

Formation à l'analyse statistique

Module 2 : statistique multidimensionnelle

- A) Classification ascendante hiérarchique
- B) Agrégation autour de centres mobiles
- C) Analyse en Composantes Principales
- D) Analyse Factorielle Discriminante
- E) Analyse des Correspondances
- F) Régression linéaire multiple

Jeu de données « morpho »

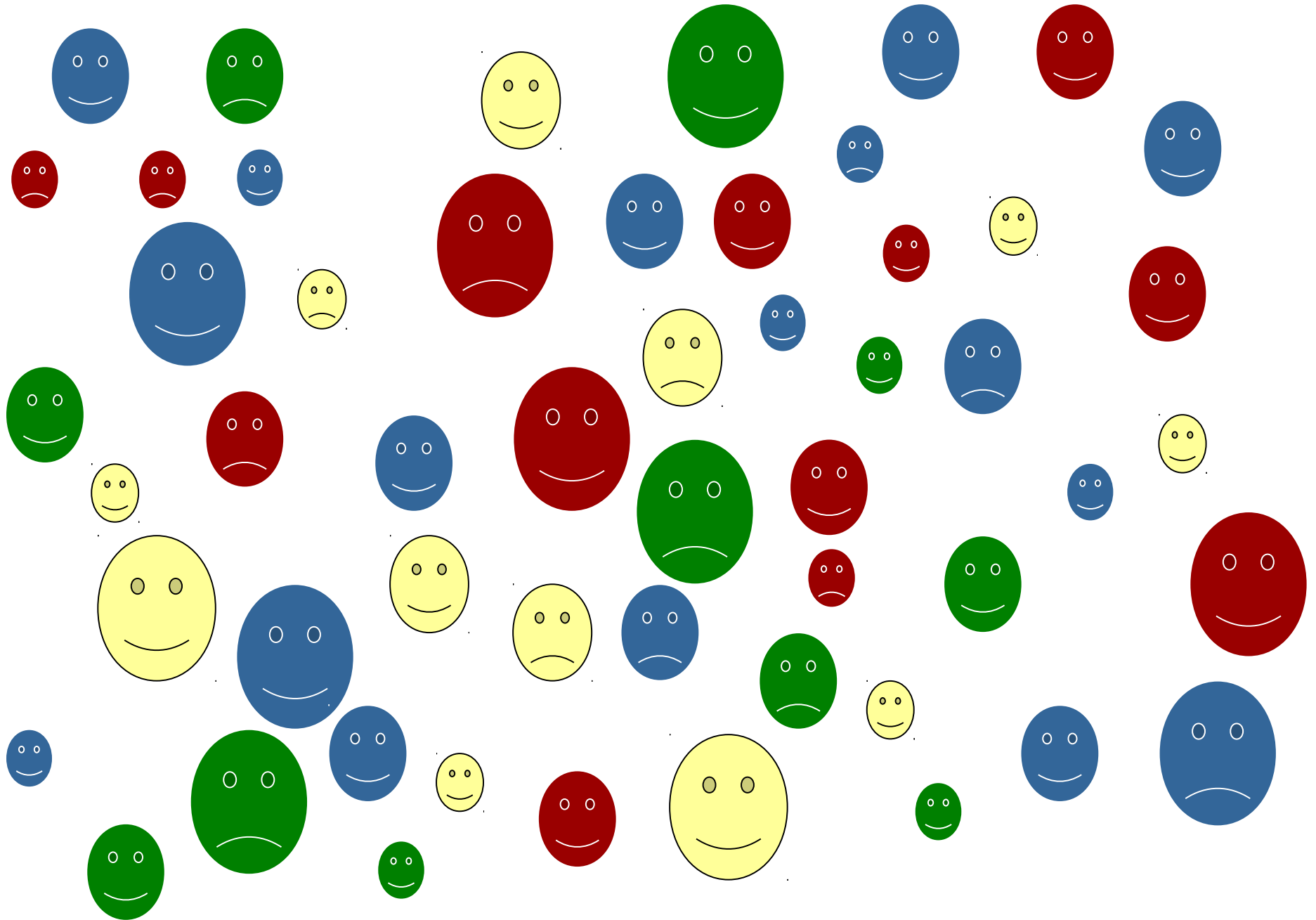
- 20 individus : 10 femmes, 10 hommes
- 5 variables :
 - V1 : tour au niveau des épaules (cm)
 - V2 : tour de poitrine (cm)
 - V3 : tour de taille (cm)
 - V4 : masse (kg)
 - V5 : taille (cm)

	V1	V2	V3	V4	V5		V1	V2	V3	V4	V5
H 1	106.2	89.5	71.5	65.6	174.0	F 1	105.0	89.0	71.2	67.3	169.5
H 2	110.5	97.0	79.0	71.8	175.3	F 2	100.2	94.1	79.6	75.5	160.0
H 3	115.1	97.5	83.2	80.7	193.5	F 3	99.1	90.8	77.9	68.2	172.7
H 4	104.5	97.0	77.8	72.6	186.5	F 4	107.6	97.0	69.6	61.4	162.6
H 5	107.5	97.5	80.0	78.8	187.2	F 5	104.0	95.4	86.0	76.8	157.5
H 6	119.8	99.9	82.5	74.8	181.5	F 6	108.4	91.8	69.9	71.8	176.5
H 7	123.5	106.9	82.0	86.4	184.0	F 7	99.3	87.3	63.5	55.5	164.4
H 8	120.4	102.5	76.8	78.4	184.5	F 8	91.9	78.1	57.9	48.6	160.7
H 9	111.0	91.0	68.5	62.0	175.0	F 9	107.1	90.9	72.2	66.4	174.0
H 10	119.5	93.5	77.5	81.6	184.0	F 10	100.5	97.1	80.4	67.3	163.8

Classification

Objectif : regrouper des objets qui se ressemblent

Classification



Classification : questions et méthodes

- Connait-on a priori le nombre de groupes ?
- Quel est le critère d'homogénéité ?
- Peut-on calculer une distance entre objets et/ou groupe d'objets ?
- ...

A) Classification hiérarchique

- sans a priori sur le nombre de groupes
- *clustering*

B) Agrégation autour de centres mobiles

- nombre de groupes connu a priori
- *classification, partitionning*

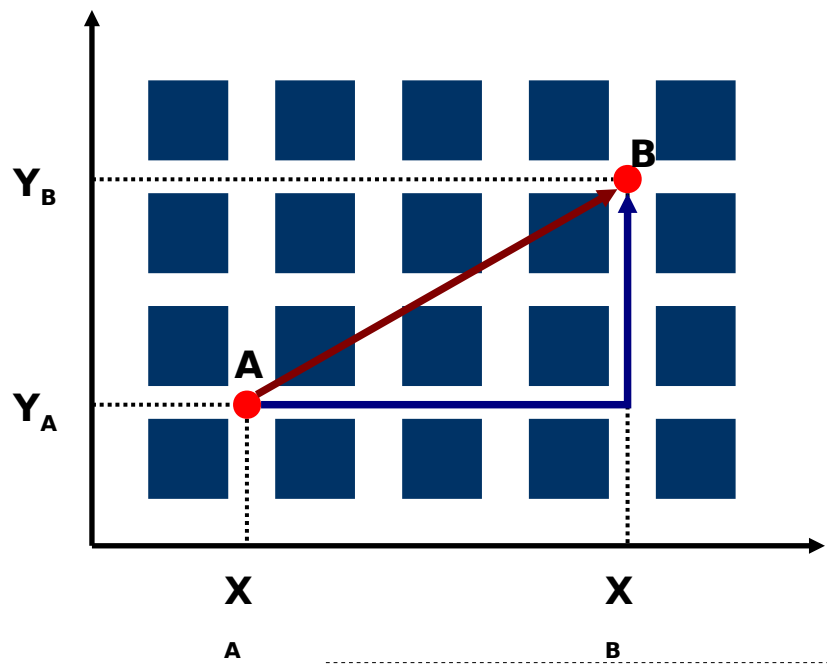
Classification ascendante hiérarchique

- Préalable : choisir
 - 1) une **distance inter-individus**
 - 2) un **critère d'agglomération** (distance entre groupes d'individus)
- Procédure itérative
 - **Début** : chaque individu est un groupe
 - **Déroulement** : regroupement des 2 objets les plus proches
 - **Fin** : une classe regroupe tous les individus
- Résultat
 - construction d'un arbre de classification (dendrogramme)

Distance inter-individus

Quelle est la distance entre les points A et B ?

→ Ça dépend de comment on se déplace de A à B.

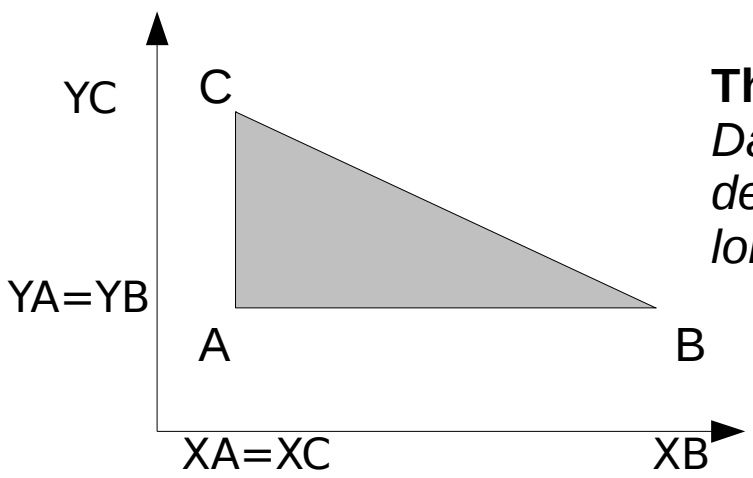


À vol d'oiseau



En contournant les immeubles

Et comment calculer la longueur du trajet ?



Théorème de Pythagore

Dans un triangle rectangle, le carré de la longueur de l'hypoténuse est égal à la somme des carrés des longueurs des 2 autres côtés.

$$BC^2 = AB^2 + AC^2$$

$$BC^2 = (X_B - X_C)^2 + (Y_C - Y_B)^2$$

Distance euclidienne

(à vol d'oiseau)

Dimension 2

X	X1	X2
Y	Y1	Y2

$$d_2(X, Y) = \sqrt{(X1 - Y1)^2 + (X2 - Y2)^2}$$

Dimension 3

X	X1	X2	X3
Y	Y1	Y2	Y3

$$d_2(X, Y) = \sqrt{(X1 - Y1)^2 + (X2 - Y2)^2 + (X3 - Y3)^2}$$

Dimension n

$$d_2(X, Y) = \sqrt{\sum_{i=1}^n (X_i - Y_i)^2} = \sqrt{(X_1 - Y_1)^2 + \dots + (X_n - Y_n)^2}$$

Détail d'un calcul
données « morpho »
(n=5)

	V1	V2	V3	V4	V5	
H1	106,2	89,5	71,5	65,6	174,0	Σ
H2	110,5	97,0	79,0	71,8	175,3	
H1-H2	-4,3	-7,5	-7,5	-6,2	-1,3	
(H1-H2) ²	18,5	56,3	56,3	38,4	1,7	
						171,1

$$d_2(H1, H2) = 13.08$$

Distance *city-block*

(en contournant les immeubles)

Dimension n

$$d_1(X,Y) = \sum_{i=1}^n |X_i - Y_i| = |X_1 - Y_1| + \dots + |X_n - Y_n|$$

Détail d'un calcul
données « morpho »
(n=5)

	V1	V2	V3	V4	V5	
H1	106,2	89,5	71,5	65,6	174,0	Σ
H2	110,5	97,0	79,0	71,8	175,3	
H1-H2	-4,3	-7,5	-7,5	-6,2	-1,3	
H1-H2	4,3	7,5	7,5	6,2	1,3	

$$d_1(H1,H2) = 26.8$$

Distances basées sur la corrélation

Plutôt que d'utiliser directement la corrélation, on préfère, dans ce contexte, transformer le coefficient pour obtenir des valeurs faibles pour des éléments proches.

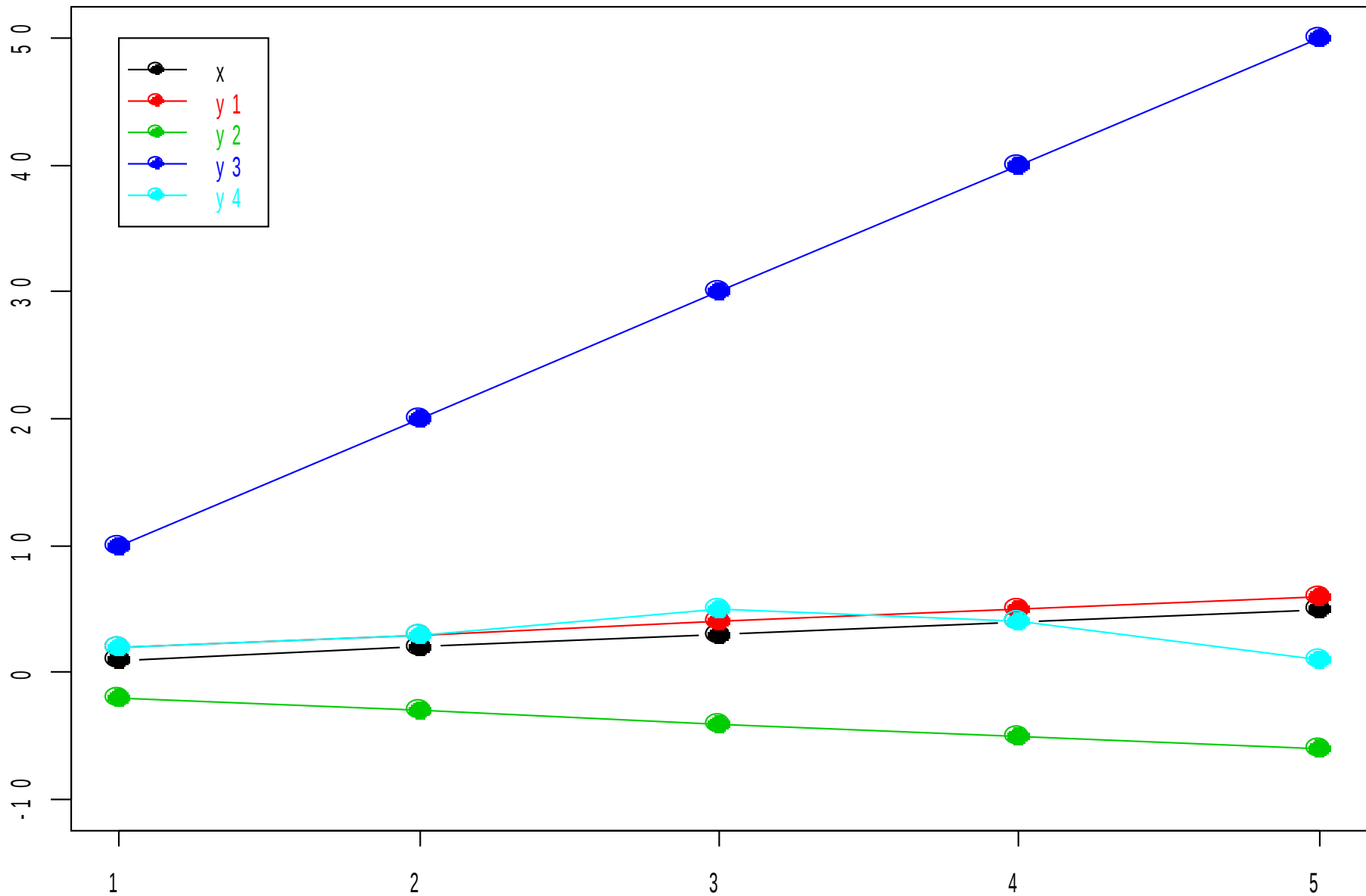
- **$1-\text{cor}(X,Y)$** : valeurs comprises entre 0 et 2
 - × Valeur proche de 0 : corrélation forte et positive
 - × Valeur proche de 1 : pas de corrélation
 - × Valeur proche de 2 : corrélation forte et négative
- **$\text{sqrt}[1-\text{cor}(X,Y)^2]$** : valeurs comprises entre 0 et 1
 - Valeur proche de 0 : corrélation forte (positive ou négative)
 - Valeur proche de 1 : pas de corrélation

Choix d'une distance

Il n'existe pas une distance meilleure que les autres ! Le choix de la distance dépend de l'objectif.

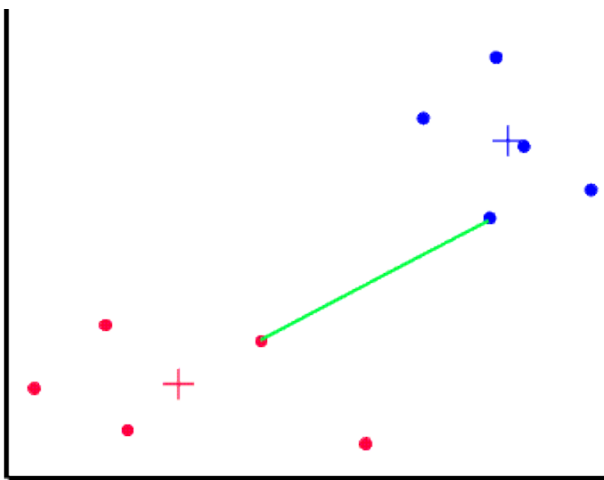
	X	Y₁	Y₂	Y₃	Y₄
	1	2	-2	10	2
	2	3	-3	20	3
	3	4	-4	30	5
	4	5	-5	40	4
	5	6	-6	50	1
Distance euclidienne	0	2	17	67	5
Corrélation 1-cor(X,Y_i)	0	0	2	0	1.1
Corrélation carrée sqrt(1-cor²(X,Y_i))	0	0	0	0	0.995

Comparaison de distances



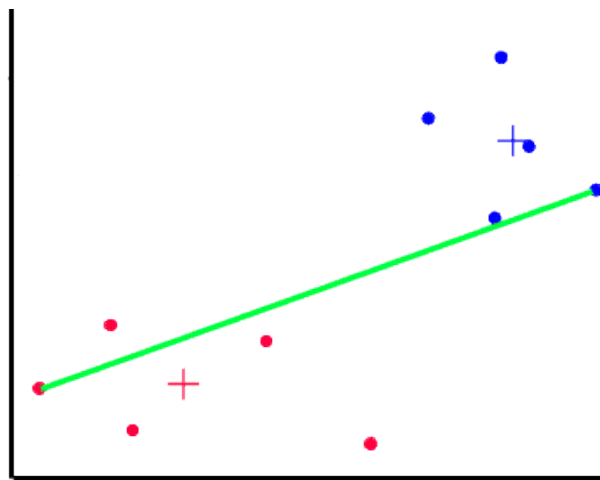
Critère d'agglomération (*linkage*)

Minimum
(*single*)



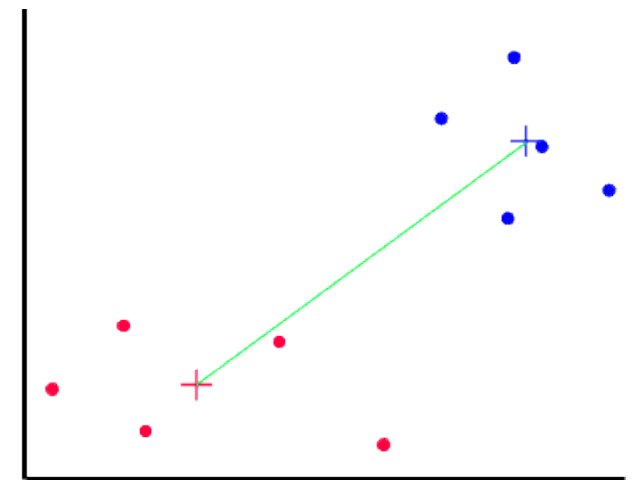
La distance entre 2 groupes est la distance entre les 2 points les plus proches.

Maximum
(*complete*)



La distance entre 2 groupes est la distance entre les 2 points les plus éloignés.

Moyen
(*average*)

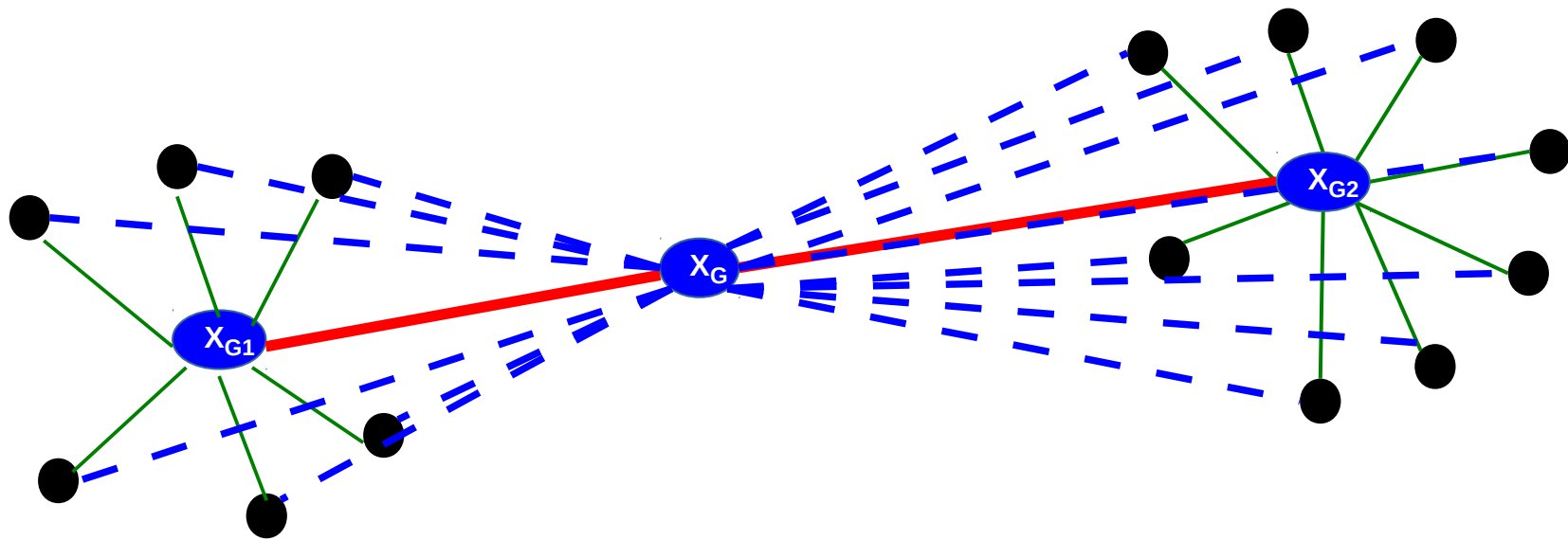


La distance entre 2 groupes est la distance entre les 2 barycentres de chaque groupe.

Critère de Ward

(décomposition de l'inertie totale)

$$\sum_{j=1}^p \sum_{i=1}^n (X_{ij} - X_G)^2 = \sum_{j=1}^p (X_{Gj} - X_G)^2 + \sum_{j=1}^p \sum_{i=1}^n (X_{ij} - X_{Gj})^2$$



L'affectation d'un point à une classe est effectuée selon la règle :

Minimiser l'inertie intra-classe \Leftrightarrow Maximiser l'inertie inter-classes

Un exemple « à la main »

Données	I1	V1	V2	V3	Effectuer la classification hiérarchique des individus avec la distance euclidienne et le critère de saut minimum.
5 individus	I2	1	2	3	
3 variables	I3	4	2	5	
	I4	4	3	7	
	I5	8	9	6	

Étape 0

Calcul des distances inter-individus

$$d(I1, I2) = \sqrt{(1-4)^2 + (2-2)^2 + (3-5)^2} = \sqrt{(3^2 + 0^2 + 2^2)} = \sqrt{13}$$

	I1	I2	I3	I4
I2	3.61			
I3	5.10	2.24		
I4	10.34	8.12	7.28	
I5	3.00	2.00	4.12	8.60

À chaque étape, regroupement des 2 objets les plus proches

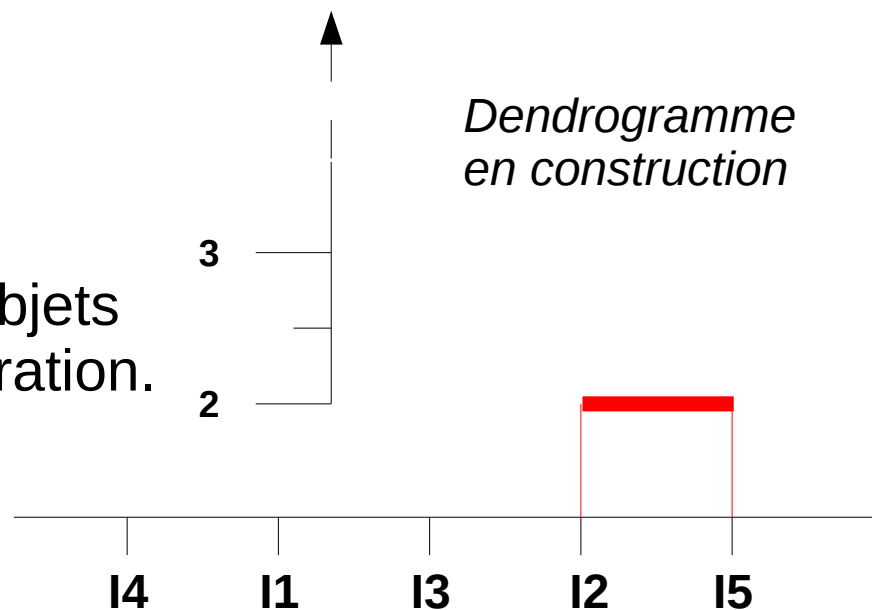
Étape 1

- Regroupement des individus I2 et I5 pour former le nœud N1 à la hauteur 2
- Avant de passer à l'étape suivante, il faut calculer les distances entre les nouveaux objets (I1, I3, I4 et **N1**) grâce au critère d'agglomération.

$$d(I1, N1) = \min\{d(I1, I2), d(I1, I5)\} = \min(3.61, 3) = 3$$

$$d(I3, N1) = \min\{d(I3, I2), d(I3, I5)\} = \min(2.24, 4.12) = 2.24$$

$$d(I4, N1) = \min\{d(I4, I2), d(I4, I5)\} = \min(8.12, 8.60) = 8.12$$



Un exemple « à la main » (2)

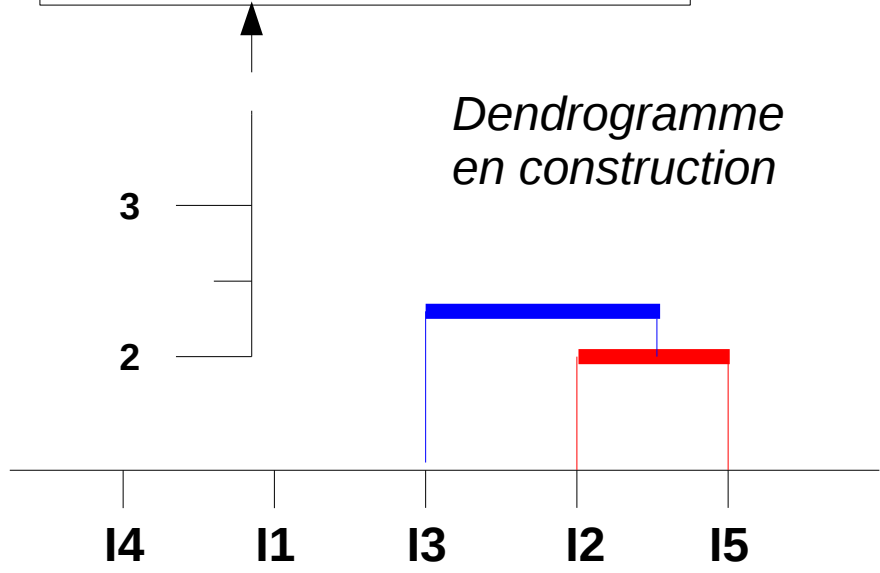
Étape 2

- Regroupement des objets I3 et N1 pour former le nœud N2 à la hauteur 2.24
- Comme précédemment, on calcule les distances entre I1, I4 et N2.

