Multivariate projection methodologies for the exploration of large biological data sets

Application in R using mixOmics

Exploration and

Integration of

Omics datasets

Plan

- Introduction
- Rappels (?)
- Exploration d'un jeu de données (ACP)
- Méthodes discriminantes (AFD, PLS-DA)
- Intégration de données (PLS, CCA, GCCA)
- Extensions *sparse*
- Extensions *multilevel*

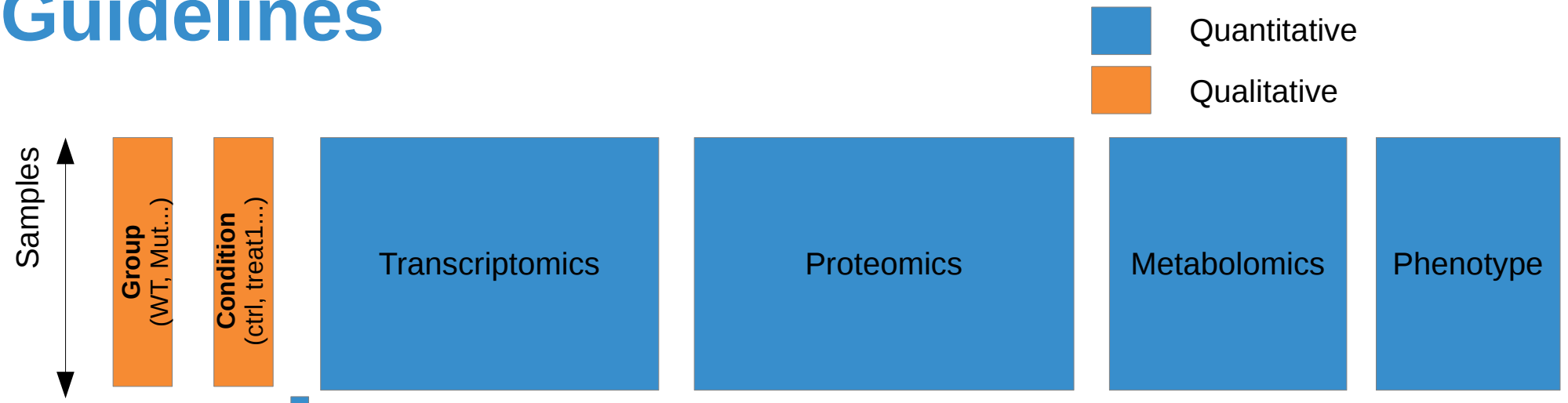
The mixOmics story

- Started with two phD projects in Université de Toulouse:
  - Ignacio González (2004-2007): rCCA
  - Kim-Anh Lê Cao (2005-2008): sPLS
- The Australian mixOmics immigration processed began in 2008 ...
  - K-A moved to UQ for a postdoc (IMB)
  - Core team established: Kim-Anh Lê Cao (FR, AUS), Ignacio González (FR), Sébastien Déjean (FR)
- First R CRAN release in May 2009
- Today
  - 4,000 downloads in 2014, 10,000 in 2015 (R CRAN unique IP adress)
  - Website: `www.mixomics.org`
  - Two web-interfaces (shiny and PHP, also Galaxy but not advertised)
  - 8 multivariate methodologies and sparse variants
  - Team: 3 core members and 4 key contributors
- 13 published articles from the team since 2008

Guidelines

- I want to explore one single data set (e.g. microarray data):
  - I would like to identify the trends or patterns in your data, experimental bias or, identify if your samples ‘naturally’ cluster according to the biological conditions: Principal Component Analysis (PCA)
- I want to want to unravel the information contained in two data sets, where two types of variables are measured on the same samples (e.g. metabolomics and transcriptomics data)
  - I would like to know if I can extract common information from the two data sets (or highlight thecorrelation between the two data sets). The total number of variables is less than the number of samples: Canonical Correlation Analysis (CCA) or Partial Least Squares (PLS) canonical mode. The total number of variables is greater than the number of samples: regularized Canonical Correlation Analysis (rCCA) or Partial Least Squares (PLS) canonical mode
- I have one single data set (e.g. microarray data) and I am interested in classifying my samples into known classes:
  - Here X = expression data and Y = vector indicating the classes of the samples. I would like to know how informative my data are to rightly classify my samples, as well as predicting the class of new samples: PLS-Discriminant Analysis (PLS-DA)
- I have one single data set (e.g. microarray data) and I have one continuous response variable or outcome for each sample. I would like to predict the response with my data:
  - Here X = expression data and Y = response vector. I would like to model a causal relationship between my data and the response vector and assess how informative my data are to predict such response: PLS-regression mode

Guidelines



- Univariate
- Bivariate: 2 quantitatives or 1 quantitative + 1 qualitative
  Correlation, statistical test
- Multivariate unsupervised
  PCA
- Multivariate supervised
  PLS-DA
- Multiblock unsupervised
  PLS (2 blocks), GCCA
- Multiblock supervised
  GCC-DA

Variance et écart-type

$$\mathrm{var}(X) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

Moyenne des carrés des écarts à la moyenne

$$\sigma(X) = \sqrt{\mathrm{var}(X)}$$

Racine carrée de la variance

Quelques propriétés de l'écart-type :

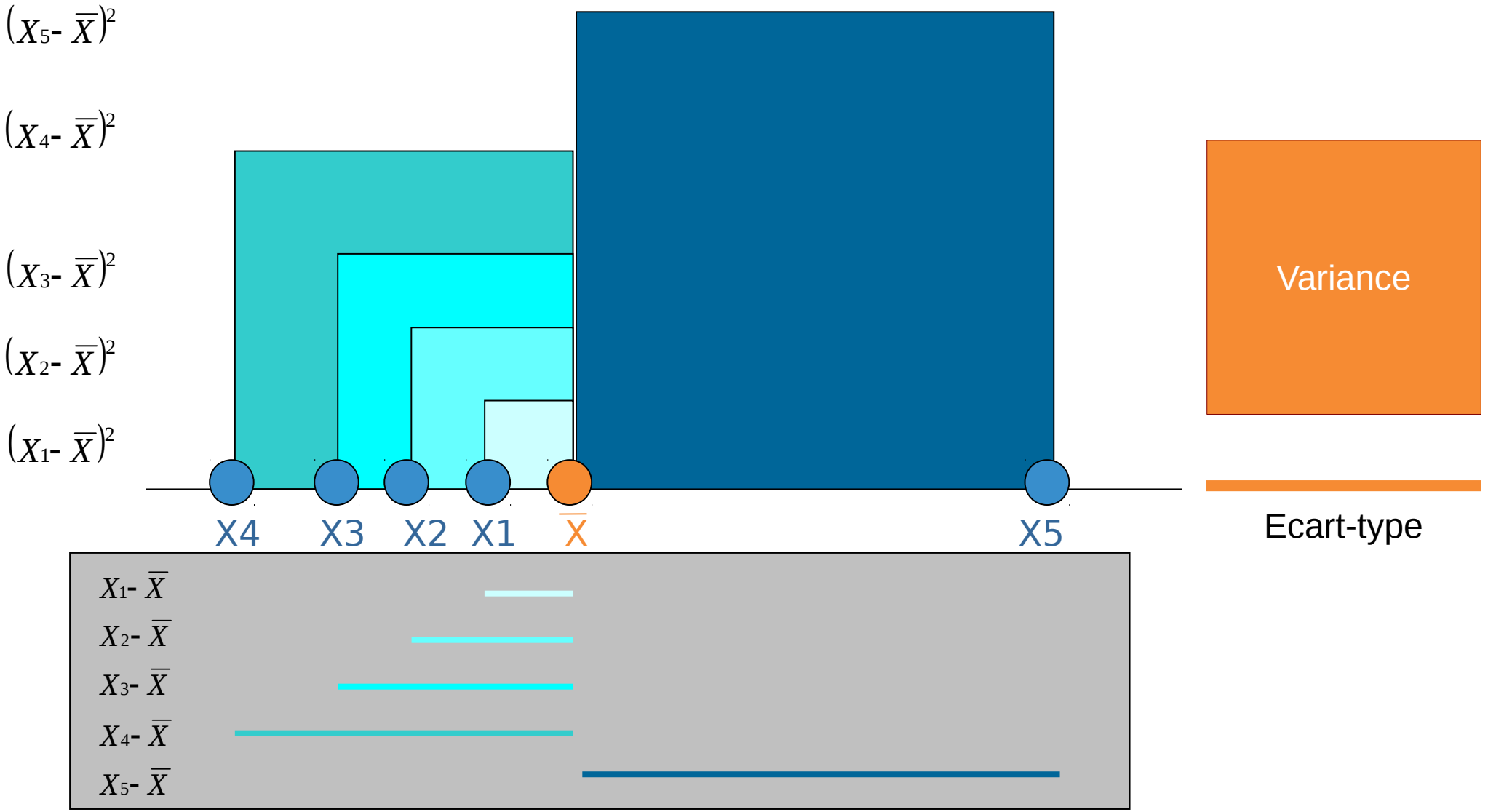
- Positif (nul si la série est constante)
- Invariant par translation
- Sensible aux valeurs extrêmes
- **De la même unité que la donnée** (et que la moyenne) :

*Si l'échantillon est constitué de mesures en **m** alors l'écart-type s'exprime également en **m** (tout comme la moyenne) ; ce qui n'est pas le cas de la variance **m²** !*

On peut ainsi additionner moyenne et écart-type (*mais pas moyenne et variance*), ce qui est fondamental pour la construction d'intervalle de confiance.

Variance et écart-type

Racine carrée de la moyenne des carrés des écarts à la moyenne



Covariance

Covariance $\mathrm{cov}(X,Y)=\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})$

cov(X,X)=var(X)

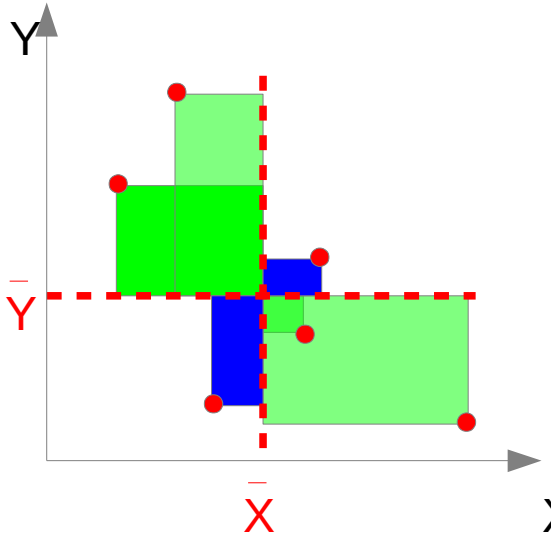
Signe du produit $(Xi-\bar{X})(Yi-\bar{Y})$

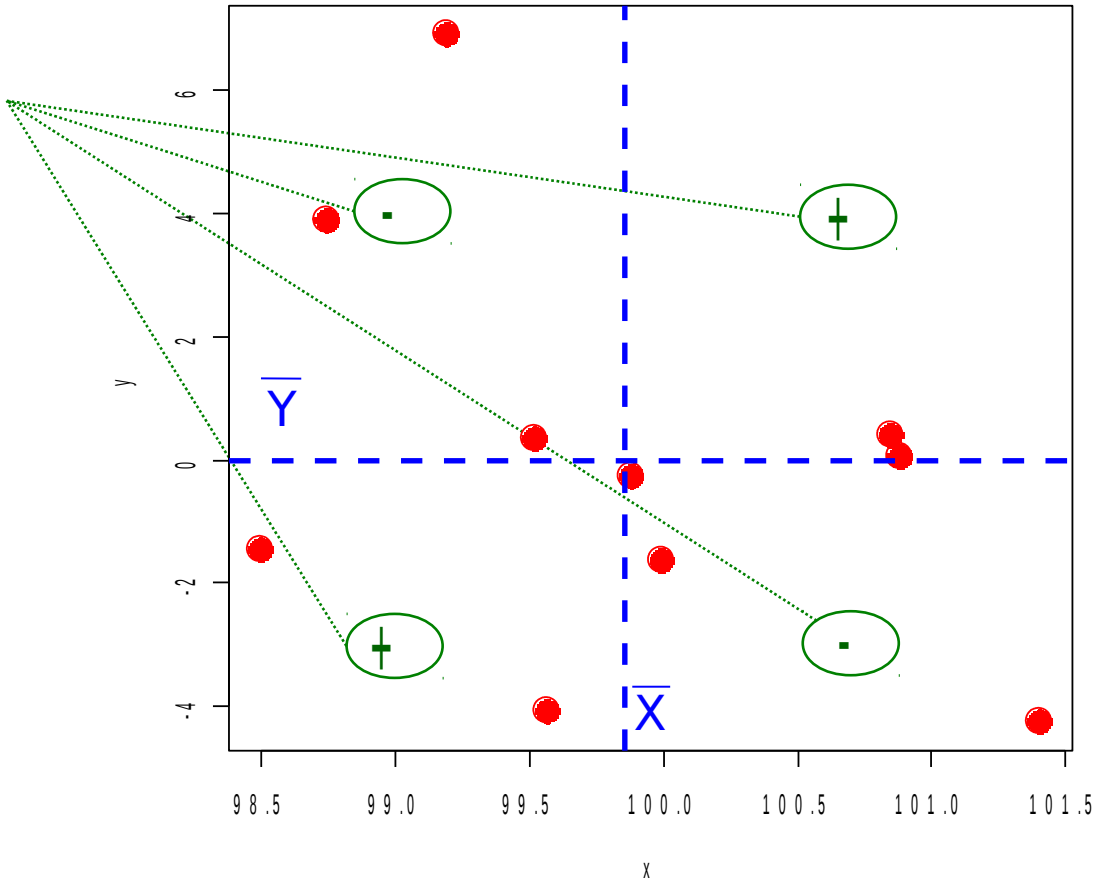
Intuitivement :

- Si les **+** l'emportent
→ liaison linéaire positive

- Si les **–** l'emportent
→ liaison linéaire négative

*Sur cet exemple : cov(X,Y)=**-1.36***





La covariance dépend des unités de mesure ➔ coefficient de corrélation

Corrélation

Quelques propriétés des coefficients de corrélation :

- Coefficient de corrélation de Pearson : **relation linéaire**
- Coefficient de corrélation de Spearman : considère les rangs, **relation monotone**
- Compris entre –1 et 1.
- Les valeurs extrêmes –1 et 1 indique des corrélations parfaites entre les 2 variables.
- Si le coefficient est positif : quand une variable est élevée, l'autre l'est également. Quand une variable est faible, l'autre l'est également.
- Si le coefficient est négatif : quand une variable est élevée (resp. faible), l'autre est faible (resp. élevée).

Combinaison linéaire

2 vecteurs

2 coefficients : c1 = 0.5 ; c2 = 2 $\quad \mathbf{W} = \begin{pmatrix} 0.5 \\ 2 \end{pmatrix}$

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

X

Combinaison linéaire des vecteurs Taille et Masse avec les coefficients c1 et c2

$$CL = 0.5 \begin{matrix} 174.0 \\ 175.3 \\ 193.5 \\ 186.5 \\ 187.2 \\ 181.5 \\ 184.0 \\ 184.5 \\ 175.0 \\ 184.0 \end{matrix} + 2 \begin{matrix} 65.6 \\ 71.8 \\ 80.7 \\ 72.6 \\ 78.8 \\ 74.8 \\ 86.4 \\ 78.4 \\ 62.0 \\ 81.6 \end{matrix} = \begin{matrix} 218.20 \\ 231.25 \\ 258.15 \\ 238.45 \\ 251.20 \\ 240.35 \\ 264.80 \\ 249.05 \\ 211.50 \\ 255.20 \end{matrix}$$

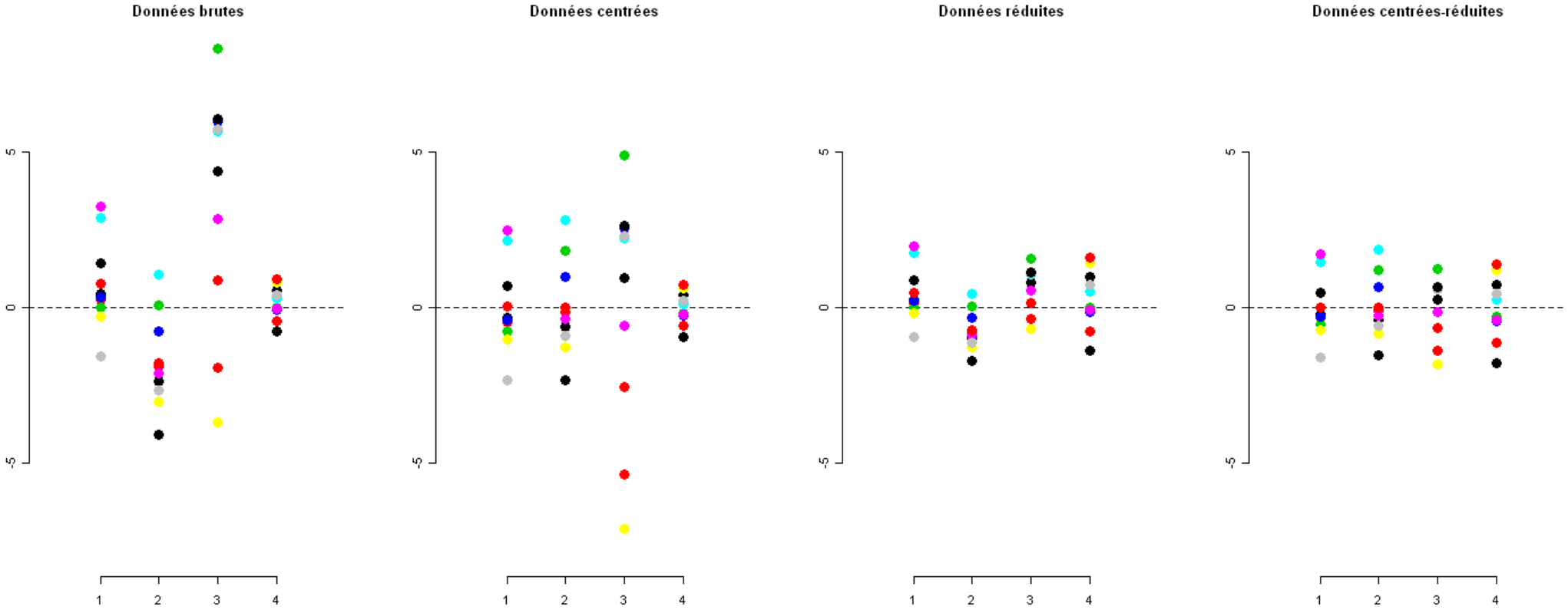
Notation matricielle : **CL** = **XW**

Exemple : une composante principale est une combinaison linéaire des variables initiales.

Centrage-réduction (*scale*)

$$Z_i = \frac{X_i - \bar{X}}{\sigma_X}$$

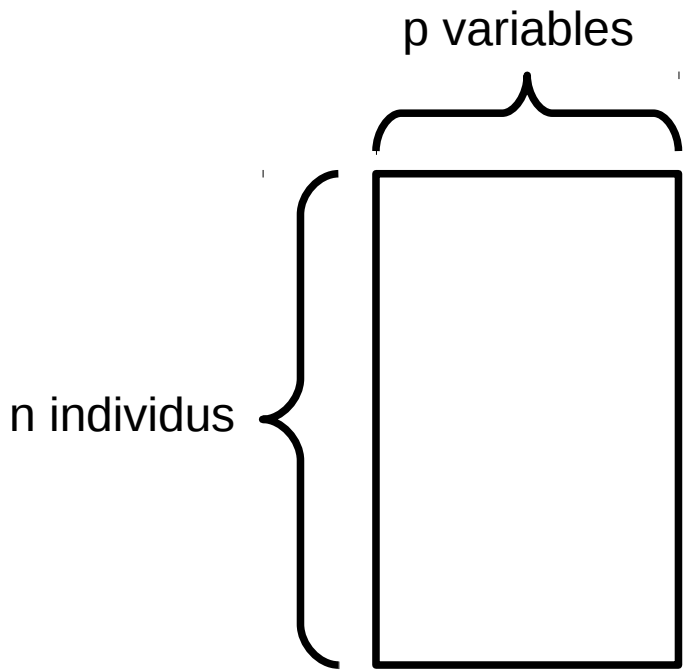
- Centrer : retrancher la moyenne
- Réduire(*) : diviser par l'écart-type (*) terminologie trompeuse : si l'écart-type est <1, la réduction dilate les données
- Permet d'exprimer des variables différentes sur une échelle commune, en les débarrassant de leurs unités physiques : les observations s'expriment en nombre d'écart-type par rapport à la moyenne.
- Après centrage-réduction, la moyenne des observations est nulle et l'écart-type vaut 1 (ainsi que la variance).
- Appelé parfois « z-transformation » ou « z-score »



Exploration d'un jeu de données

Analyse en Composantes Principales

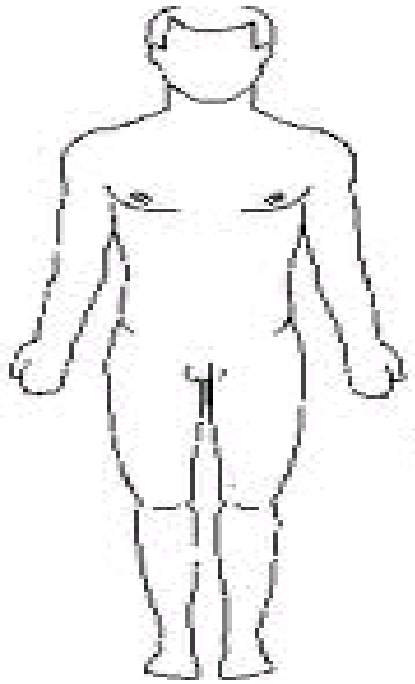
Objectif : décrire sans a priori un tableau de données constitué exclusivement de variables quantitatives.



