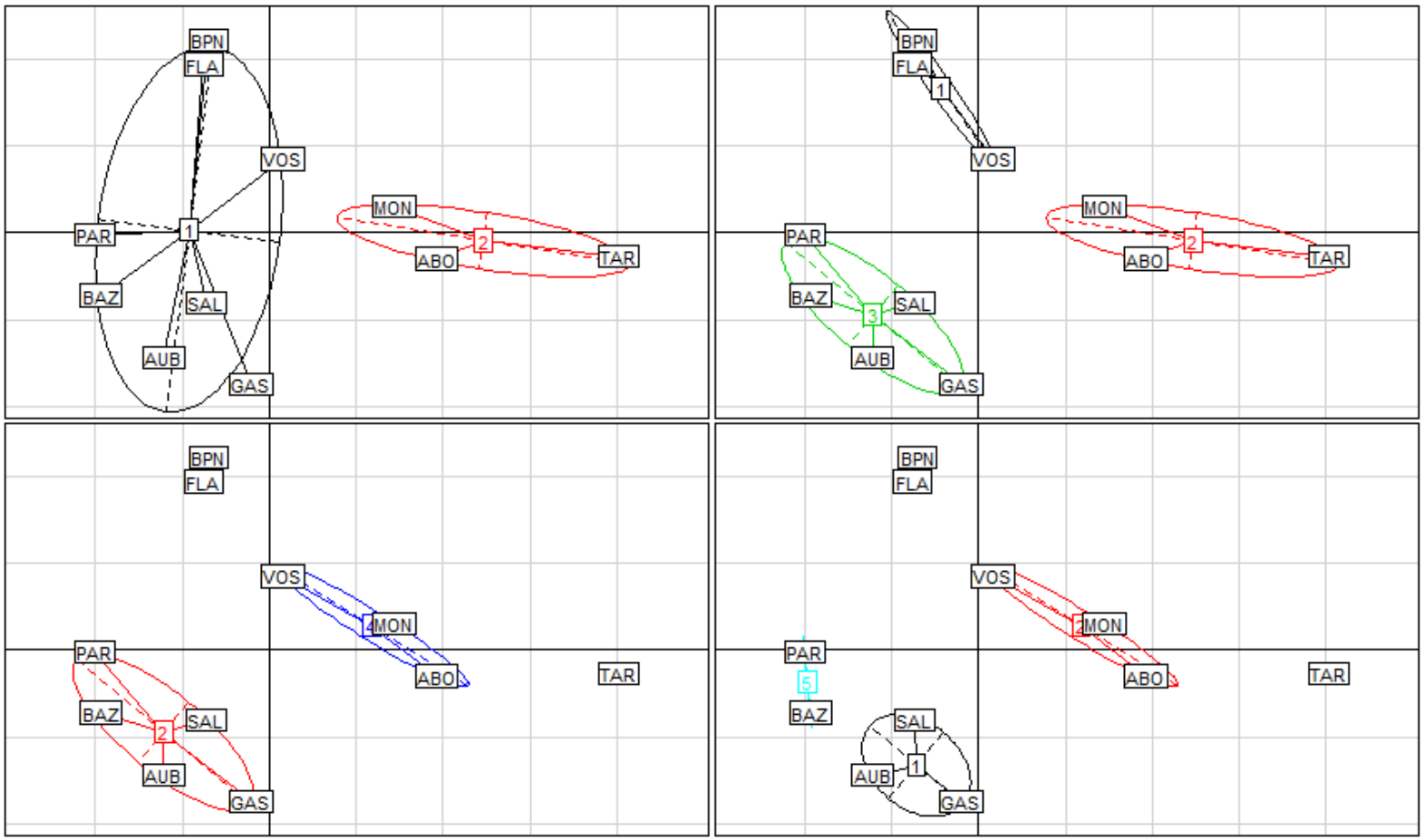


Figure 6.1 – 1 : Etapes de l'algorithme des centres mobiles

Lebart et al,2006

Partitionnement
Agrégation autour de centres mobiles (kmeans)



Conclusion I
Analyses factorielles

Analyses simples

ACP, Correspondances, Correspondances multiples, Hill et Smith

Analyses multitables

Analyse Factorielle Multiple, Analyse triadique partielle, RLQ, Analyse des corrélations canoniques, STATIS, Coinertie, Coinertie multiple,

Information exogène

Analyse entre classes, ACP sur variables instrumentales

Conclusion II
Classification

Classification hiérarchique

Un grand nombre de

- Distances
- Méthodes
- Représentations

Partitionnement

adaptée aux grands tableaux

Méthodes complémentaires

Les techniques de classification peuvent compléter et nuancer les résultats des analyses factorielles (Lebart et al, 2006)

- Compréhension de la structure des données
- Aide dans l'interprétation

References / Extensions R

- Armatte, M. (2008). Histoire et préhistoire de l'analyse des données par J P Benzécri. Un cas de généalogie rétrospective. *Journal Electronique d'Histoire des Probabilités et de la Statistique, 24:2, décembre 2008*
- Benzecri J.P. (1976,1977). Histoire et préhistoire de l'analyse des données. *Cahiers de l'Analyse des Données, 1 à 4, 1976; 1, 1977*.
- Bressoux, P. (2010). Modélisation statistique appliquée aux sciences sociales. *de boeck*.
- Dray, S. et Dufour, A-B. (2007). The ade4 package: implementing the duality diagram for ecologists. *Journal of Statistical Software*, 22(4).
- Escofier B., et Pagès, J.(1998). Analyses factorielles simples et multiples. *Dunod*.
- Lebart, L., Piron, M, Morineau, A. (2006). Statistique exploratoire multidimensionnelle. *Dunod*.
- Legendre P., Legendre L. (2012). Numerical ecology. *Elsevier*.
- Le Roux B., Rouanet, H, 2004. Geometric Data Analysis. *Kluwer*
- Salmi B. et al (2010). Multivariate analysis to compare pig meat quality traits according to breed and rearing system .*Proceedings of the 9th WCGALP, Leipzig, August 1-6, 2010, 442*

- ade4. *http://pbil.univ-lyon1.fr/ade4/*
- FactomineR *http://factominer.free.fr/*
- vegan *http://cran.r-project.org/web/packages/vegan/*