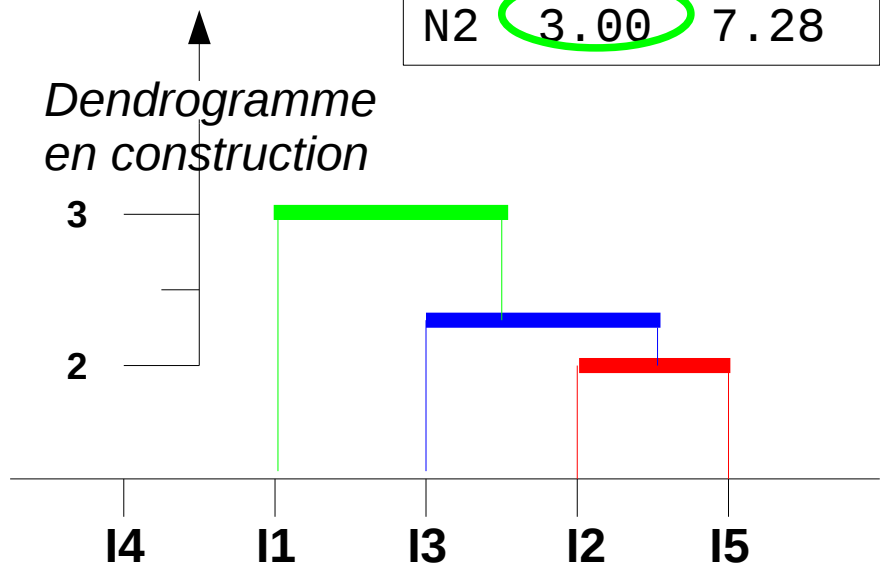
$$d(I1, N2) = \min\{d(I1, I3), d(I1, N1)\} = \min(5.10, 3) = 3$$

$$d(I4, N2) = \min\{d(I4, I3), d(I4, N1)\} = \min(7.28, 8.12) = 7.28$$

	I1	I3	I4
I3	5.10		
I4	10.34	7.28	
N1	3.00	2.24	8.12



	I1	I4
I4	10.34	
N2	3.00	7.28



Étape 3

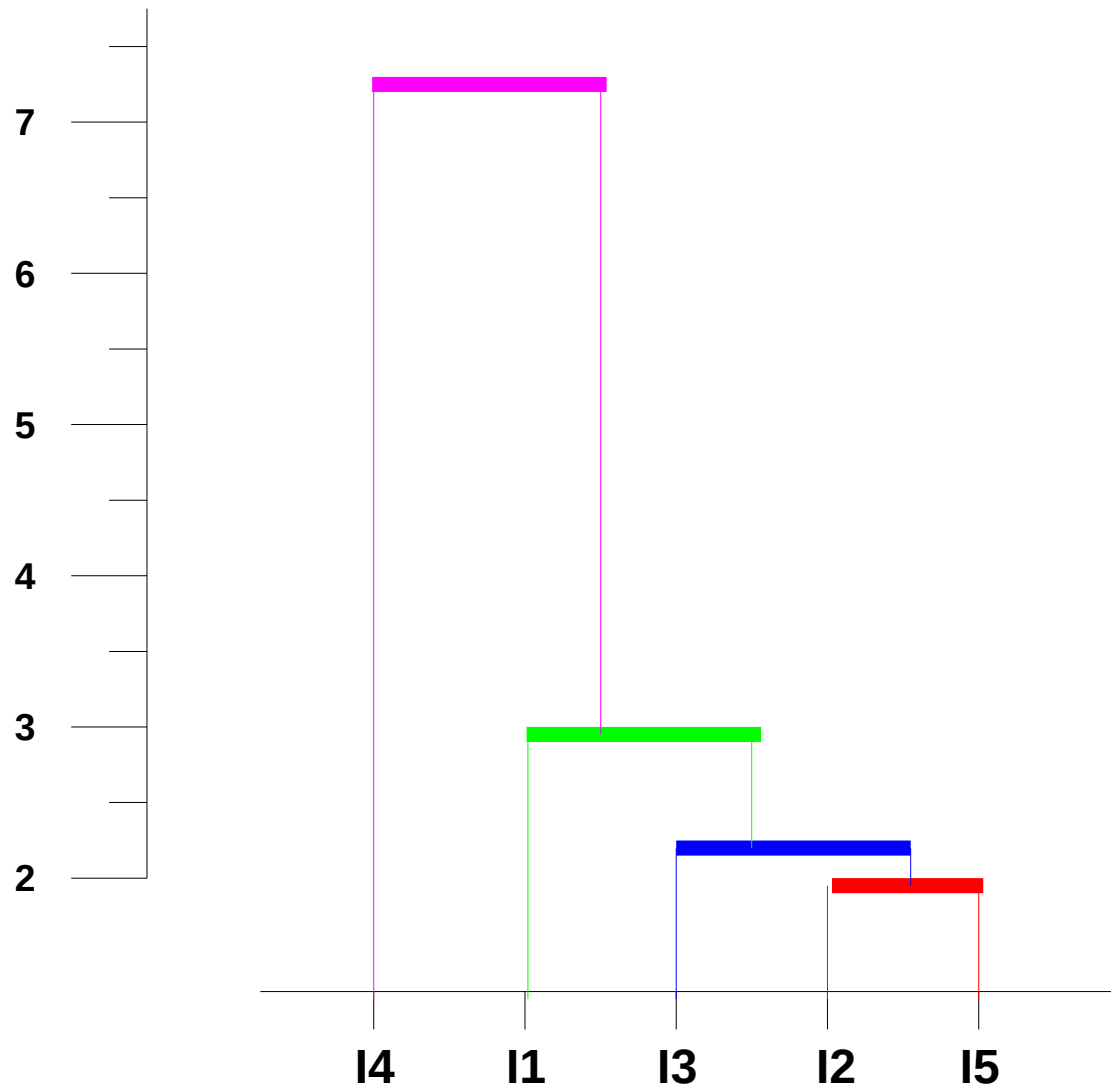
- Regroupement des objets I1 et N2 pour former le nœud N3 à la hauteur 3.
- Il reste à calculer la distance entre I4 et N3.

$$d(I4, N3) = \min\{d(I4, I1), d(I4, N2)\} = \min(10.34, 7.28) = 7.28$$

Un exemple « à la main » (3)

Étape 4 (fin de l'algorithme)

- Regroupement des objets I4 et N3 pour former le nœud terminal N4 à la hauteur 7.28

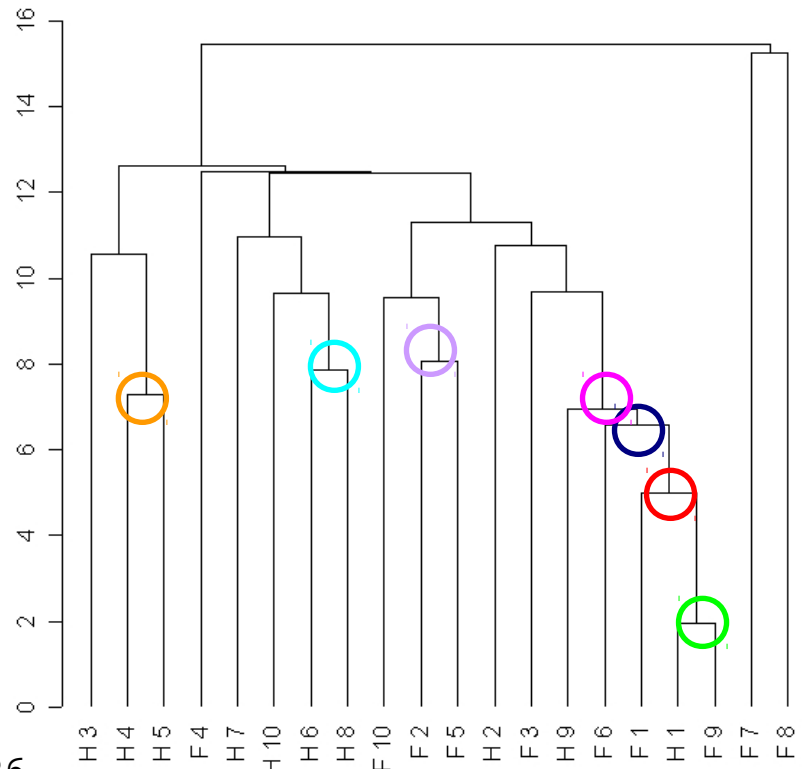


Dendrogramme « morpho »

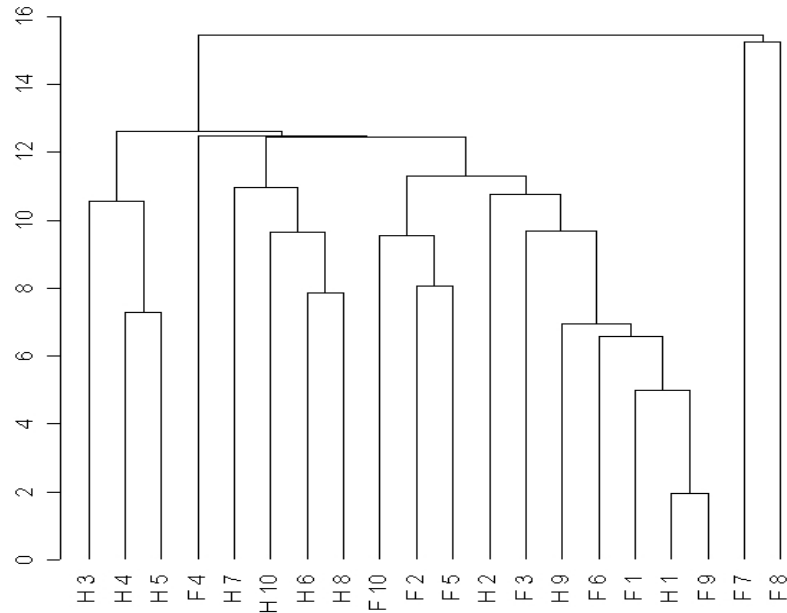
- Distance euclidienne
- Saut minimum



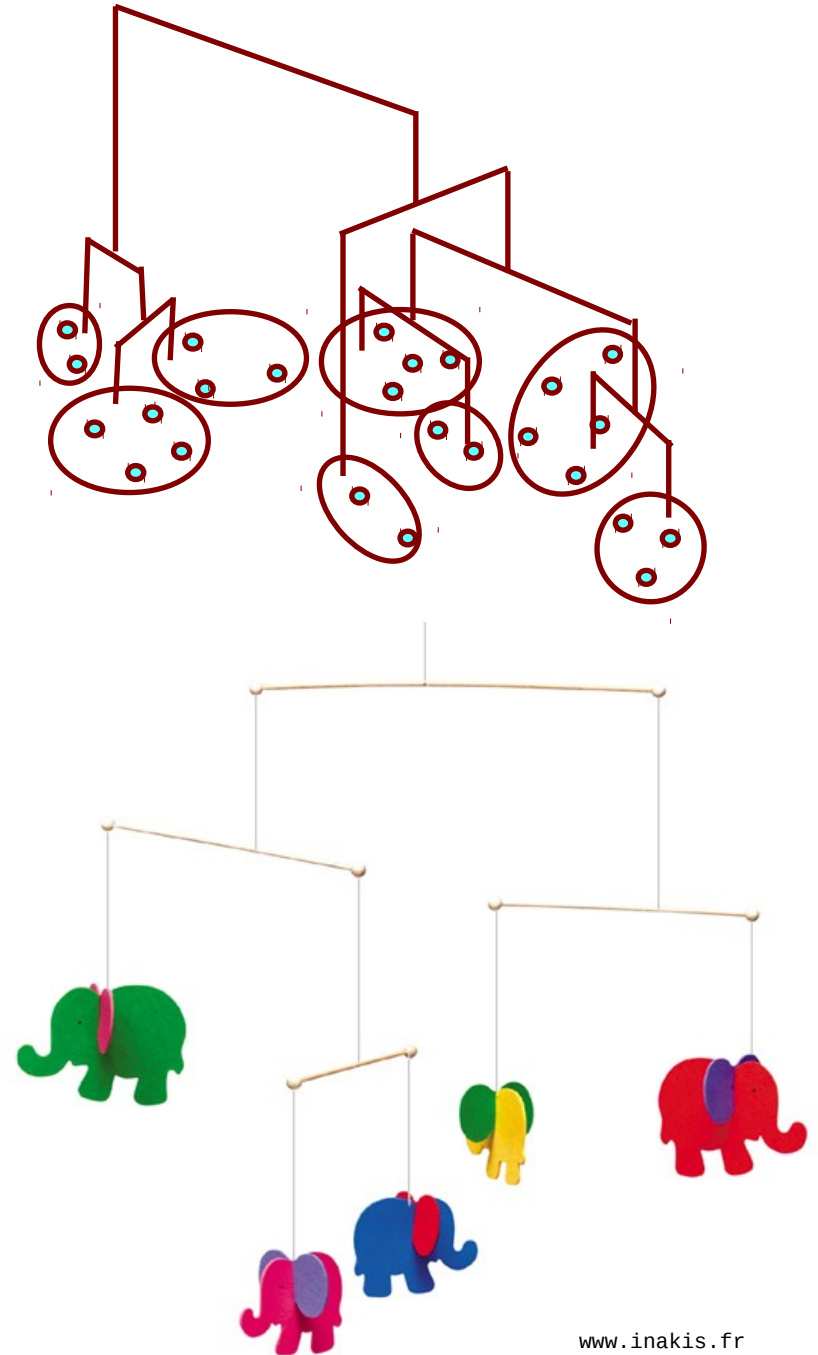
H 2	13																																							
H 3	30	21																																						
H 4	17	13	16																																					
H 5	22	14	11	7																																				
H 6	24	12	14	17	15																																			
H 7	35	24	17	26	20	14																																		
H 8	26	16	13	18	14	8	11																																	
H 9	7	16	31	20	25	24	36	25																																
H10	24	16	13	18	14	11	15	10	25																															
F 1	5	14	33	21	25	26	37	28	10	26																														
F 2	20	19	37	27	29	30	38	33	26	31	17																													
F 3	10	14	30	17	21	26	36	29	17	27	10	15																												
F 4	14	19	40	28	32	29	40	32	14	33	12	19	18																											
F 5	25	21	38	30	31	29	36	34	30	32	22	8	20	23																										
F 6	8	11	25	14	17	20	30	20	10	18	10	21	13	18	26																									
F 7	18	29	47	33	39	40	51	42	18	41	16	27	21	15	33	24																								
F 8	31	44	61	46	52	55	66	57	32	55	30	39	34	30	45	37	15																							
F 9	2	11	29	16	21	22	33	24	7	23	5	20	10	14	25	7	19	33																						
F10	17	16	36	24	27	27	37	31	21	32	14	10	11	14	13	20	23	36	16																					
	H1	H2	H3	H4	H5	H6	H7	H8	H9	H10	F1	F2	F3	F4	F5	F6	F7	F8	F9																					



Dendrogramme

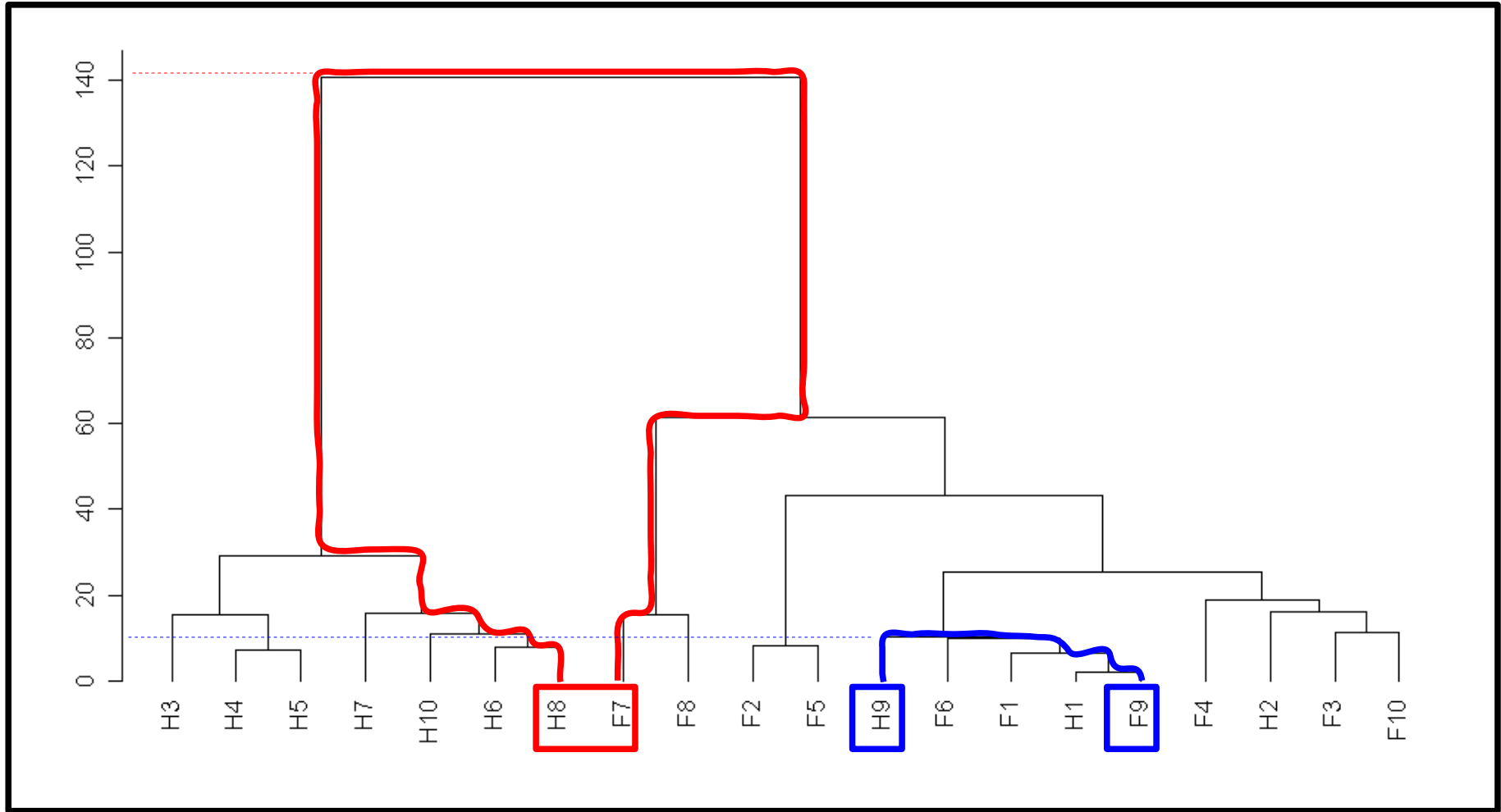


Un dendrogramme doit être vu comme un mobile pour bébé (le truc que l'on pend au dessus d'un berceau et dont les branches peuvent tourner). Ainsi, les « distances » entre les individus ne sont valables qu'en remontant les branches pas selon leur proximité « visuelle ».



Exemple « morpho »

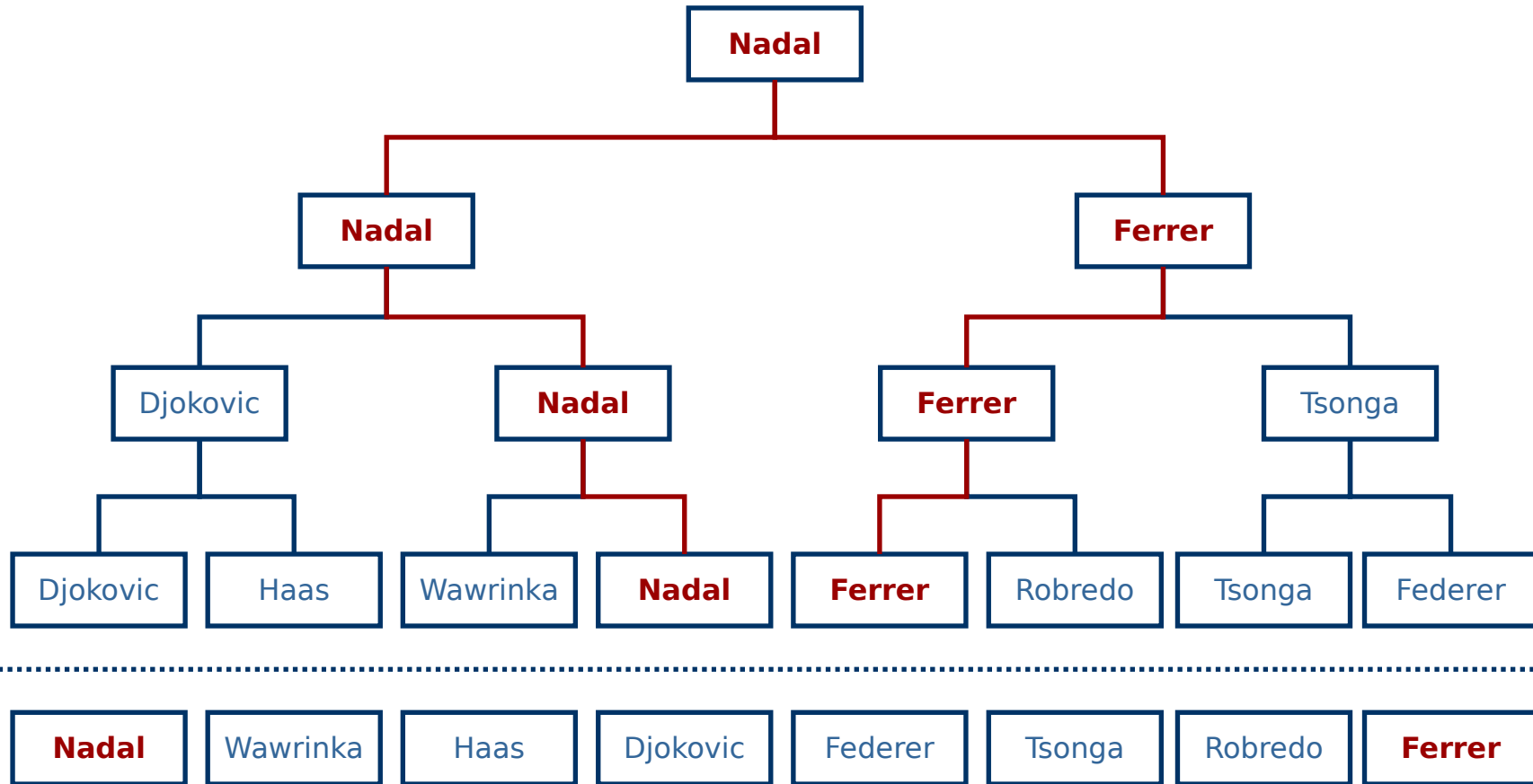
- Distance euclidienne
- Saut de Ward



Dendrogramme

Encore un coup pour les sportifs !

Tableau masculin du tournoi de Roland Garros 2013 (à partir des 1/4 de finales)

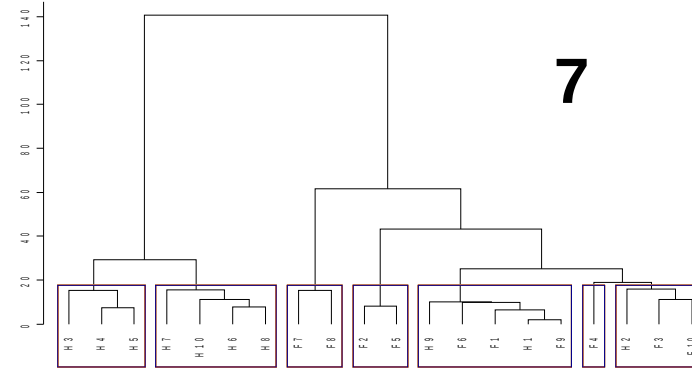
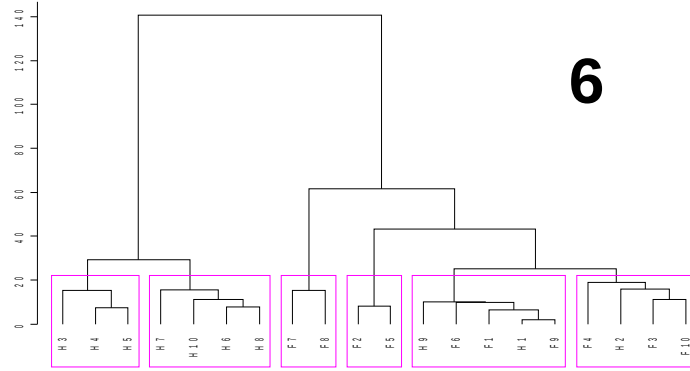
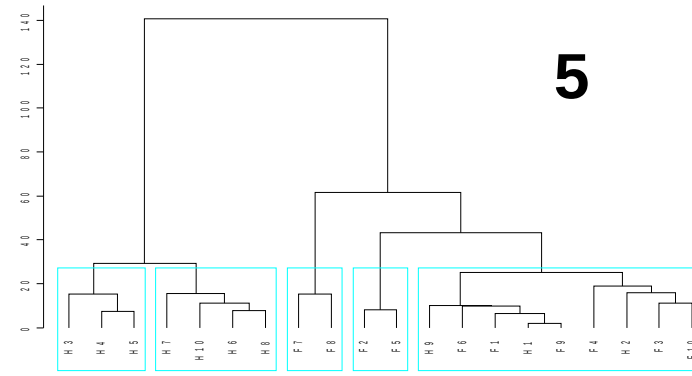
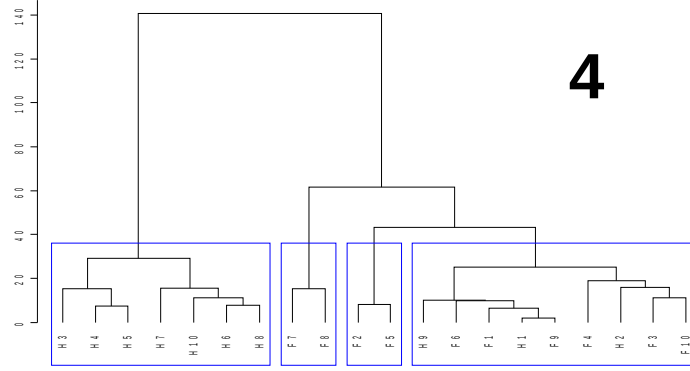
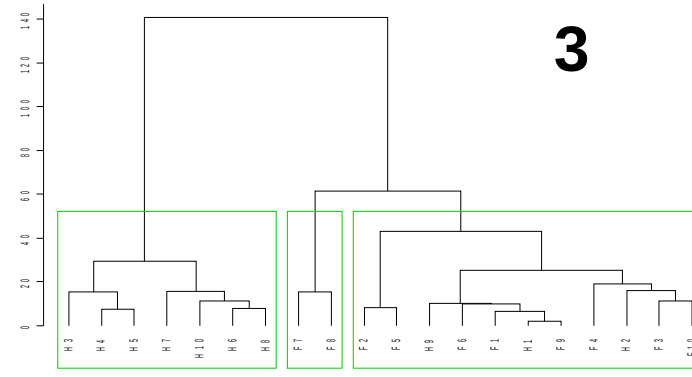
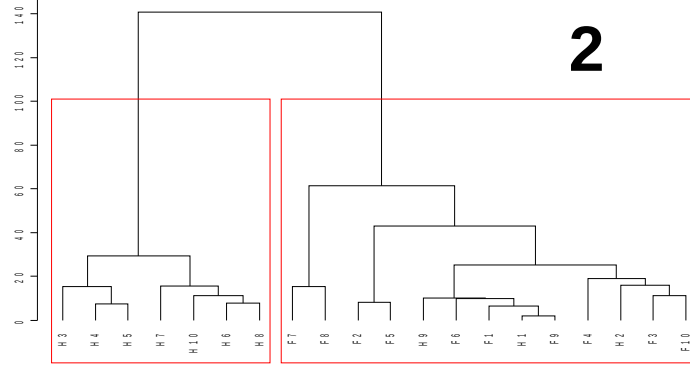


Les rencontres successives ne dépendent pas de l'ordre des joueurs sur le tableau initial. Ferrer et Nadal ne pouvaient se rencontrer qu'en finale quelle que soit leur position « visuelle » dans le tableau initial.

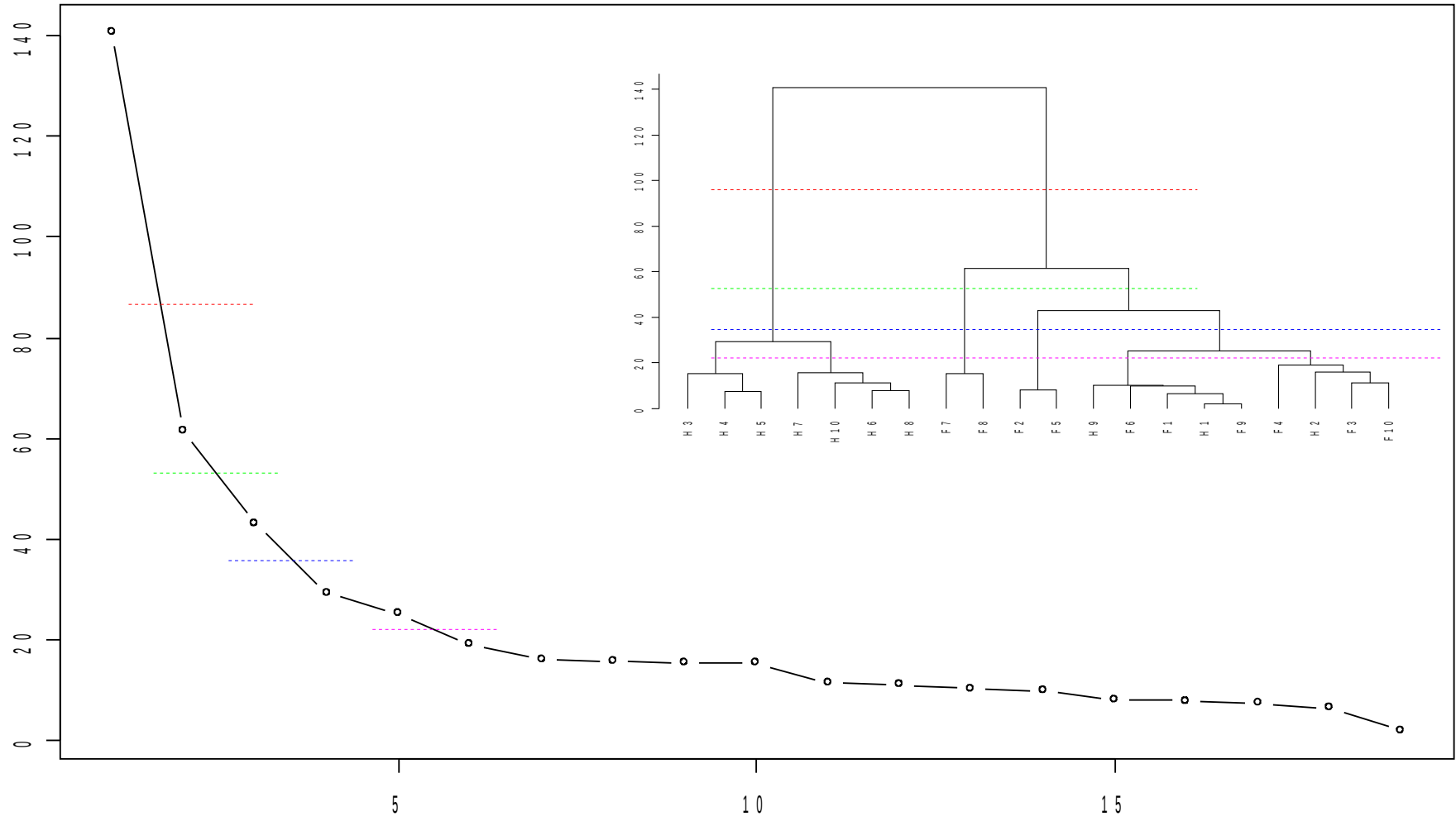
Interprétation d'un dendrogramme

- Un dendrogramme fournit une classification des éléments lorsque l'on se donne une « hauteur de coupe » de l'arbre.
- Plus l'arbre est coupé « bas » (proche des éléments initiaux) plus la classification obtenue est fine.
- Une hauteur de coupe est pertinente si elle se trouve entre 2 noeuds dont les hauteurs sont « relativement » éloignées.