Un jeu de données

- 20 individus
- 5 variables :

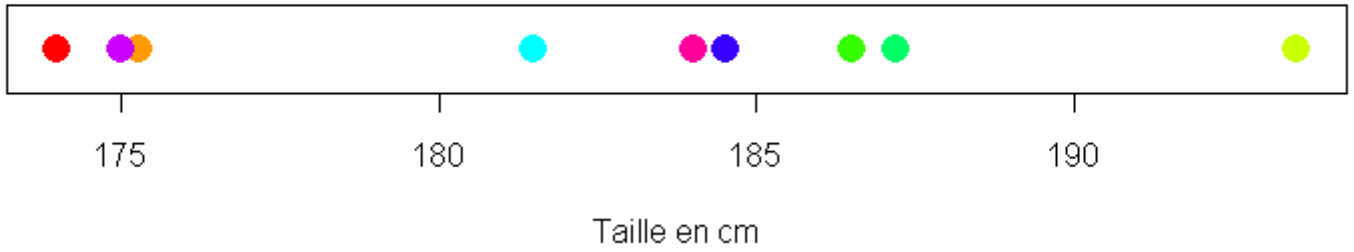
V1 : tour au niveau des épaules (cm)
V2 : tour de poitrine (cm)
V3 : tour de taille (cm)
V4 : masse (kg)
V5 : taille (cm)



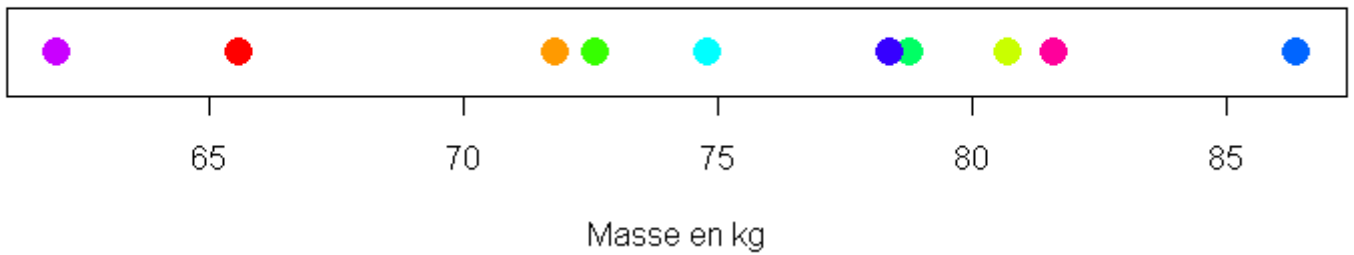
	V1	V2	V3	V4	V5
H 1	106.2	89.5	71.5	65.6	174.0
H 2	110.5	97.0	79.0	71.8	175.3
H 3	115.1	97.5	83.2	80.7	193.5
H 4	104.5	97.0	77.8	72.6	186.5
H 5	107.5	97.5	80.0	78.8	187.2
H 6	119.8	99.9	82.5	74.8	181.5
H 7	123.5	106.9	82.0	86.4	184.0
H 8	120.4	102.5	76.8	78.4	184.5
H 9	111.0	91.0	68.5	62.0	175.0
H 10	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

Représentation graphique 1D

Taille : 174.0 175.3 193.5 186.5 187.2 181.5 184.0 184.5 175.0 184.0

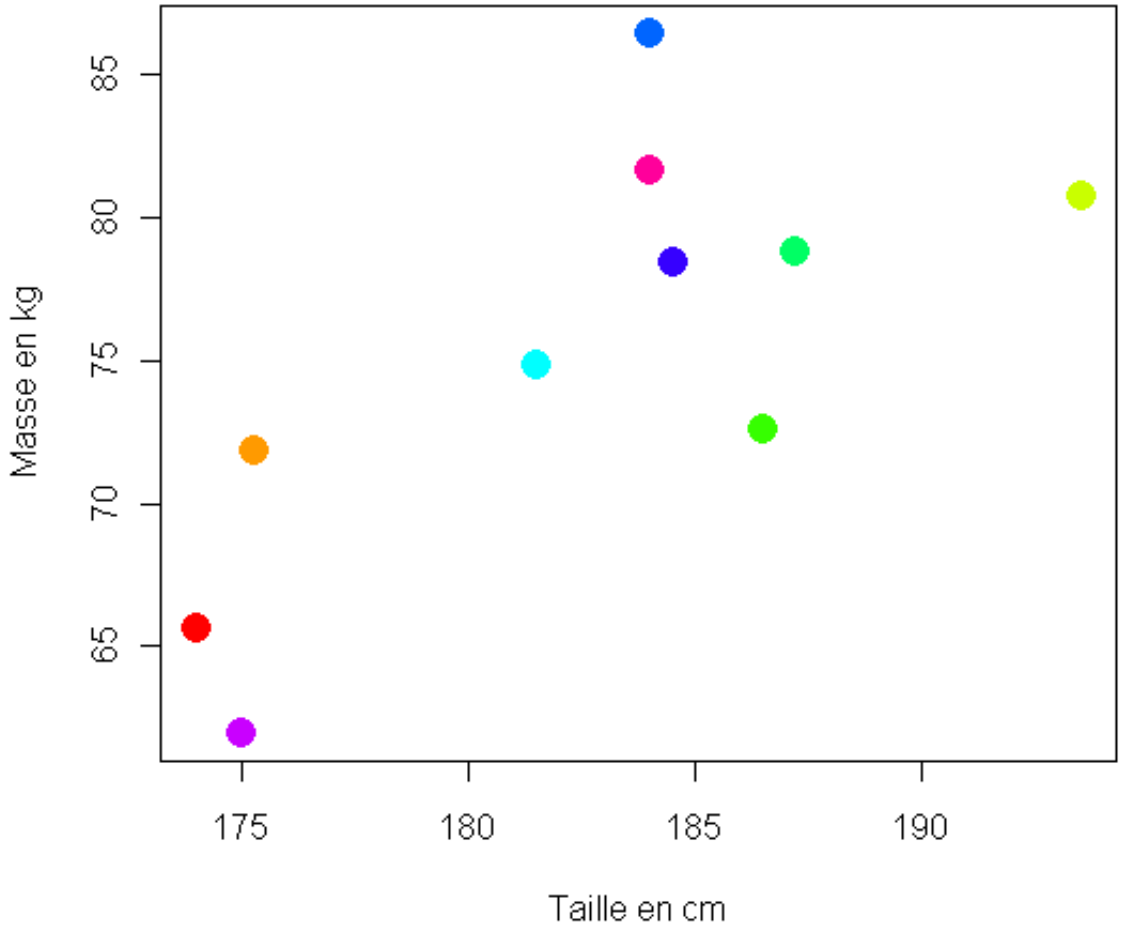


Masse : 65.6 71.8 80.7 72.6 78.8 74.8 86.4 78.4 62.0 81.6



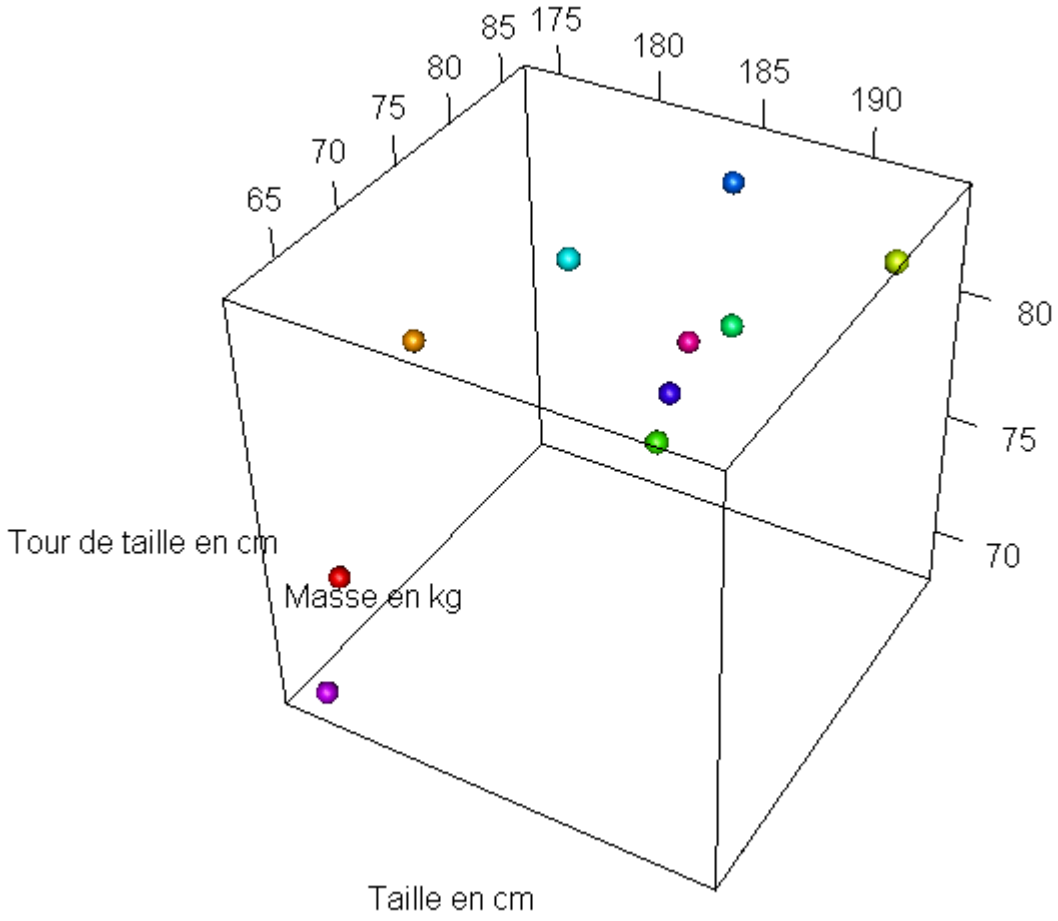
Représentation graphique 2D

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

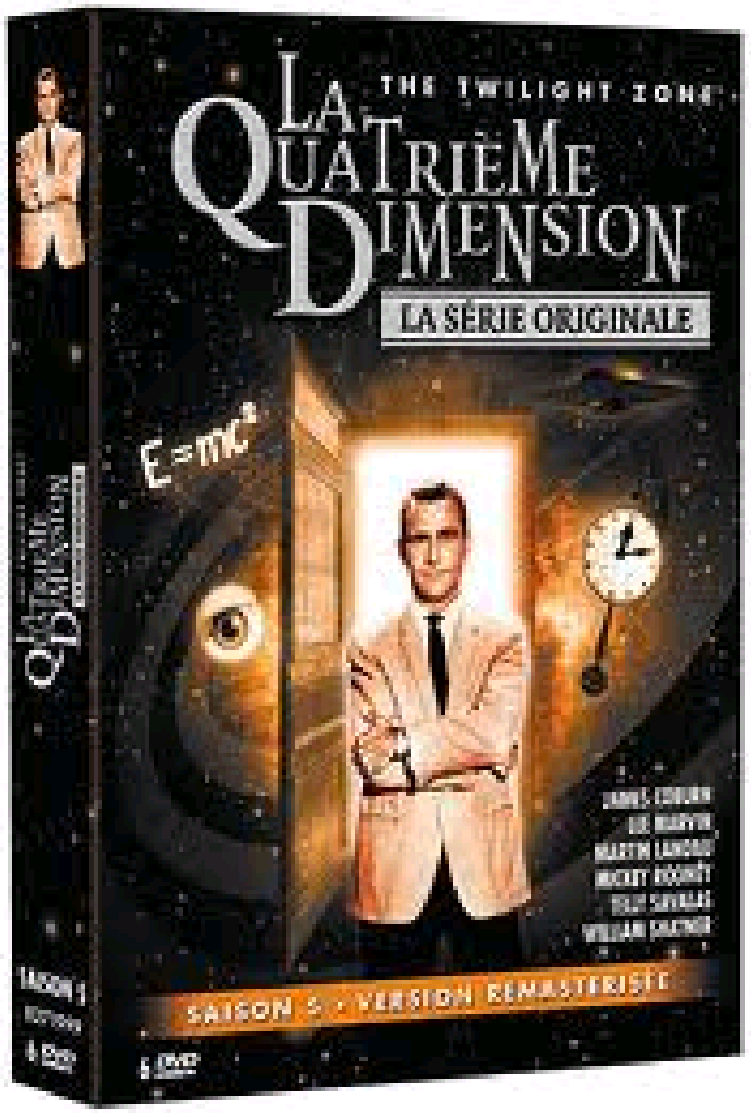


Représentation graphique 3D

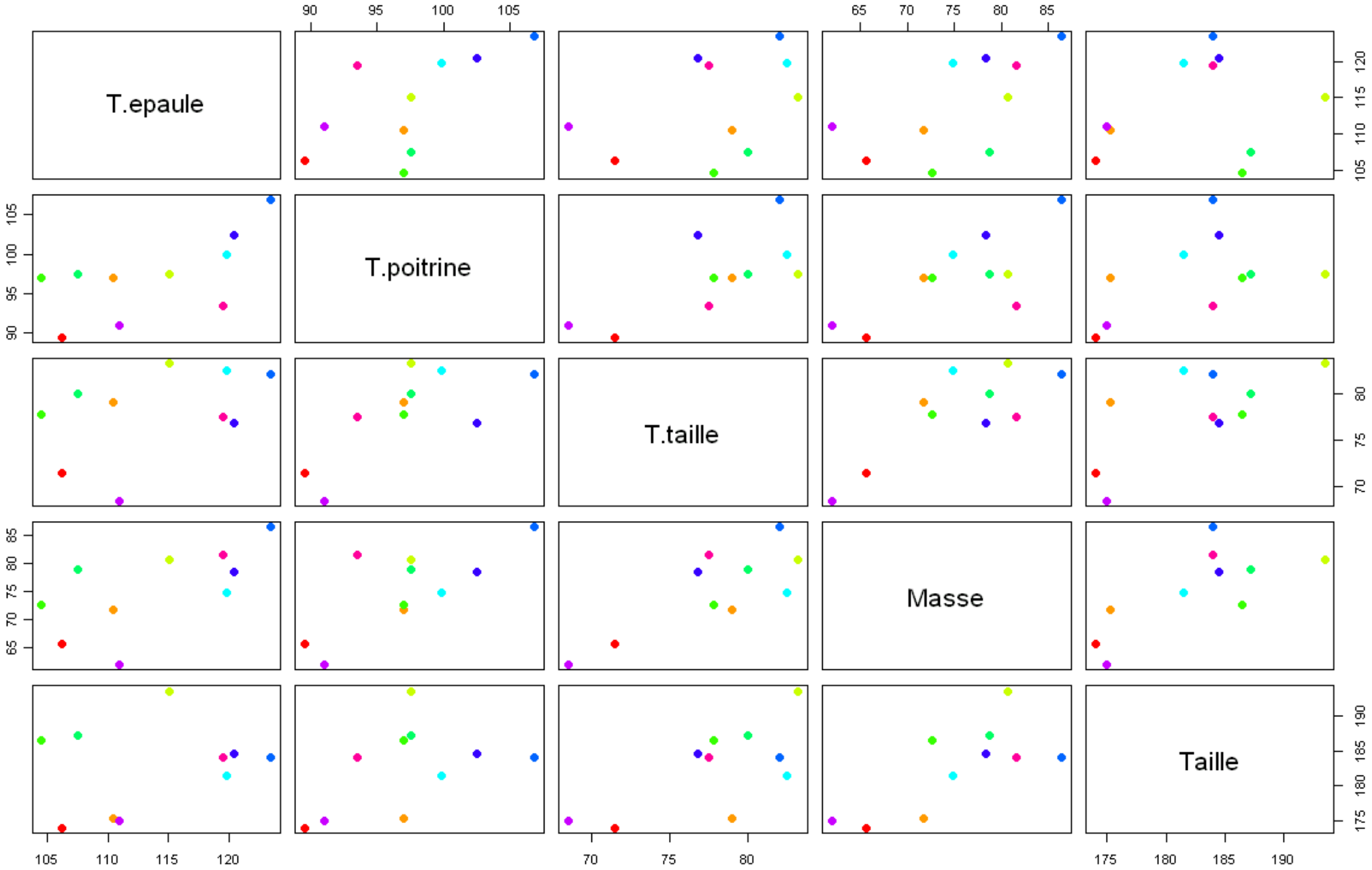
Taille :	174.0	175.3	193.5	186.5	187.2	181.5	184.0	184.5	175.0	184.0
Masse :	65.6	71.8	80.7	72.6	78.8	74.8	86.4	78.4	62.0	81.6
T. Taille :	71.5	79.0	83.2	77.8	80.0	82.5	82.0	76.8	68.5	77.5

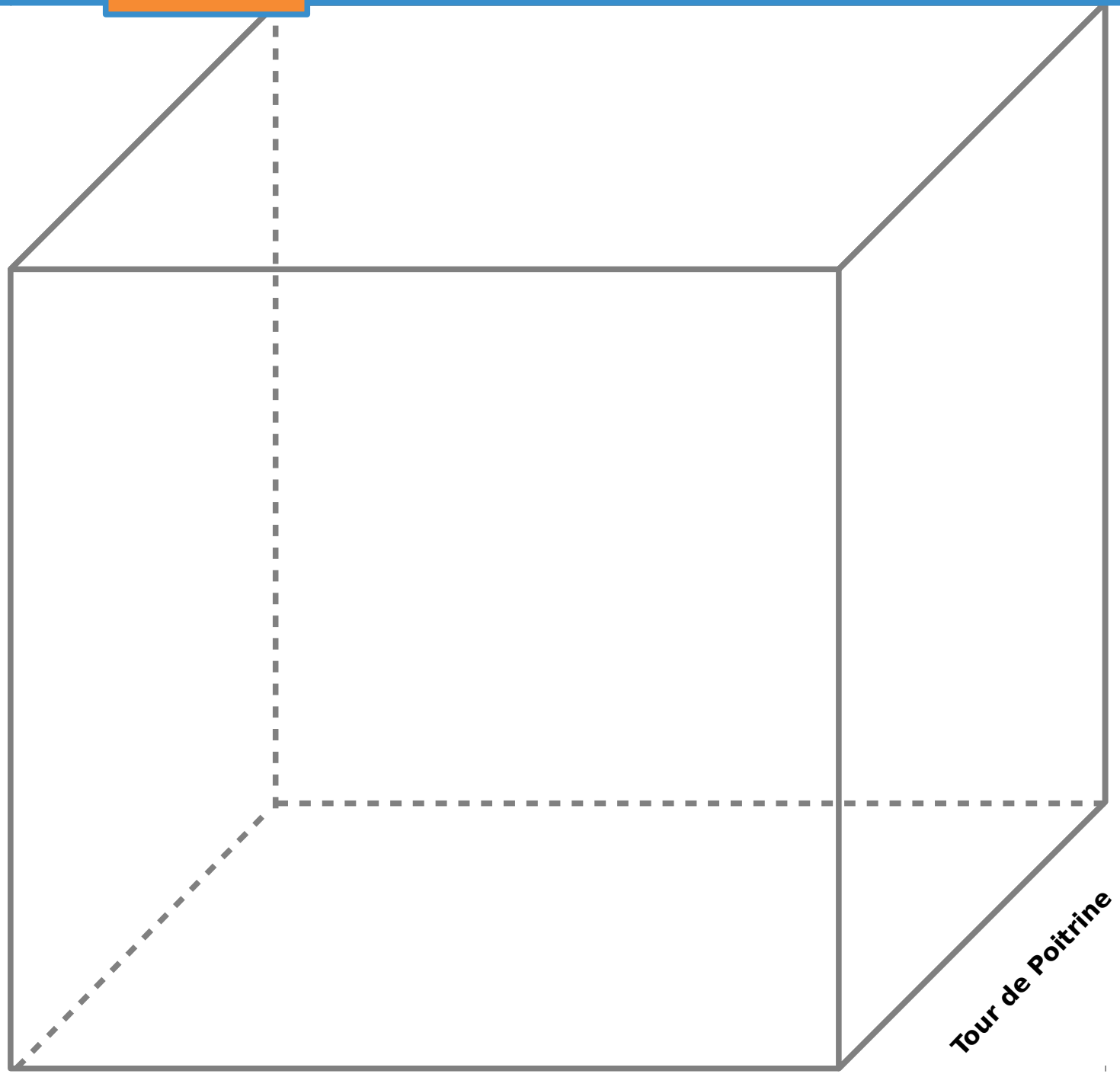


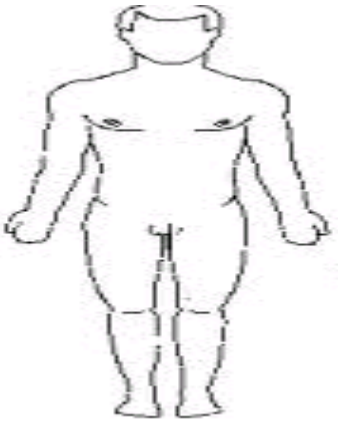
4D ?

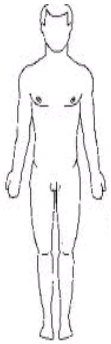

THE TWILIGHT ZONE
LA QUATRIÈME DIMENSION
LA SÉRIE ORIGINALE
E=mc²
SAISON 5 · VERSION REMASTERISÉE

Alternative à la 4D (ou plus)




Tour d'Épaule
Tour de Taille
Tour de Poitrine



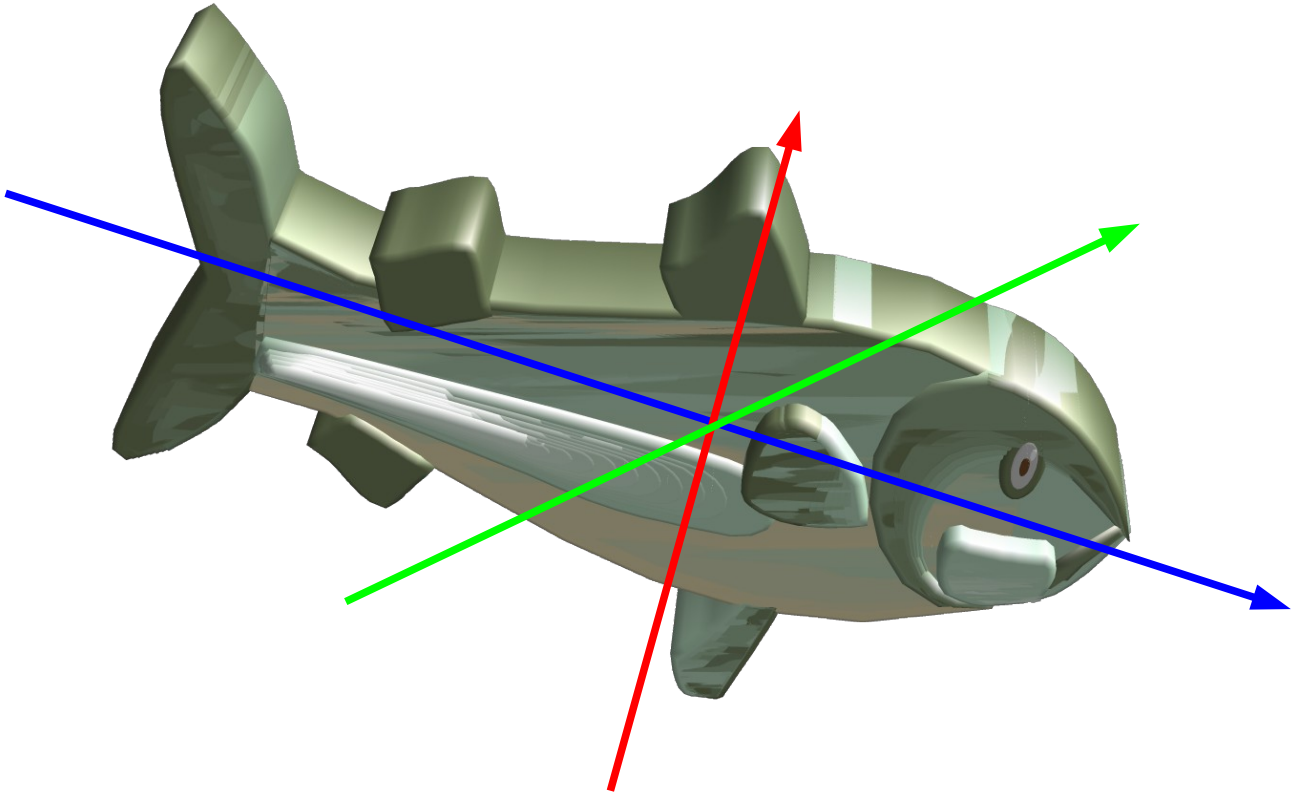


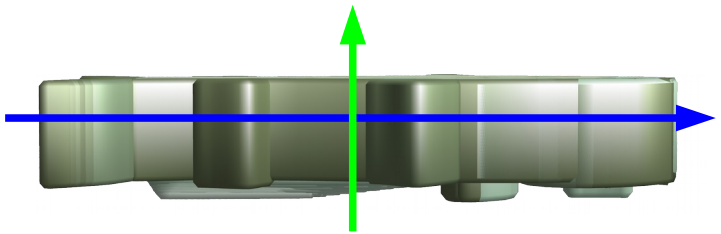
Tour d'Épaule

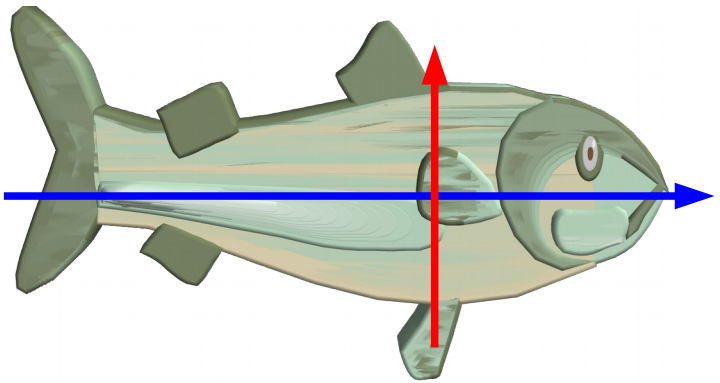
Tour de Taille

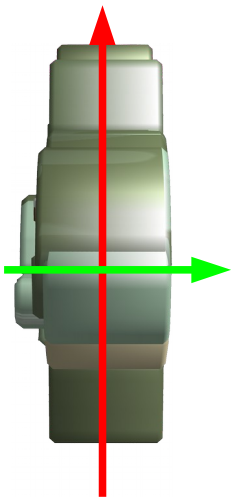
Tour de Poitrine

1ère Composante Principale : « costauditude »

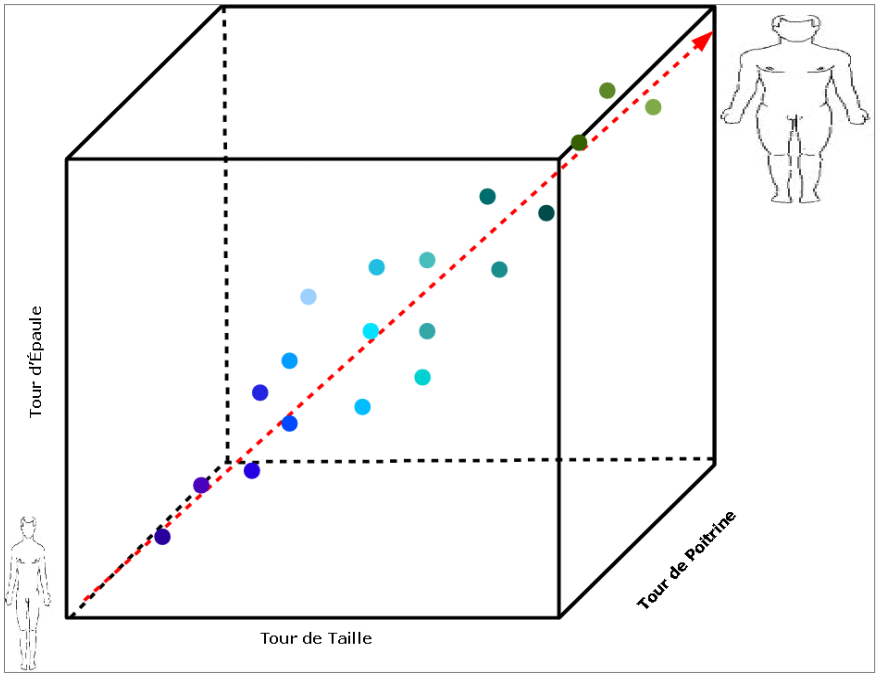




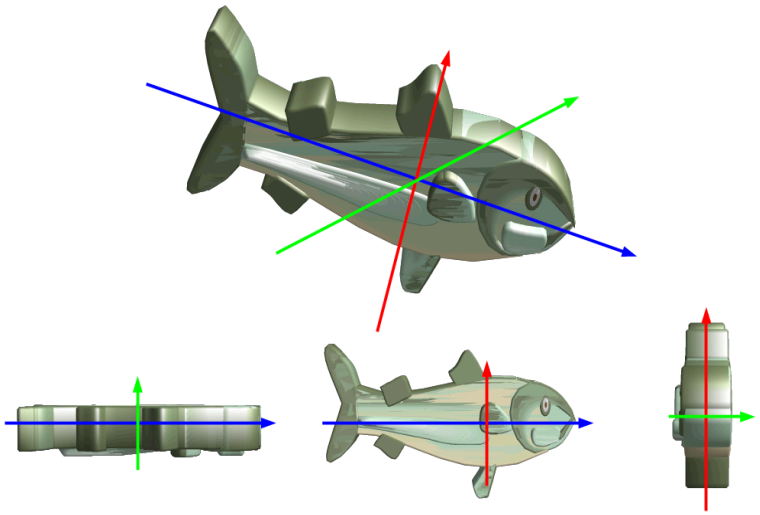




Commentaires



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 3ème 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.

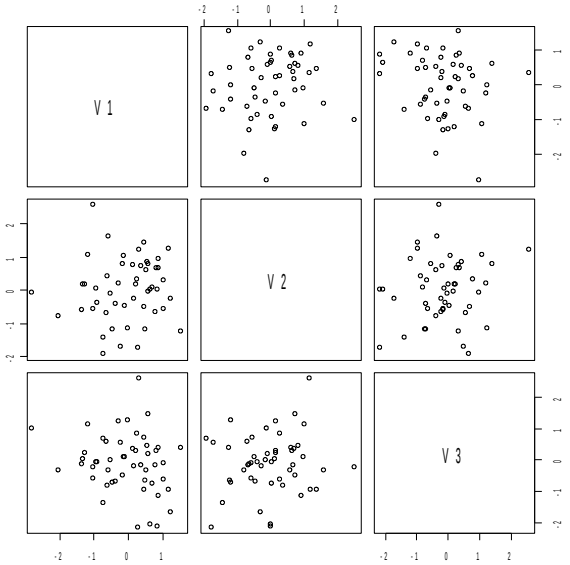
- Le principe de l'ACP est de trouver un axe (la première composante principale), issu d'une combinaison linéaire des variables initiales, tel que la variance du nuage autour de cet axe soit maximale. Et de réitérer ce processus dans des directions orthogonales pour déterminer les composantes principales suivantes.

- Du point de vue des variables, l'ACP permet de conserver au mieux la structure de corrélation entre les variables initiales.

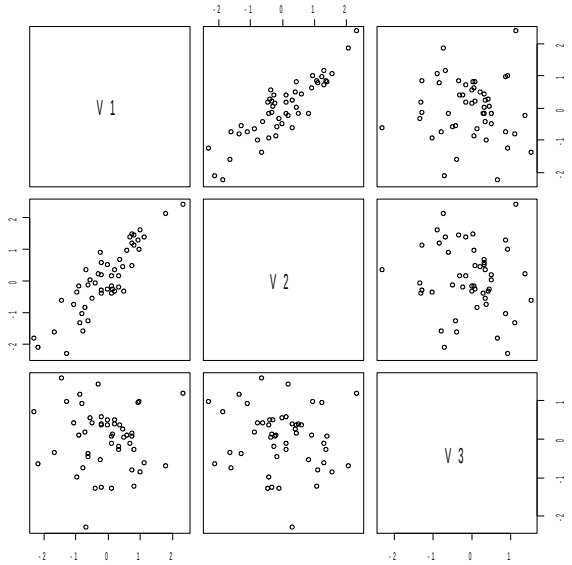
ACP : exemples simulés

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

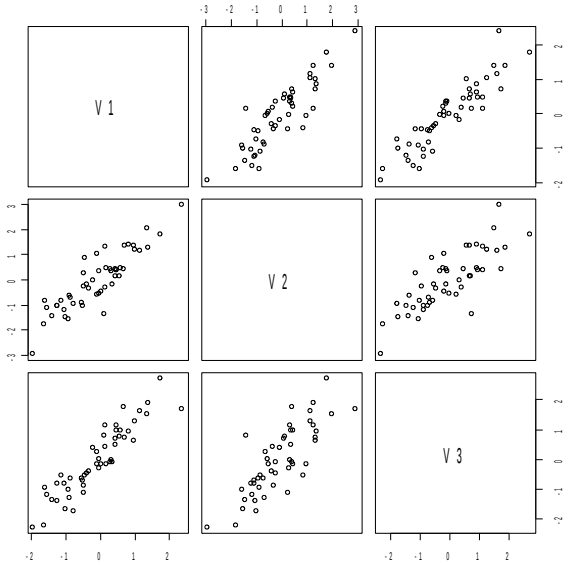
Cas 1)
{V1} - {V2} - {V3}



Cas 2)
{V1 - V2} - {V3}



Cas 3)
{V1 - V2 - V3}



Matrices de corrélation

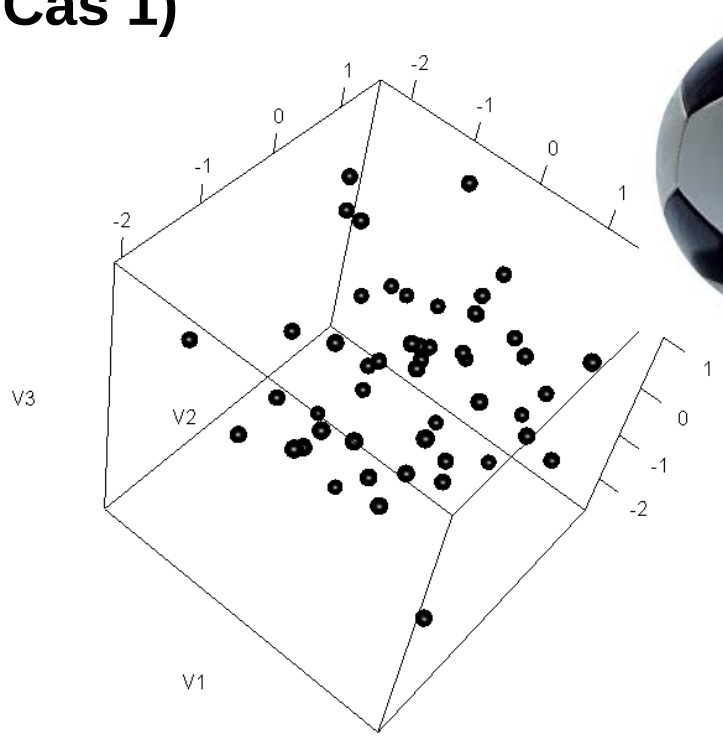
1)	V1	V2	V3
V1	1.0	-0.10	0.00
V2	-0.1	1.00	-0.12
V3	0.0	-0.12	1.00

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

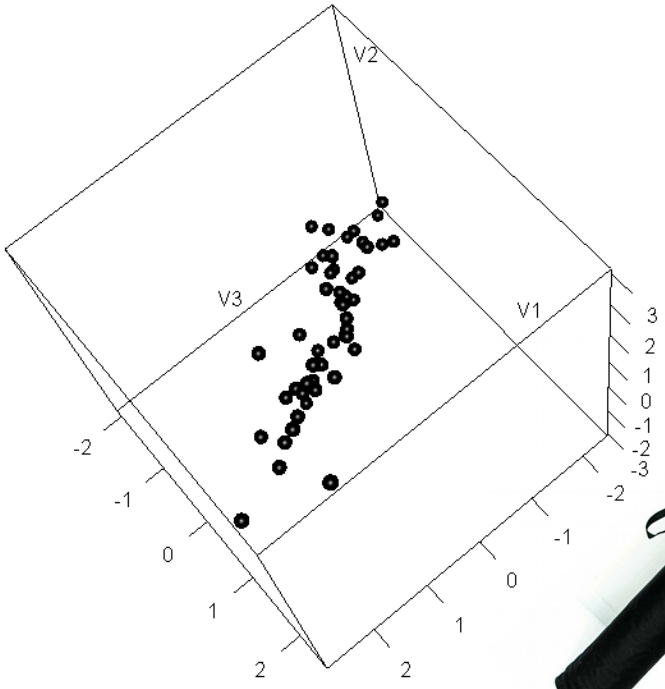
3)	V1	V2	V3
V1	1.00	0.88	0.92
V2	0.88	1.00	0.81
V3	0.92	0.81	1.00

ACP : exemples simulés

Cas 1)



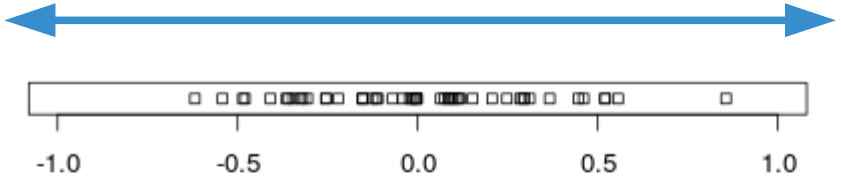
Cas 2)

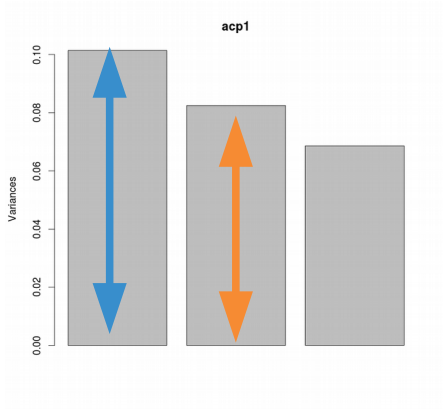


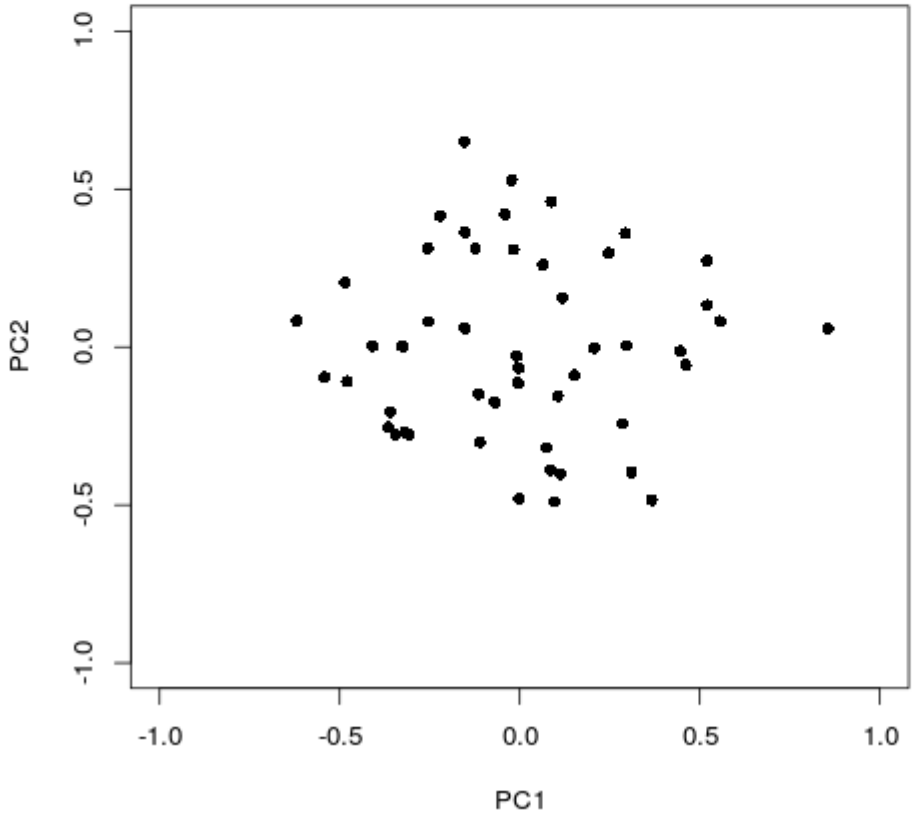
Cas 3)

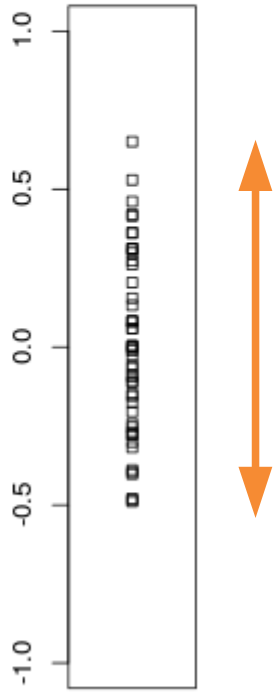
Représentation des individus

1)