Nombre de groupes



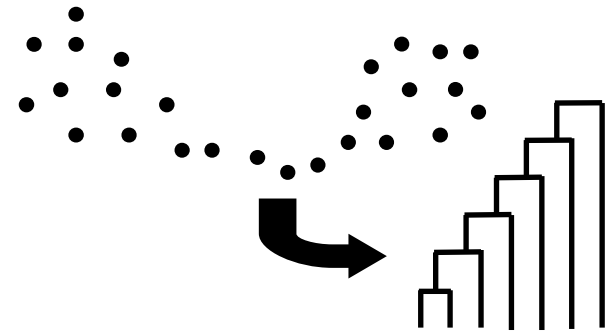
Aide : représentation de la hauteur des noeuds



Remarques

- **aucune connaissance a priori** sur le phénomène étudié n'est requise
- réflexion préalable sur le **choix de la distance** et du **critère d'agglomération**. Il est même vivement conseillé de procéder à plusieurs classifications et de s'intéresser aux choses qui changent

- critère de saut minimum \Rightarrow effet de chaîne
(dendrogramme en forme d'escalier)

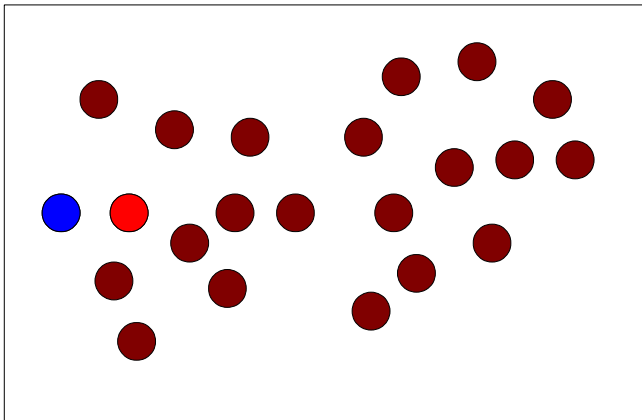


- le résultat est « **figé** ». Une fois effectué, un regroupement ne peut plus être défait (Quand on dit oui, c'est pour la vie !)
- Attention à l'interprétation d'un dendrogramme (**proximité « visuelle » !!!**)

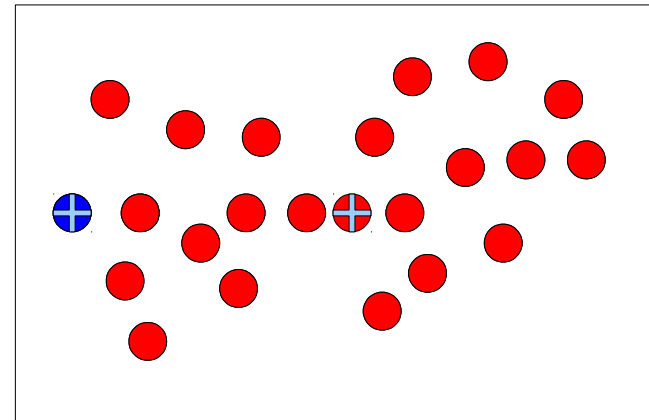
Agrégation autour de centres mobiles

- **Préalable** : déterminer le nombre de groupes (k) soit par une connaissance a priori du phénomène étudié, soit par une autre méthode (classification hiérarchique par exemple)
- **Procédure itérative**
 - **Début** : k centres (tirés aléatoirement ou imposés)
 - **Déroulement** :
 - tous les individus sont affectés au centre le plus proche
 - les centres de chaque groupe sont recalculés
 - **Fin** : les individus ne changent pas de groupe entre 2 étapes successives
- **Résultat**
 - répartition des individus en k groupes

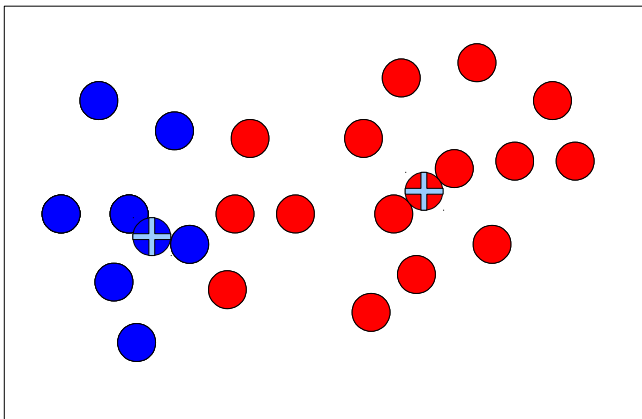
Centres mobiles : principe (k=2)



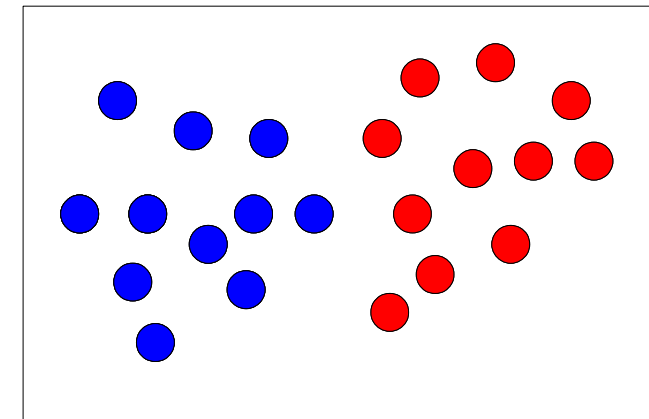
Tirage aléatoire de 2 centres initiaux



Affectation de chaque point au centre le plus proche et calcul des nouveaux centres



Affectation et calcul des nouveaux centres

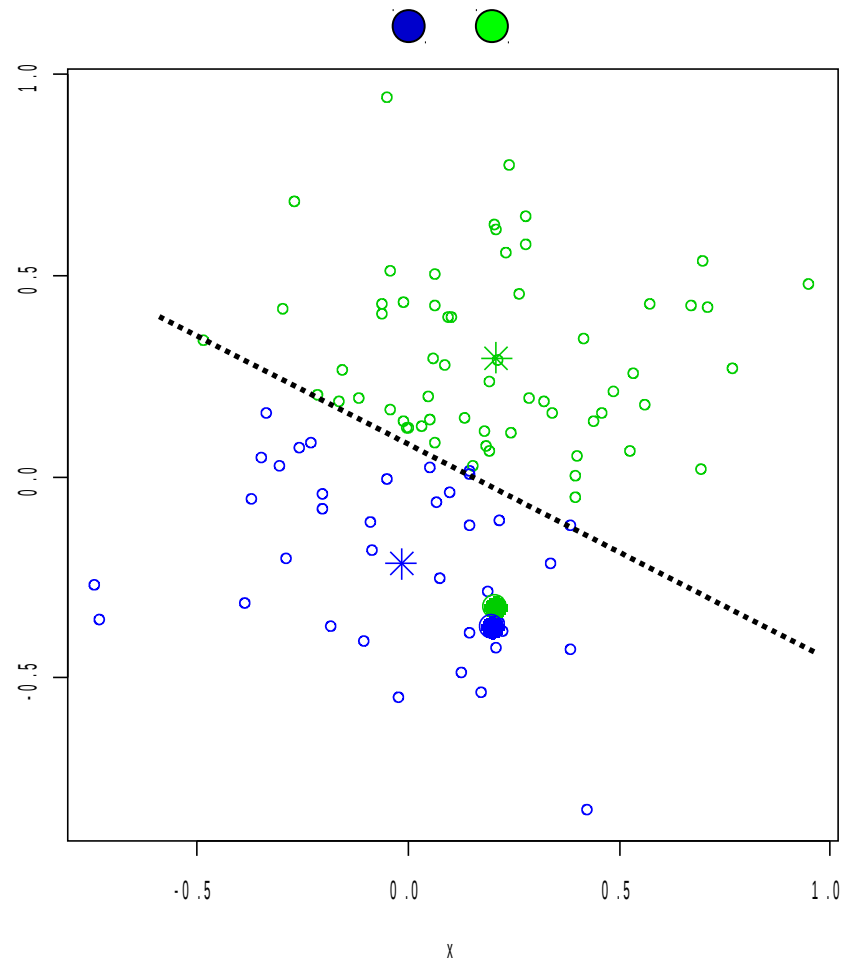
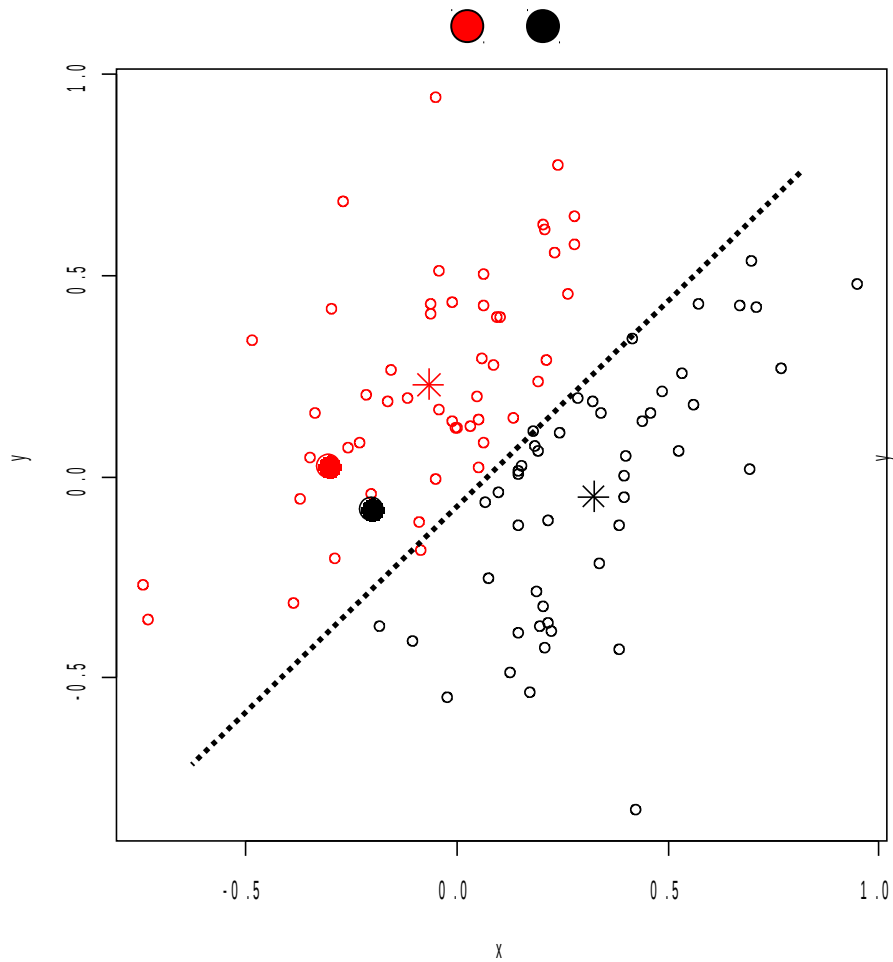


Dernière affectation

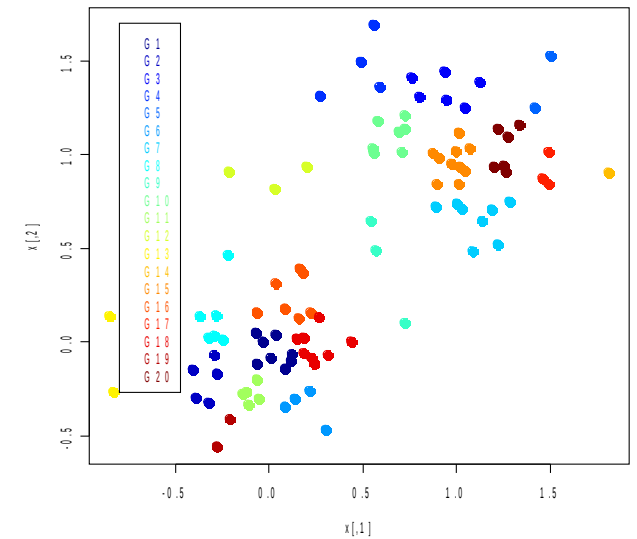
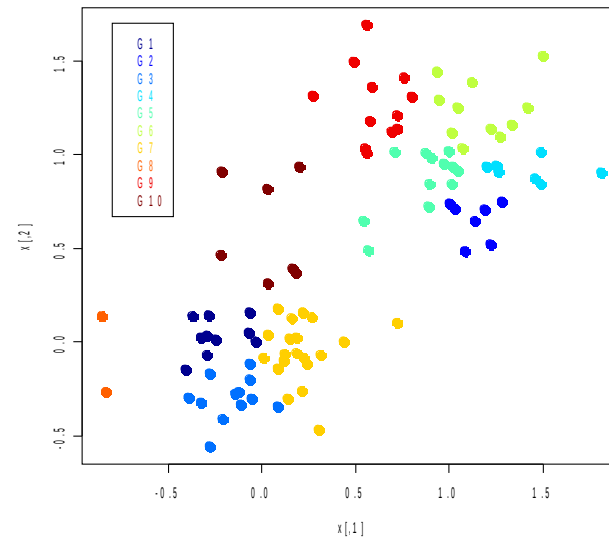
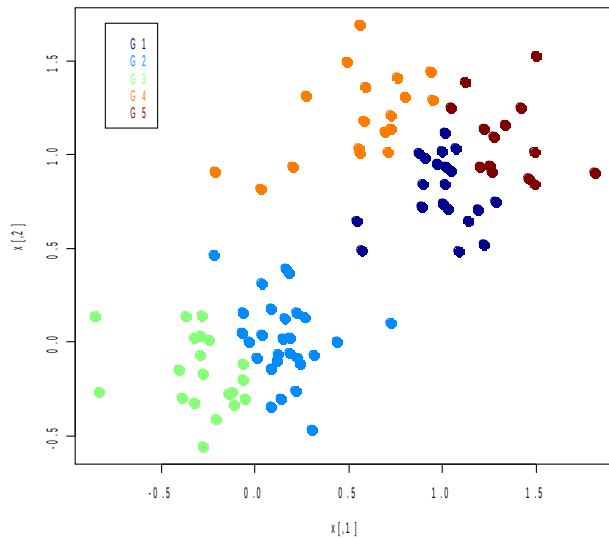
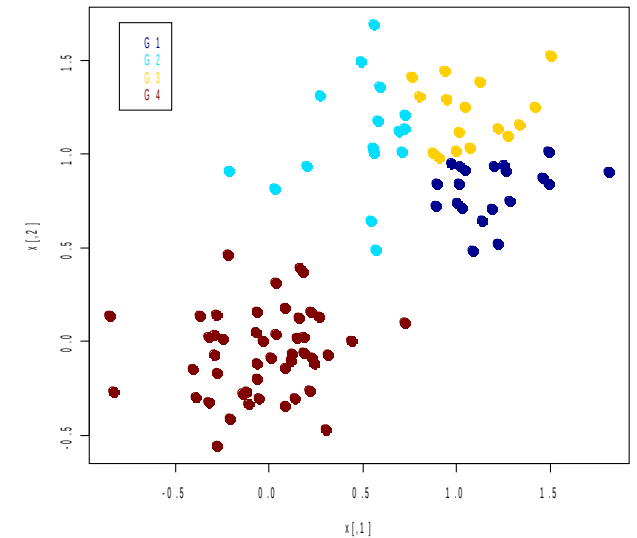
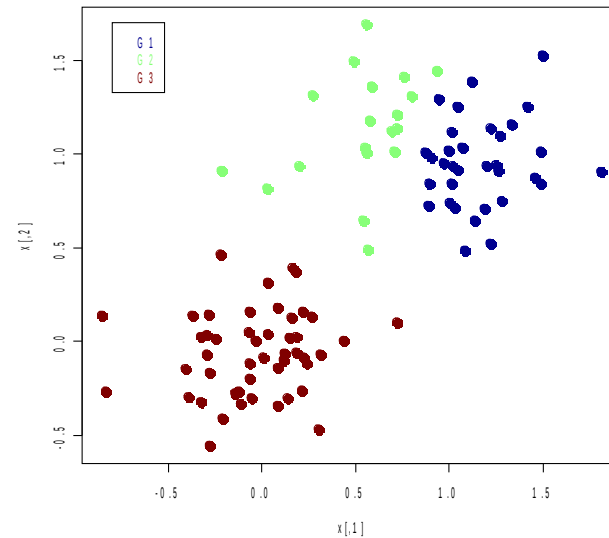
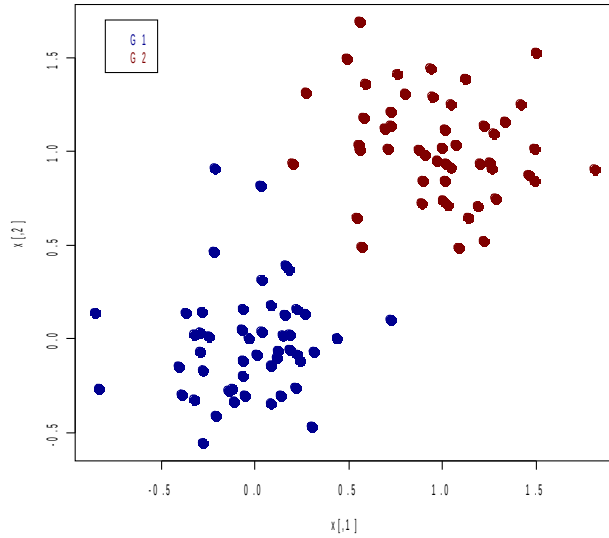
La méthode des « nuées dynamiques » est une variante de celle-ci. Elle consiste à déplacer le centre de gravité de la classe après chaque affectation d'un nouveau point.

Initialisation de l'algorithme

Influence du **tirage aléatoire** des centres initiaux



Détermination du nombre de groupes



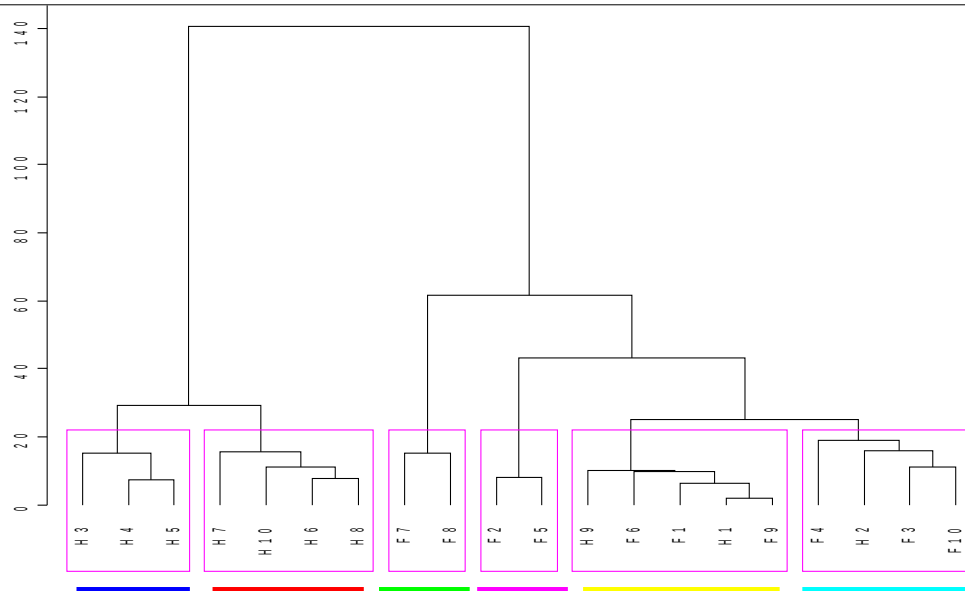
Exemple « morpho »

Résultats comparatifs de 6 exécutions d'un algorithme d'agrégation autour de centres mobiles ($k=6$) : 1-5) initialisation aléatoire, *) centres initiaux issus de la classification hiérarchique.

	H1	H2	H3	H4	H5	H6	H7	H8	H9	H10	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10
1)	1	4	5	4	4	6	6	6	1	5	1	3	1	1	3	1	2	2	1	3
2)	6	3	2	3	2	5	5	5	6	5	6	4	6	6	4	6	1	1	6	4
3)	2	1	4	1	1	4	3	4	2	4	2	5	2	2	5	2	6	6	2	5
4)	4	2	1	2	2	1	1	1	4	1	4	5	4	3	5	4	6	6	4	5
5)	2	4	6	4	6	3	3	3	2	3	2	1	2	2	1	2	5	5	2	1
*)	1	1	3	3	3	4	4	4	1	4	1	5	1	2	5	1	6	6	1	5

Rappel des 6 groupes issus de la classification hiérarchique :

- **H3 - H4 - H5**
- **H7 - H10 - H6 - H8**
- **F7 - F8**
- **F2 - F5**
- **H9 - F6 - F1 - H1 - F9**
- **F4 - H2 - F3 - F10**

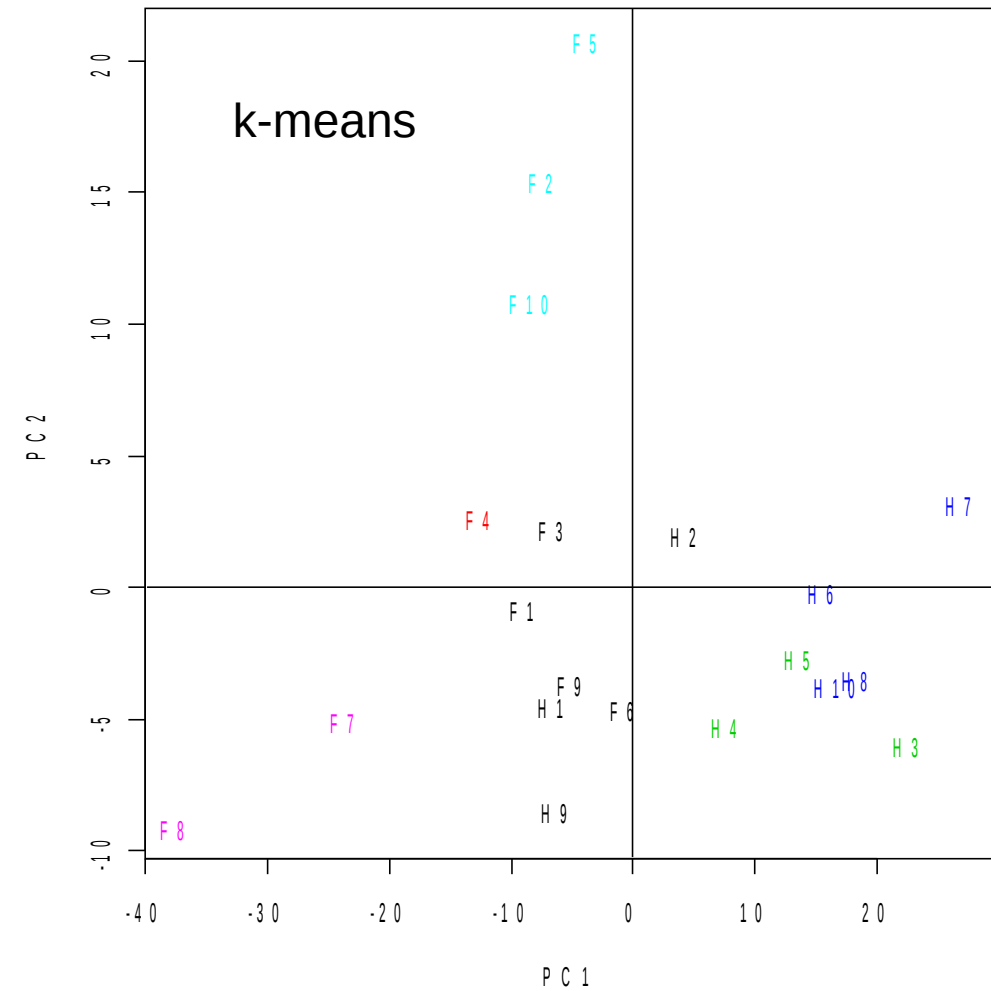
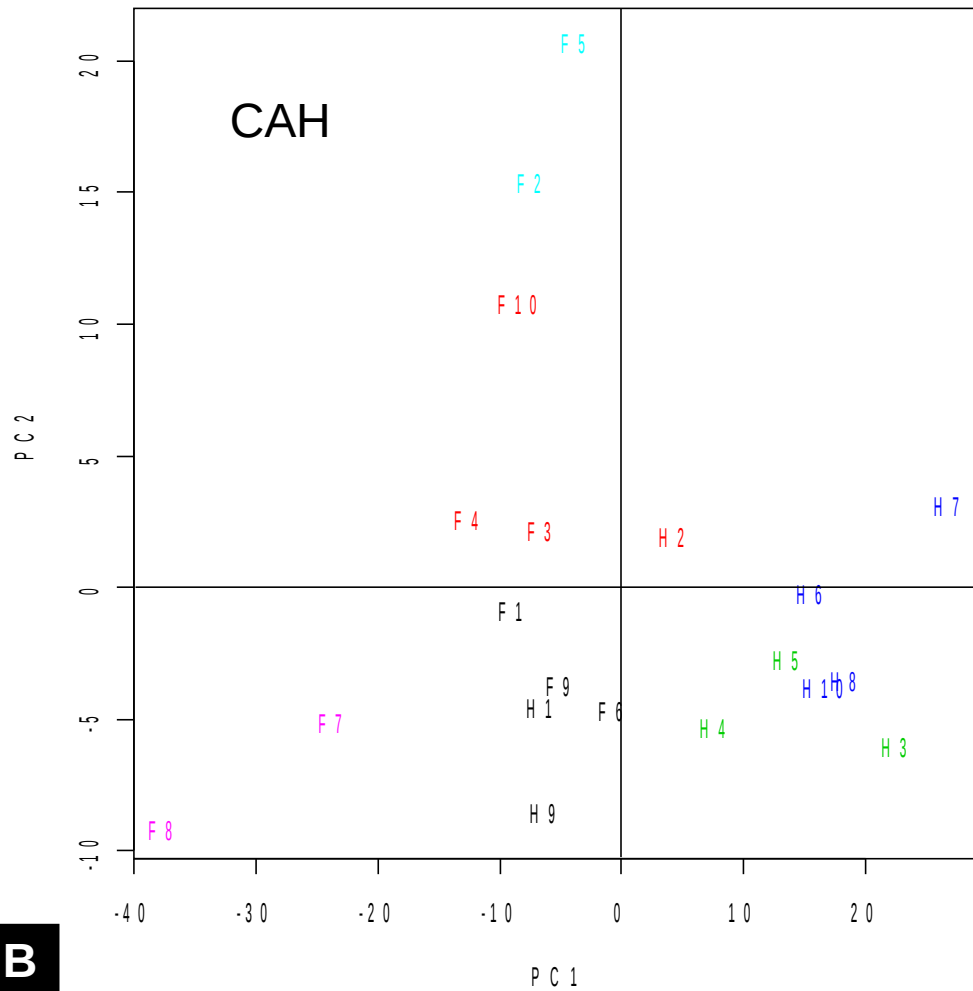


Exemple « morpho » (suite)

Bilan des changements de groupe après exécution de l'algorithme k-means initialisé aux centres des classes issus de la classification hiérarchique.

	k-means					
CAH	1	2	3	4	5	6
1	5	0	0	0	0	0
2	2	1	0	0	1	0
3	0	0	3	0	0	0
4	0	0	0	4	0	0
5	0	0	0	0	2	0
6	0	0	0	0	0	2

Vue des changements sur le premier plan principal ; une couleur caractérise un groupe.