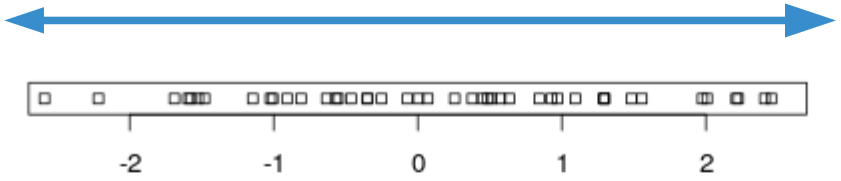


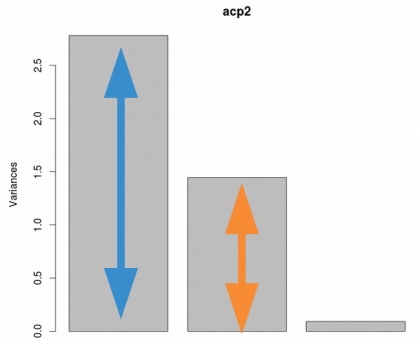


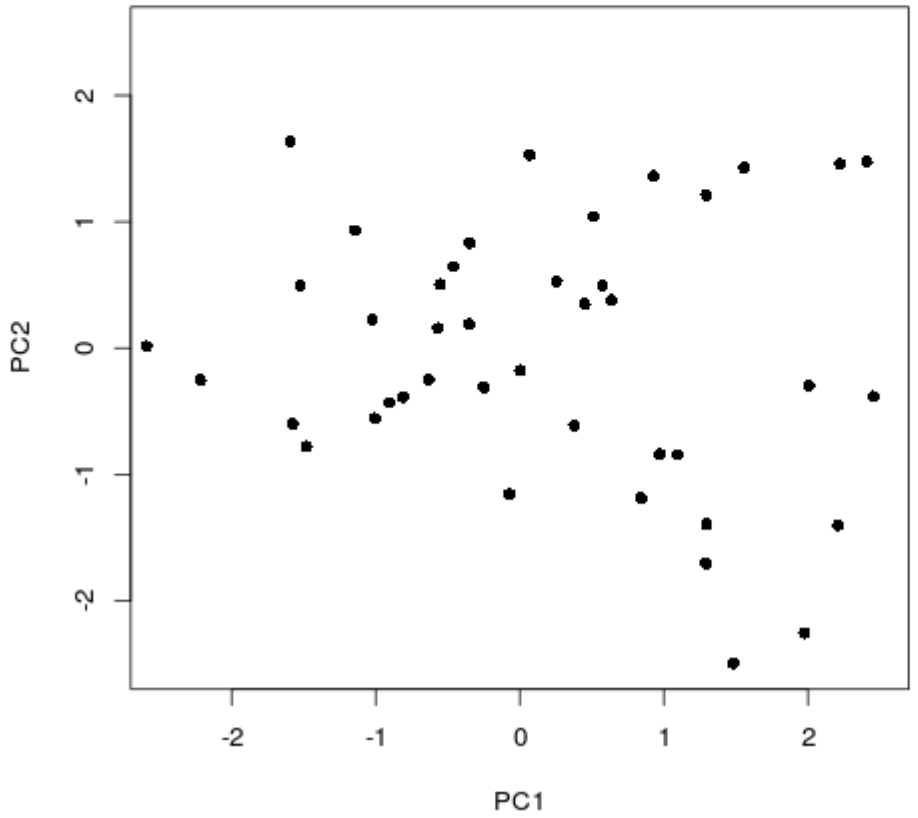


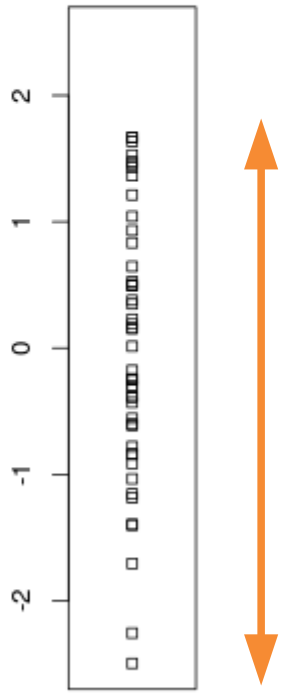
Représentation des individus

2)



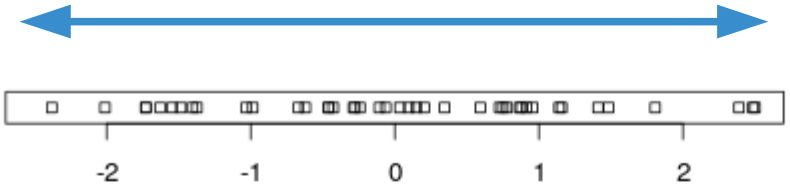


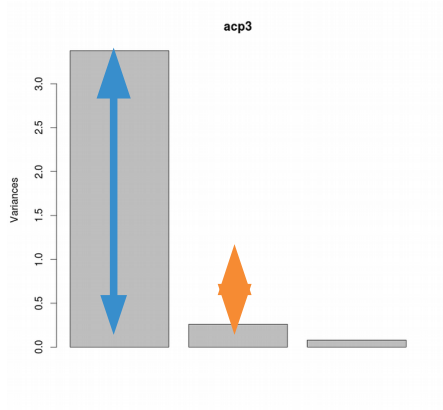


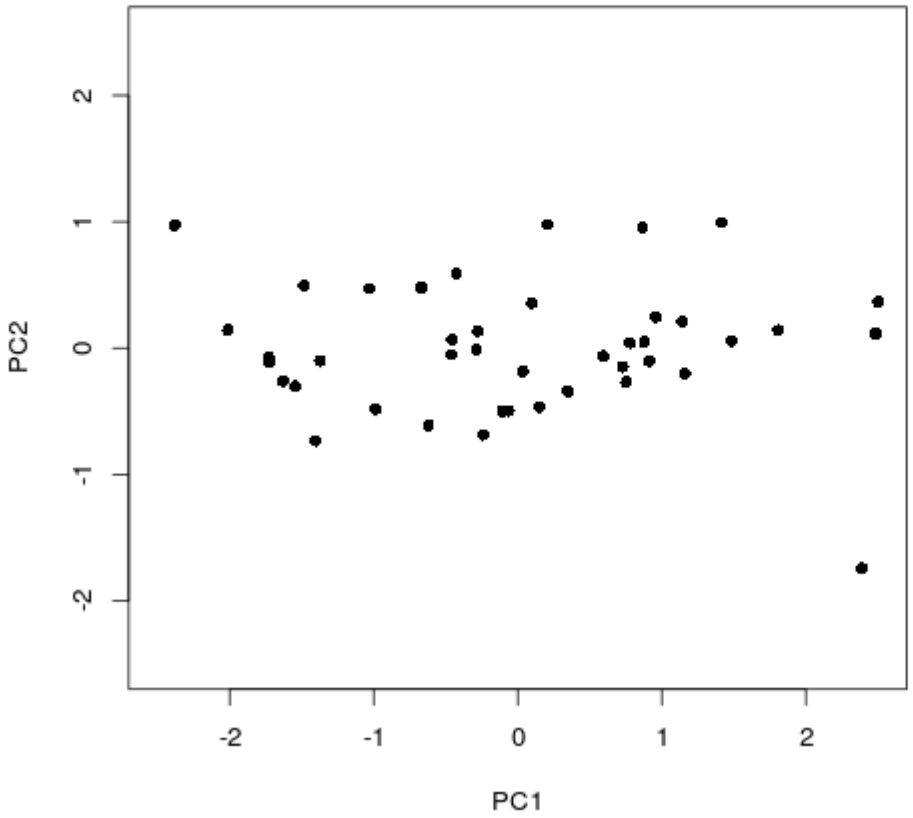


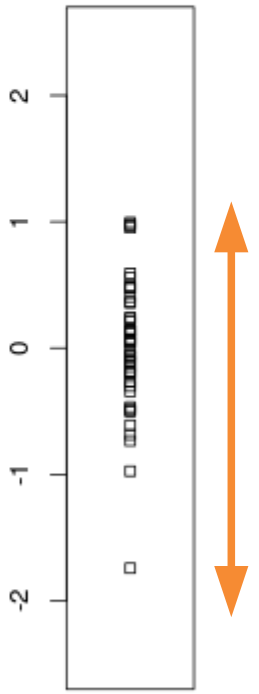
Représentation des individus

3)

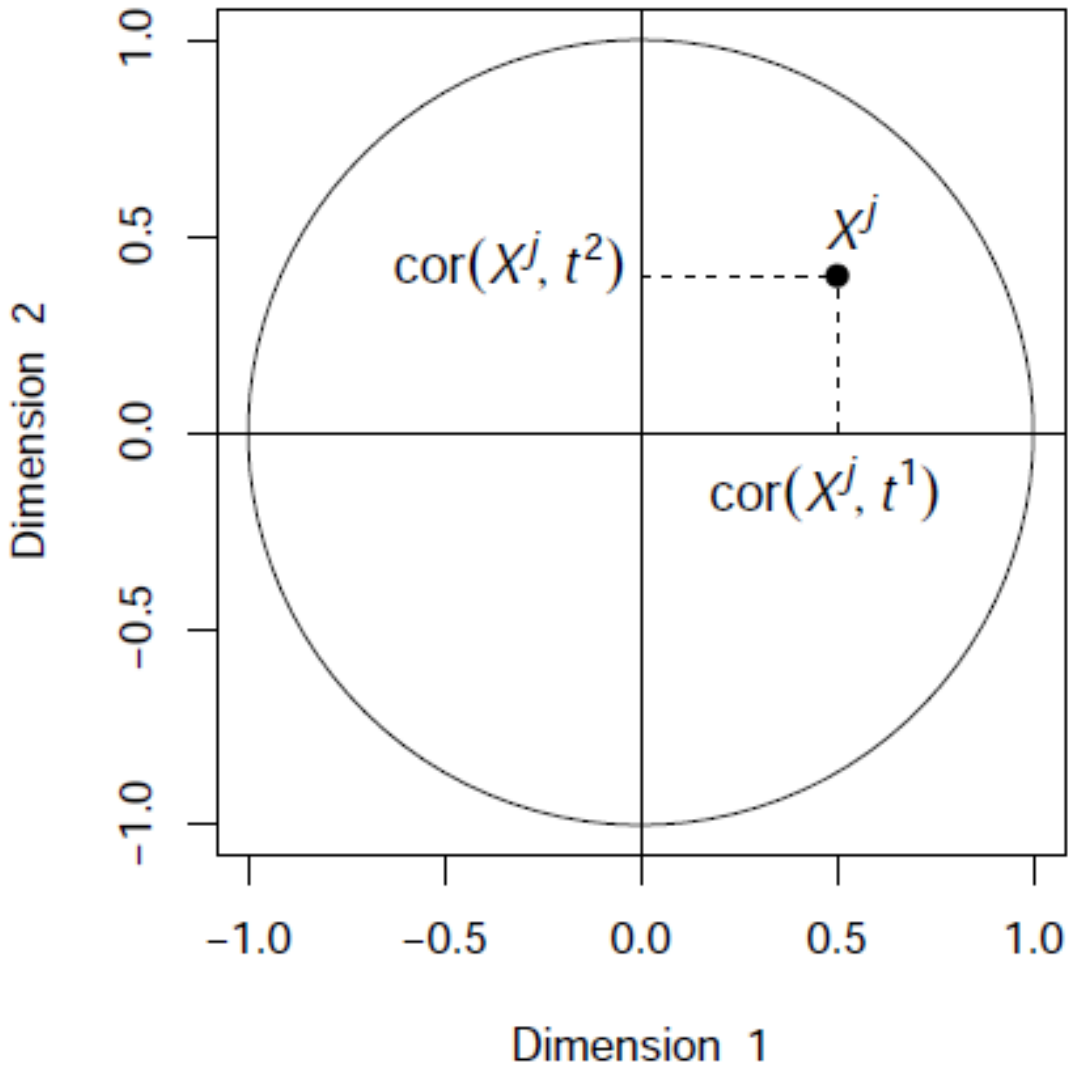








Représentation des variables

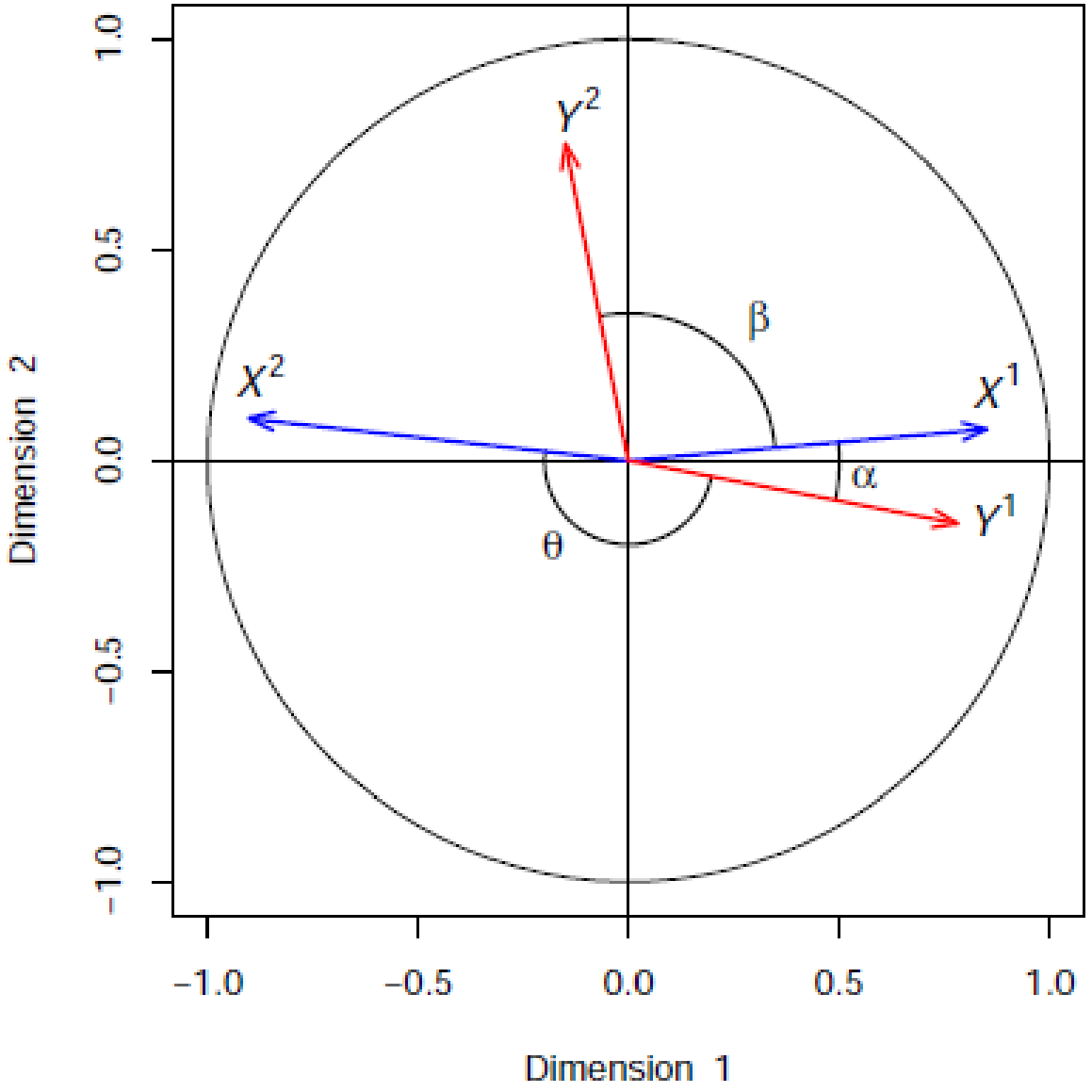


La coordonnée d'une variable X^j sur une composante t^i est donnée par la corrélation entre cette variable et t^i.

Représentation des variables

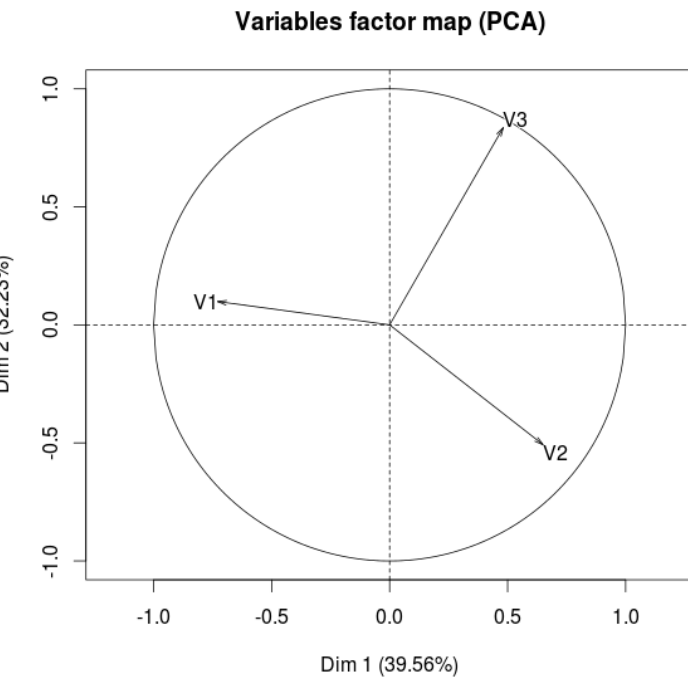
La corrélation entre deux variables est :

- positive si l'angle est aigu cos(α) > 0
- négatif si l'angle est obtus cos(θ) < 0
- nul si les vecteurs sont perpendiculaires cos(β)≈0

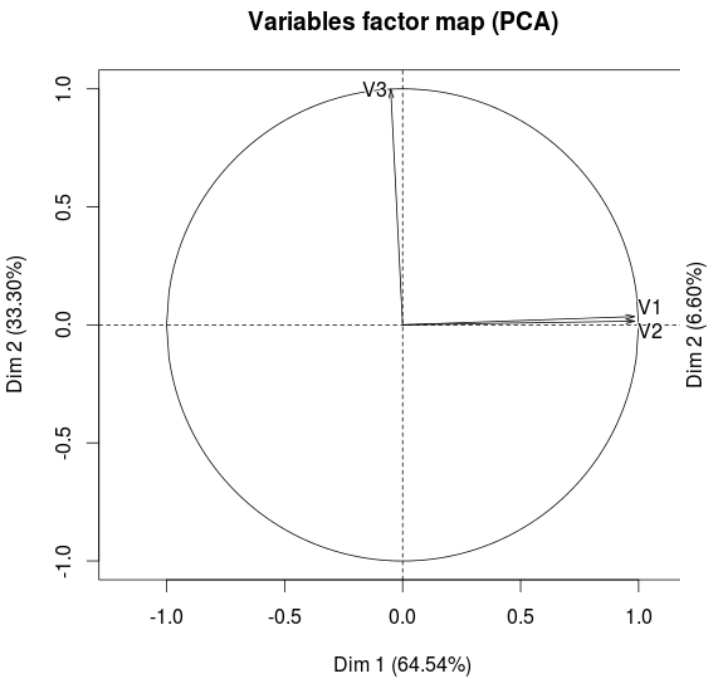


Représentation des variables

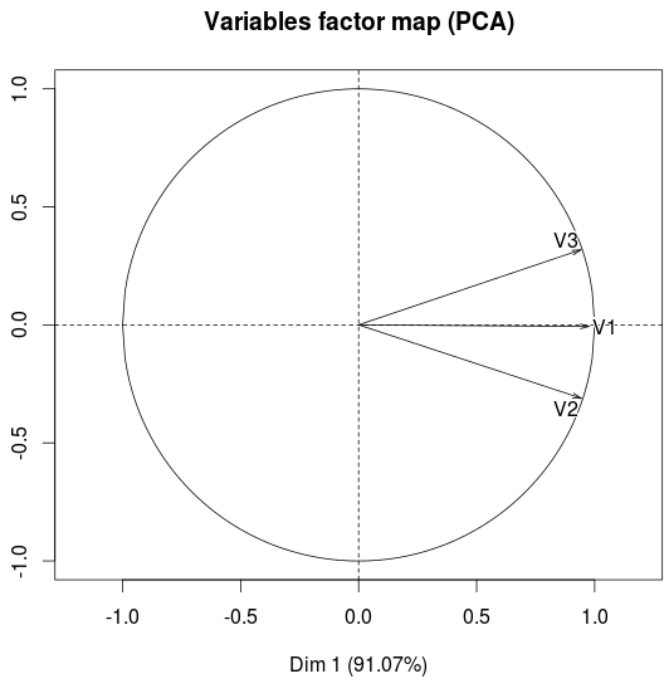
1)



2)



3)



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

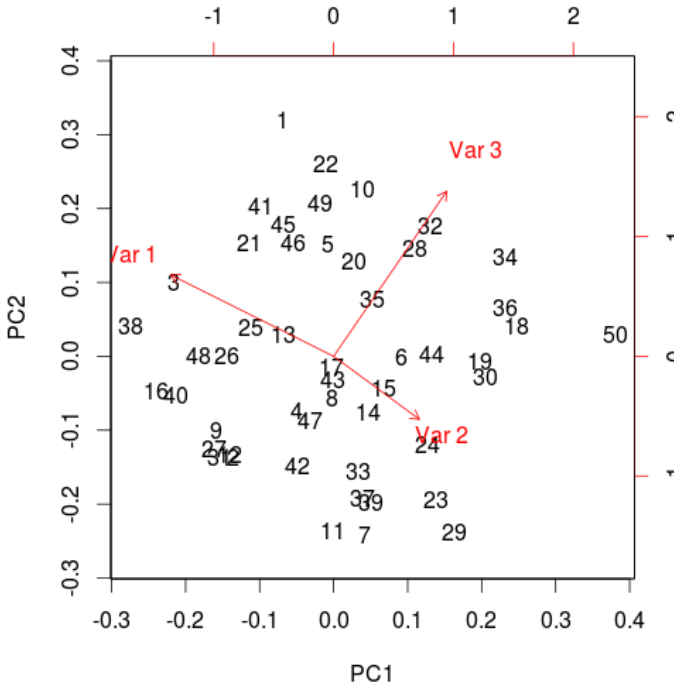
2)	V1	V2	V3
V1	1.00	0.88	-0.05
V2	0.88	1.00	-0.11
V3	-0.05	-0.11	1.00

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

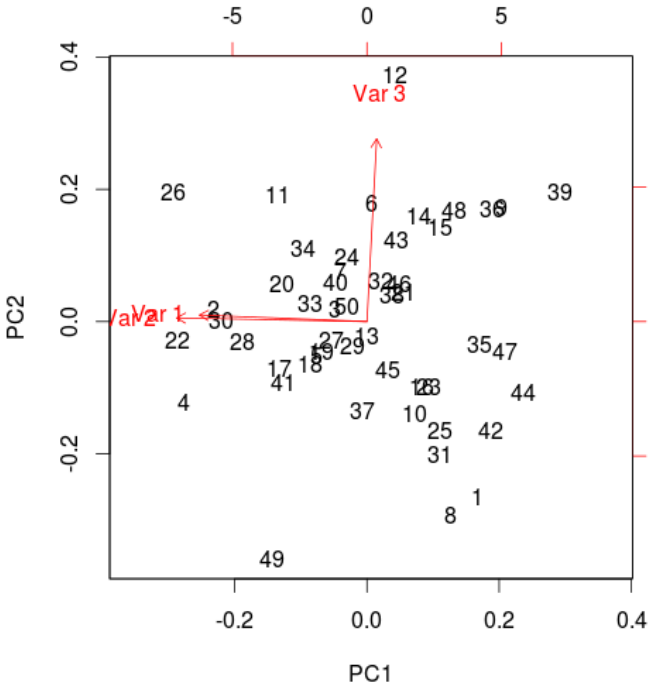
Biplot

Représentation simultanée des individus et des variables

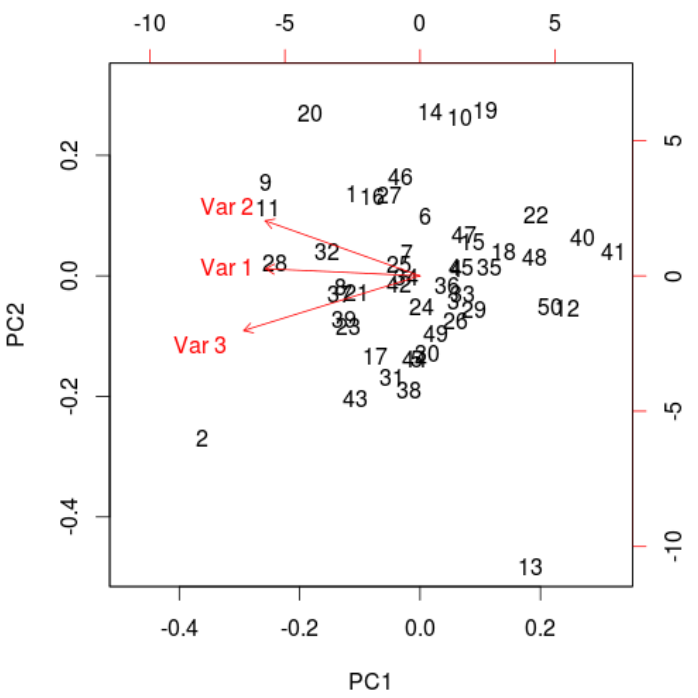
1)



2)



3)

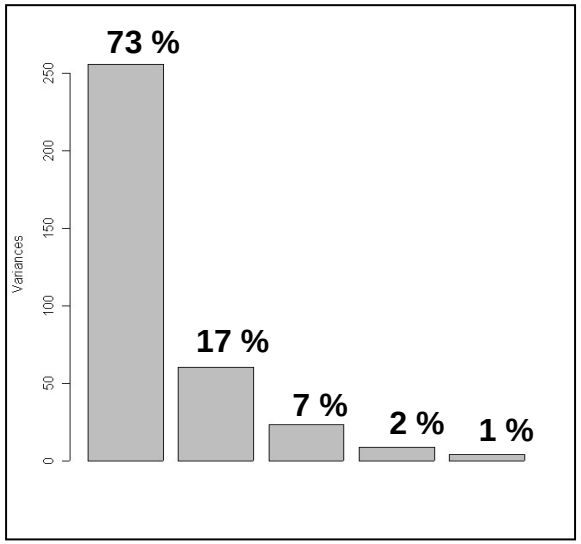


ACP en pratique

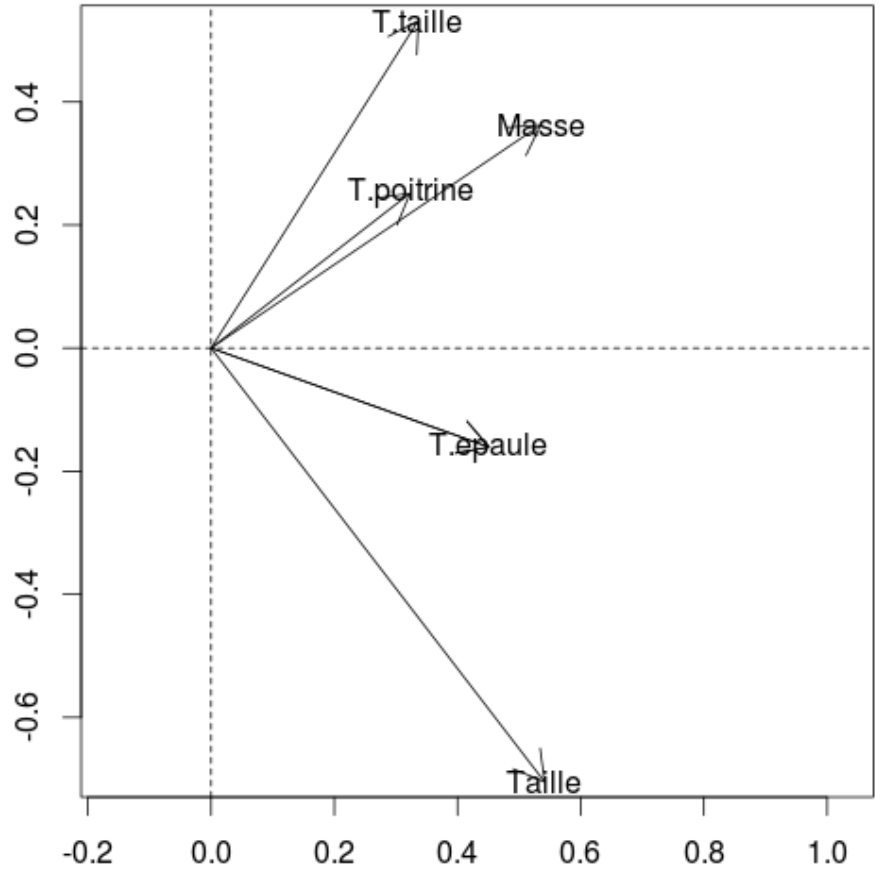
- Conséquences d'une réduction éventuelle des données :
  - sans réduction : une variable à forte variance va «tirer» tout l'effet de l'ACP
  - avec réduction : une variable qui n'est qu'un bruit va se retrouver avec une variance apparente égale à une variable informative
- Gestion (et imputation) de données manquantes : utilisation de l'algorithme NIPALS (nécessite « beaucoup » de composantes)

The best thing to do about missing data is not to have any. Gertrude Cox

Exemple « morpho »



Représentation des variables

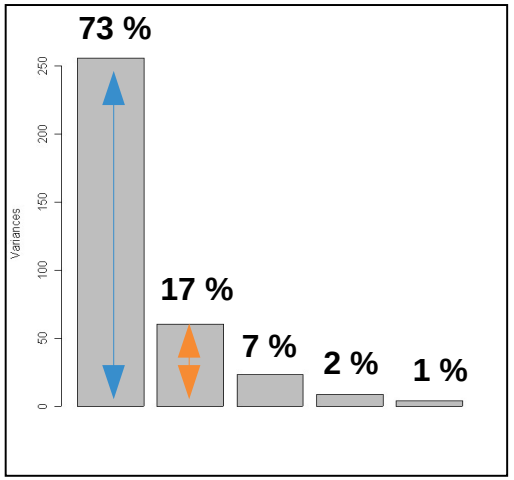


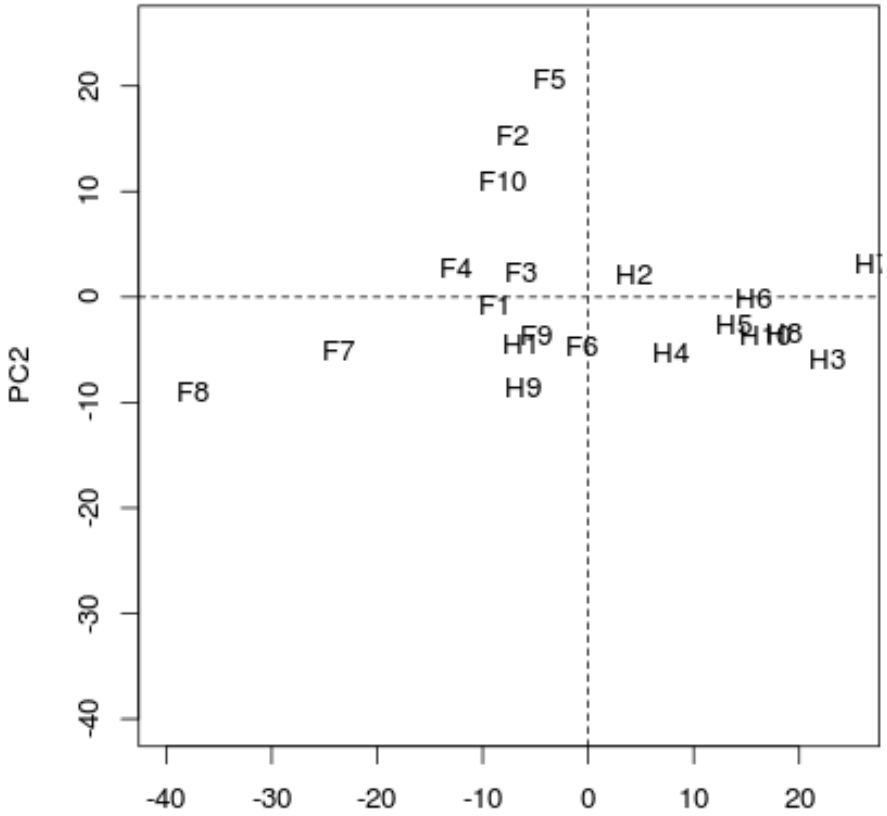
- 90% de l'information expliquée par les 2 premières CP
- le passage de 5 à 2 dimensions se fait en « perdant » 10% d'information
- Axe 1 « gabarit » : séparation des grands gabarit (valeurs élevées pour les 5 variables) à droite et des petits à gauche
- Axe 2 « embonpoint » : en bas, variables liées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

Exemple « morpho »

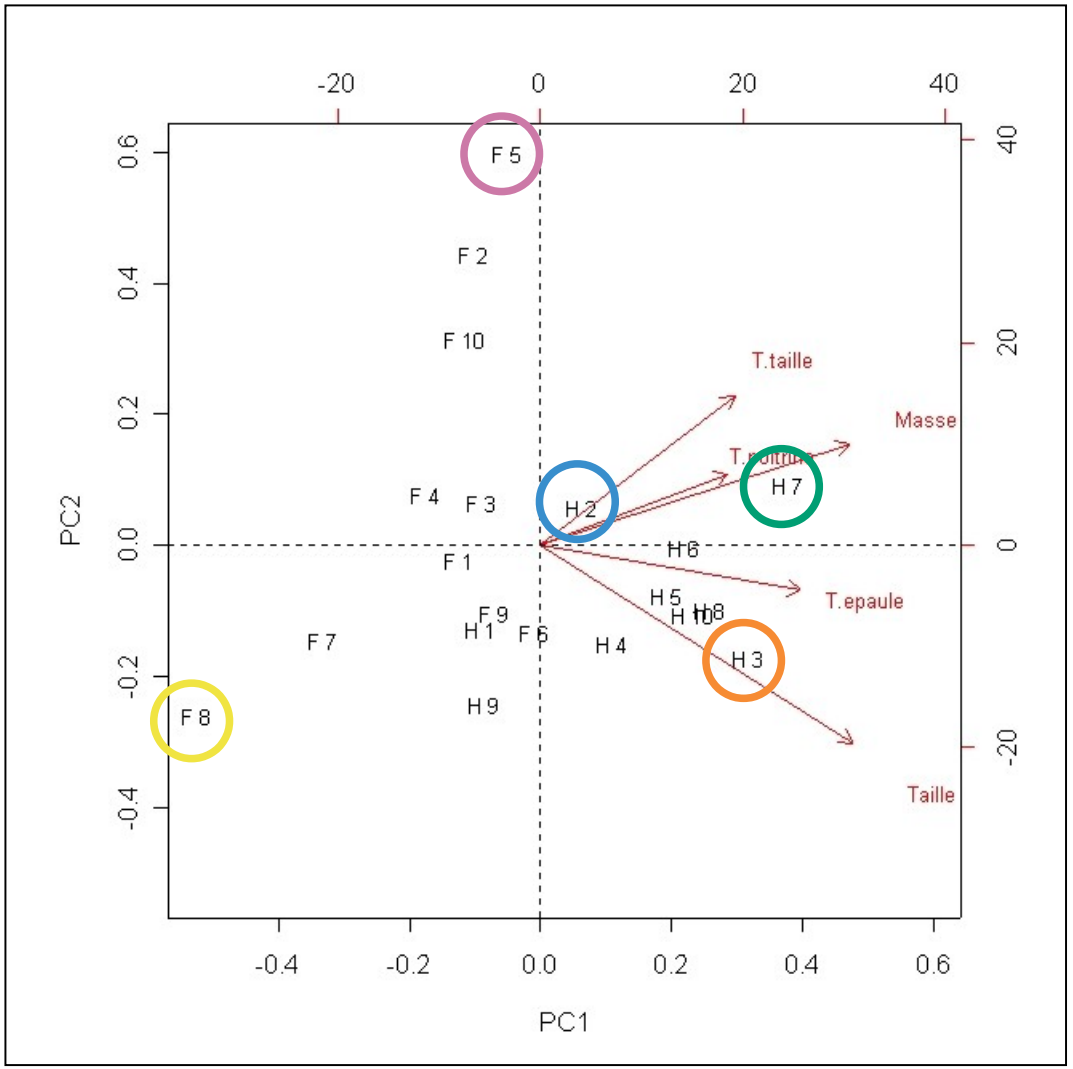




Représentation des individus

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



Exemple « morpho »

Les données

	T.ep	T.po	T.ta	Ma	Tai
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

Matrice de covariance

	T.epaule	T.poitrine	T.taille	Masse	Taille
T.epaule	**68.64**	37.74	28.08	55.32	61.19
T.poitrine	37.74	**37.51**	33.90	45.70	32.40
T.taille	28.08	33.90	**50.77**	56.58	27.70
Masse	55.32	45.70	56.58	**85.71**	59.52
Taille	61.19	32.40	27.70	59.52	**109.31**

68.64 + 37.51 + 50.77 + 85.71 + 109.31 = 351.94

Les données projetées sur les composantes principales

	PC1	PC2	PC3	PC4	PC5
H1	-6.50	-4.48	-0.37	-1.03	1.27
H2	4.40	2.04	0.81	1.87	1.38
H3	22.66	-5.94	-6.18	0.11	1.97
H4	7.78	-5.24	-8.38	4.10	-1.74
H5	13.73	-2.67	-8.02	0.82	-2.15
H6	15.67	-0.15	4.49	2.33	4.40
H7	26.99	3.19	6.29	0.04	-3.08
H8	18.41	-3.43	5.63	1.09	-1.96
H9	-6.25	-8.48	4.97	0.79	1.86
H10	16.78	-3.67	1.99	-7.08	1.22
F1	-8.83	-0.78	0.28	-3.02	0.07
F2	-7.28	15.41	-2.31	-3.00	-2.35
F3	-6.45	2.25	-7.60	0.95	1.15
F4	-12.51	2.68	8.91	4.27	-1.53
F5	-3.65	20.76	-0.30	-2.45	1.99
F6	-0.63	-4.62	0.34	-3.46	-2.80
F7	-23.61	-5.07	2.20	1.19	-1.15
F8	-37.50	-9.07	-1.33	-1.89	-0.02
F9	-4.98	-3.61	0.33	-0.50	1.02
F10	-8.24	10.89	-1.74	4.86	0.44

Matrice de covariance

	PC1	PC2	PC3	PC4	PC5
PC1	**255.66**	0.00	0.00	0.00	0.00
PC2	0.00	**60.18**	0.00	0.00	0.00
PC3	0.00	0.00	**23.48**	0.00	0.00
PC4	0.00	0.00	0.00	**8.61**	0.00
PC5	0.00	0.00	0.00	0.00	**4.01**

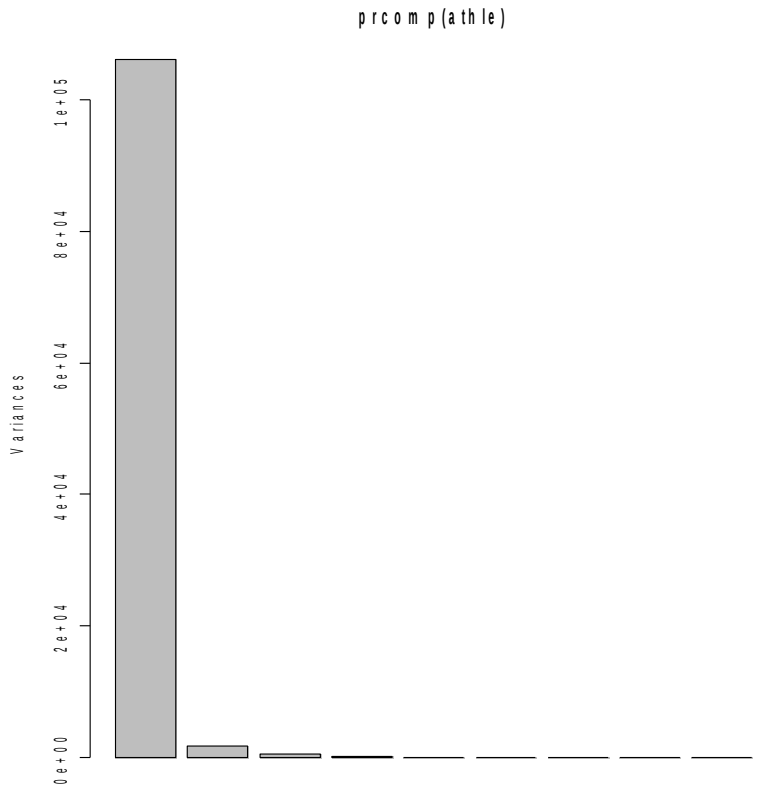
255.66 + 60.18 + 23.48 + 8.61 + 4.01 = 351.94

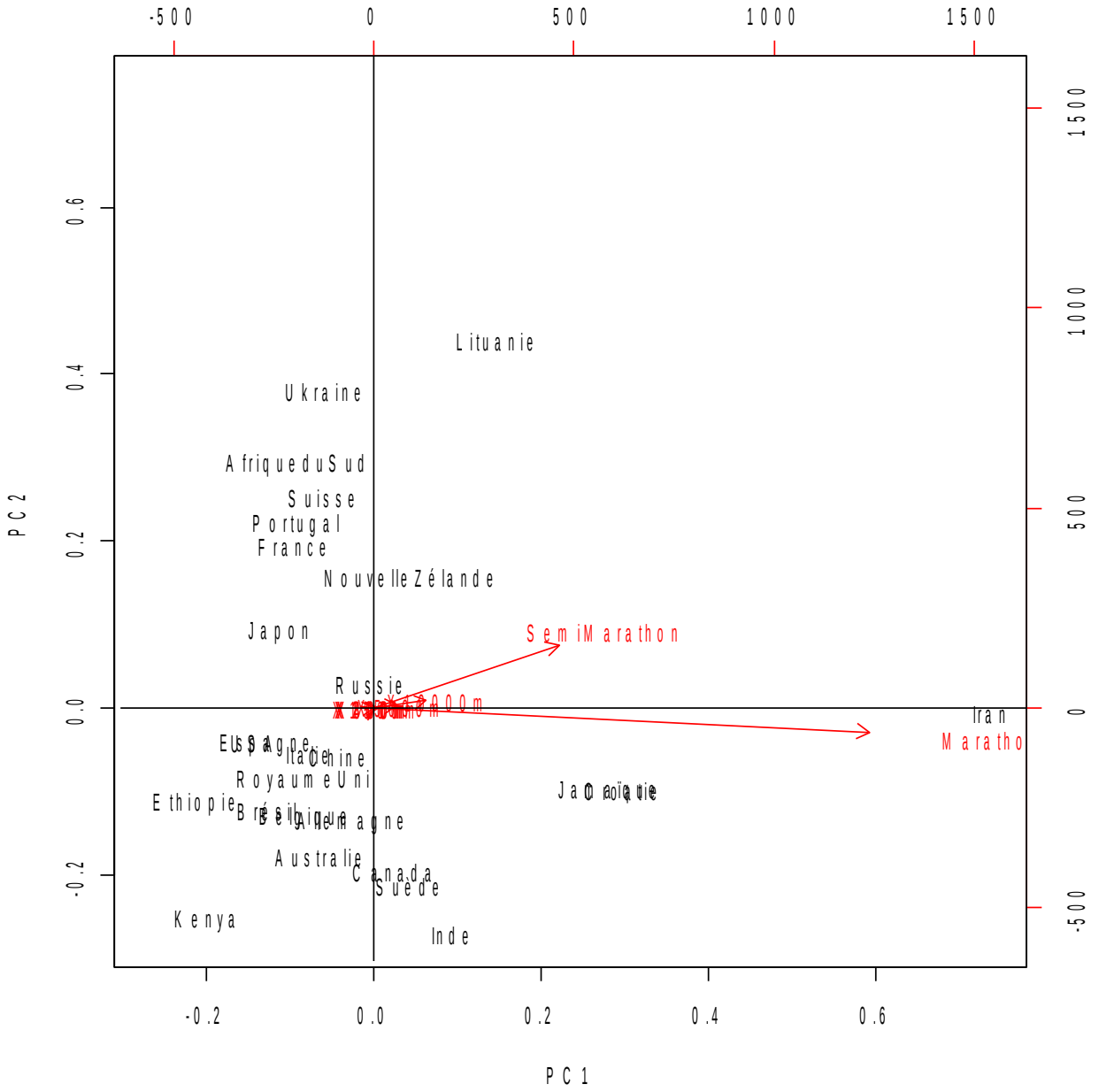
Exemple « athlé »