Remarques

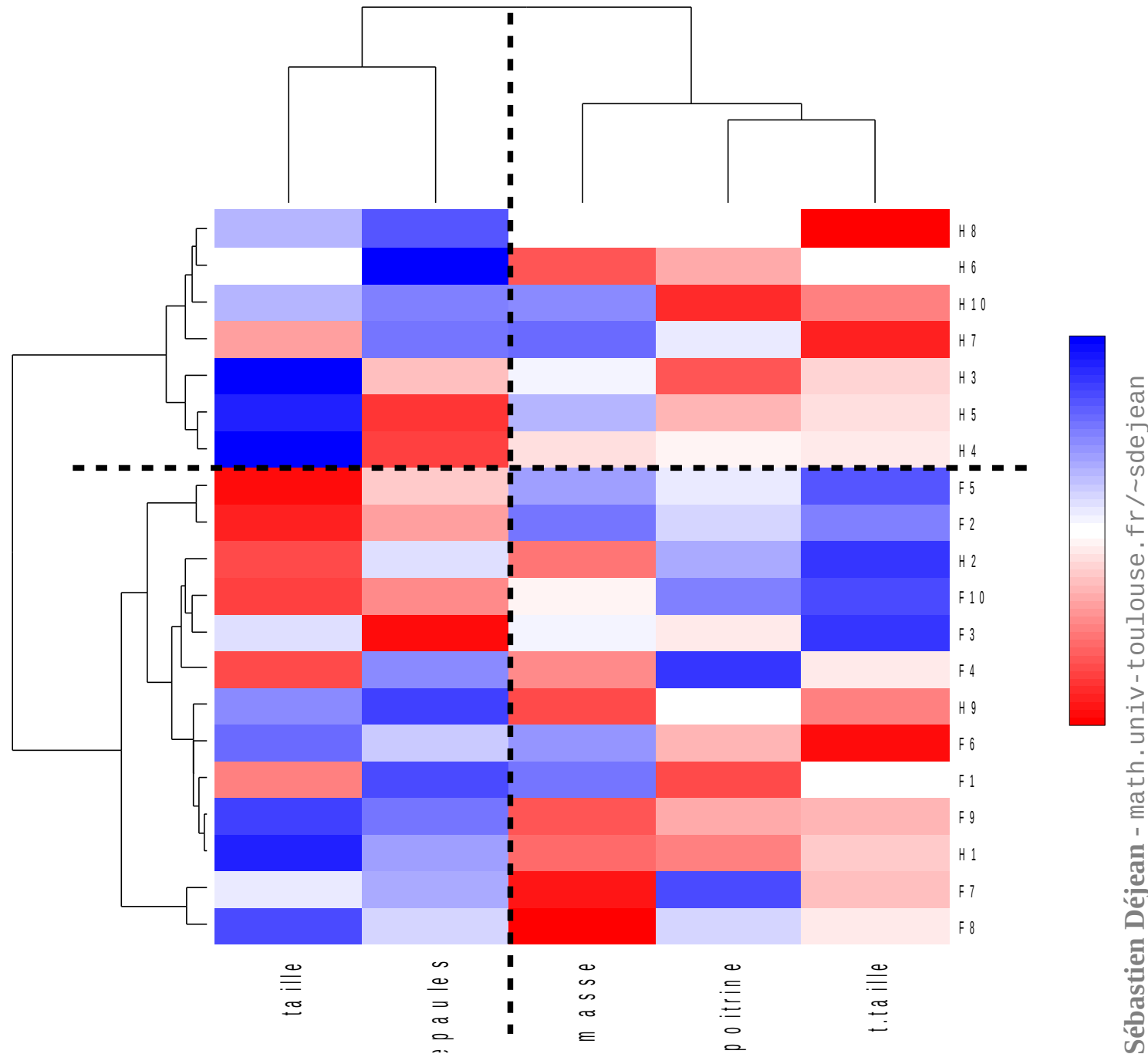
- nécessité de déterminer préalablement le nombre de groupes
- si les centres initiaux sont tirés aléatoirement, des exécutions successives de l'algorithme peuvent donner des résultats différents. Face à cela, on peut :
 - imposer les centres initiaux (par exemple, les barycentres des classes issues d'une classification hiérarchique)
 - s'intéresser aux formes fortes : quels que soient les centres, certains points se retrouvent toujours dans le même groupe
- un algorithme de type k-means exécuté après une classification hiérarchique permet de « stabiliser » la classification obtenue : l'algorithme k-means va permettre à des points de changer de groupe au gré de l'évolution de l'algorithme (ce que ne permet pas la classification hiérarchique, *Quand on dit oui...*)

« Double » classification

- Classification des individus et des variables de l'étude
- Représentation *heatmap* : les dendrogrammes sont représentés autour d'une image de la matrice des données
- L'interprétation du graphique obtenu vise à mettre en relation la classification des individus et des variables

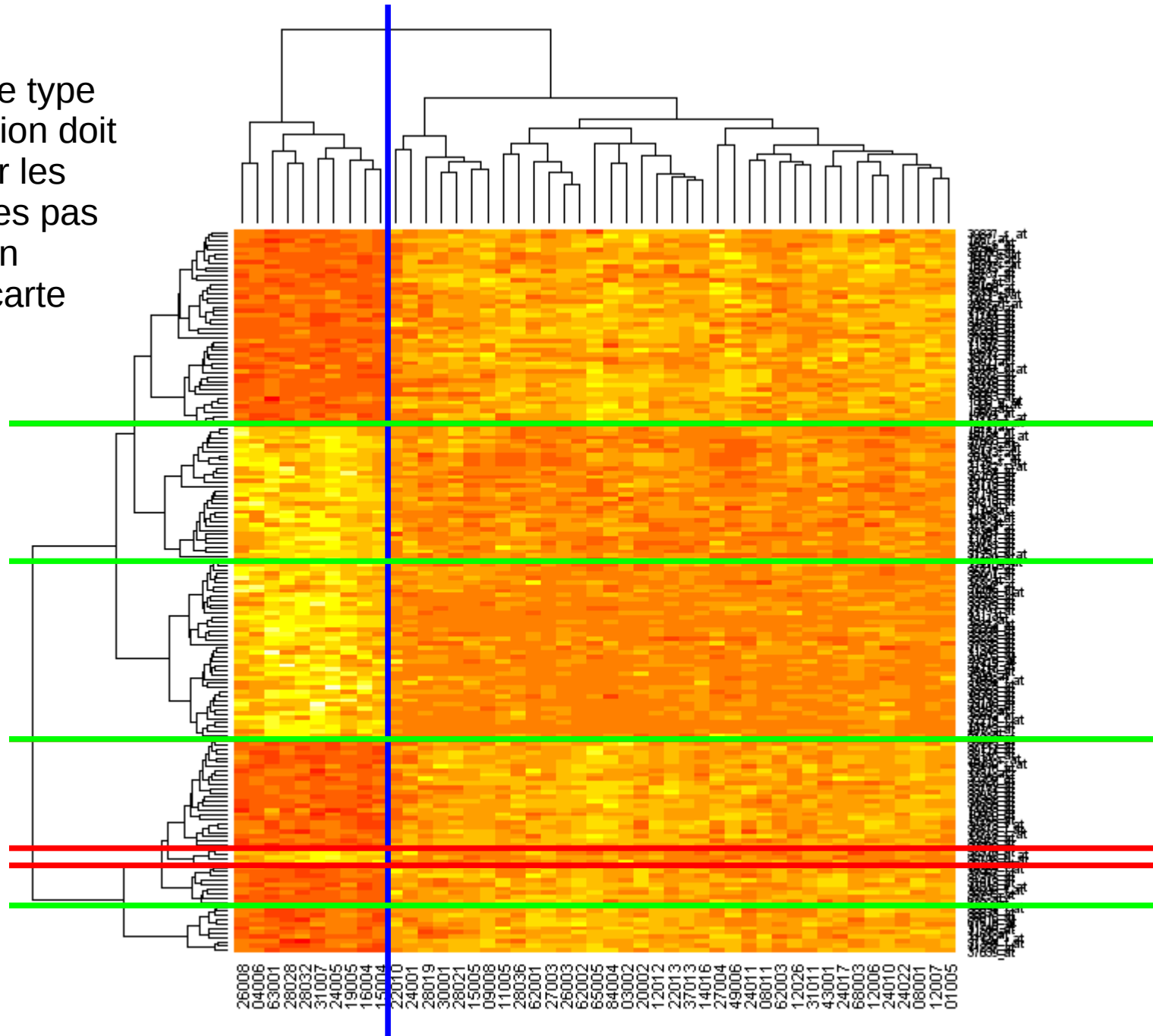
Heatmap « morpho »

- Nuancier de couleur : du rouge (valeurs faibles) vers le bleu (valeurs élevées) en passant par le blanc.
- Les données ont été centrées en colonnes pour éviter de distinguer la variable « taille » dont les valeurs sont d'un autre ordre de grandeur que les 4 autres.
- Chaque groupe d'individus peut être caractérisé par les variables morphologiques

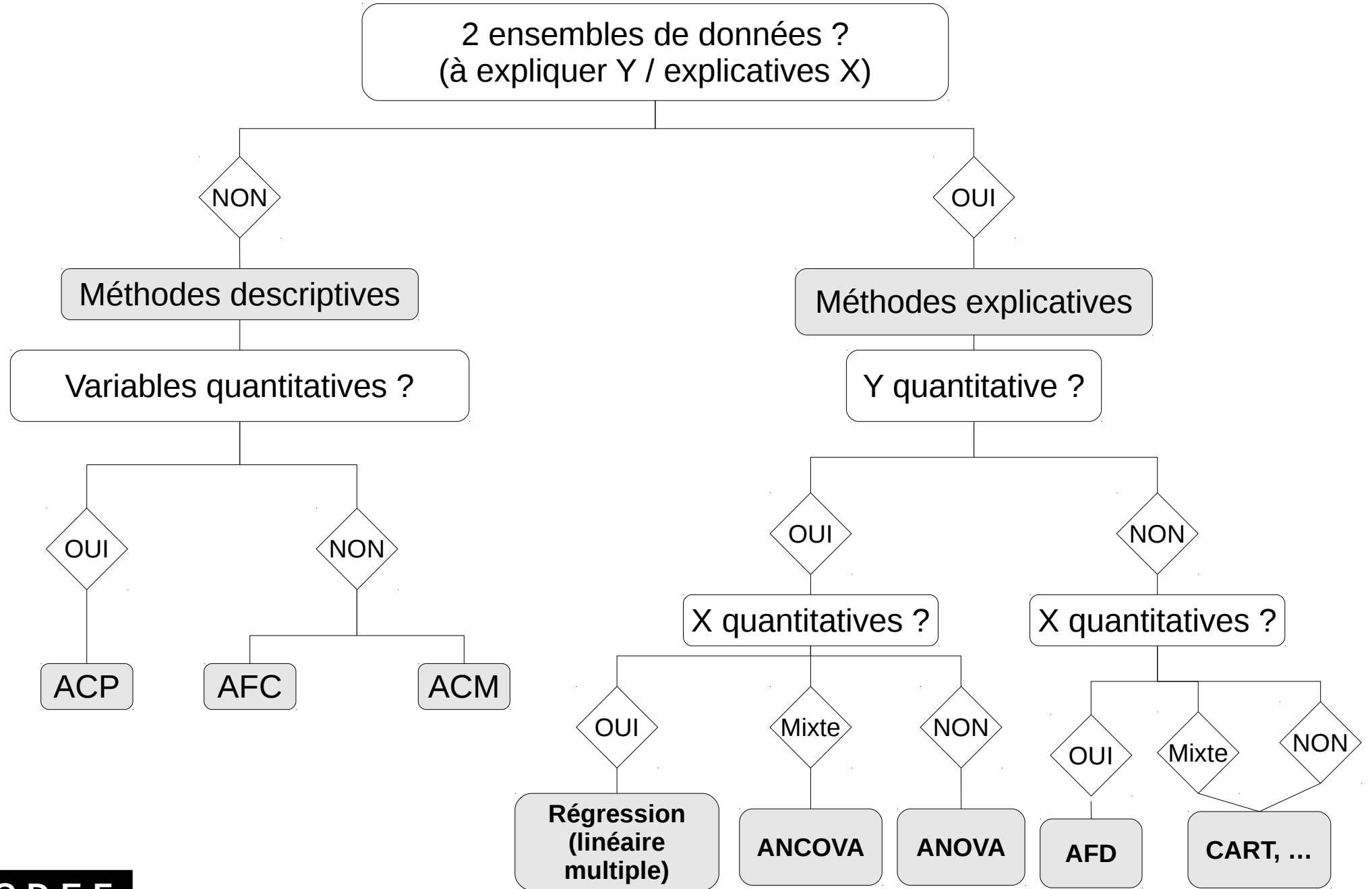


Interprétation d'une *heatmap*

L'analyse de ce type de représentation doit être guidée par les dendrogrammes pas par l'impression visuelle de la carte centrale.

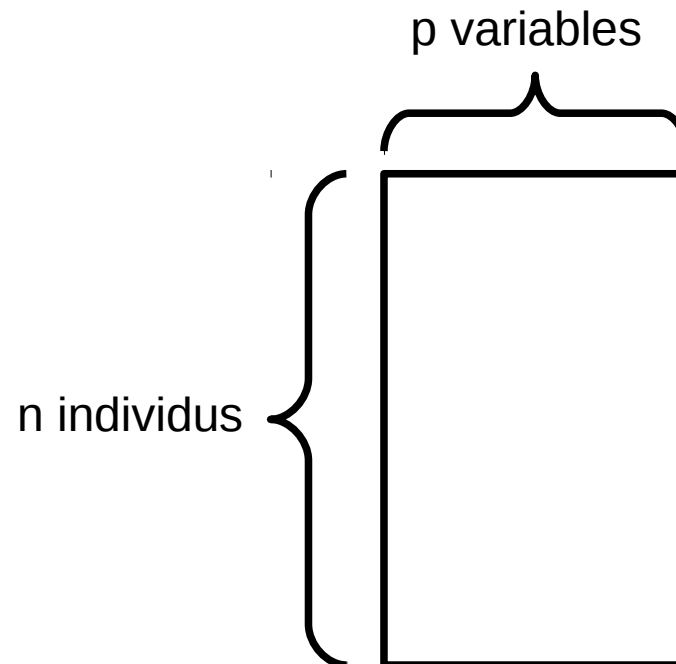


Panorama de quelques méthodes multidimensionnelles

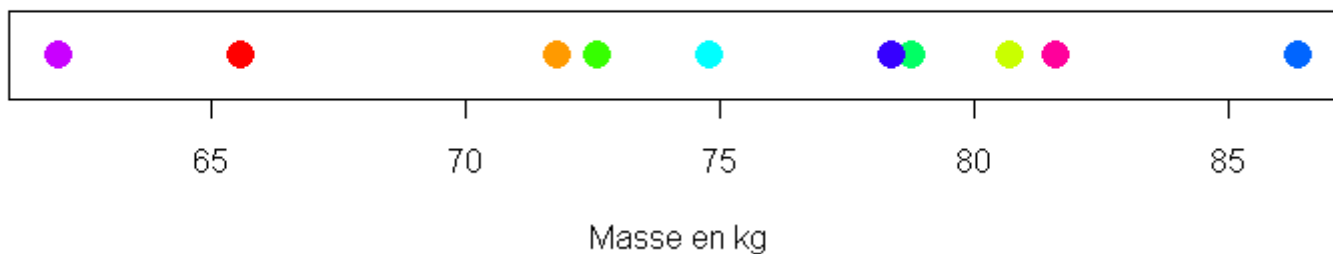
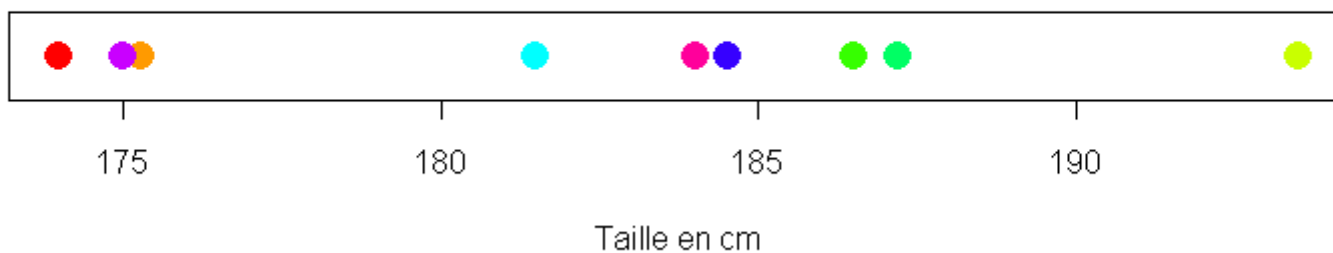


Analyse en Composantes Principales

Objectif : décrire sans a priori un tableau de données constitué exclusivement de variables **quantitatives**.

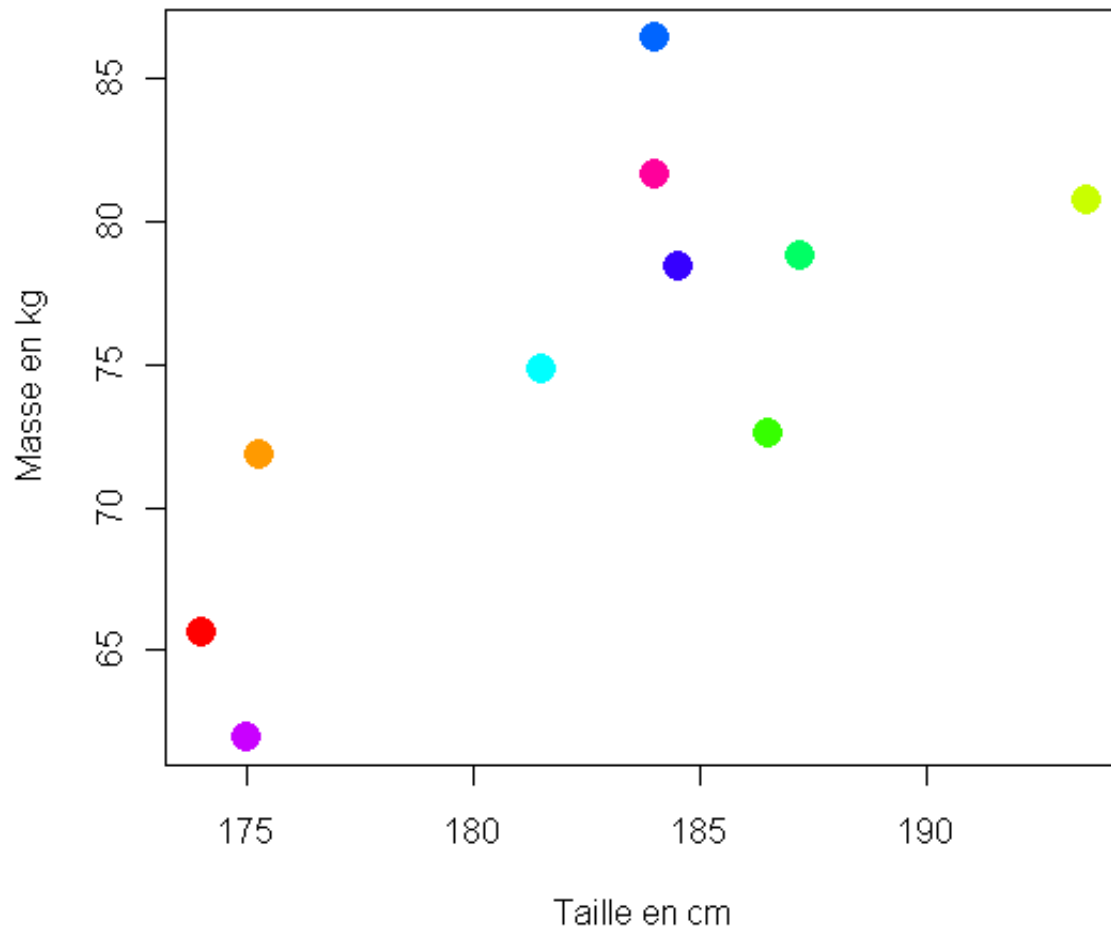


Représentation graphique 1D



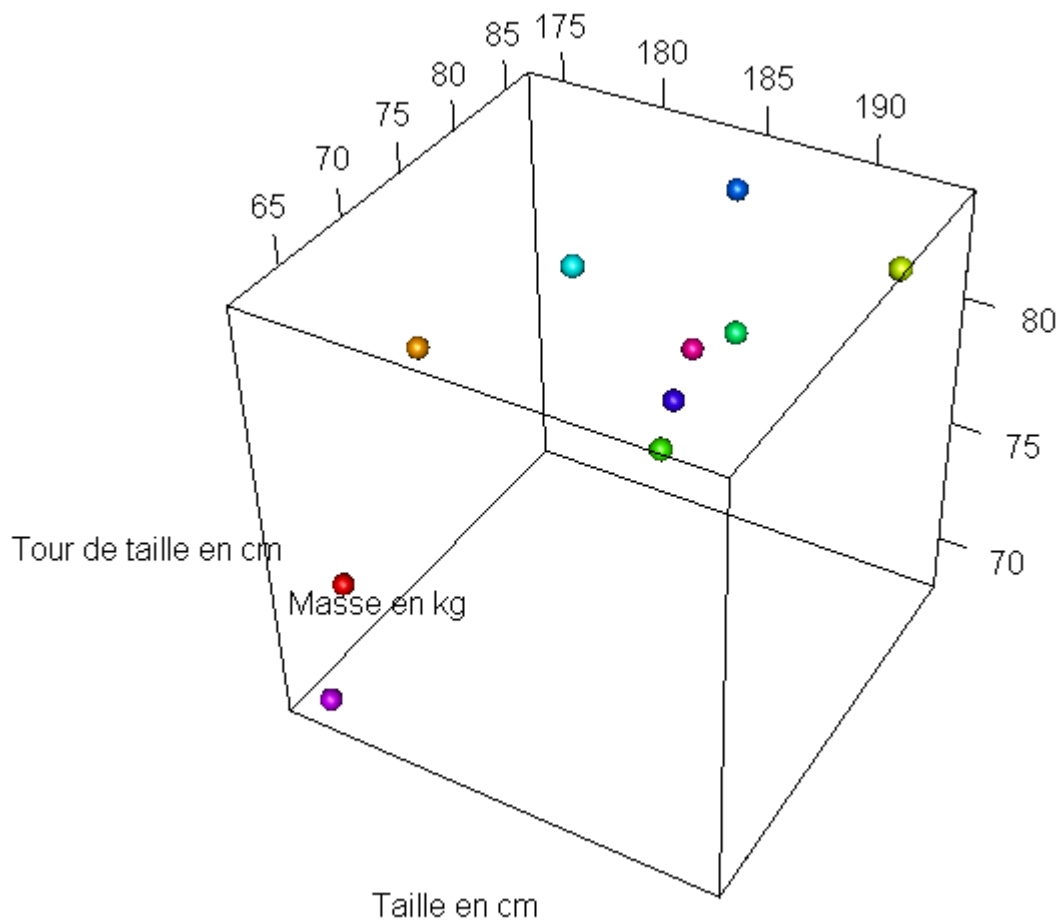
Représentation graphique 2D

Taille :	174.0	175.3	193.5	186.5	187.2	181.5	184.0	184.5	175.0	184.0
Masse :	65.6	71.8	80.7	72.6	78.8	74.8	86.4	78.4	62.0	81.6

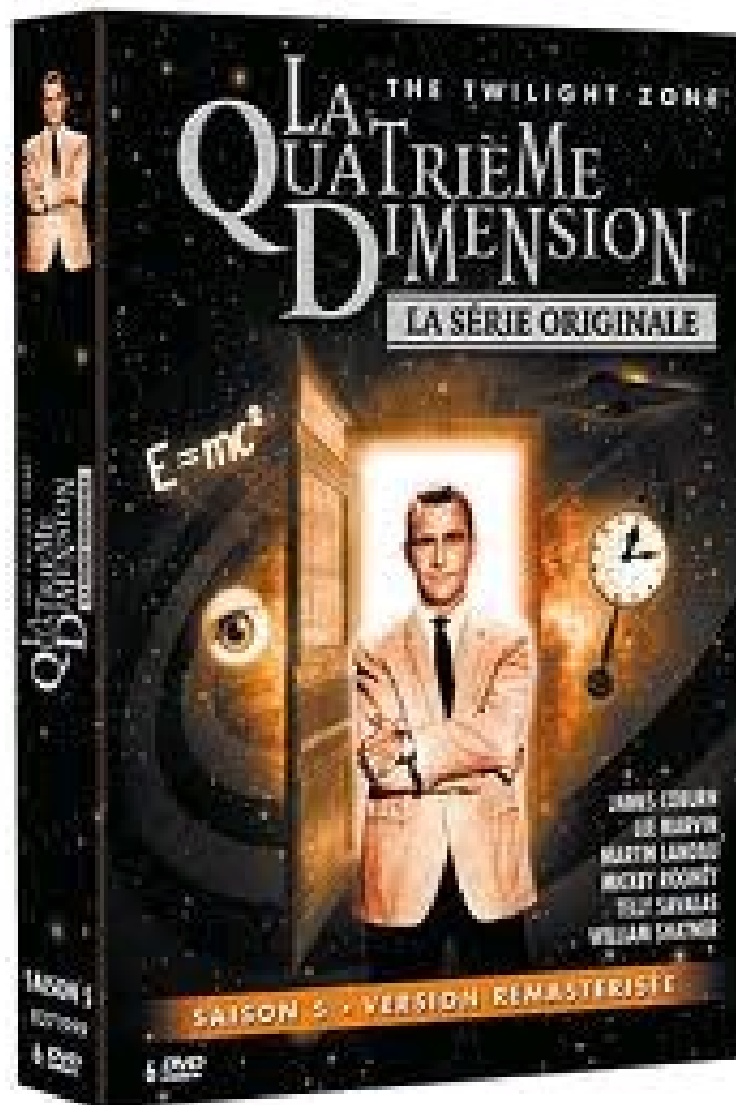


Représentation graphique 3D

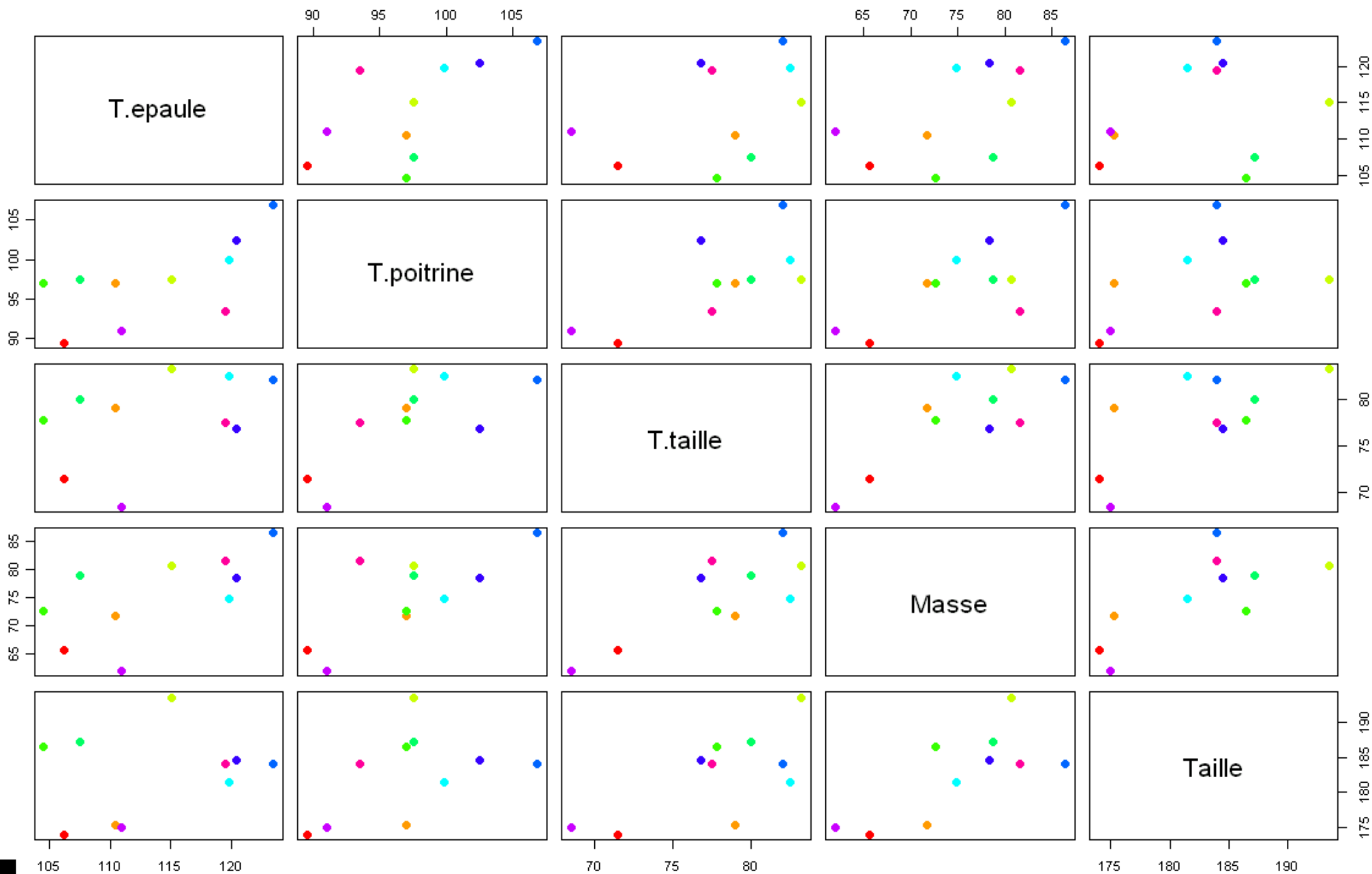
	●	●	●	●	●	●	●	●	●	
Taille :	174.0	175.3	193.5	186.5	187.2	181.5	184.0	184.5	175.0	184.0
Masse :	65.6	71.8	80.7	72.6	78.8	74.8	86.4	78.4	62.0	81.6
T. Taille :	71.5	79.0	83.2	77.8	80.0	82.5	82.0	76.8	68.5	77.5



Représentation graphique 4D

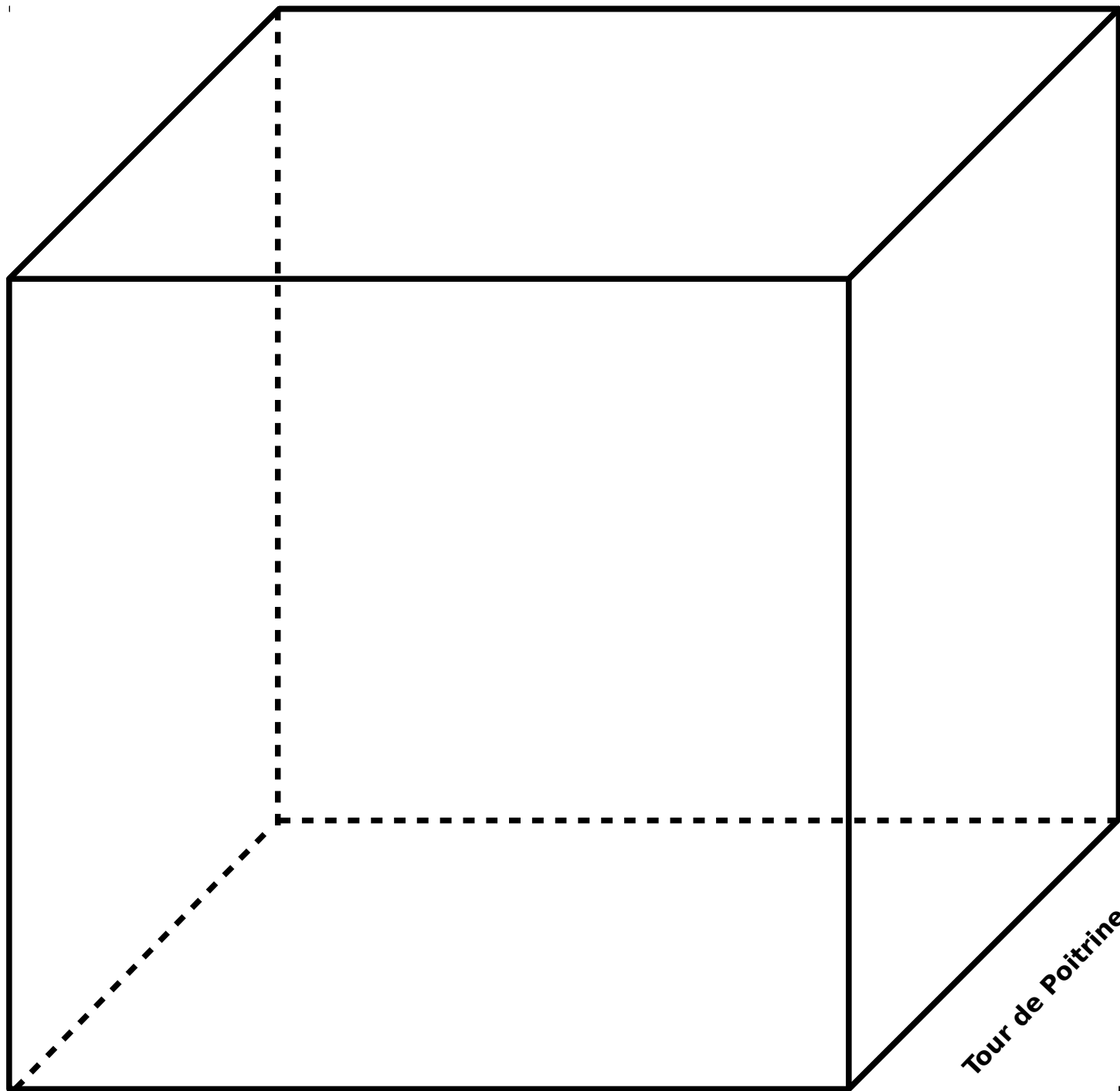


Alternative à la 4D (ou plus)