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

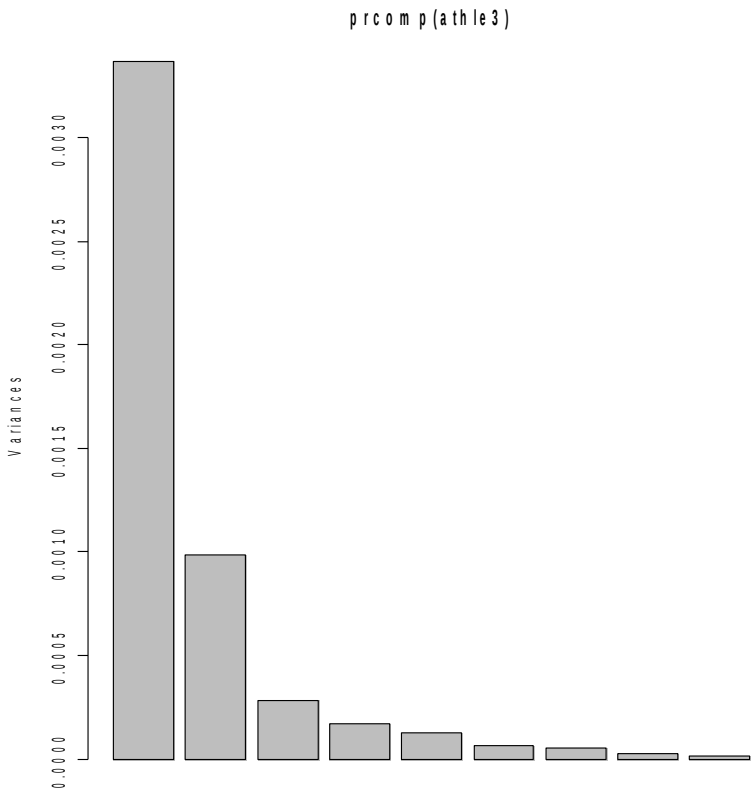
Exemple « athlé »

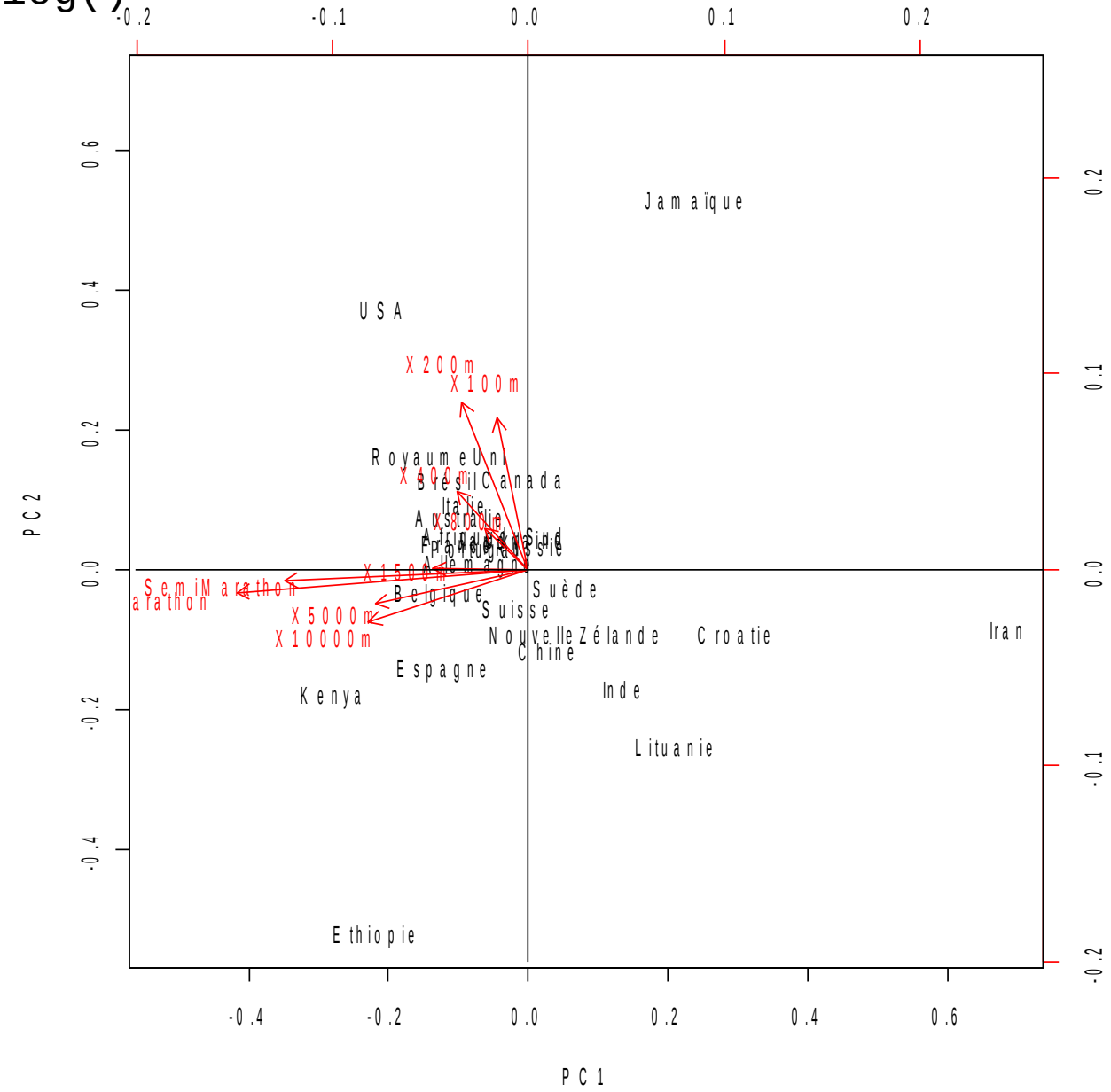




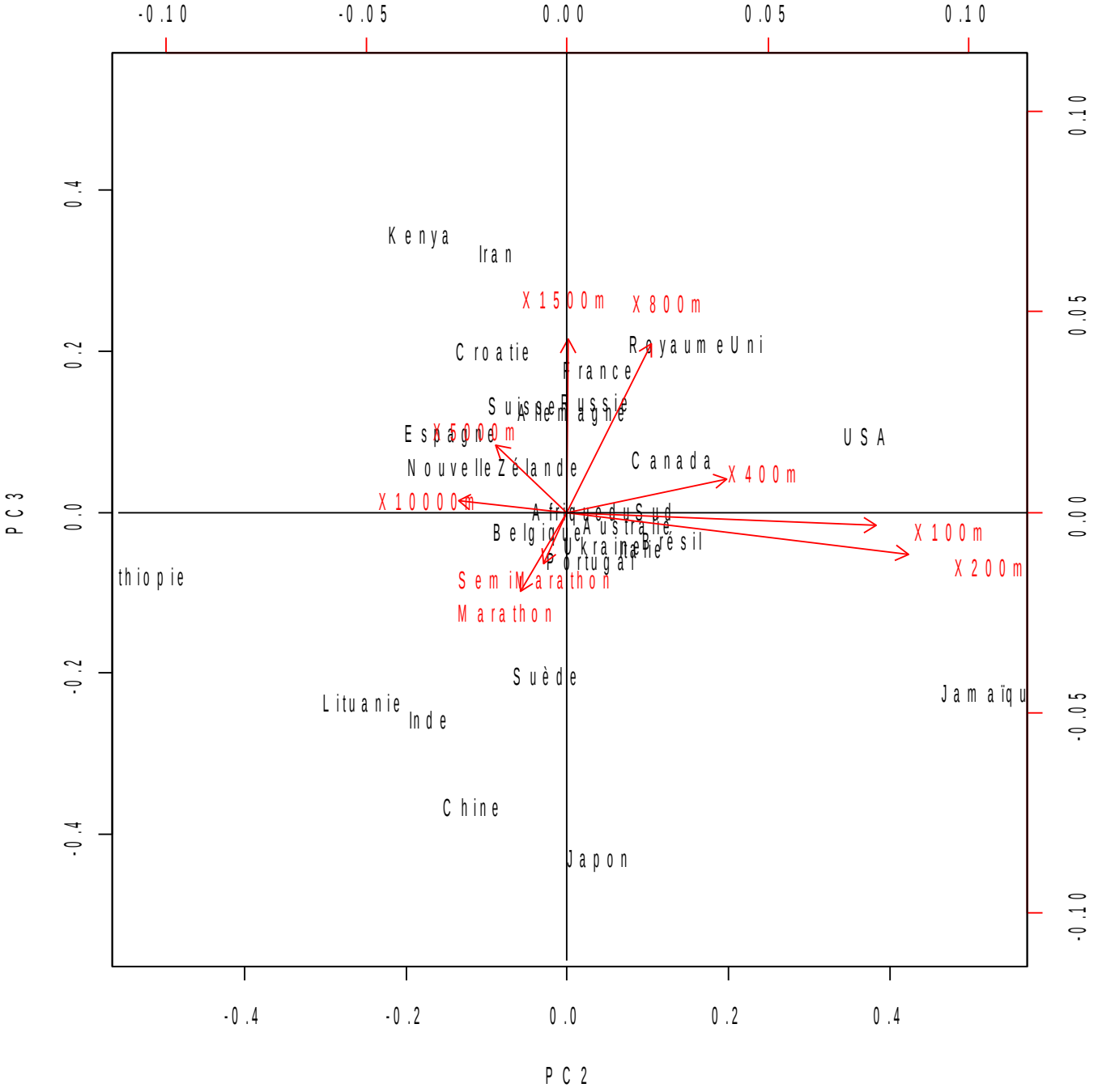
Exemple « athlé »

Après conversion des données en `-log()`





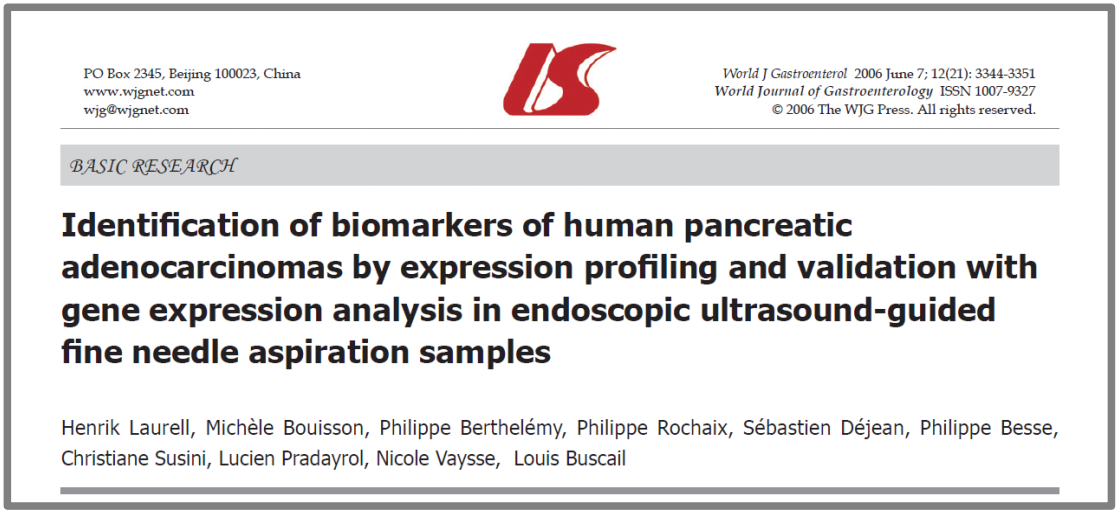
Exemple « athlé »



Exemple « transcriptome 1 »

Expression de **868** gènes mesurée sur **22** échantillons :

- Lignées pancréatiques (7 échantillons) : ASPC1, Bx-PC3, Capan 1, Capan 2, Mia-PaCa2, NP 29, Panc1 ;
- Lignées coliques (5 échantillons) : CaCo2, HCT116, HT29, SW480, SW620 ;
- Lignée leucémique (1 échantillon) : K562 ;
- Pièces tumorales (6 échantillons) : PT1, PT2, PT3, PT4, PT5, PT6 ;
- Pancréas normal (3 échantillons) : PancNorm1, PancNorm2, PancNorm3 ;



PO Box 2345, Beijing 100023, China
www.wjgnet.com
wjg@wjgnet.com

World J Gastroenterol 2006 June 7; 12(21): 3344-3351
World Journal of Gastroenterology ISSN 1007-9327

BASIC RESEARCH

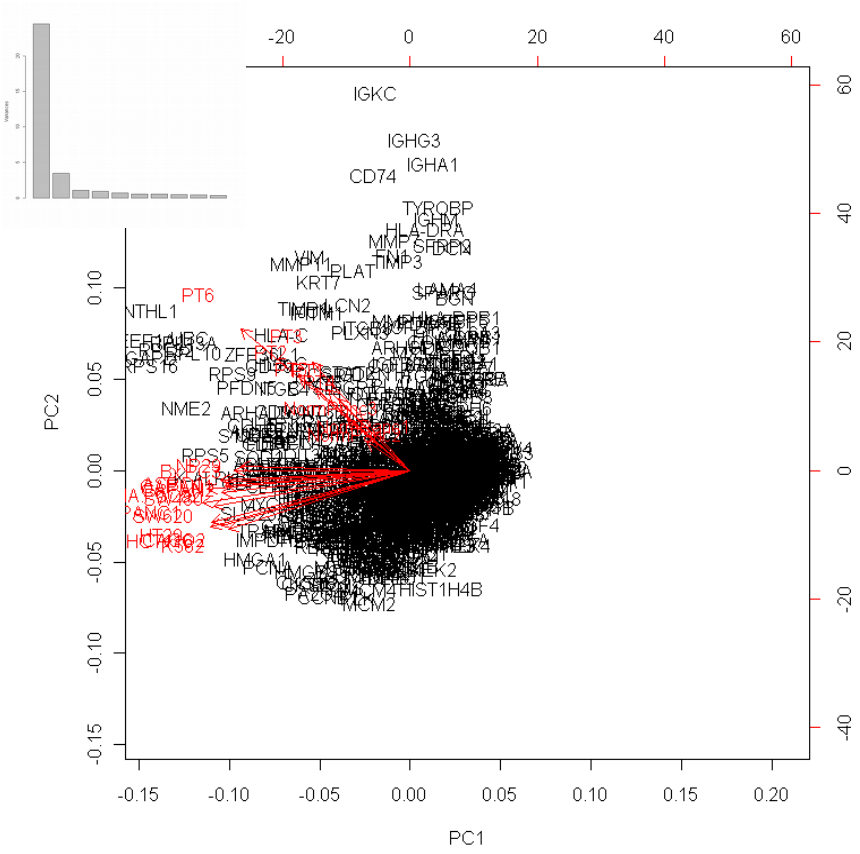
Identification of biomarkers of human pancreatic adenocarcinomas by expression profiling and validation with gene expression analysis in endoscopic ultrasound-guided fine needle aspiration samples

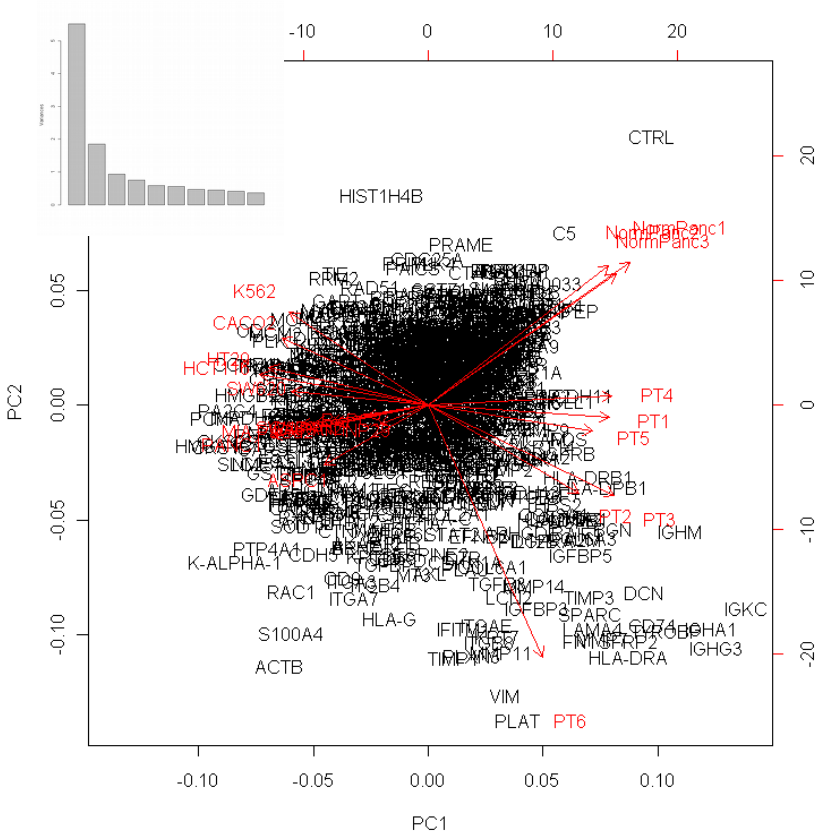
Henrik Laurell, Michèle Bouisson, Philippe Berthelémy, Philippe Rochaix, Sébastien Déjean, Philippe Besse, Christiane Susini, Lucien Pradayrol, Nicole Vaysse, Louis Buscail

Extrait des données

	ASPC1	Bx-PC3	CAPAN1	CAPAN2	NP29	PANC1	MIA-PaCa2	PT1	PT2	PT3	PT4	PT5	PT6	CACO2	....
MAPRE1	1,838	1,736	1,523	2,062	1,353	2,488	2,319	-0,133	0,086	0,555	-0,036	0,238	1,279	2,551	
VIL2	1,458	1,687	1,429	0,788	0,605	0,736	2,243	0,02	0,745	0,25	-0,267	0,19	1,606	0,999	
NME2	3,82	4,452	4,966	4,719	4,031	4,912	5,252	2,958	3,167	3,11	2,743	2,327	3,641	4,141	
NME1	1,819	2,069	3,088	2,648	2,346	3,609	2,85	0,489	1,423	0,53	0,616	0,877	1,353	2,485	
MARK3	0,962	0,363	0,933	1,082	0,446	1,108	0,786	0,004	-0,045	-0,289	0,134	0,193	0,585	1,101	
JUN	2,157	1,417	0,887	-0,204	1,402	1,898	3,404	2,877	2,151	3,219	0,591	2,398	3,606	-0,054	
MYC	2,852	2,965	3,32	2,69	2,997	2,009	3,856	0,376	0,941	1,981	1,225	1,582	1,274	3,028	
FOSL1	2,342	1,996	2,233	1,345	1,963	3,229	3,36	-0,065	0,171	0,812	0,596	-0,774	-0,216	-1,167	
JUNB	-0,486	-0,046	-0,179	-0,649	-0,035	-0,757	-0,642	0,399	0,499	0,56	0,368	-0,3	1,208	-1,231	
AXL	0,741	1,194	-0,433	-0,513	0,326	1,353	1,358	0,018	1,122	0,358	0,501	0,362	1,281	-1,012	
ERBB3	2,733	2,499	2,727	2,35	2,503	1,29	3,142	1,555	0,928	2,503	0,619	0,443	1,877	2,449	
FLT1	2,023	2,674	3,294	3,043	2,686	3,287	3,762	2,178	1,402	2,282	0,679	0,94	2,677	2,436	
...															