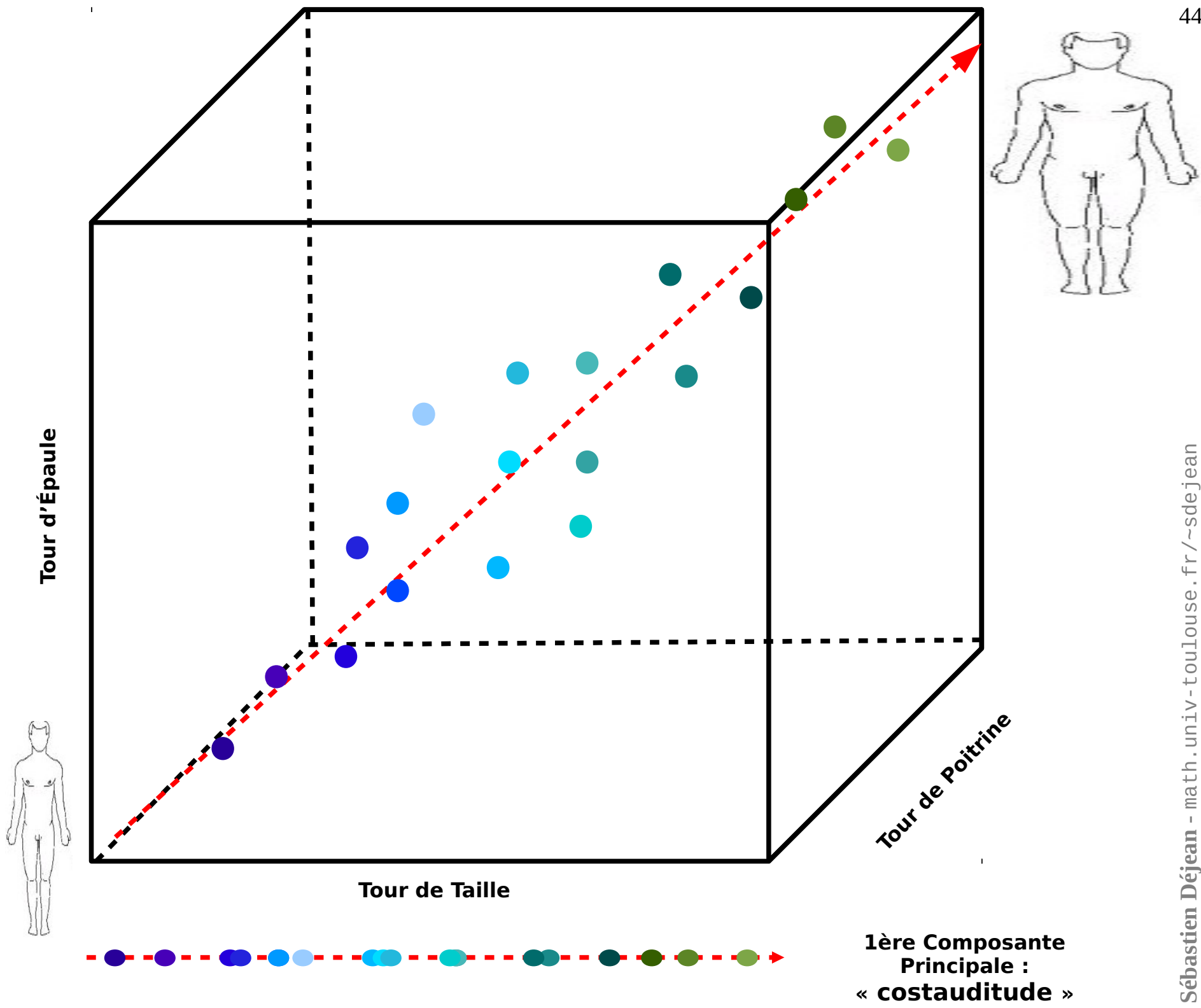


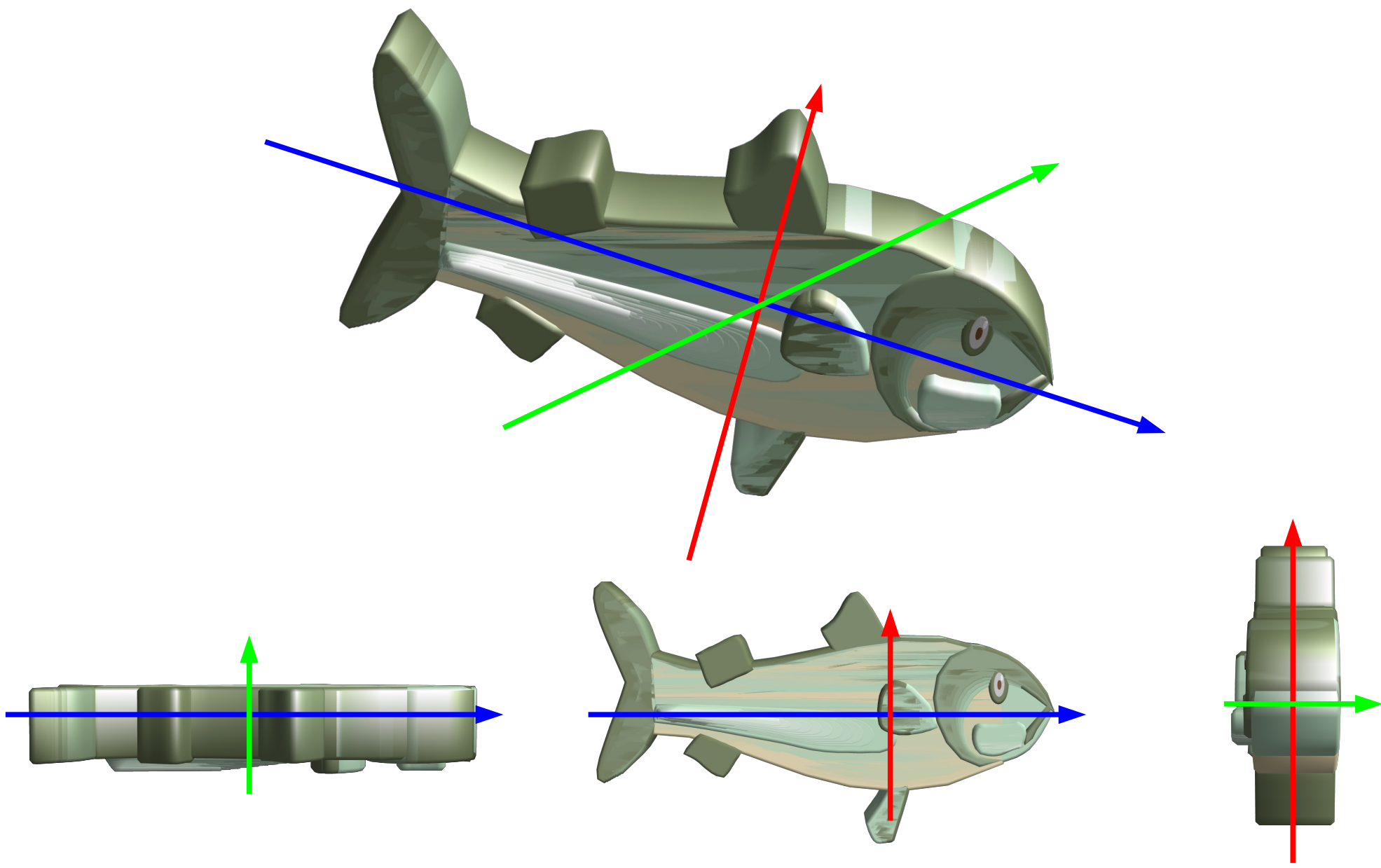
Tour d'Épaule



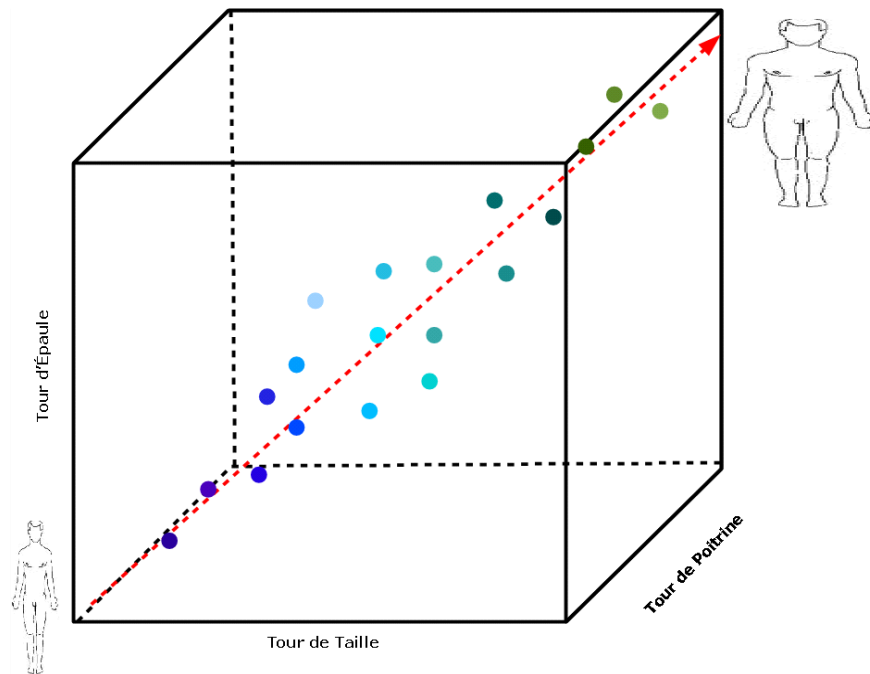
Tour de Taille

Tour de Poitrine

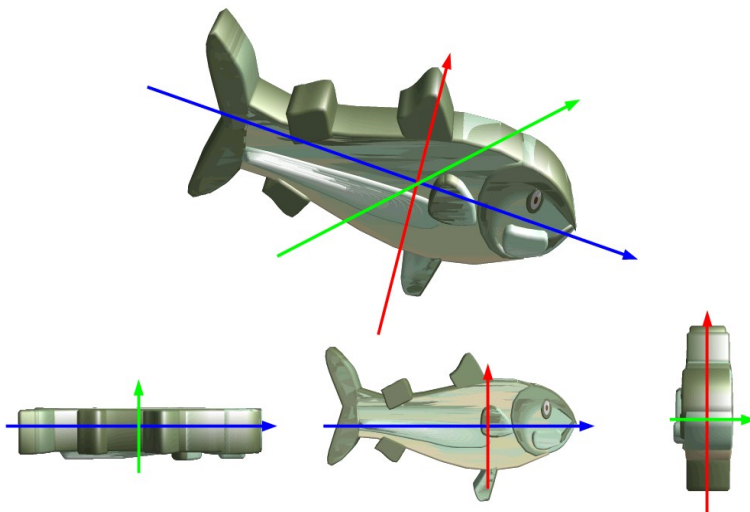




Commentaires



Les variables morphologiques recueillies présentent des **corrélations importantes**. On peut en effet supposer qu'une personne ayant un tour d'épaule important a également un tout de poitrine élevé (sauf exceptions...). Dans ces conditions, l'information apportée par les différentes variables est **redondante**. Graphiquement, sur les 3 variables (« Tour des épaules », « Tour de poitrine » et « Tour de taille »), cela se traduit par des zones vides de points dans le cube. Une variable unique calculée comme **combinaison** de ces 3 variables (représentée par la flèche en pointillés) suffirait à représenter les individus avec une **perte d'information minimale** car tous les points sont relativement proches de ce nouvel axe qui est la première composante principale.



Parmi les projections possibles en 2D, toutes ne permettent pas de reconnaître aussi facilement l'objet représenté. Parmi les 3 projections proposées, l'image du centre est la plus **fidèle** à l'original. Nous n'avons aucun mal à reconnaître l'objet initial car la projection s'est faite sur le plan formé par les 2 directions selon lesquelles l'objet initial **s'étale le plus** (grande variabilité). L'information apportée par la 3^{ème} dimension est minimale et sa perte n'est pas préjudiciable à la reconnaissance de l'objet.

Autrement dit...

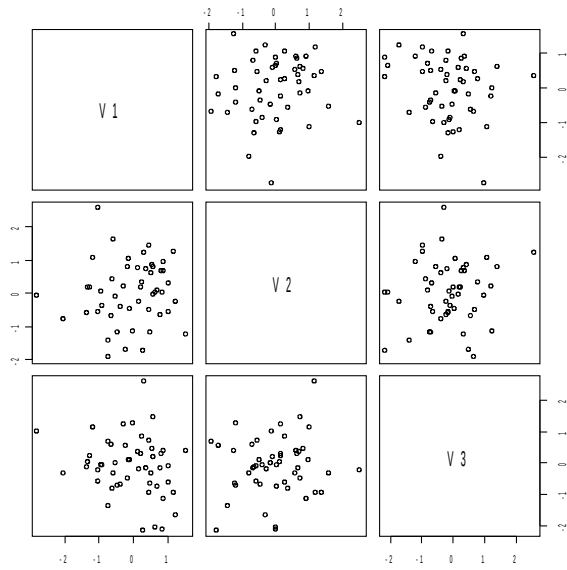
- L'ACP permet de déterminer les espaces de dimension inférieure à l'espace initial sur lesquels la projection du nuage de points initial soit **la moins déformée possible**, autrement dit celle qui conserve **le plus d'information** c'est-à-dire de **variabilité**.
- Le principe de l'ACP est de trouver un axe (la première composante principale), issu d'une **combinaison linéaire des variables initiales**, tel que la **variance** du nuage autour de cet axe soit **maximale**. Et de réitérer ce processus dans des directions orthogonales pour déterminer les composantes principales suivantes.
- Du point de vue des variables, l'ACP permet de conserver au mieux la **structure de corrélation** entre les variables initiales.

ACP : exemples simulés

Tableau de données : 50 individus, 3 variables (V1 – V2 - V3)

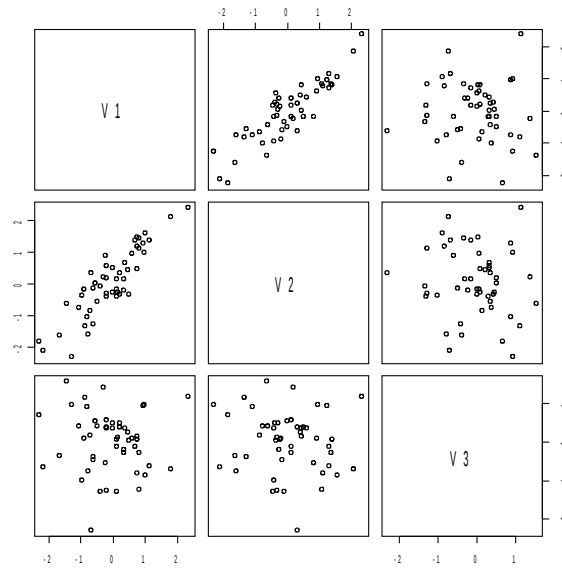
Cas 1)

{V1} - {V2} - {V3}



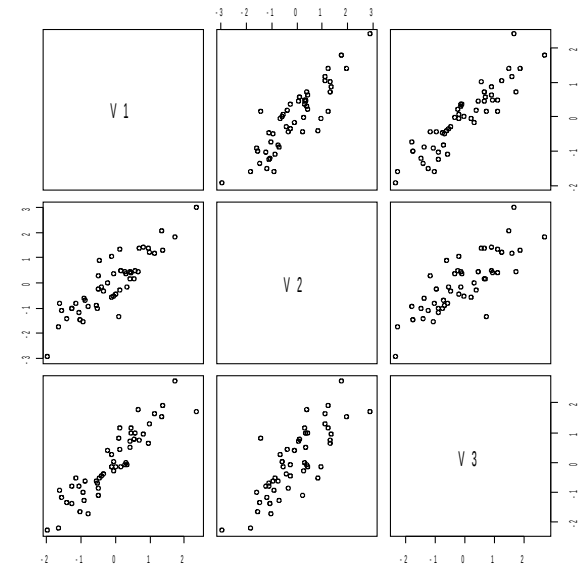
Cas 2)

{V1 - V2} - {V3}



Cas 3)

{V1 - V2 - V3}



Matrices de corrélation

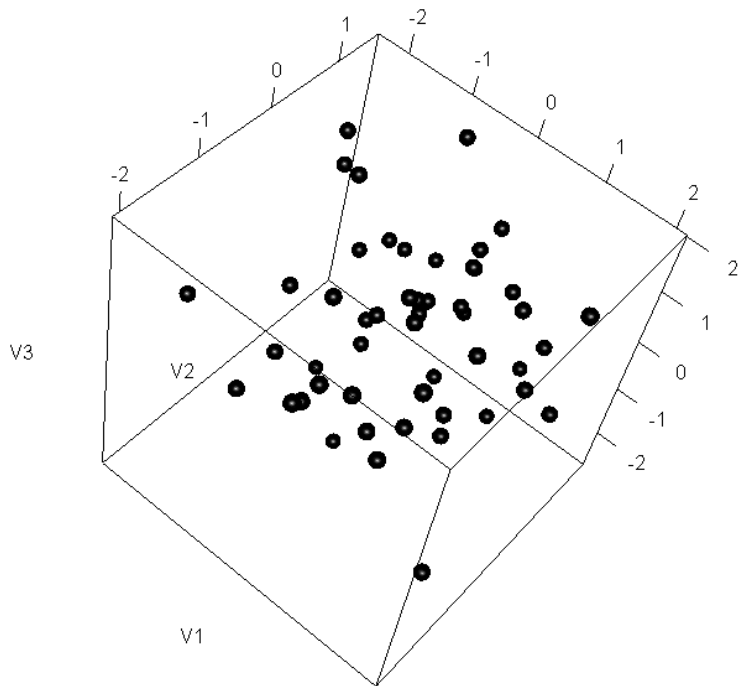
1)	V1	V2	V3
V1	1.0	-0.10	0.00
V2	-0.1	1.00	-0.12
V3	0.0	-0.12	1.00

2)	V1	V2	V3
V1	1.00	0.88	-0.05
V2	0.88	1.00	-0.11
V3	-0.05	-0.11	1.00

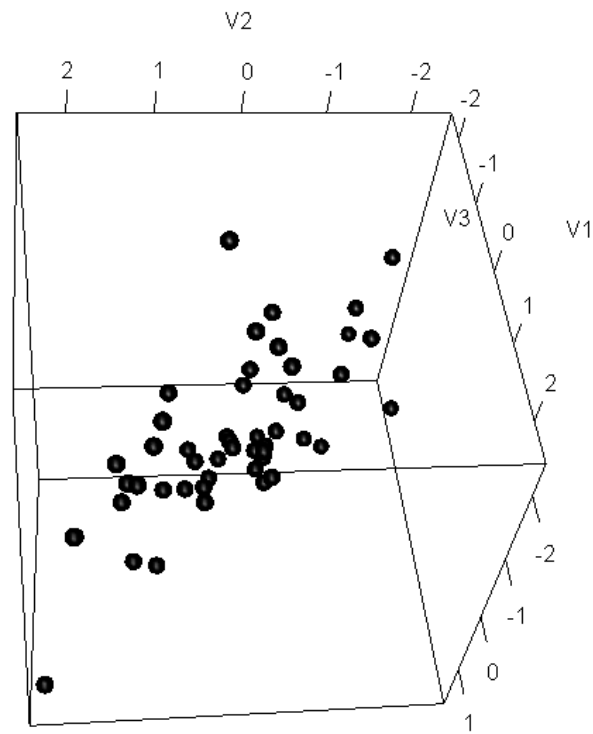
3)	V1	V2	V3
V1	1.00	0.88	0.92
V2	0.88	1.00	0.81
V3	0.92	0.81	1.00

Exemple : vues en 3D

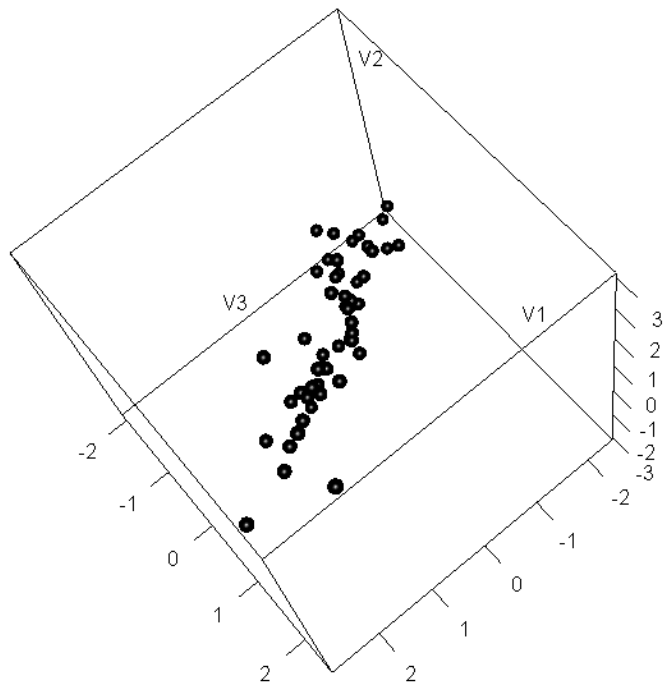
Cas 1)



Cas 2)



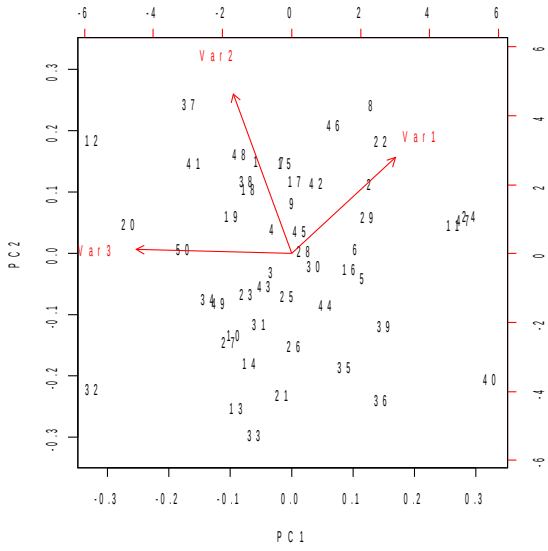
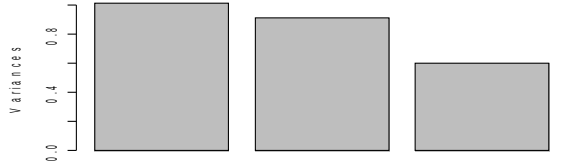
Cas 3)



Exemple : éboulis des valeurs propres et biplot

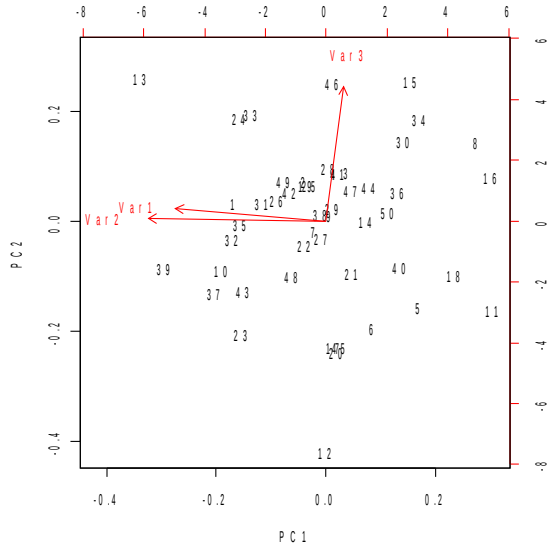
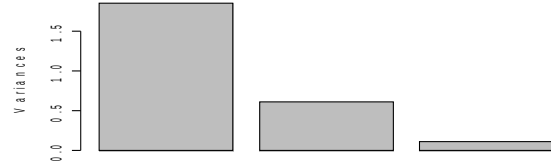
1)

ac p1



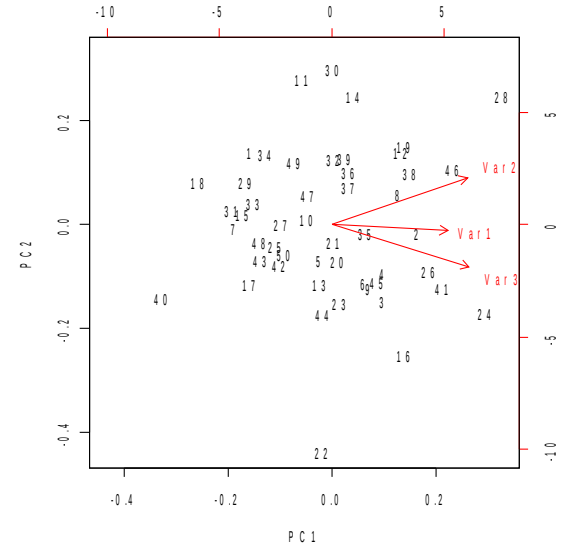
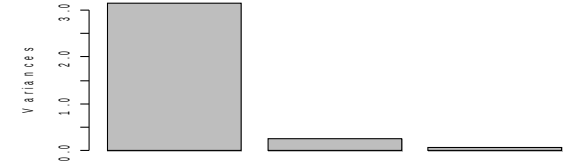
2)

ac p2

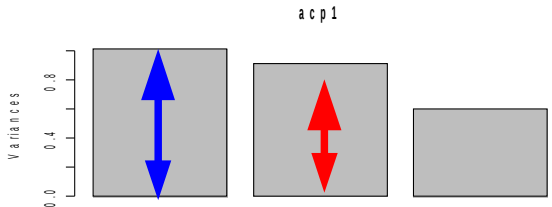
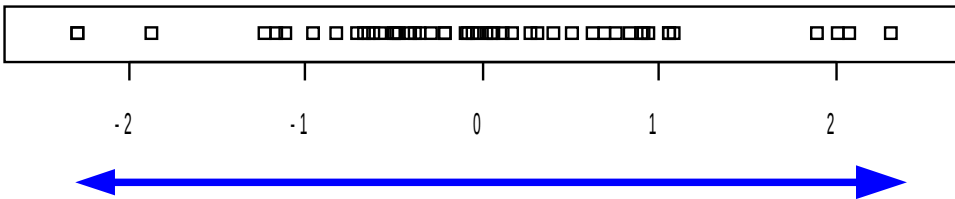


3)

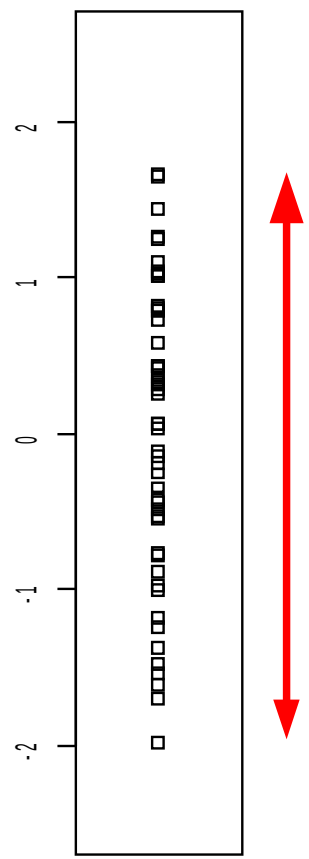
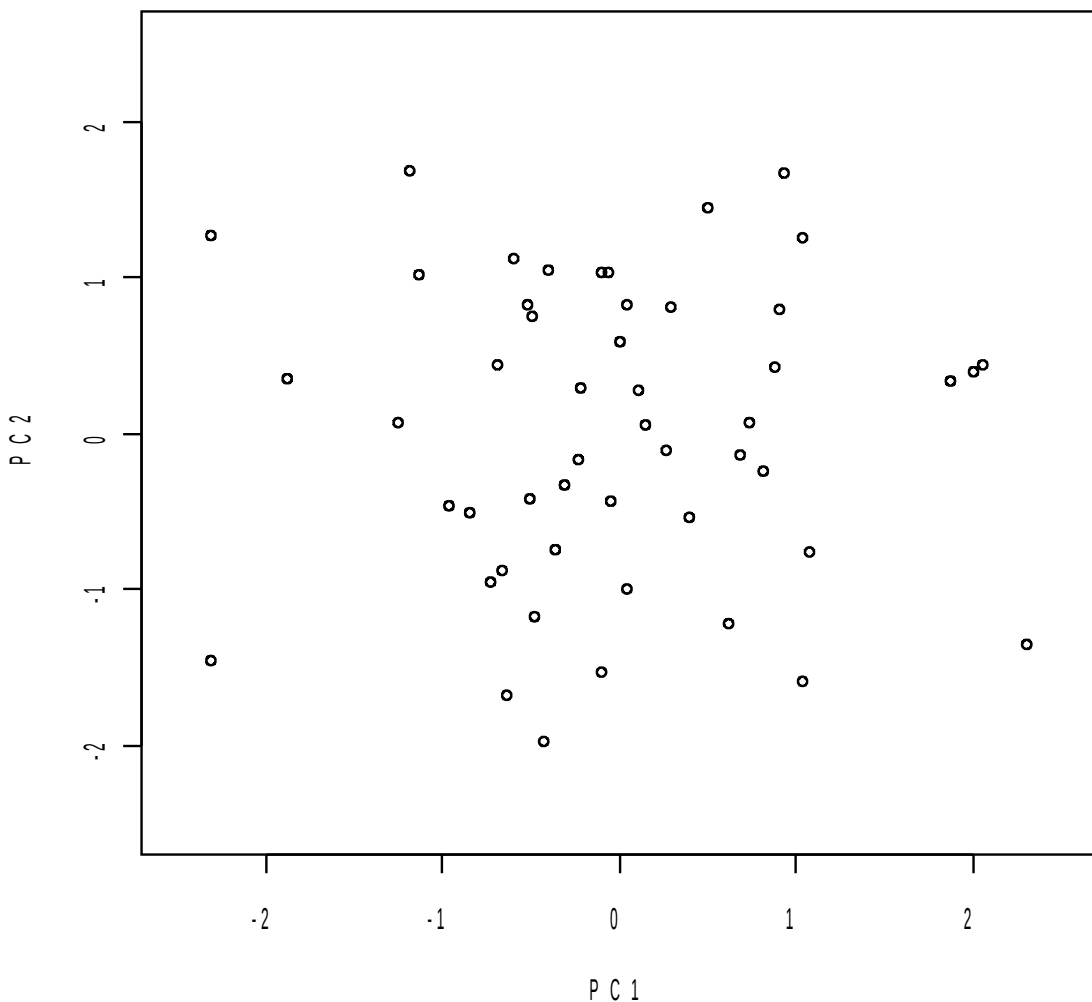
ac p3