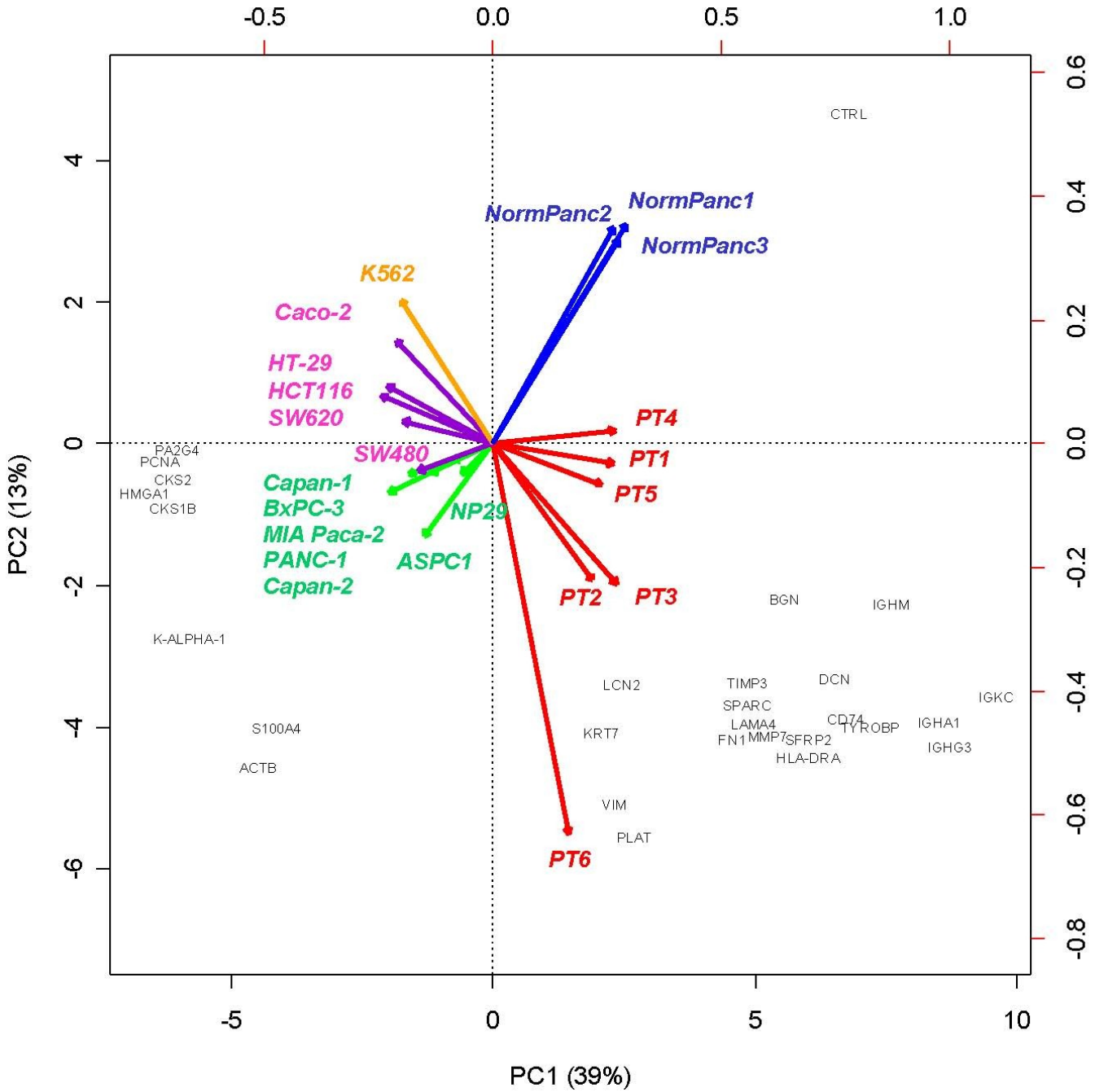
Exemple « transcriptome 1 »





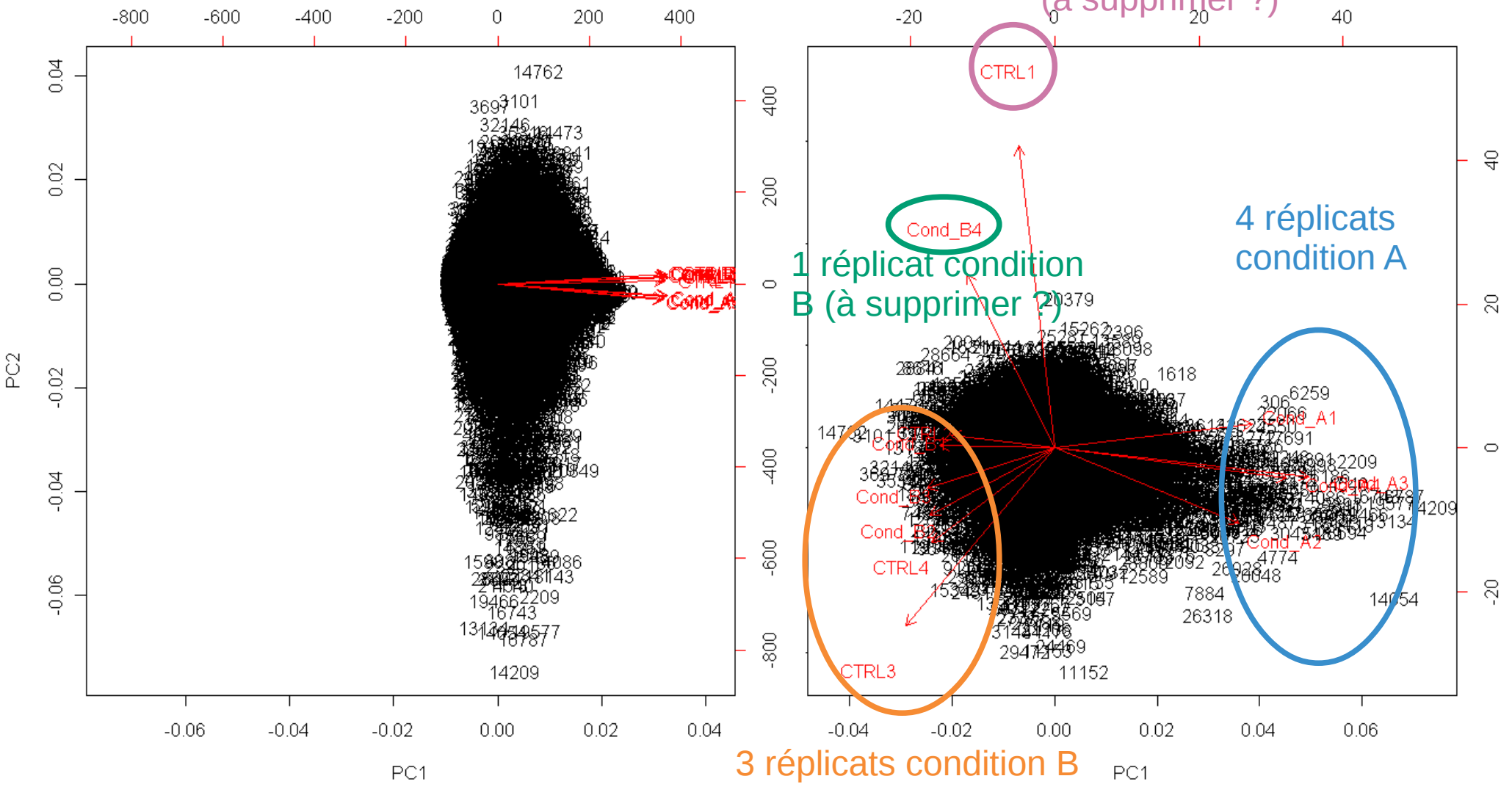
Avec réduction

Exemple « transcriptome 1 »



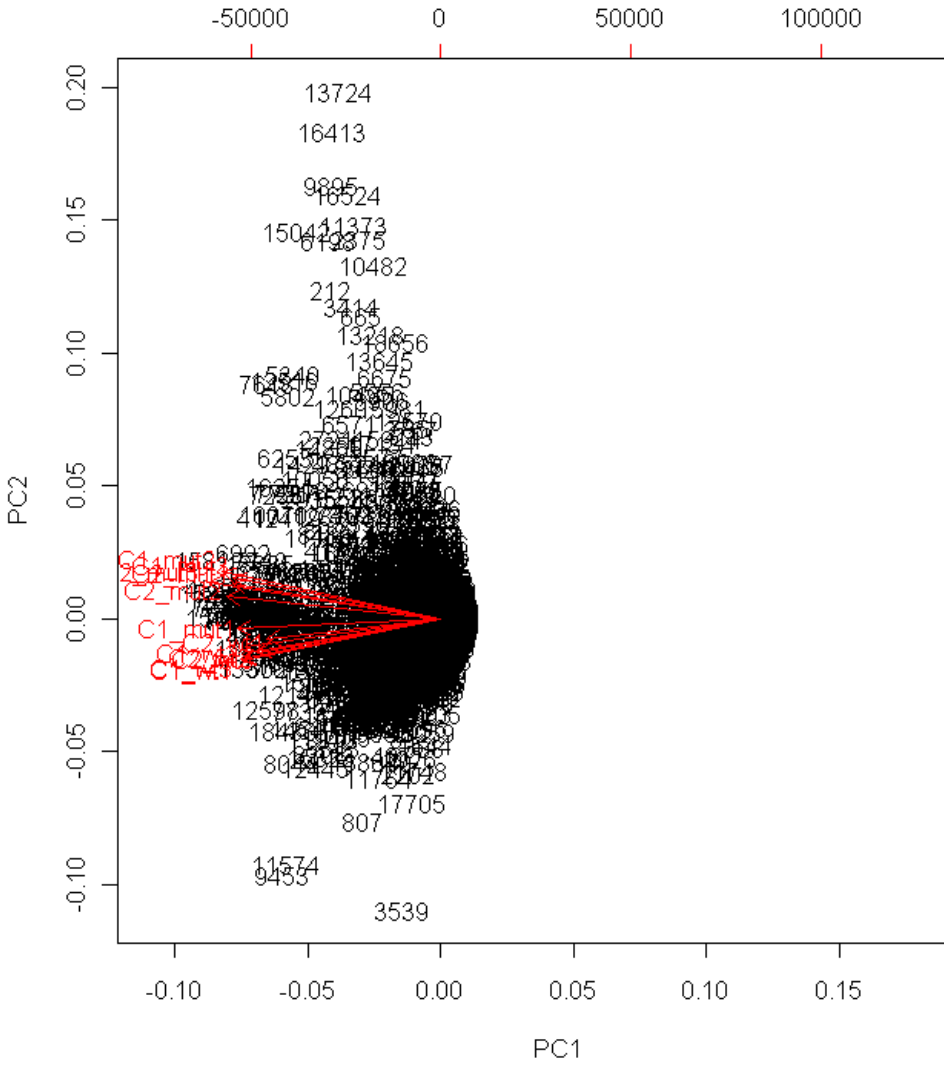
Exemple « transcriptome 2 »

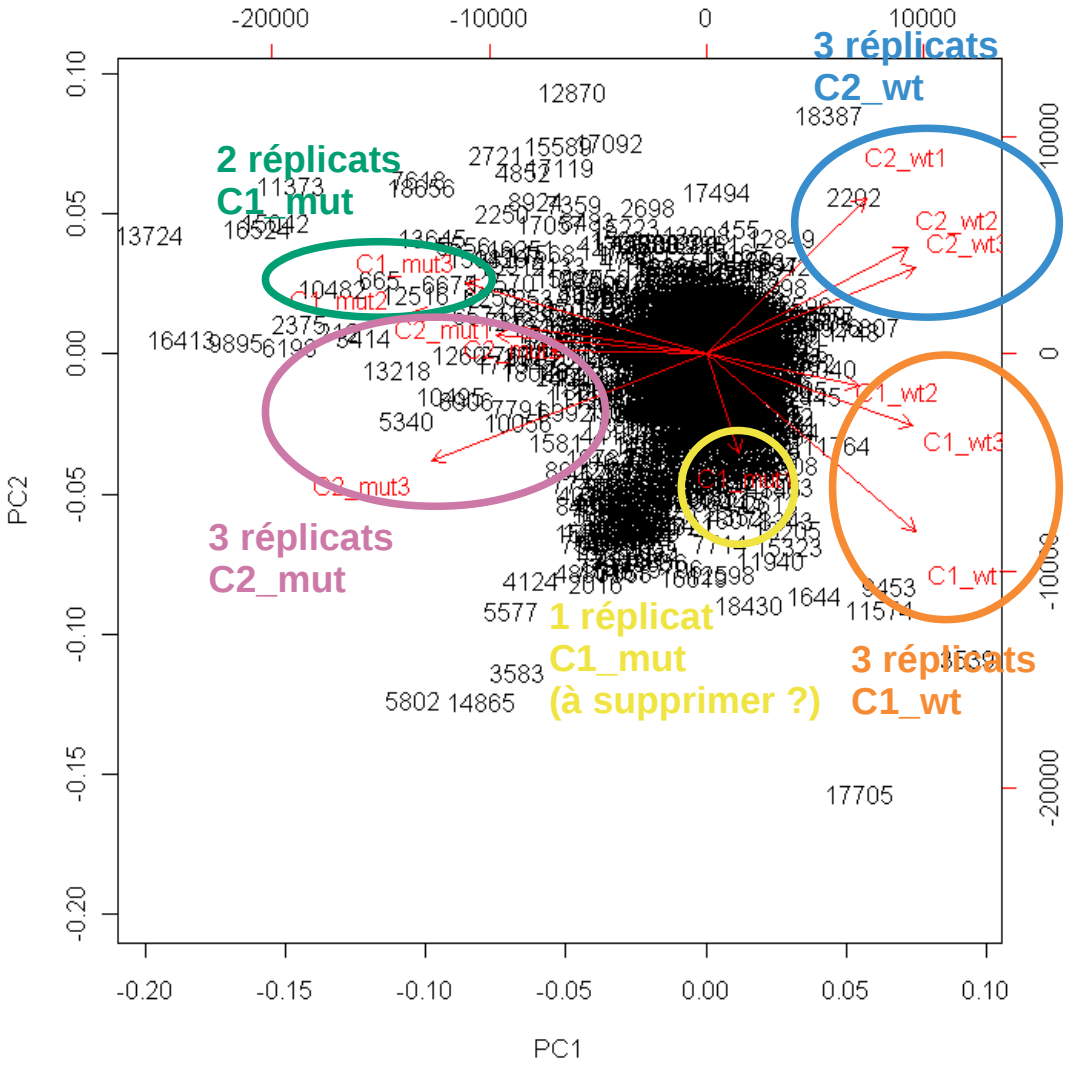
3 conditions, 4 réplicats, 38000 gènes, puce Affymetrix



Exemple « transcriptome 3 »

4 conditions (2 traitements * 2 génotypes), 3 réplicats, 20000 gènes, puce Affymetrix

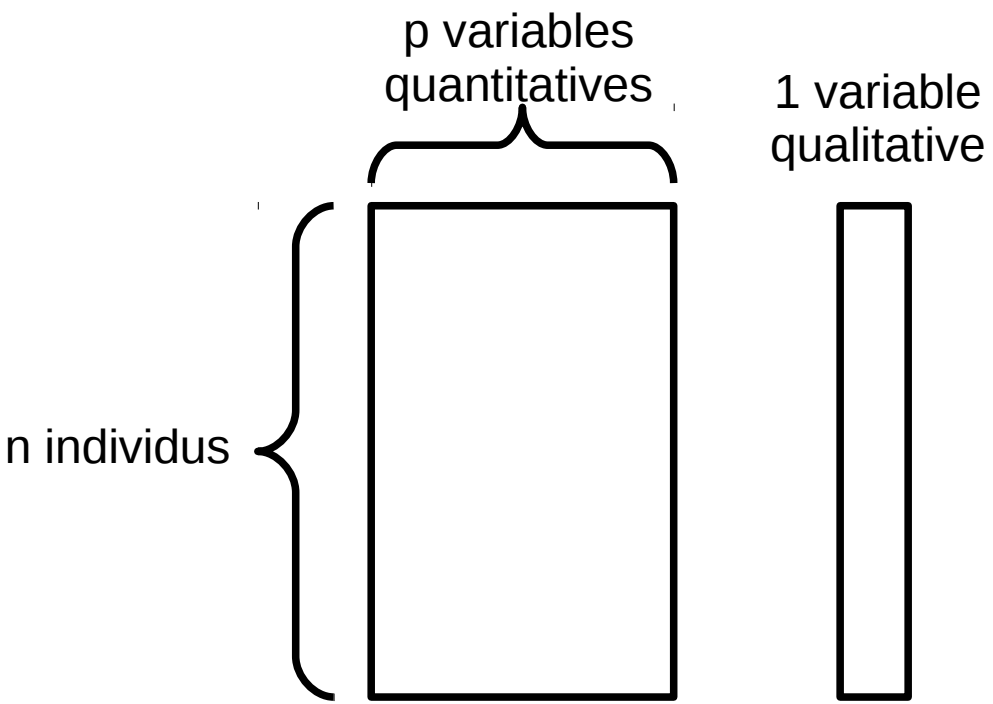




Méthodes discriminantes

Analyse Factorielle Discriminante (AFD)

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.



AFD : exemple simulé

Tableau de données

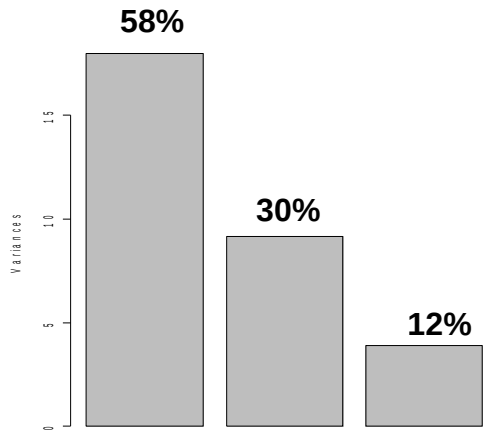
- 50 individus, 4 variables
- 3 quantitatives V1 – V2 – V3
- 1 qualitative Groupe à 2 modalités A et B

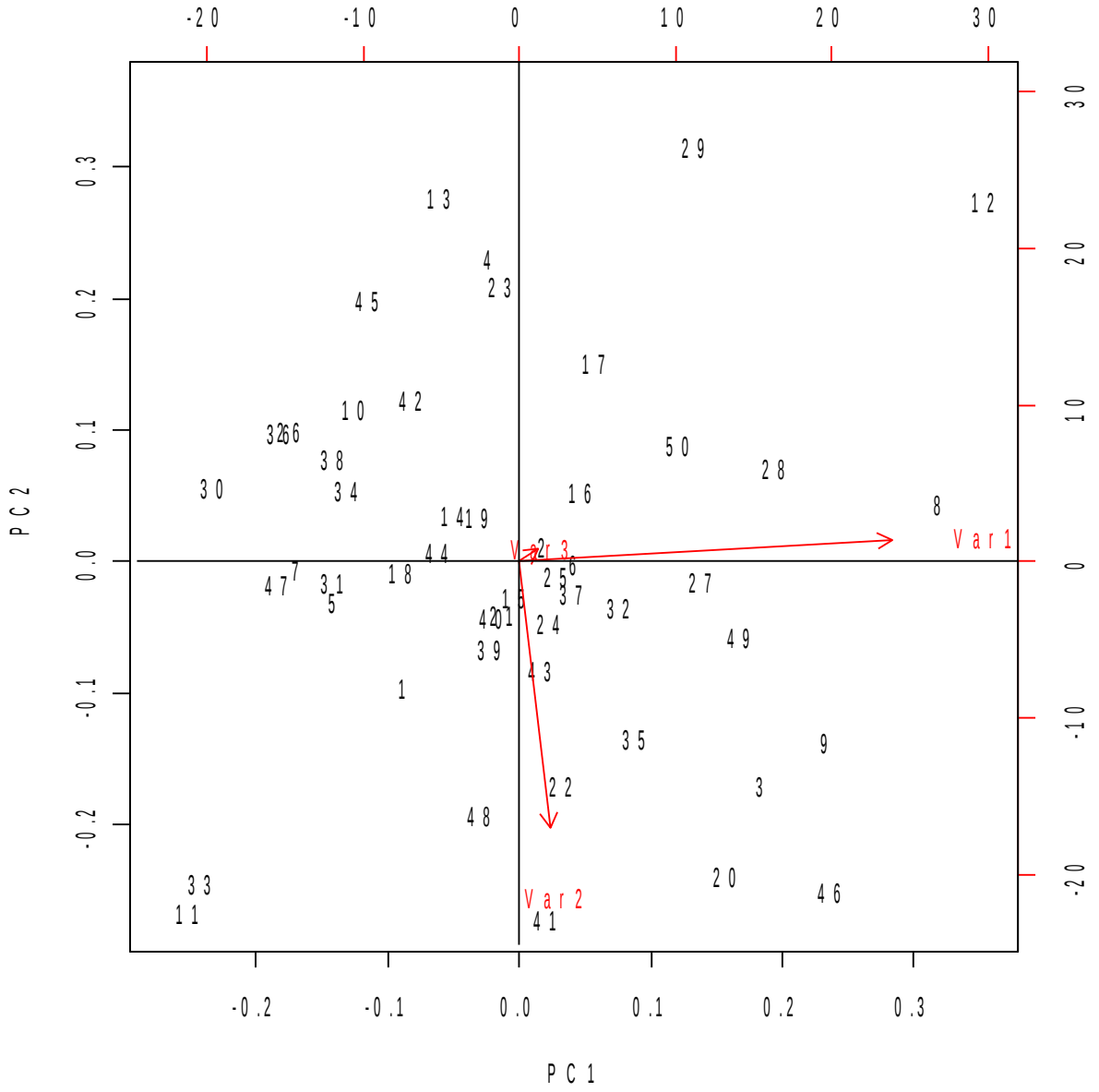
Peut-on trouver un espace de représentation qui sépare au mieux les individus du groupe A de ceux du groupe B ?

	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
48	-0.19	3.95	-1.89	B
49	5.92	1.54	-1.72	B
50	4.82	-1.70	-2.41	B

AFD : exemple simulé

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

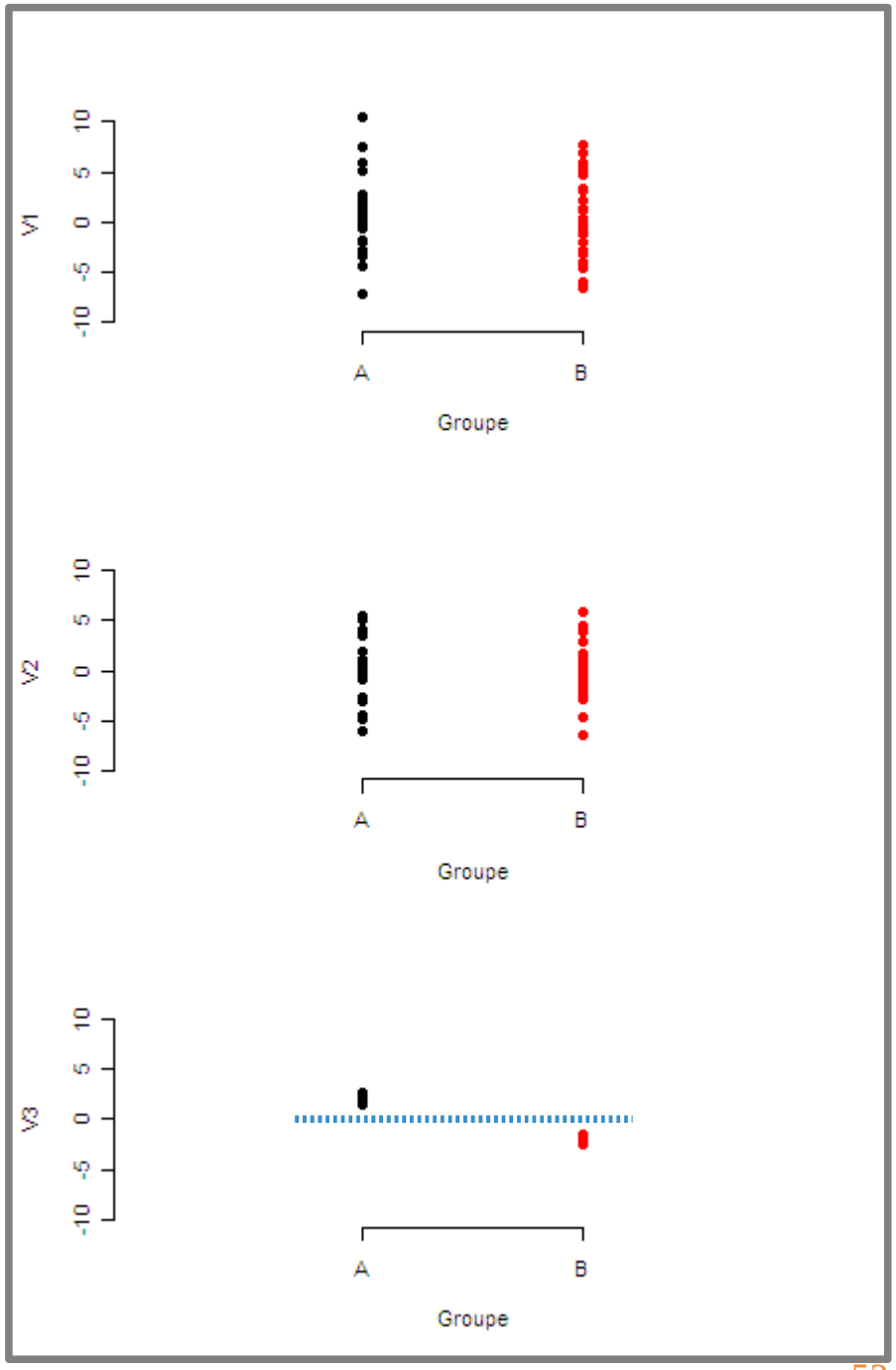




- Les 3 CP sont clairement identifiées respectivement aux 3 variables initiales V1-V2-V3.
- La plus grande part de la variabilité des données est expliquée par V1, puis V2 et enfin V3.