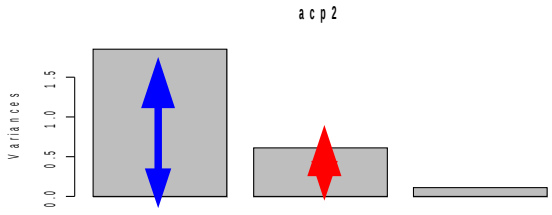
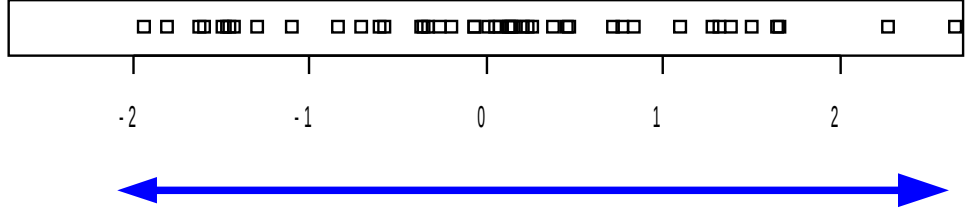
Exemple : représentation des individus



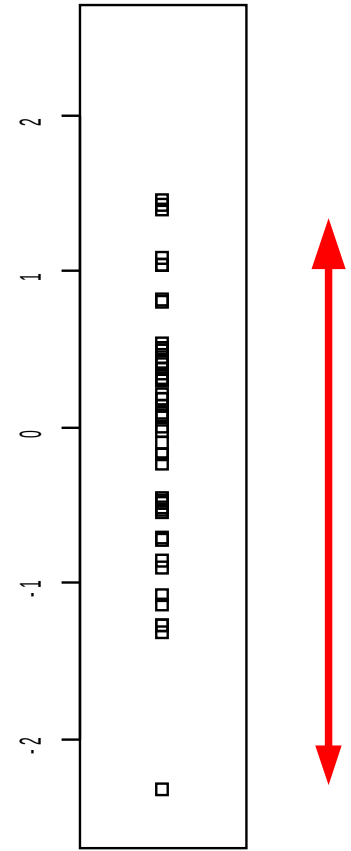
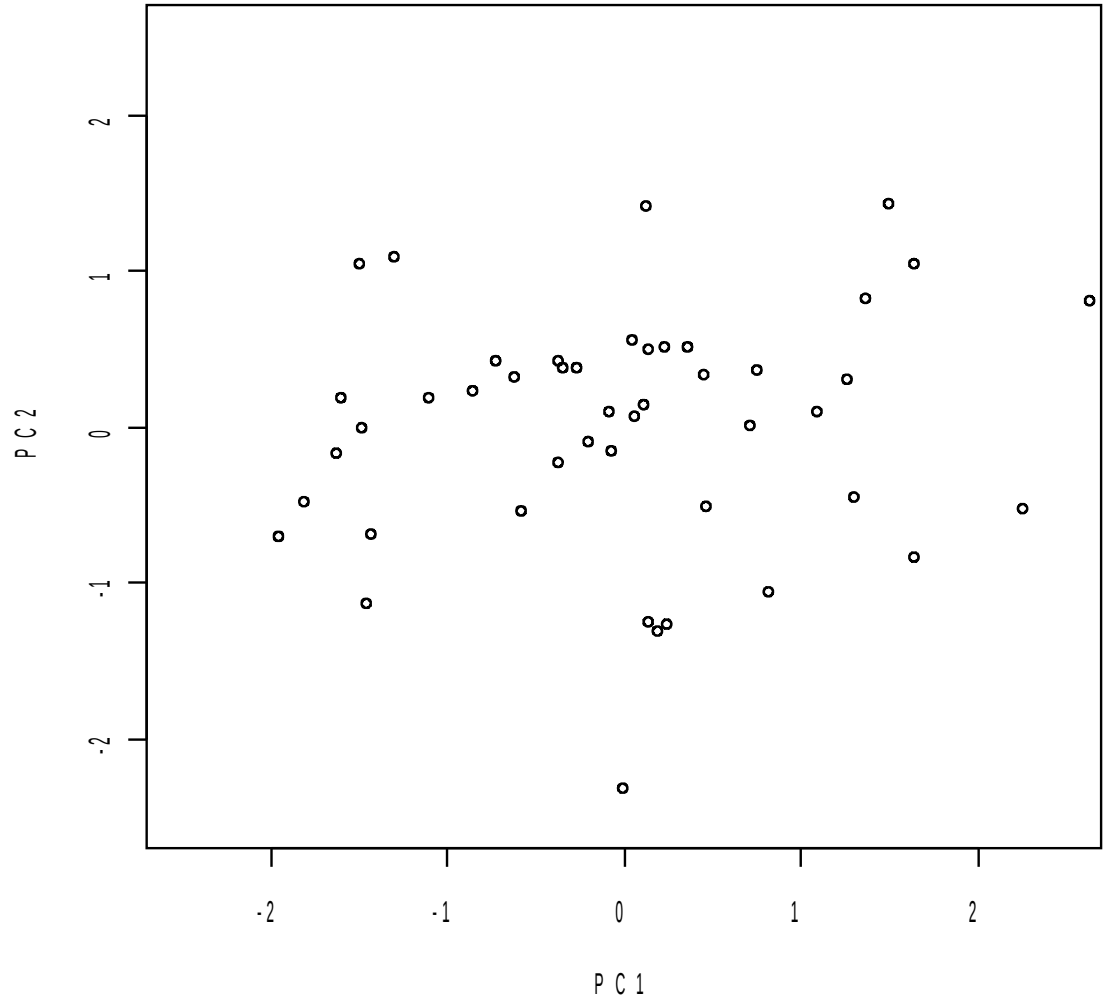
1)



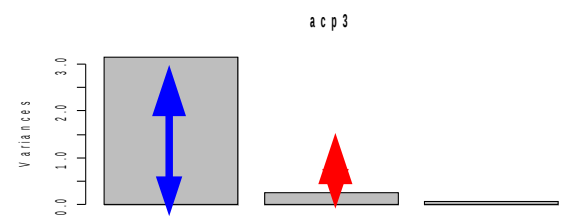
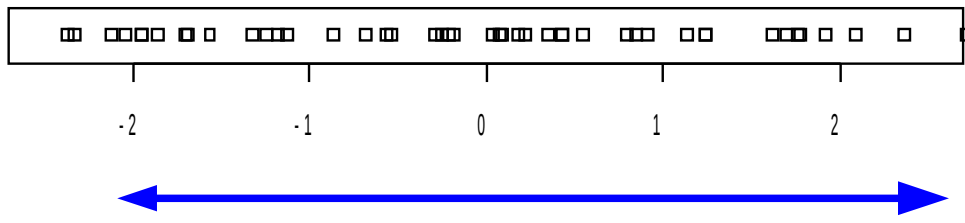
Exemple : représentation des individus



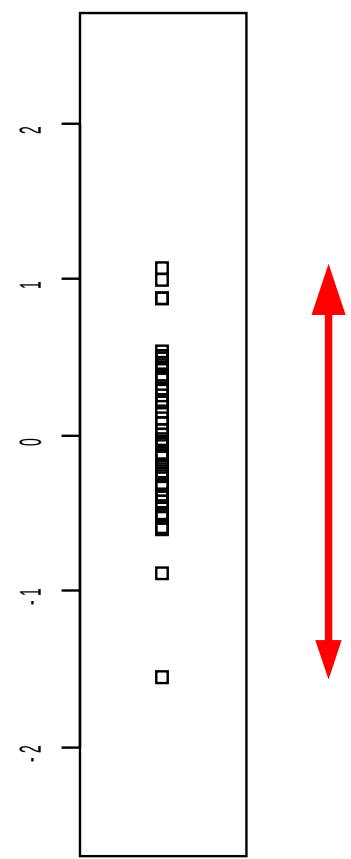
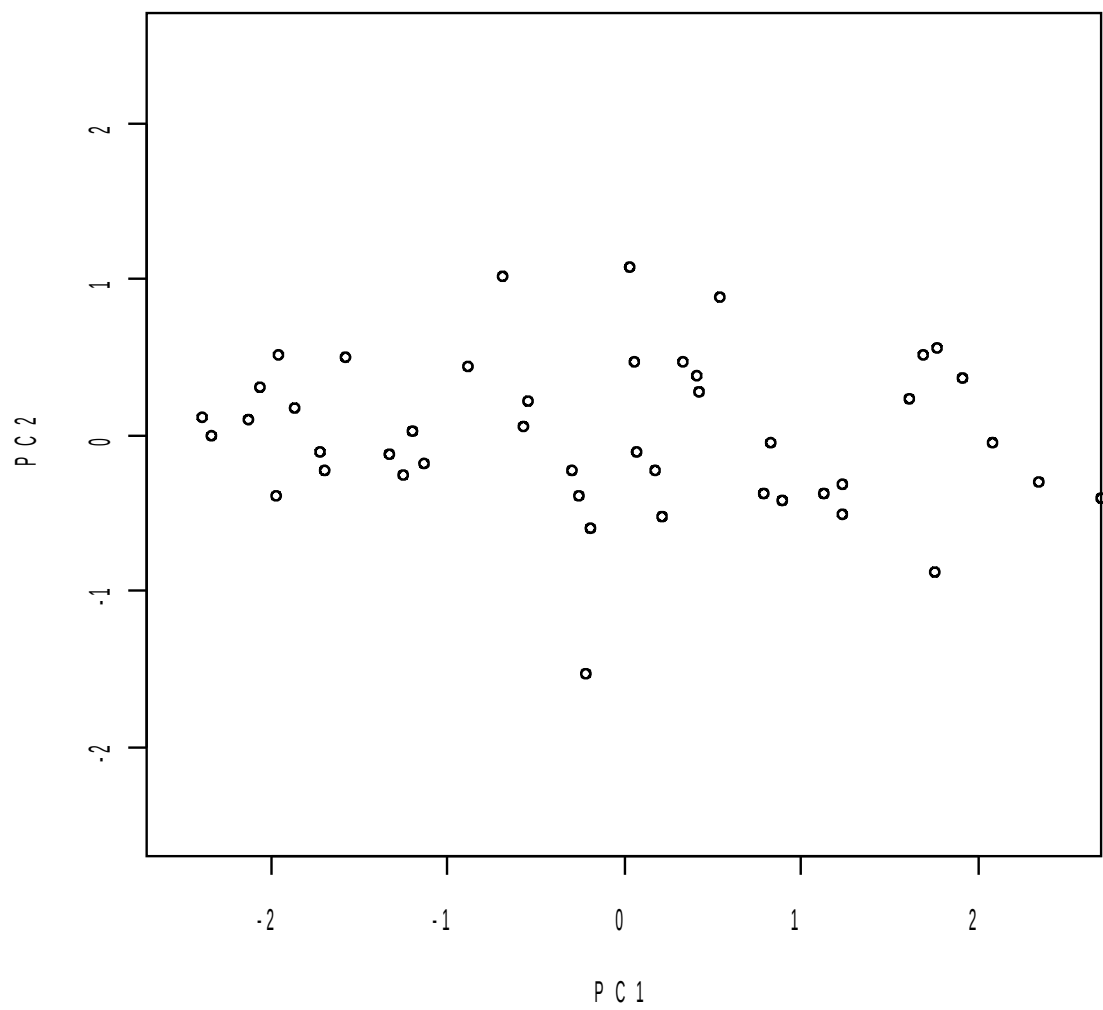
2)



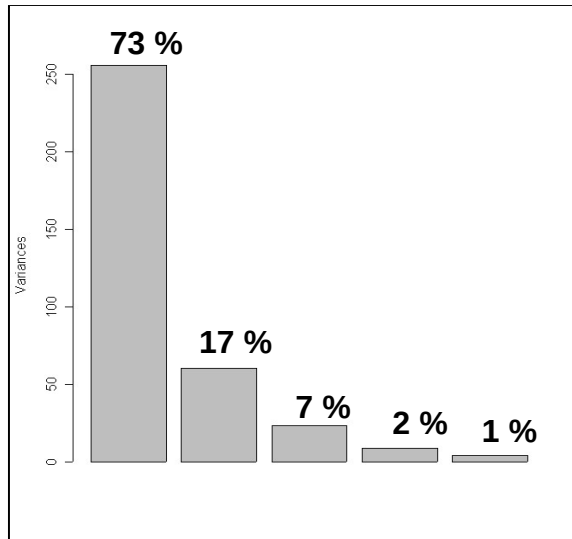
Exemple : représentation des individus



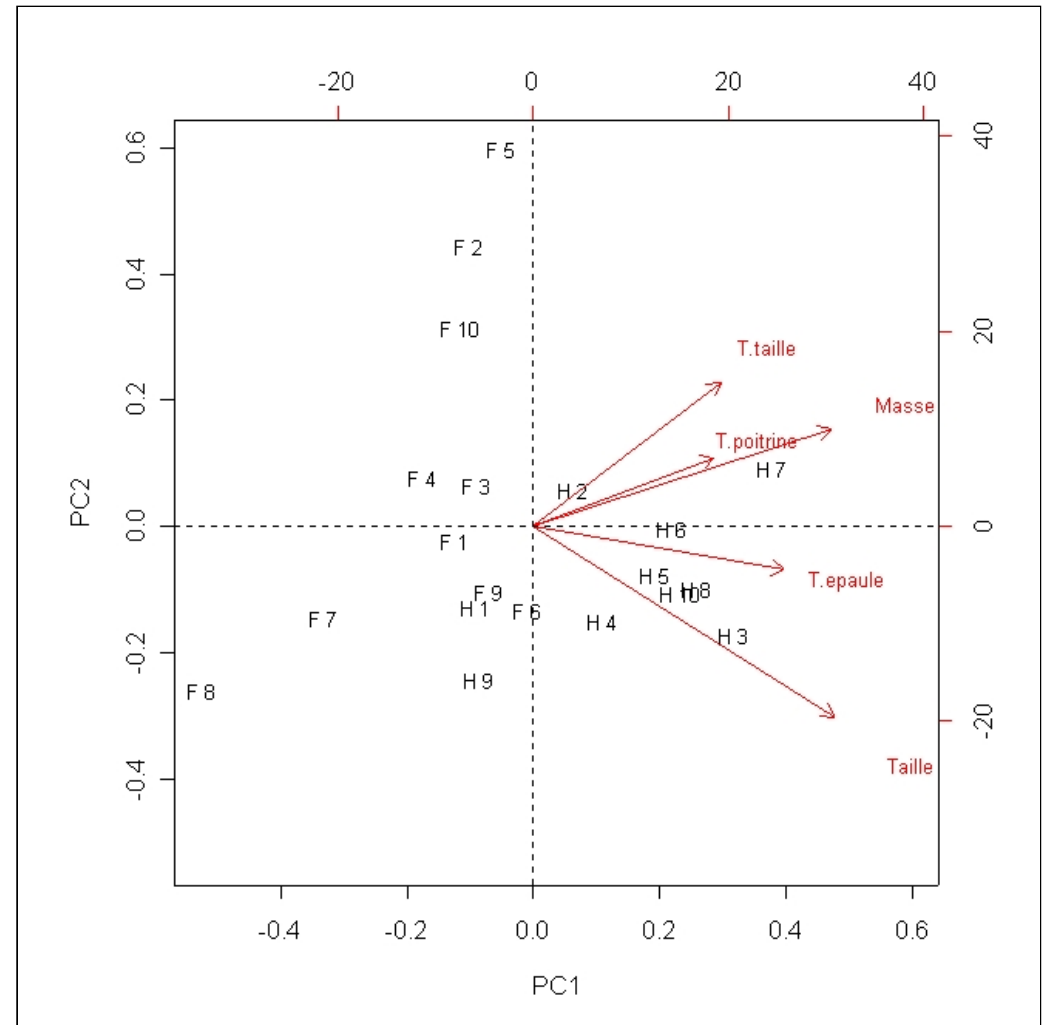
3)



ACP : exemple « morpho »



- 90% de l'information expliquée par les 2 premières CP
- le passage de 5 à 2 dimensions se fait en « perdant » 10% d'information
- Axe 1 « gabarit » : séparation des grands gabarit (valeurs élevées pour les 5 variables) à droite et des petits à gauche
- Axe 2 « embonpoint » : en bas, variables liées à la taille et à la carrure, en haut, masse et tour de taille / poitrine

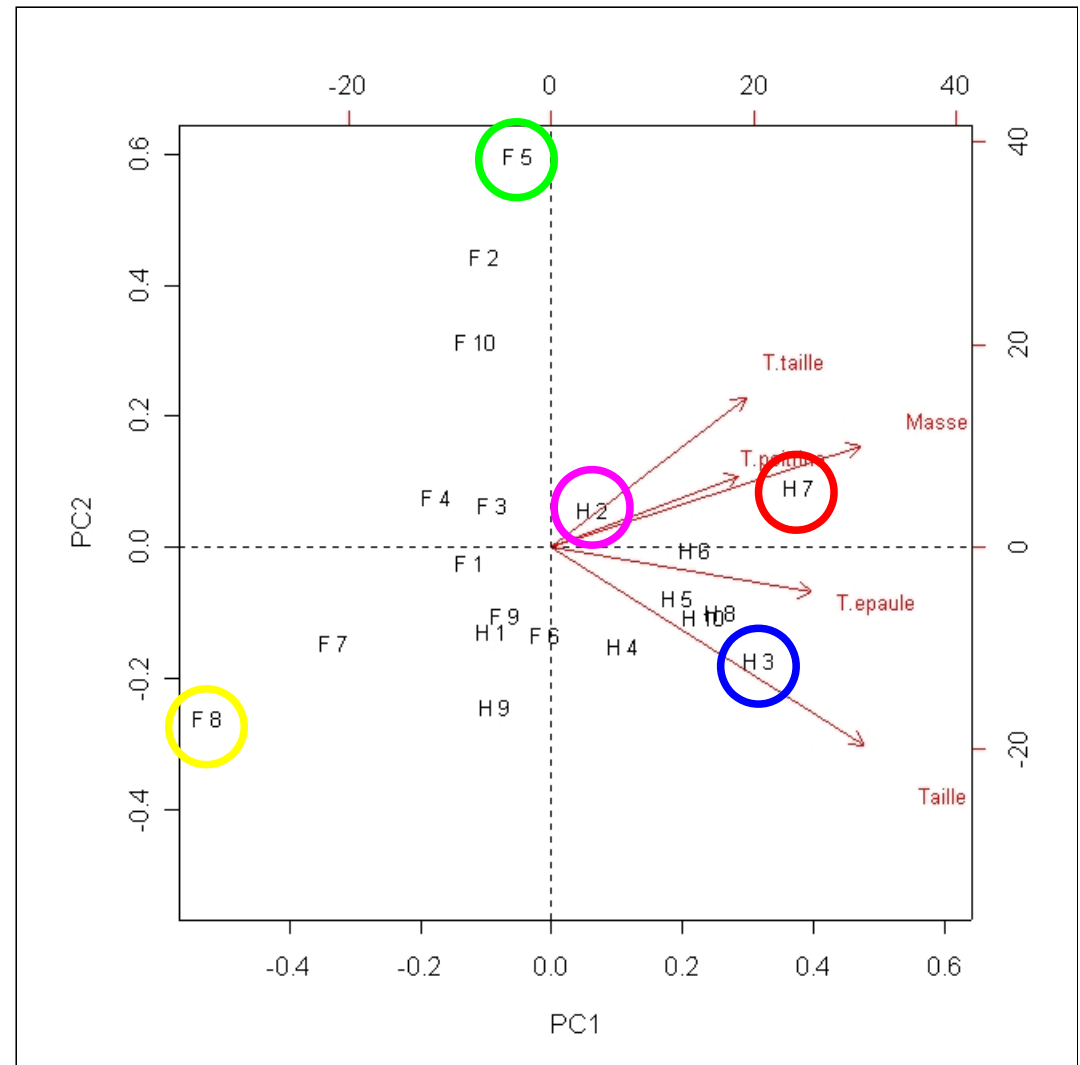


Matrice des corrélations

	T.ep	T.p	T.t	M	T
T.ep	1.00	0.74	0.48	0.72	0.71
T.p	0.74	1.00	0.78	0.81	0.51
T.t	0.48	0.78	1.00	0.86	0.37
M	0.72	0.81	0.86	1.00	0.61
T	0.71	0.51	0.37	0.61	1.00

ACP : exemple « morpho »

	T.ep	T.p	T.t	M	T
H 1	106.2	89.5	71.5	65.6	174.0
H 2	110.5	97.0	79.0	71.8	175.3
H 3	115.1	97.5	83.2	80.7	193.5
H 4	104.5	97.0	77.8	72.6	186.5
H 5	107.5	97.5	80.0	78.8	187.2
H 6	119.8	99.9	82.5	74.8	181.5
H 7	123.5	106.9	82.0	86.4	184.0
H 8	120.4	102.5	76.8	78.4	184.5
H 9	111.0	91.0	68.5	62.0	175.0
H 10	119.5	93.5	77.5	81.6	184.0
F 1	105.0	89.0	71.2	67.3	169.5
F 2	100.2	94.1	79.6	75.5	160.0
F 3	99.1	90.8	77.9	68.2	172.7
F 4	107.6	97.0	69.6	61.4	162.6
F 5	104.0	95.4	86.0	76.8	157.5
F 6	108.4	91.8	69.9	71.8	176.5
F 7	99.3	87.3	63.5	55.5	164.4
F 8	91.9	78.1	57.9	48.6	160.7
F 9	107.1	90.9	72.2	66.4	174.0
F 10	100.5	97.1	80.4	67.3	163.8



Individu moyen (origine du repère) :

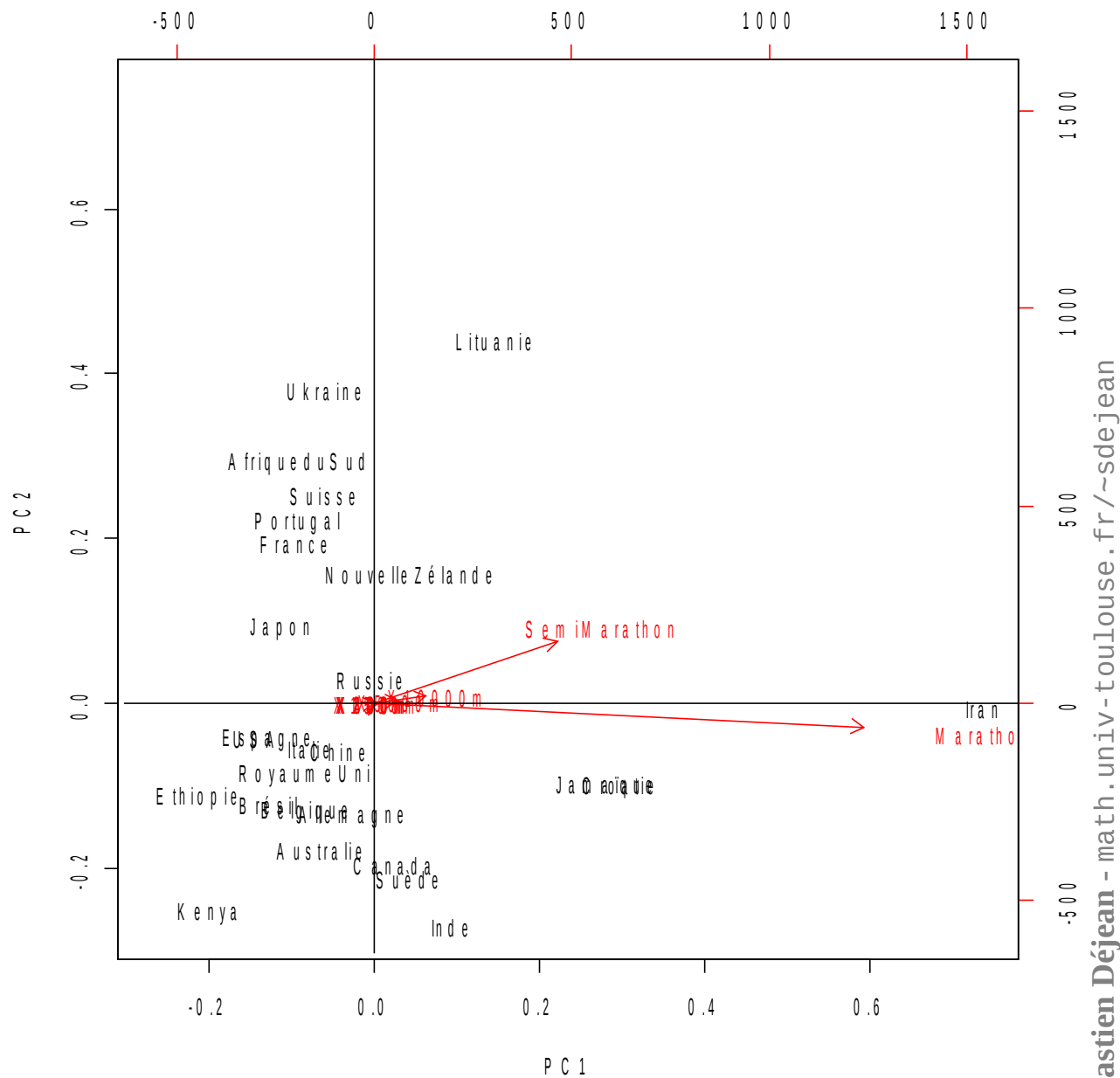
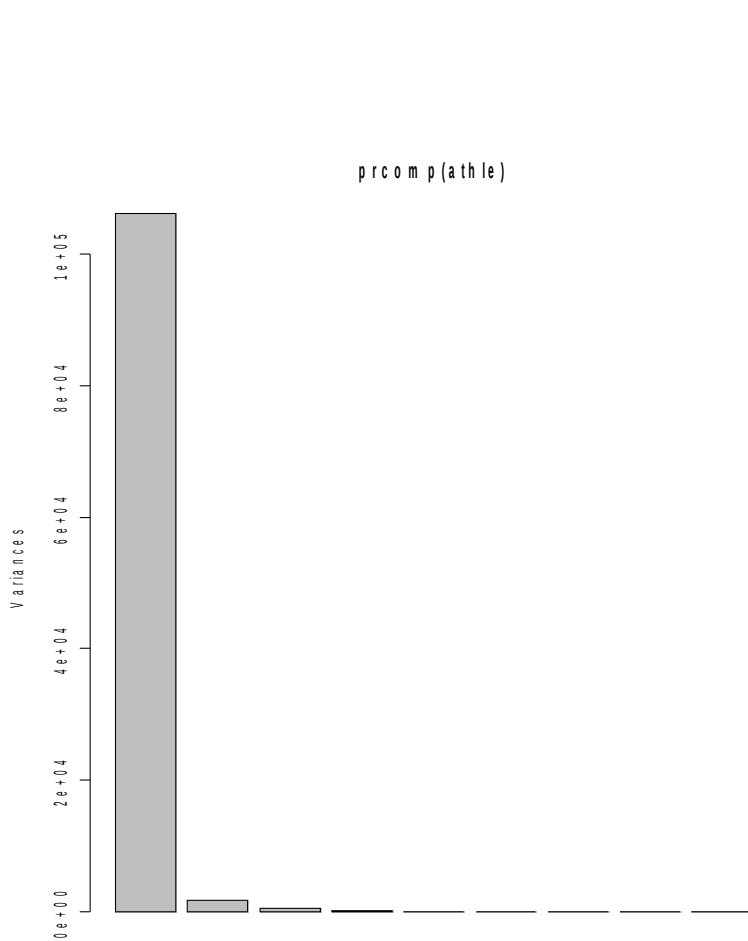
108.1 94.2 75.4 70.6 174.4

Données : record d'athlétisme

Records nationaux (en secondes) de quelques épreuves d'athlétisme

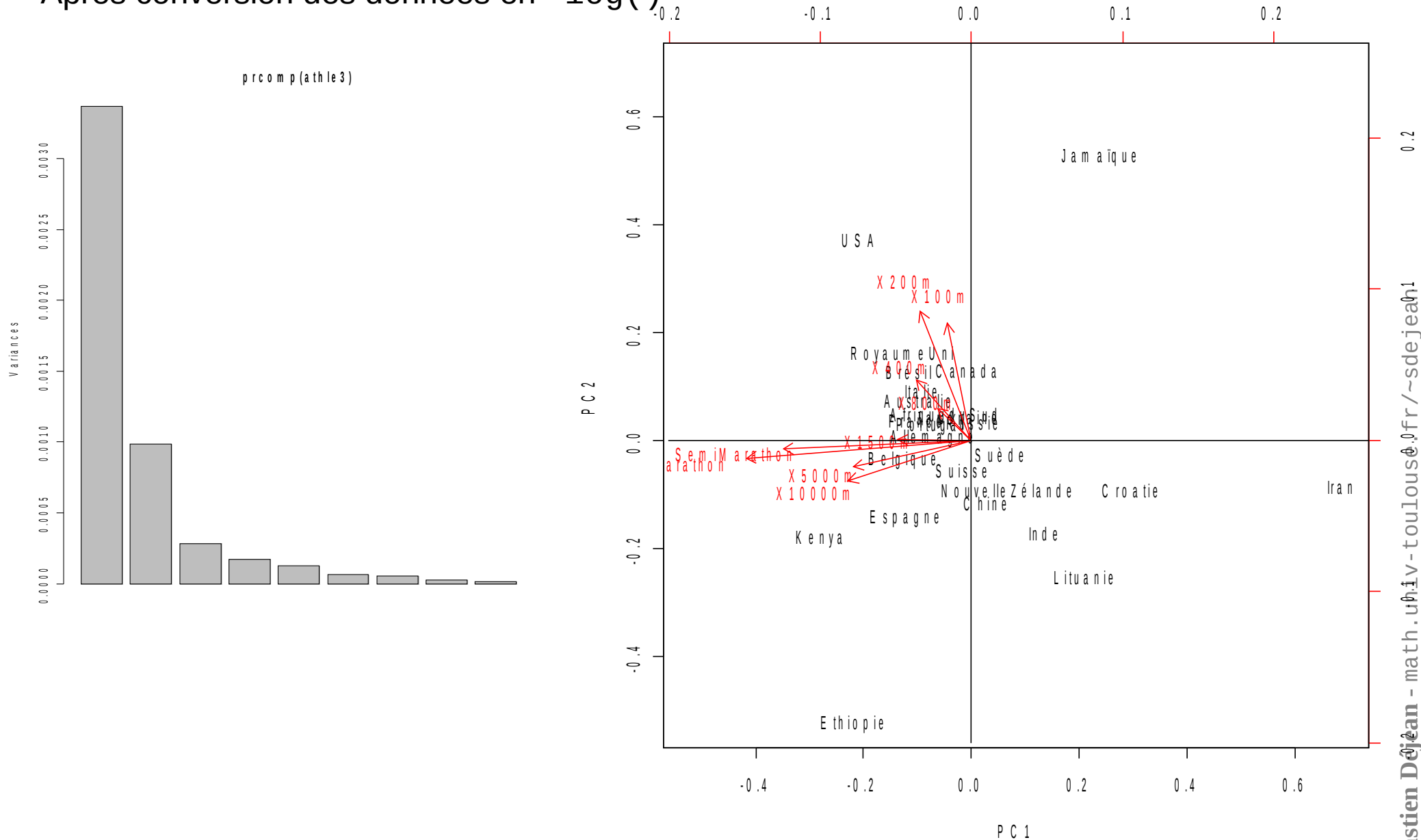
	100m	200m	400m	800m	1500m	5000m	10000m	SemiMarathon	Marathon
Australie	9.93	20.06	44.38	104.40	211.96	775.76	1649.73	3602	7671
Belgique	10.02	20.19	44.78	103.86	214.13	769.71	1612.30	3605	7640
Brésil	10.00	19.89	44.29	101.77	213.25	799.43	1648.12	3573	7565
RoyaumeUni	9.87	19.87	44.36	101.73	209.67	780.41	1638.14	3609	7633
Canada	9.84	20.17	44.44	103.68	211.71	793.96	1656.01	3650	7809
Chine	10.17	20.54	45.25	106.44	216.49	805.14	1670.00	3635	7695
Croatie	10.25	20.76	45.64	104.07	213.30	817.76	1704.32	3827	8225
Ethiopie	10.50	21.08	45.89	106.08	211.13	757.35	1577.53	3535	7439
France	9.99	20.16	44.46	103.15	208.98	778.83	1642.78	3658	7596
Allemagne	10.06	20.20	44.33	103.65	211.58	774.70	1641.53	3634	7727
Inde	10.30	20.73	45.48	105.77	218.00	809.70	1682.89	3672	7920
Iran	10.29	21.11	46.37	104.74	218.80	833.40	1762.65	4103	8903
Italie	10.01	19.72	45.19	103.17	212.78	785.59	1636.50	3620	7642
Jamaïque	9.58	19.19	44.49	105.21	219.19	813.10	1712.44	3816	8199
Japon	10.00	20.03	44.78	106.18	217.42	793.20	1655.09	3625	7576
Kenya	10.26	20.43	44.18	102.01	206.34	759.74	1587.85	3513	7467
Lituanie	10.33	20.88	45.73	106.64	220.90	797.90	1651.50	3851	7955
NouvelleZélande	10.11	20.42	46.09	104.30	212.17	790.19	1661.95	3732	7815
Portugal	9.86	20.01	46.11	104.91	210.07	782.86	1632.47	3665	7596
Russie	10.10	20.23	44.60	102.47	212.28	791.99	1673.12	3675	7747
AfriqueduSud	10.06	20.11	44.59	102.69	213.56	794.16	1649.94	3678	7593
Espagne	10.14	20.59	44.96	103.83	208.95	782.54	1634.44	3592	7562
Suède	10.18	20.30	44.56	105.54	216.49	797.59	1675.74	3655	7838
Suisse	10.16	20.41	44.99	102.55	211.75	787.54	1673.16	3686	7643
Ukraine	10.07	20.00	45.11	105.08	210.33	790.78	1679.80	3711	7635
USA	9.69	19.32	43.18	102.60	209.30	776.27	1633.98	3583	7538

ACP "brute"

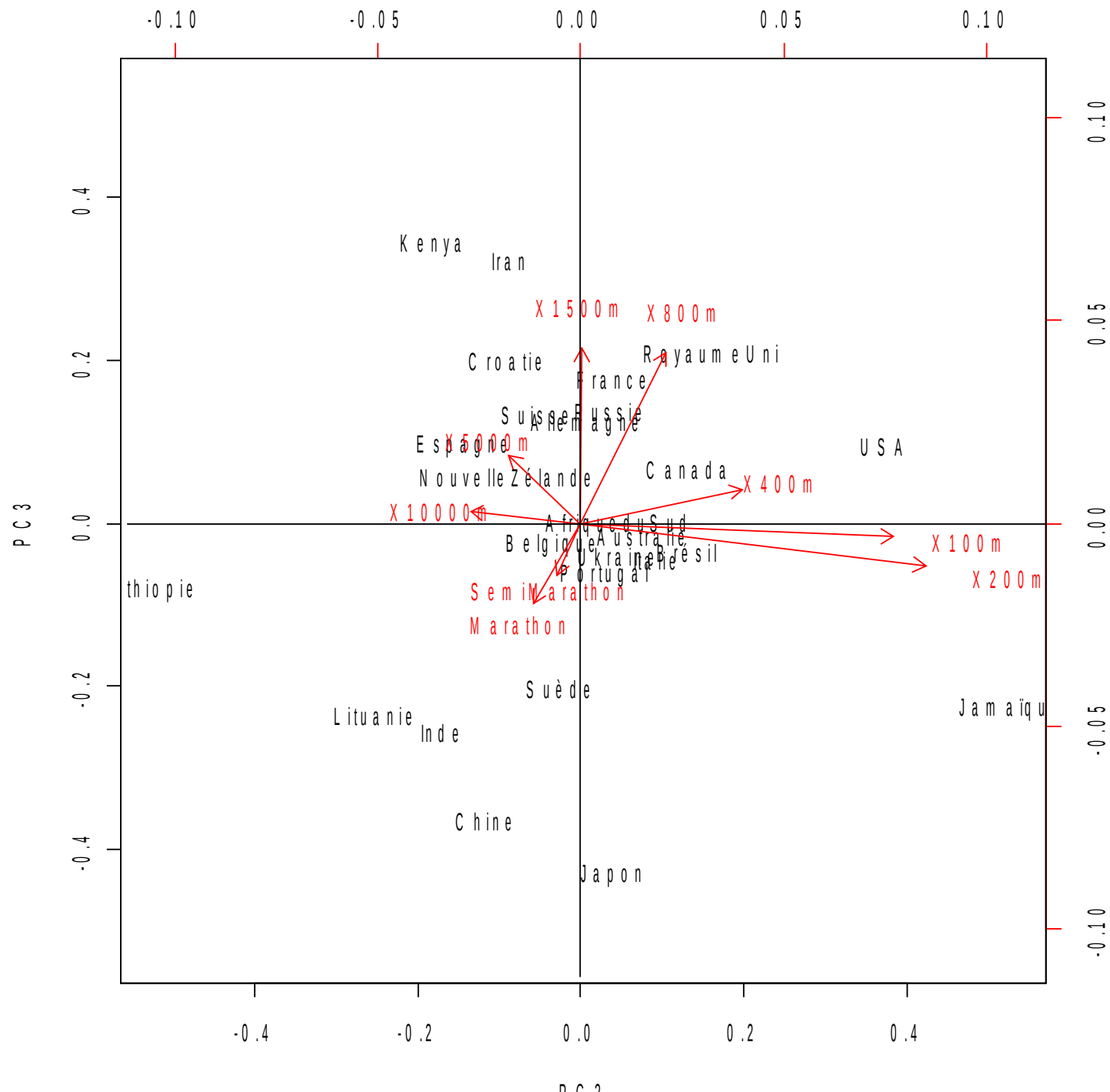


ACP après pré-traitement

Après conversion des données en $-\log()$

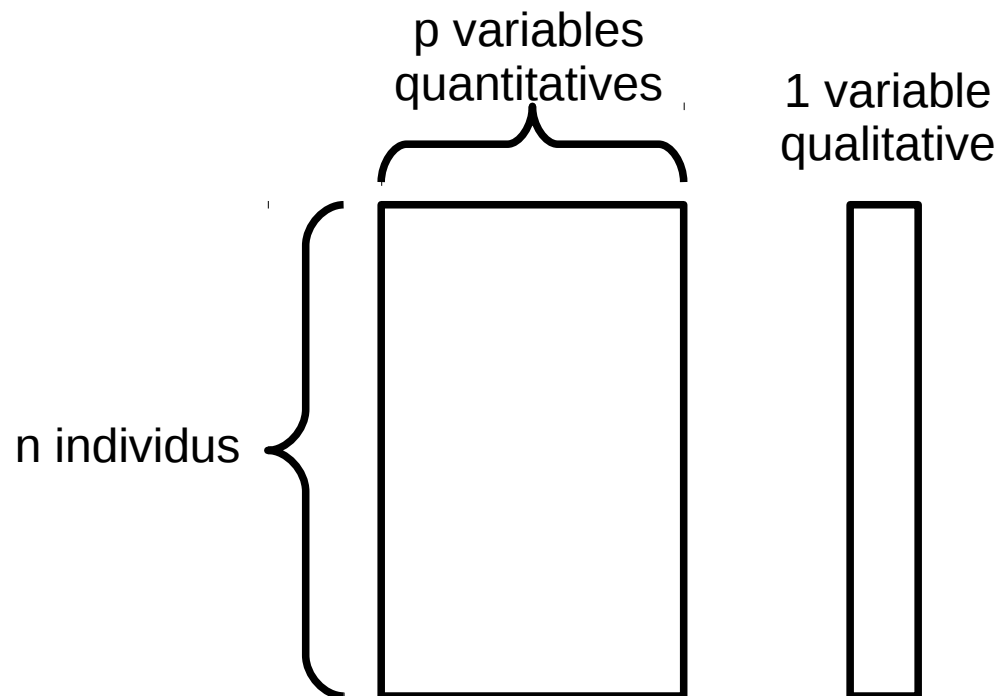


Plan PC2-PC3



Analyse Factorielle Discriminante

Objectif : décrire un tableau de données constitué de variables **quantitatives** et d'une variable **qualitative** en cherchant à afficher distinctement les différentes modalités de la variable qualitative.



Analyse Factorielle Discriminante

	V1	V2	V3	V4	V5	S
I 1	106.2	89.5	71.5	65.6	174.0	M
I 2	110.5	97.0	79.0	71.8	175.3	M
I 3	115.1	97.5	83.2	80.7	193.5	M
I 4	104.5	97.0	77.8	72.6	186.5	M
I 5	107.5	97.5	80.0	78.8	187.2	M
I 6	119.8	99.9	82.5	74.8	181.5	M
I 7	123.5	106.9	82.0	86.4	184.0	M
I 8	120.4	102.5	76.8	78.4	184.5	M
I 9	111.0	91.0	68.5	62.0	175.0	M
I 10	119.5	93.5	77.5	81.6	184.0	M
I 11	105.0	89.0	71.2	67.3	169.5	F
I 12	100.2	94.1	79.6	75.5	160.0	F
I 13	99.1	90.8	77.9	68.2	172.7	F
I 14	107.6	97.0	69.6	61.4	162.6	F
I 15	104.0	95.4	86.0	76.8	157.5	F
I 16	108.4	91.8	69.9	71.8	176.5	F
I 17	99.3	87.3	63.5	55.5	164.4	F
I 18	91.9	78.1	57.9	48.6	160.7	F
I 19	107.1	90.9	72.2	66.4	174.0	F
I 20	100.5	97.1	80.4	67.3	163.8	F

Peut-on trouver un espace de représentation qui sépare au mieux les hommes des femmes en fonction des 5 variables morphologiques recueillies ?

AFD : exemple simulé

	V1	V2	V3	Groupe
1	-2.02	1.93	2.09	A
2	1.37	-0.12	2.01	A
3	6.02	4.15	1.77	A
4	0.50	-4.84	2.63	A
5	-3.46	0.40	2.04	A
6	2.03	0.22	2.09	A
7	-4.27	-0.19	1.84	A
8	10.44	-0.08	1.43	A
9	7.53	3.55	1.59	A
10	-2.75	-2.69	2.06	A
11	-7.16	5.18	2.00	A
12	11.82	-4.89	2.25	A
13	-0.52	-5.94	2.05	A
14	-0.62	-0.77	1.97	A
15	0.67	0.64	1.76	A
16	2.34	-0.93	1.74	A
17	2.79	-2.98	2.07	A
18	-1.87	0.05	2.02	A
19	-0.09	-0.69	2.32	A
20	5.07	5.57	2.08	A
21	0.38	0.90	1.69	A
22	1.50	3.79	1.96	A
23	0.78	-4.40	1.81	A
24	1.40	1.16	2.13	A
25	1.64	0.38	1.77	A
26	-4.00	-2.60	-1.95	B
27	5.15	0.59	-1.94	B
28	6.98	-1.14	-2.17	B
29	5.57	-6.49	-2.15	B
30	-5.84	-1.83	-1.82	B
31	-3.20	-0.07	-2.14	B
32	3.20	0.87	-1.50	B
33	-6.63	4.56	-1.92	B
34	-2.80	-1.53	-1.70	B
35	3.43	2.98	-2.14	B
36	-4.24	-2.61	-2.18	B
37	2.20	0.55	-1.89	B
38	-3.07	-2.07	-1.97	B
39	0.26	1.30	-1.85	B
40	0.32	0.79	-1.78	B
41	1.14	5.79	-1.64	B
42	-1.21	-2.88	-1.50	B
43	1.38	1.71	-2.11	B
44	-0.80	-0.38	-1.99	B
45	-2.04	-4.60	-2.00	B
46	7.67	5.84	-2.09	B
47	-4.50	-0.15	-1.85	B

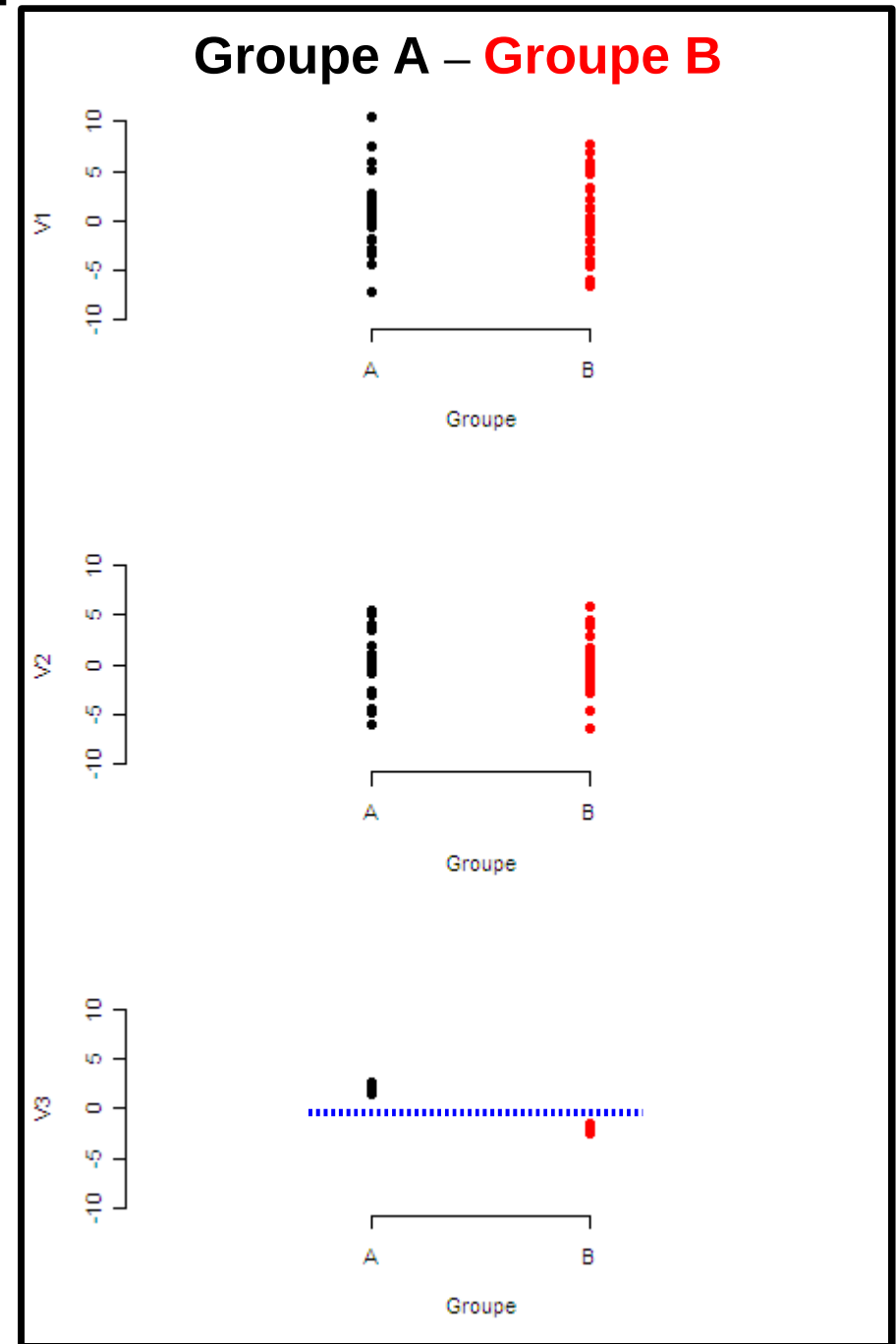
Tableau de données

- 50 individus, 4 variables
- 3 quantitatives V1 – V2 – V3
- 1 qualitative Groupe à 2 modalités A et B

Peut-on trouver un espace de représentation qui sépare au mieux les individus du groupe A de ceux du groupe B ?

AFD : exemple simulé

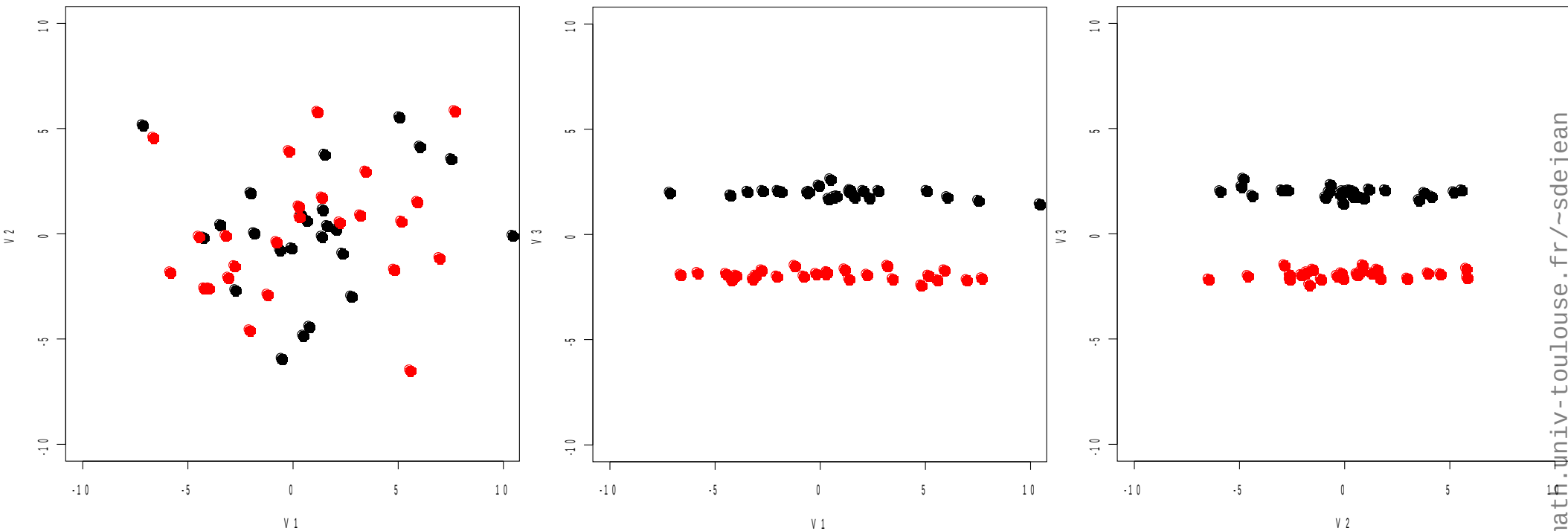
Représentation des 50 individus selon les 3 variables séparément avec couleur selon la modalité de la variable qualitative.



On voit bien que la variable V3 joue un rôle prépondérant dans la discrimination des 2 groupes.

AFD : exemple simulé

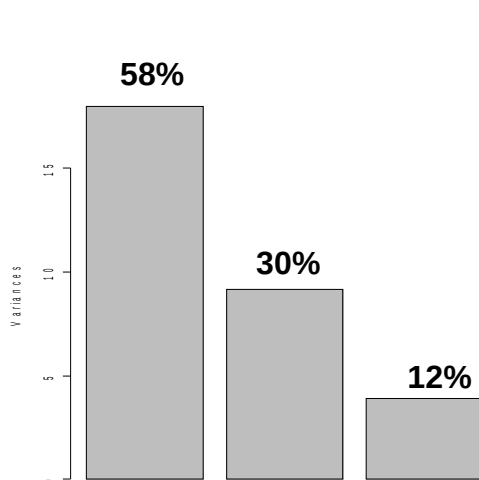
Autres représentations croisant les variables 2 à 2.



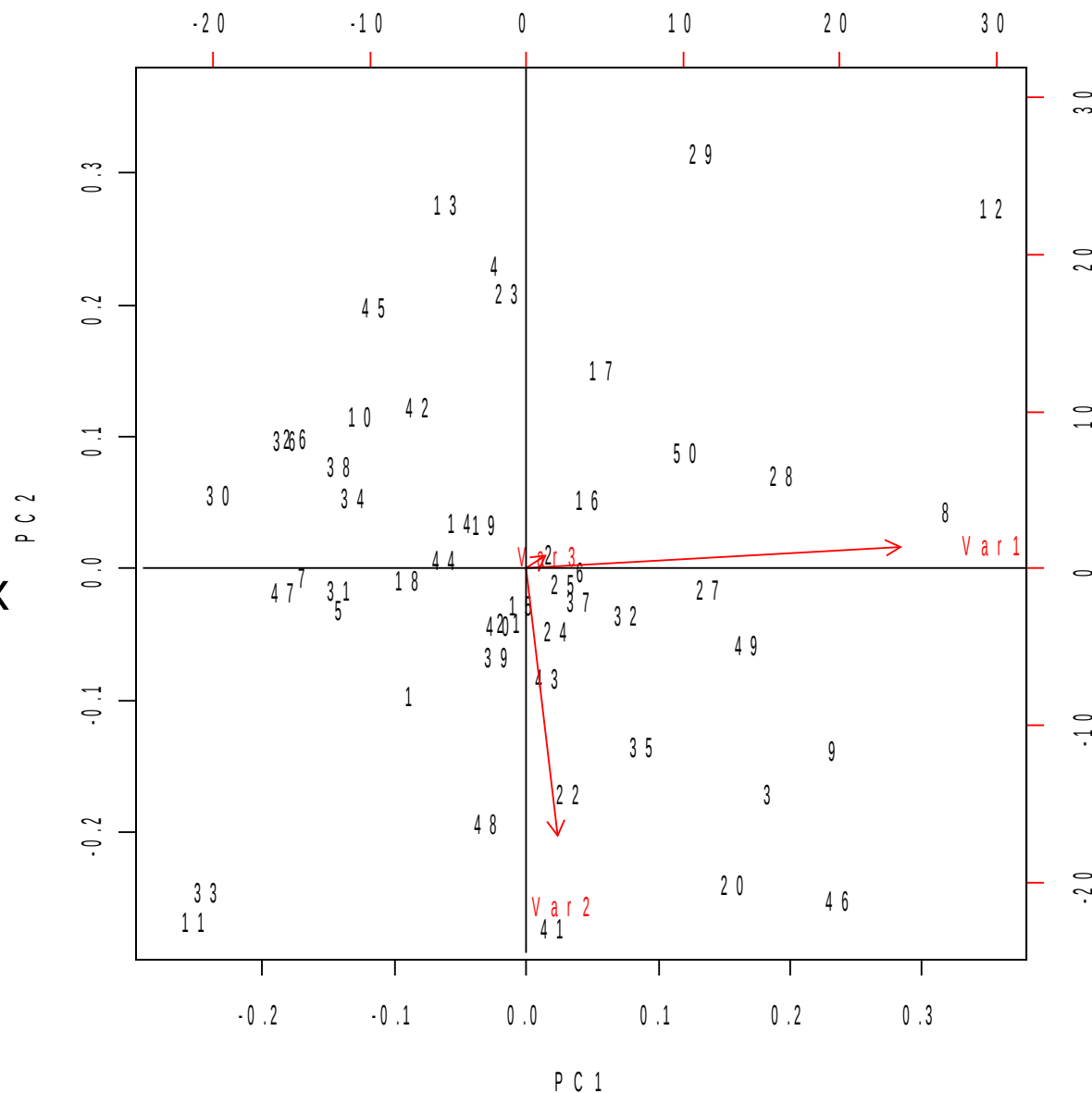
Le caractère discriminant de V3 est encore clairement mis en évidence.

AFD : exemple simulé

Résultat d'une **ACP** appliquée sur les données (**sans prise en compte de la variable qualitative**).

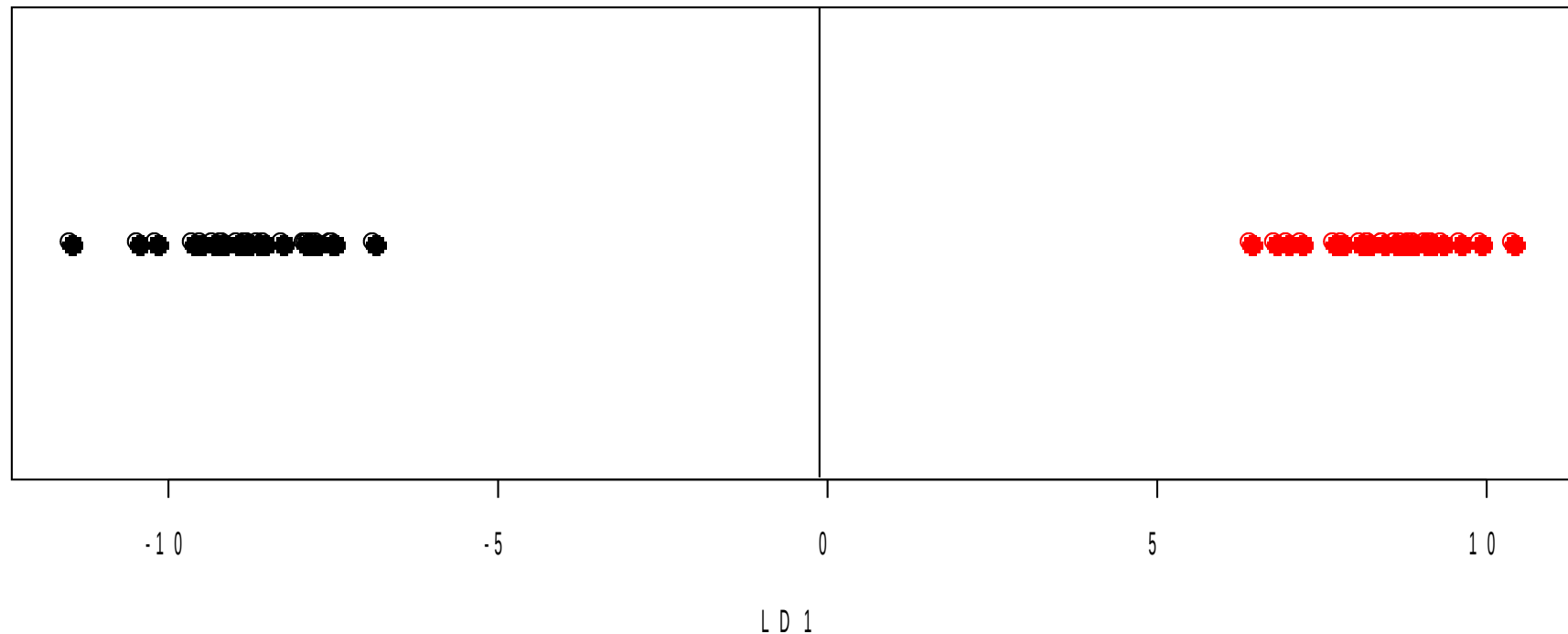


- Les 3 CP sont clairement identifiées respectivement aux 3 variables initiales V1-V2-V3.
- La plus grande part de la variabilité des données est expliquée par V1, puis V2 et enfin V3.
- Le caractère discriminant de V3 n'est pas pris en compte dans cet analyse.



AFD : exemple simulé

Résultat d'une AFD



- 2 modalités → 1 variable discriminante (1 axe de représentation)
- Combinaison linéaire des variables initiales :

$$\text{LD1} = -0.058 * V1 - 0.028 * V2 - \mathbf{4.41} * V3$$
- La variable discriminante LD1 correspond à l'opposé de la variable V3.

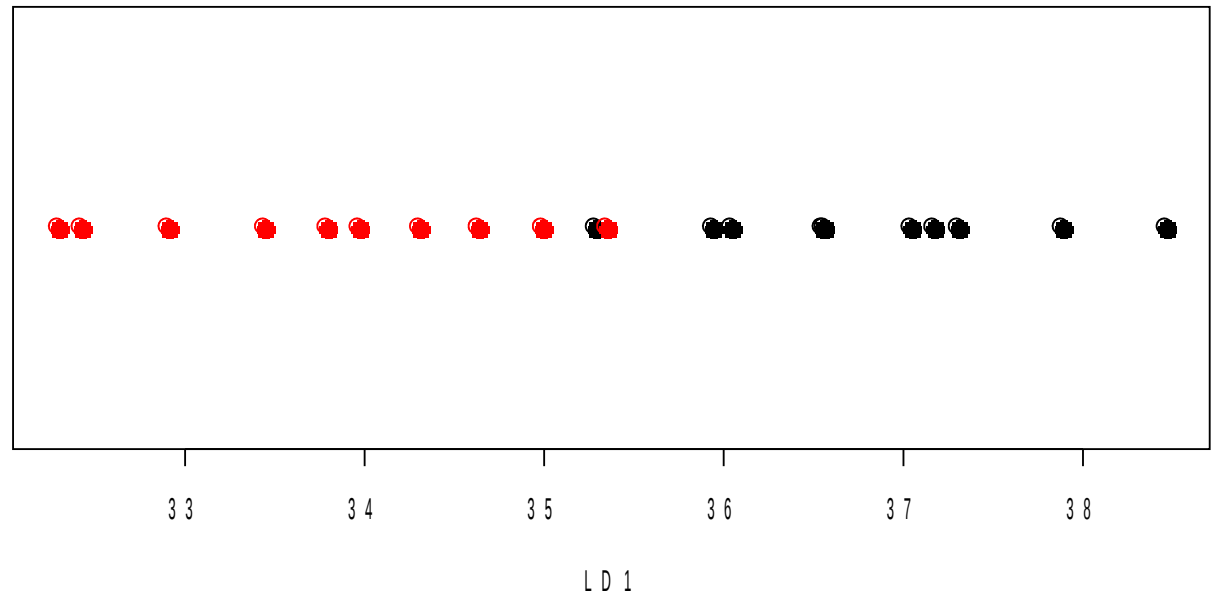
AFD : exemple « morpho »

Barycentres des 2 groupes

	t.e	t.p	t.t	m	t	LD1
F	102.31	91.15	72.82	65.88	166.17	33.81
H	113.80	97.23	77.88	75.27	182.55	36.82

Coefficients of linear discriminants:

	LD1
t.epaules	0.12
t.poitrine	-0.022
t.taille	0.11
Masse	-0.11
Taille	0.14



Sur ces données, la discrimination H / F se fait essentiellement selon les variables Taille et Masse.

AFD : principe

- L'AFD est équivalente à une ACP sur les barycentres des groupes définis par les modalités de la variable qualitative de l'étude
- On recherche ainsi un espace de petite dimension dans lequel les barycentres sont le plus écartés possibles (affichant une variabilité maximale)
- Dans le cas $k=2$, le sous-espace de représentation est nécessairement de dimension 1 (une droite)

AFD décisionnelle

- Pour un individu supplémentaire, connaissant les variables quantitatives, le problème « décisionnel » consiste à l'affecter à une des classes définies par la variable qualitative
- Règle simple : affecter le nouveau point à la classe dont le barycentre est le plus proche (il existe d'autres règles plus sophistiquées...)
- Application : credit scoring, diagnostic, contrôle qualité...

Analyse Factorielle des Correspondances

- Deux variables qualitatives
- Information recueillie sur N individus
- Présentation dans un tableau de contingence
- Exemple : données caith (aide de la fonction `corresp()` du package MASS de R)

		Hair colour				
		fair	red	medium	dark	black
Eye colour	blue	326	38	241	110	3
	light	688	116	584	188	4
	medium	343	84	909	412	26
	dark	98	48	403	681	85

5387 habitants dans la région de Caithness (Écosse), répartis selon la couleur de leurs yeux et de leurs cheveux.

AFC : préliminaire

- Test du χ^2 : Couleurs des yeux et des cheveux sont-elles indépendantes ?
- En cas de test non significatif, l'AFC n'a pas vraiment de sens
- Exemple caith

Valeurs observées

	fair	red	medium	dark	black
blue	326	38	241	110	3
light	688	116	584	188	4
medium	343	84	909	412	26
dark	98	48	403	681	85

Valeurs théoriques sous H0

	fair	red	medium	dark	black
blue	194	38	285	185	16
light	427	84	627	408	35
medium	479	94	704	458	39
dark	355	70	522	340	29

p-value < 2.2e-16

On rejette H0, on peut supposer que les couleurs des yeux et des cheveux ne sont pas indépendants

L'AFC va permettre d'étudier plus finement les relations entre les modalités des 2 variables qualitatives

AFC : profils lignes et colonnes

	fair	red	medium	dark	black	somme
Blue	326	38	241	110	3	718
Light	688	116	584	188	4	1580
Medium	343	84	909	412	26	1774
Dark	98	48	403	681	85	1315
Somme	1455	286	2137	1391	118	5387

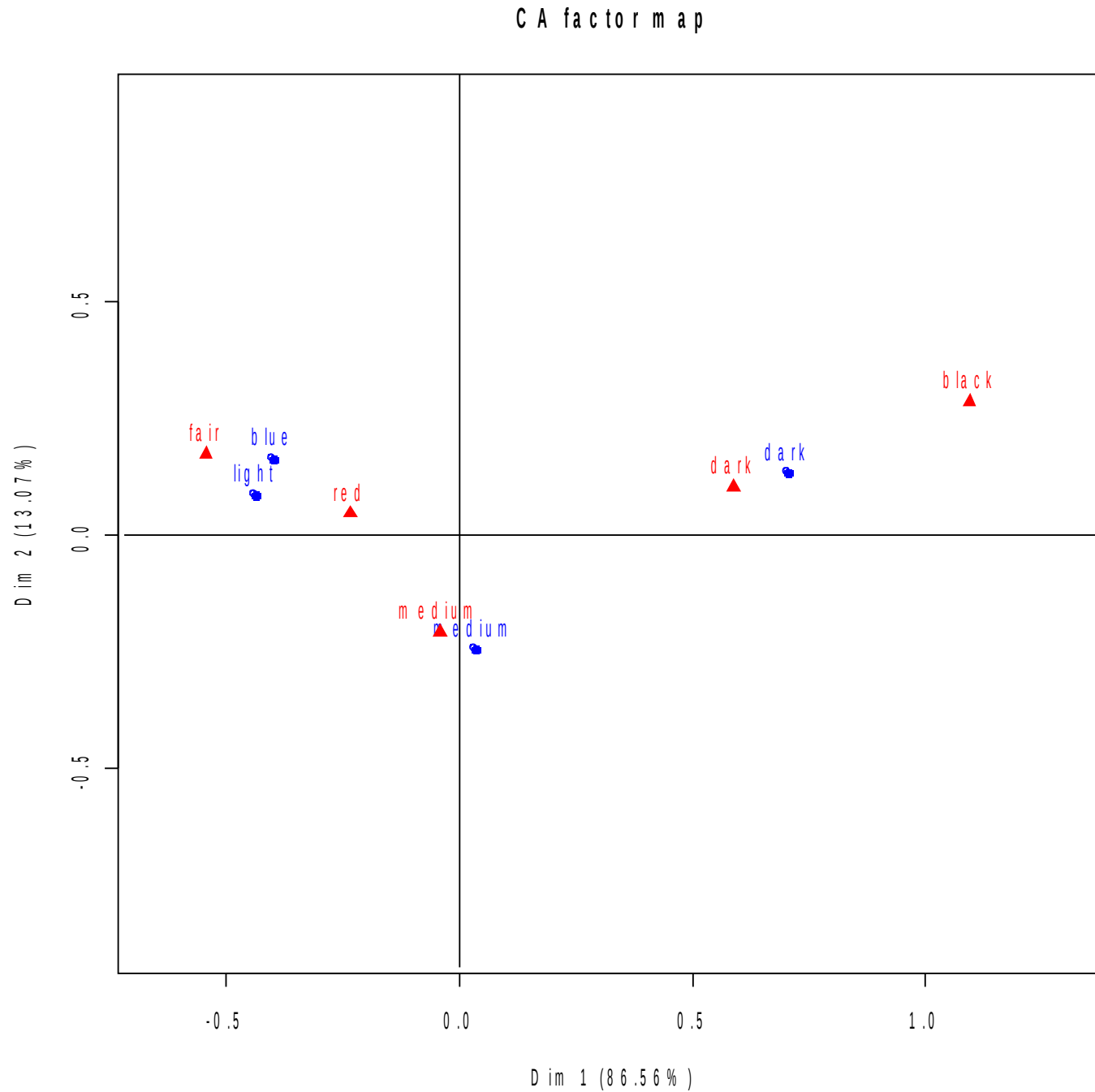
Profils lignes

	fair	red	medium	dark	black	somme
Blue	0.454	0.053	0.336	0.153	0.004	1
light	0.435	0.073	0.370	0.119	0.003	1
medium	0.193	0.047	0.512	0.232	0.015	1
dark	0.075	0.037	0.306	0.518	0.065	1

Profils colonnes

	fair	red	medium	dark	black
blue	0.224	0.133	0.113	0.079	0.025
light	0.473	0.406	0.273	0.135	0.034
medium	0.236	0.294	0.425	0.296	0.220
dark	0.067	0.168	0.189	0.490	0.720
somme	1	1	1	1	1

AFC : représentation graphique



Analyse des Correspondances Multiples

- Plusieurs (>2) variables qualitatives
- Méthode adaptée au traitement d'enquêtes
- Attention à la présentation des données
- ACM = AFC sur un tableau de Burt ou sur un tableau disjonctif complet
- Interprétation des résultats parfois (souvent) délicate

ACM : exemple (1)

45 observations (films), 3 variables qualitatives (distributeur – 6 modalités, note de la critique – 3 modalités, type – 4 modalités)

STAR WARS: Episode 1	20th. Century Fox	>=7	Action
Titanic	20th. Century Fox	>=7	Autres
There's Something About Mary	20th. Century Fox	>=7	Comédie
Independence Day	20th. Century Fox	<6	Action
True Lies	20th. Century Fox	[6;7[Action
Madamme Doubtfire	20th. Century Fox	[6;7[Comédie
Home Alone II	20th. Century Fox	<6	Comédie
Home Alone	20th. Century Fox	[6;7[Comédie
The Sixth Sense	Buena Vista	>=7	Autres
Toy Story 2	Buena Vista	>=7	Familial
Austin Powers	Buena Vista	[6;7[Comédie
Air Force One	Buena Vista	[6;7[Action
Armageddon	Buena Vista	<6	Action
Formigez	Buena Vista	>=7	Familial
The Waterboy	Buena Vista	<6	Comédie
101 Dalmates	Buena Vista	[6;7[Familial
Toy Story I	Buena Vista	>=7	Familial
Pocahontas	Buena Vista	<6	Familial
The Santa Clause	Buena Vista	<6	Action
Aladin	Buena Vista	>=7	Familial
The Beast and the beauty	Buena Vista	>=7	Familial
Pretty Woman	Buena Vista	[6;7[Comédie
Mission Impossible	Paramount Pictures	[6;7[Action
The Lion King	Paramount Pictures	>=7	Familial
The Firm	Paramount Pictures	[6;7[Autres
Ghost	Paramount Pictures	[6;7[Comédie
Teenager mutant ninja turtles	Paramount Pictures	<6	Action
Sleepless in Seattle	Tristar Films	[6;7[Comédie
A few good man	Tristar Films	>=7	Autres
Terminator II	Tristar Films	>=7	Action
City Slickers	Tristar Films	[6;7[Comédie
FALTA	Universal Pictures	<6	Action
Jurassic Park II	Universal Pictures	[6;7[Comédie
Apollo XIII	Universal Pictures	>=7	Autres
The Flintstones	Universal Pictures	<6	Comédie
Jurassic Park	Universal Pictures	[6;7[Action
The Matrix	Warner Bross	>=7	Action
Twister	Warner Bross	[6;7[Action
Batman Forever	Warner Bross	<6	Action
ace ventura	Warner Bross	<6	Comédie
Forrest gump	Warner Bross	>=7	Autres
The fugitive	Warner Bross	>=7	Action
Batman Returns	Warner Bross	[6;7[Action

Tableau de Burt

Tableau 13x13 (nombre total de modalités)

	DISTRIBU.20th	DISTRIBU.BV	DISTRIBU.PP	DISTRIBU.TF	DISTRIBU.UP	DISTRIBU.WB	CRI_C	CRI_B	CRI_A	TYPE.Act	TYPE.Aut	TYPE.Com	TYPE.Fam
DISTRIBU.20th	8	0	0	0	0	0	3	2	3	3	1	4	0
DISTRIBU.BV	0	14	0	0	0	0	4	4	6	3	1	3	7
DISTRIBU.PP	0	0	5	0	0	0	3	1	1	2	1	1	1
DISTRIBU.TF	0	0	0	4	0	0	2	0	2	1	1	2	0
DISTRIBU.UP	0	0	0	0	5	0	2	2	1	2	1	2	0
DISTRIBU.WB	0	0	0	0	0	9	4	2	3	7	1	1	0
CRI_C	3	4	3	2	2	4	18	0	0	8	1	8	1
CRI_B	2	4	1	0	2	2	0	11	0	6	0	4	1
CRI_A	3	6	1	2	1	3	0	0	16	4	5	1	6
TYPE.Act	3	3	2	1	2	7	8	6	4	18	0	0	0
TYPE.Aut	1	1	1	1	1	1	1	0	5	0	6	0	0
TYPE.Com	4	3	1	2	2	1	8	4	1	0	0	13	0
TYPE.Fam	0	7	1	0	0	0	1	1	6	0	0	0	8

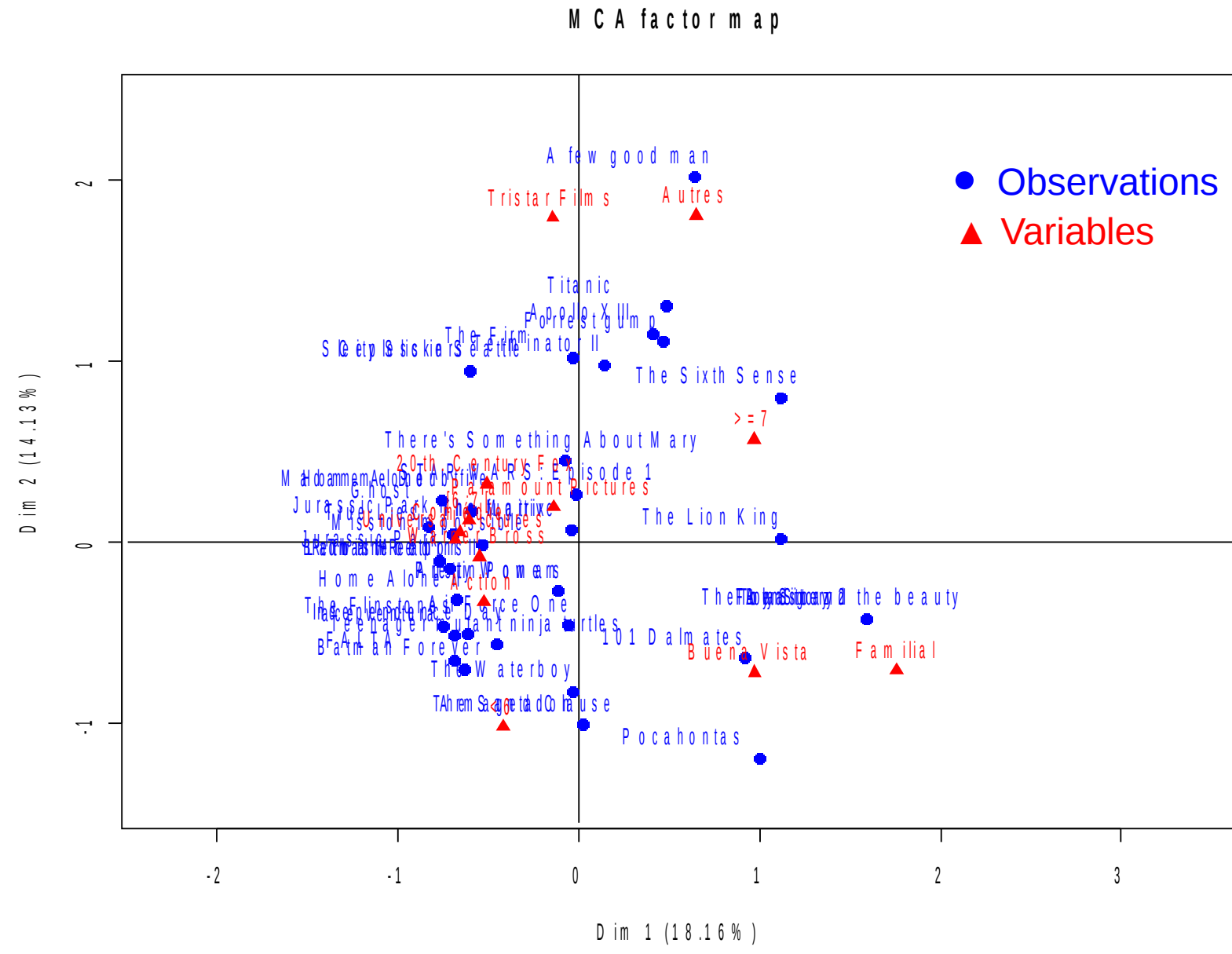
Tableau disjonctif complet

Tableau 45x13

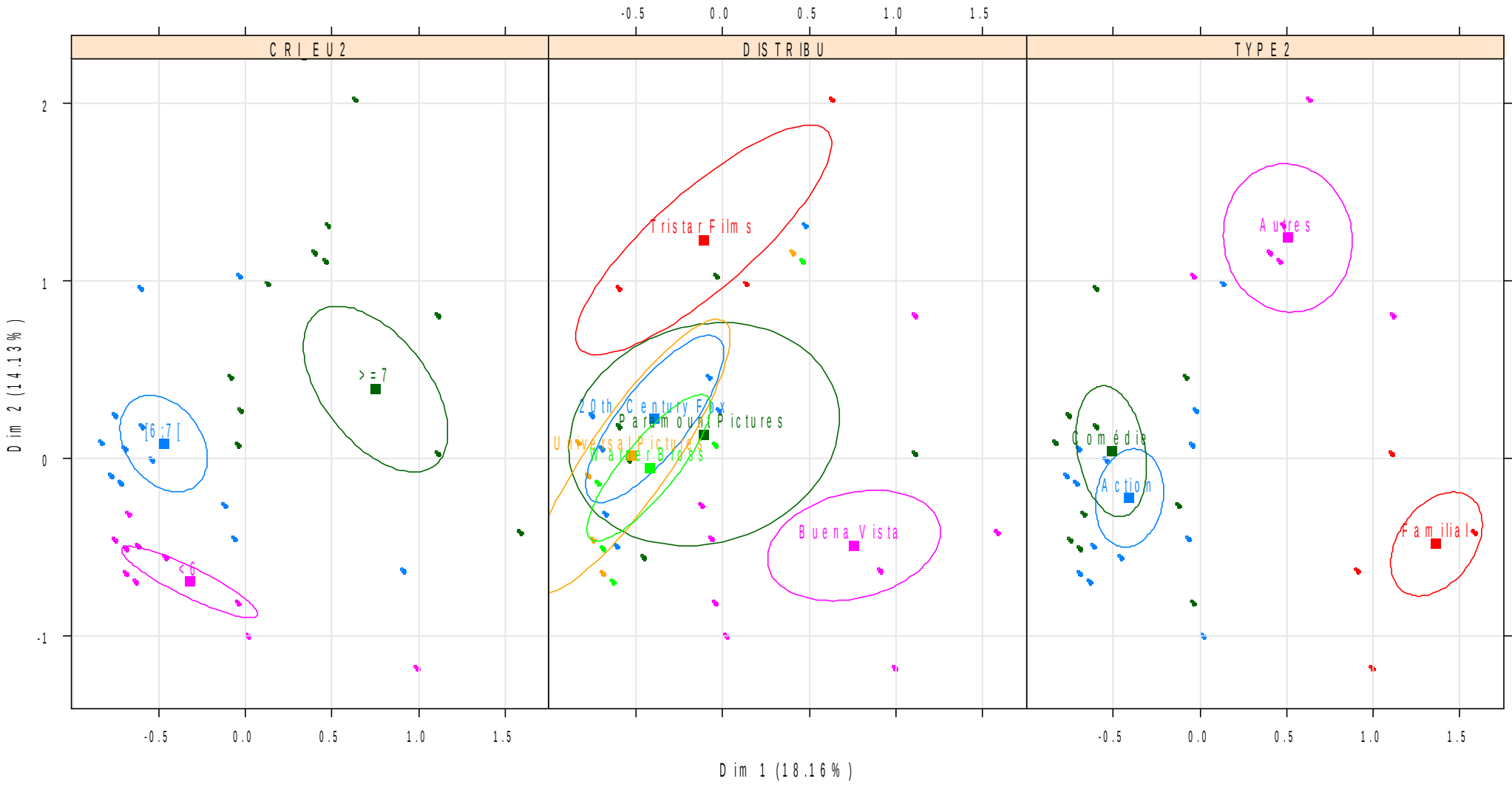
(nombre d'observations x nombre total de modalités)

	DISTRIBU.20th	DISTRIBU.BV	DISTRIBU.PP	DISTRIBU.TF	DISTRIBU.UP	DISTRIBU.WB	CRI_C	CRI_B	CRI_A	TYPE.Act	TYPE.Aut	TYPE.Com	TYPE.Fam
STAR WARS: Episode 1	1	0	0	0	0	0	0	0	1	1	0	0	0
Titanic	1	0	0	0	0	0	0	0	1	0	1	0	0
There's Something About Mary	1	0	0	0	0	0	0	0	1	0	0	1	0
Independence Day	1	0	0	0	0	0	0	1	0	1	0	0	0
True Lies	1	0	0	0	0	0	1	0	0	1	0	0	0
Madamme Doubtfire	1	0	0	0	0	0	1	0	0	0	0	1	0
Home Alone II	1	0	0	0	0	0	0	1	0	0	0	1	0
Home Alone	1	0	0	0	0	0	1	0	0	0	0	1	0
The Sixth Sense	0	1	0	0	0	0	0	0	1	0	1	0	0
Toy Story 2	0	1	0	0	0	0	0	0	1	0	0	0	1
Austin Powers	0	1	0	0	0	0	1	0	0	0	0	1	0
Air Force One	0	1	0	0	0	0	1	0	0	1	0	0	0
Armageddon	0	1	0	0	0	0	0	1	0	1	0	0	0
Formigez	0	1	0	0	0	0	0	0	1	0	0	0	1
The Waterboy	0	1	0	0	0	0	0	1	0	0	0	1	0
101 Dalmates	0	1	0	0	0	0	1	0	0	0	0	0	1
Toy Story I	0	1	0	0	0	0	0	0	1	0	0	0	1
Pocahontas	0	1	0	0	0	0	0	1	0	0	0	0	1
The Santa Clause	0	1	0	0	0	0	0	1	0	1	0	0	0
Aladin	0	1	0	0	0	0	0	0	1	0	0	0	1
The Beast and the beauty	0	1	0	0	0	0	0	0	1	0	0	0	1
Pretty Woman	0	1	0	0	0	0	1	0	0	0	0	1	0
Mission Impossible	0	0	1	0	0	0	1	0	0	1	0	0	0
The Lion King	0	0	1	0	0	0	0	0	1	0	0	0	1
The Firm	0	0	1	0	0	0	1	0	0	0	1	0	0
Ghost	0	0	1	0	0	0	1	0	0	0	0	1	0
Teenager mutant ninja turtles	0	0	1	0	0	0	0	1	0	1	0	0	0
Sleepless in Seattle	0	0	0	1	0	0	1	0	0	0	0	1	0
A few good man	0	0	0	1	0	0	0	0	1	0	1	0	0
Terminator II	0	0	0	1	0	0	0	0	1	1	0	0	0
City Slickers	0	0	0	1	0	0	1	0	0	0	0	1	0
FALTA	0	0	0	0	1	0	0	1	0	1	0	0	0
Jurassic Park II	0	0	0	0	1	0	1	0	0	0	0	1	0
Apollo XIII	0	0	0	0	1	0	0	0	1	0	1	0	0
The Flintstones	0	0	0	0	1	0	0	1	0	0	0	1	0
Jurassic Park	0	0	0	0	1	0	1	0	0	1	0	0	0
The Matrix	0	0	0	0	0	1	0	0	1	1	0	0	0
Twister	0	0	0	0	0	1	1	0	0	1	0	0	0
Batman Forever	0	0	0	0	0	1	0	1	0	1	0	0	0
ace ventura	0	0	0	0	0	1	0	1	0	0	0	1	0
Forrest gump	0	0	0	0	0	1	0	0	1	0	1	0	0
The fugitive	0	0	0	0	0	1	1	0	1	1	0	0	0
Batman Returns	0	0	0	0	0	1	1	0	0	1	0	0	0

ACM : représentation graphique



ACM : représentation graphique



ACM : exemple (2)



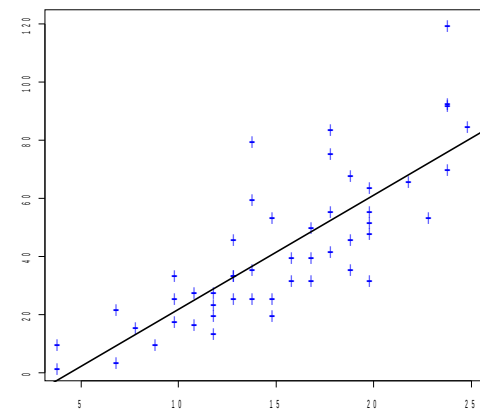
- Données qualitatives recueillies sur 1453 crânes de grands singes.
- Absence ou présence de 41 caractères discrets
- Source : J. Braga (1995), Définition de certains caractères discrets crâniens chez *Pongo*, *Gorilla* et *Pan*. Perspectives taxonomiques et phylogénétiques. *Thèse de doctorat, Université de Bordeaux I*.
- Extrait des données

Singe	TAX	CNHY-D	CNHY-G	PRPA-D	PRPA-G	SPAP	PRST-D	PRST-G	PTPE-D	PTPE-G	PTPA-D
1	PTT	1	1	0	0	0	0	0	0	1	1
1	PTT	1	1	NA	NA	0	0	0	0	0	0
1	PTT	1	1	0	0	0	0	0	0	0	1
1	PTT	0	1	0	0	0	0	0	0	1	1
1	PTT	0	1	0	0	0	0	0	1	1	0
1	PTT	0	1	0	0	0	0	0	1	1	1
1	PTT	1	1	0	0	0	0	0	0	1	1
...
2	PPA	0	1	1	1	0	1	1	1	1	1
2	PPA	1	1	0	0	0	1	1	0	0	0
2	PPA	1	1	0	0	0	0	1	1	1	0
2	PPA	1	0	1	0	0	1	1	0	0	0
2	PPA	1	1	0	0	0	0	0	0	0	1

Régression linéaire multiple

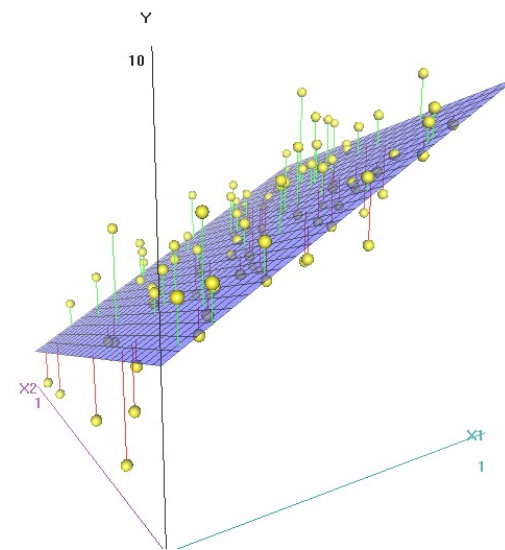
Régression linéaire simple :

- **1** variable à expliquer (Y) par **1** variable explicative (X)
- trouver les paramètres a_0 (ordonnée à l'origine) et a_1 (pente) de la **droite** qui passe « au mieux » dans le nuage de points de Y en fonction de X
- $Y = a X + b$



Régression linéaire « double » :

- **1** variable à expliquer (Y) par **2** variables explicatives (X1 et X2)
- trouver les paramètres a_0 , a_1 et a_2 du **plan** qui passe « au mieux » dans le nuage de points de Y en fonction de X1 et X2
- $Y = a_0 + a_1 X_1 + a_2 X_2$



Régression linéaire multiple :

- **1** variable à expliquer (Y) par **p** variables explicatives (X1, ... Xp)
- trouver les paramètres a_0 , a_1 , ... a_p de l'**hyperplan** qui passe « au mieux » dans le nuage de points de Y en fonction de X1, X2, ... et Xp
- $Y = a_0 + a_1 X_1 + a_2 X_2 + \dots + a_p X_p$

La représentation graphique est « moins évidente » pour $p > 2$

Régression linéaire multiple

- Estimation des paramètres \hat{a}_i (formule matricielle)
- Valeurs ajustées

$$\hat{Y}_i = \hat{a}_0 + \sum_{i=1}^p \hat{a}_i X_i = \hat{a}_0 + \hat{a}_1 X_1 + \hat{a}_2 X_2 + \dots + \hat{a}_p X_p$$

- Qualité du modèle

(part de variance expliquée par le modèle)

$$R^2 = \frac{\text{var}(\hat{Y}_i)}{\text{var}(Y_i)}$$

- Sélection de variables (choix de modèles) : R^2 ajusté, C_p de Mallows, algorithmes *forward*, *backward*, *stepwise*...

Régression linéaire multiple

Exemple : Modèle de régression pour la masse d'individus en fonction du tour d'épaules, du tour de poitrine, du tour de taille et de la taille

Modèle à 4 variables explicatives

Estimation des paramètres

Coefficients:
 (Intercept) t.epaules t.poitrine t.taille taille
 -53.52049 0.34221 -0.03813 0.87249 0.14319

Adéquation du modèle

Residual standard error: 3.669 on 15 degrees of freedom
 Multiple R-squared: 0.876, Adjusted R-squared: 0.8429
 F-statistic: 26.49 on 4 and 15 DF, p-value: 1.203e-06

Residuals:
 Min 1Q Median 3Q Max
 -6.8373 -1.8649 -0.5518 2.0784 5.9578

Test sur les coefficients

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-53.52049	16.22046	-3.300	0.004863	**
t.epaules	0.34221	0.19245	1.778	0.095638	.
t.poitrine	-0.03813	0.29808	-0.128	0.899919	
t.taille	0.87249	0.19486	4.477	0.000443	***
taille	0.14319	0.11476	1.248	0.231251	

Modèle à 2 variables explicatives

Estimation des paramètres

Coefficients:
 (Intercept) t.epaules t.taille
 -43.4172 0.4523 0.8643

Adéquation du modèle

Residual standard error: 3.63 on 17 df
 Multiple R-squared: 0.8625, Adj. R2: 0.8463
 F-statistic: 53.31 on 2 and 17 DF, p-val: 4.745e-06

Residuals:
 Min 1Q Median 3Q Max
 -7.2664 -1.6180 -0.1873 2.0431 5.7792

Test sur les coefficients

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-43.4172	11.6497	-3.727	0.00168	**
t.epaules	0.4523	0.1143	3.958	0.00102	**
t.taille	0.8643	0.1329	6.505	5.39e-06	***

Quelques ouvertures

- Positionnement multidimensionnel : analyse de tableaux de distances
- Analyse canonique, régression PLS : recherche de relations entre 2 ensembles de variables quantitatives
- MANOVA : plusieurs variables quantitatives à expliquer par plusieurs variables qualitatives
- Réseaux de neurones, *Support Vector Machine*, Classification and Regression Tree (CART), forêts aléatoires...