AFD : exemple simulé

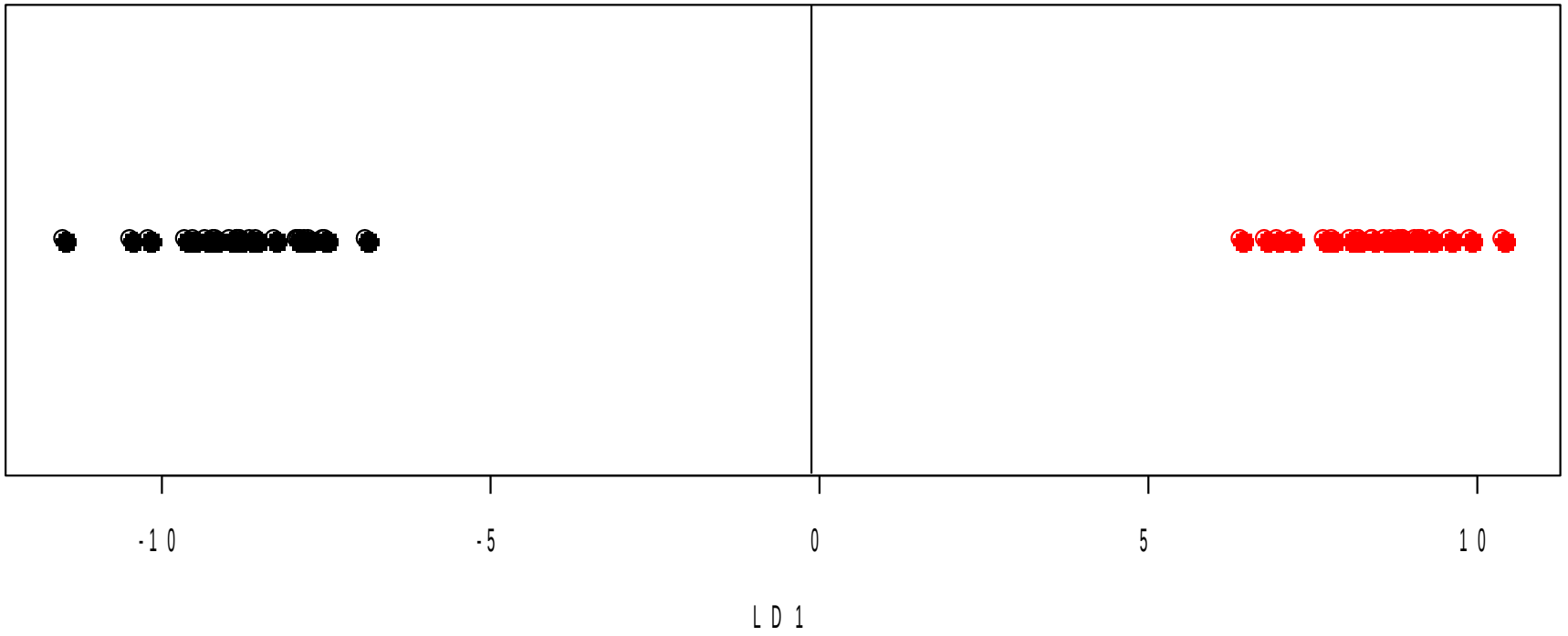
Représentation des 50 individus selon les 3 variables séparément avec couleur selon la modalité de la variable qualitative.



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

AFD : exemple simulé

Résultat d'une AFD



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

$$LD1 = -0.058 * V1 - 0.028 * V2 \mathbf{- 4.41} * V3$$

- La variable discriminante LD1 correspond à l'opposé de la variable V3.

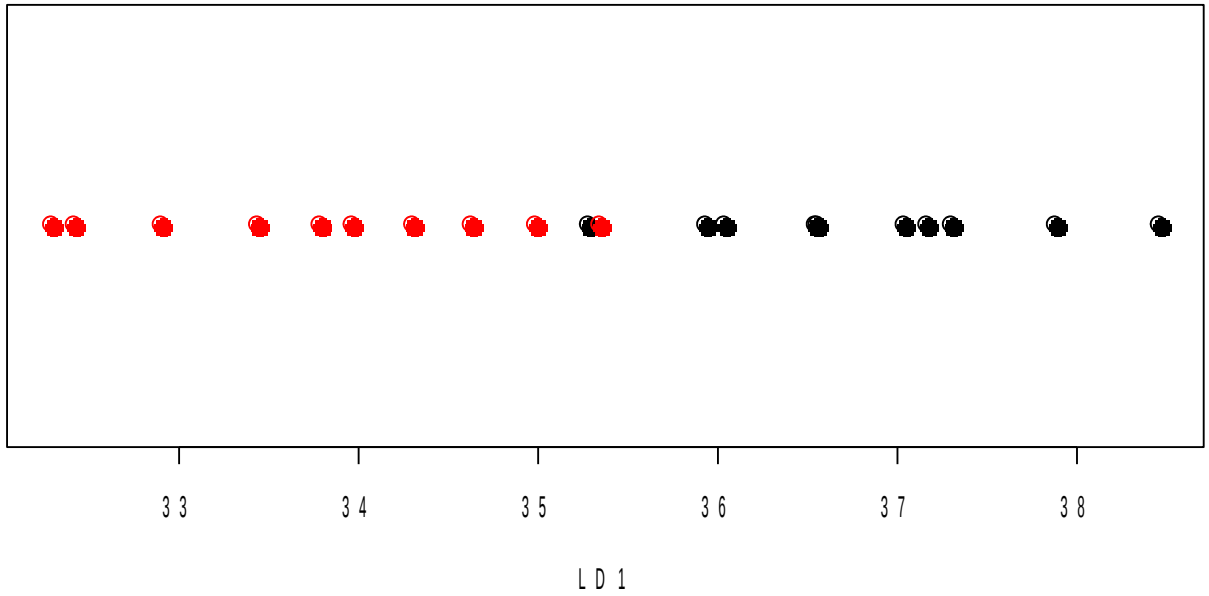
Exemple « morpho »

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...

Exemple « iris »

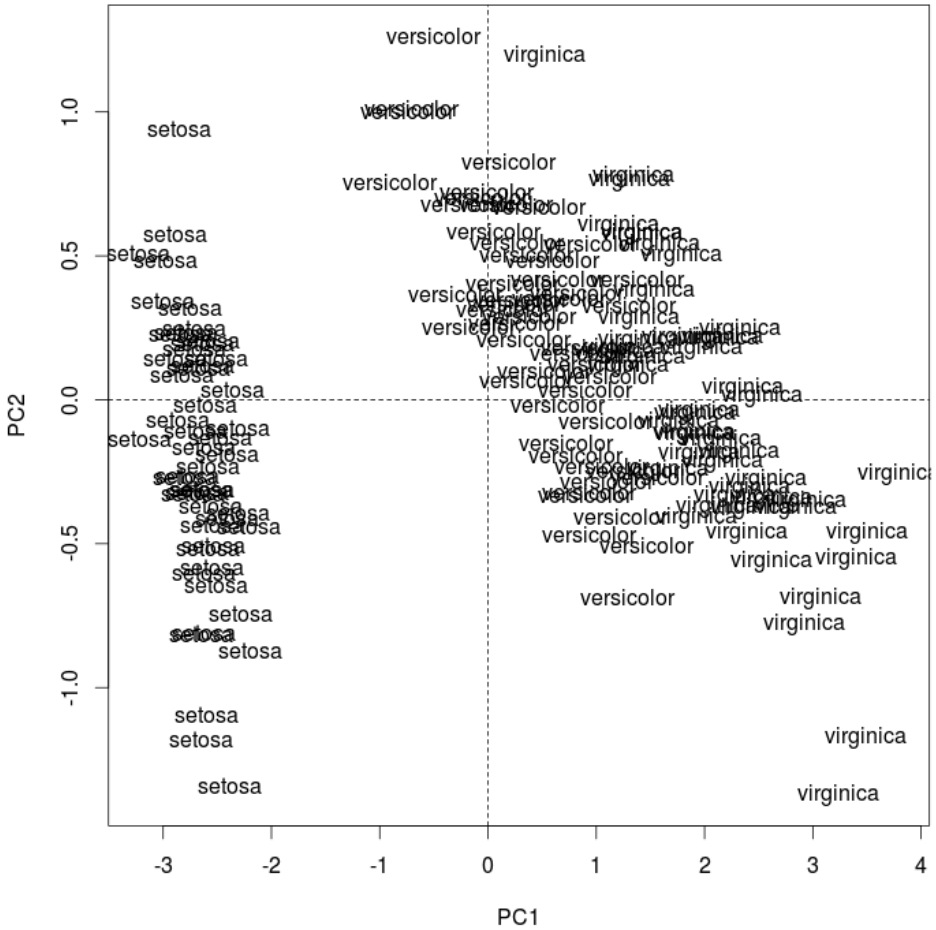
	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
45	5.1	3.8	1.9	0.4	setosa
46	4.8	3.0	1.4	0.3	setosa
47	5.1	3.8	1.6	0.2	setosa
48	4.6	3.2	1.4	0.2	setosa
49	5.3	3.7	1.5	0.2	setosa
50	5.0	3.3	1.4	0.2	setosa
51	7.0	3.2	4.7	1.4	versicolor
52	6.4	3.2	4.5	1.5	versicolor
53	6.9	3.1	4.9	1.5	versicolor
54	5.5	2.3	4.0	1.3	versicolor
55	6.5	2.8	4.6	1.5	versicolor
95	5.6	2.7	4.2	1.3	versicolor
96	5.7	3.0	4.2	1.2	versicolor
97	5.7	2.9	4.2	1.3	versicolor
98	6.2	2.9	4.3	1.3	versicolor
99	5.1	2.5	3.0	1.1	versicolor
100	5.7	2.8	4.1	1.3	versicolor
101	6.3	3.3	6.0	2.5	virginica
102	5.8	2.7	5.1	1.9	virginica
103	7.1	3.0	5.9	2.1	virginica
104	6.3	2.9	5.6	1.8	virginica
105	6.5	3.0	5.8	2.2	virginica
145	6.7	3.3	5.7	2.5	virginica
146	6.7	3.0	5.2	2.3	virginica
147	6.3	2.5	5.0	1.9	virginica
148	6.5	3.0	5.2	2.0	virginica
149	6.2	3.4	5.4	2.3	virginica
150	5.9	3.0	5.1	1.8	virginica

This famous (Fisher's or Anderson's) iris data set gives the measurements in centimeters of the variables sepal length and width and petal length and width, respectively, for 50 flowers from each of 3 species of iris. The species are Iris setosa, versicolor, and virginica.

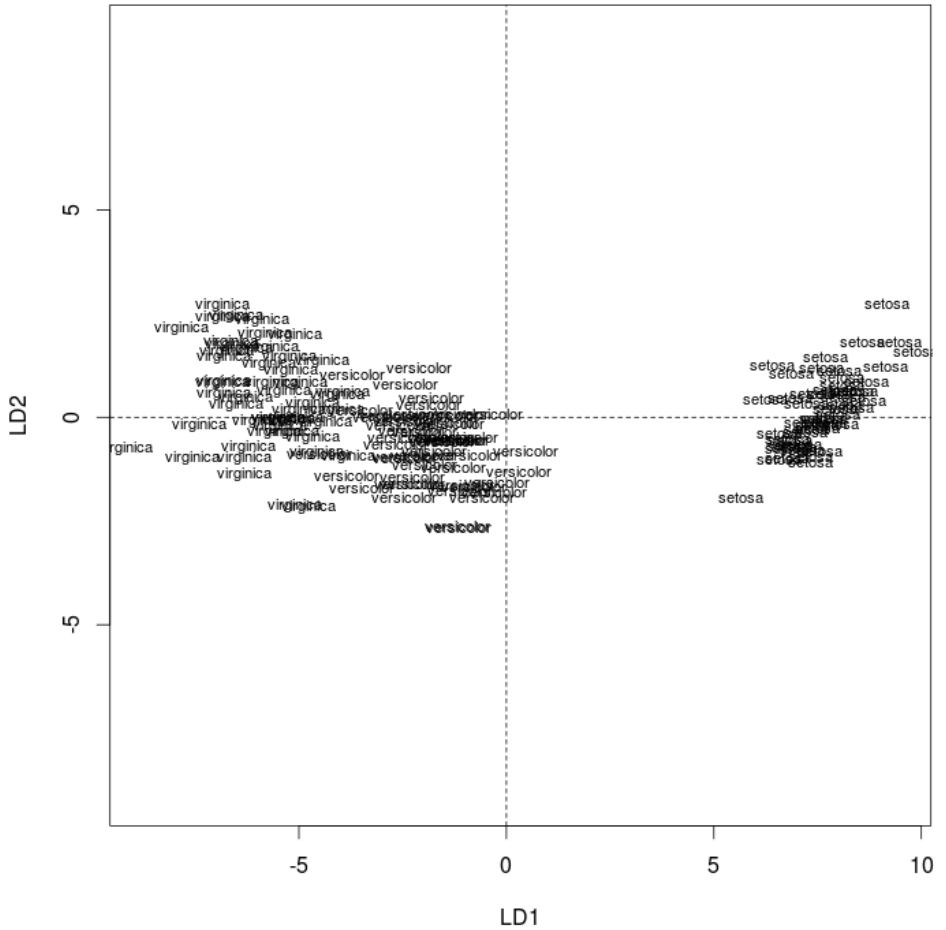
R documentation

Exemple « iris »

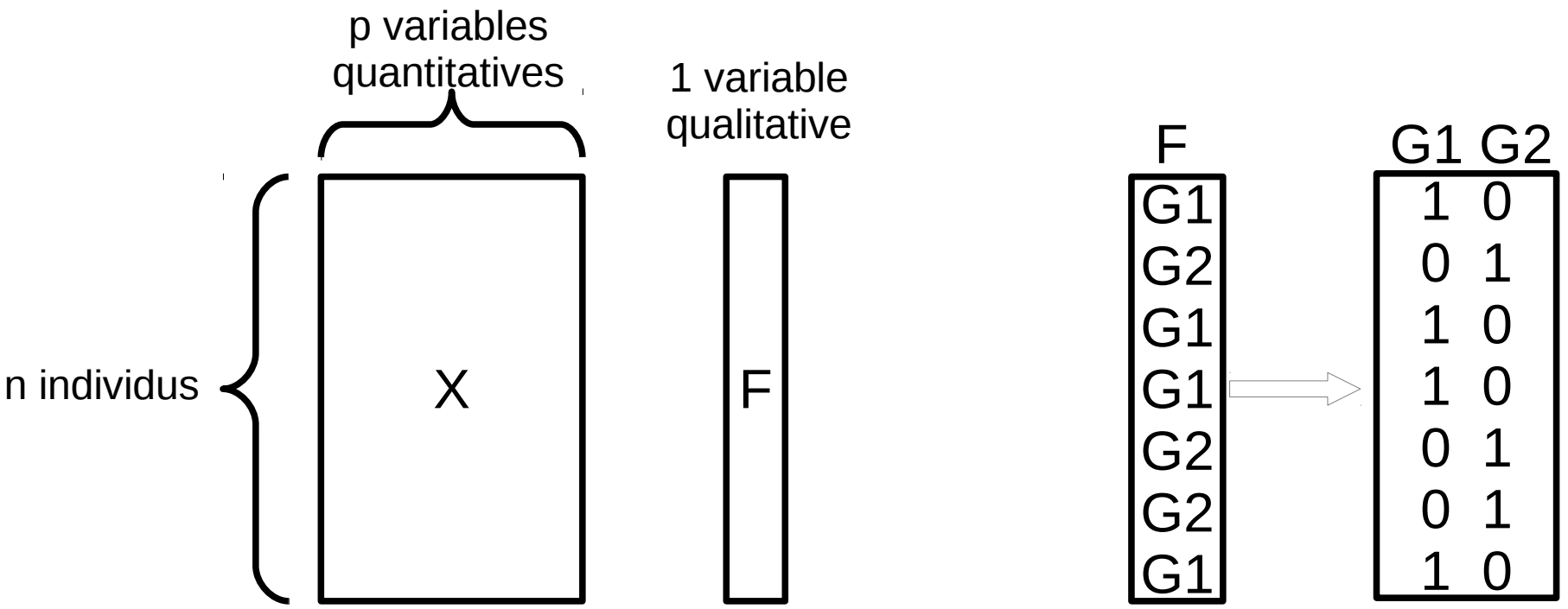
ACP



AFD

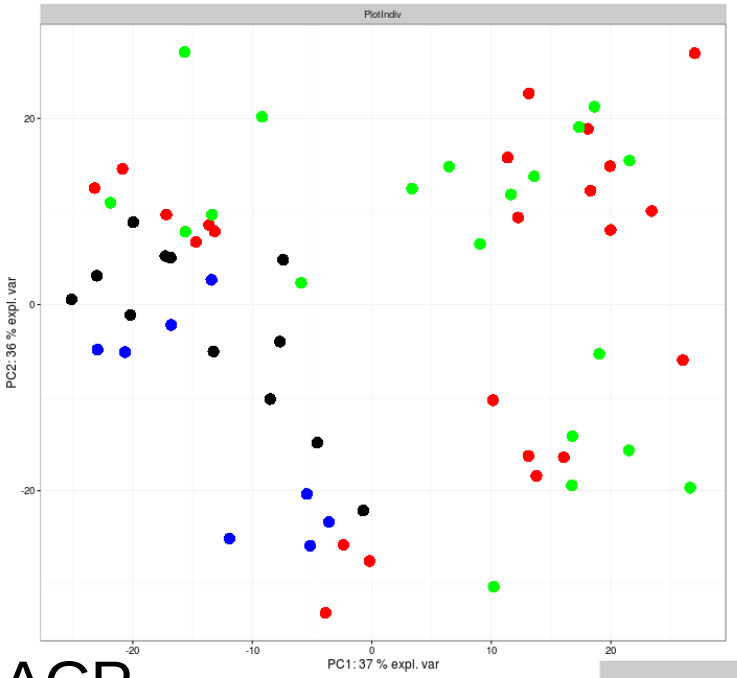


Analyse Discriminante PLS (PLS-DA)

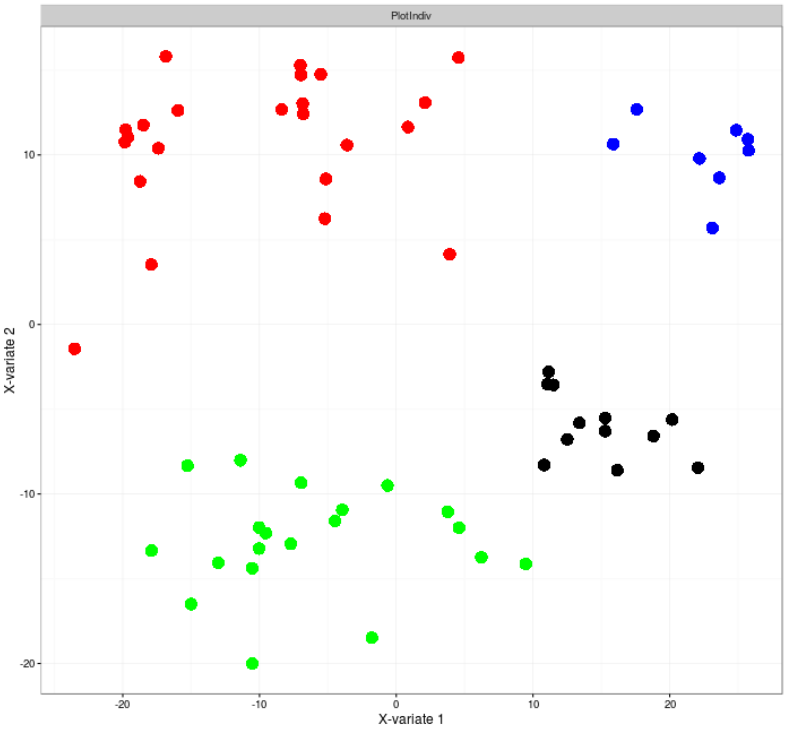


La régression PLS peut s'appliquer aussi face à un problème de discrimination. Dans ce cas (PLS-DA), la variable qualitative à expliquer est convertie en une matrice d'indicatrice.

Comparaison ACP-PLSDA

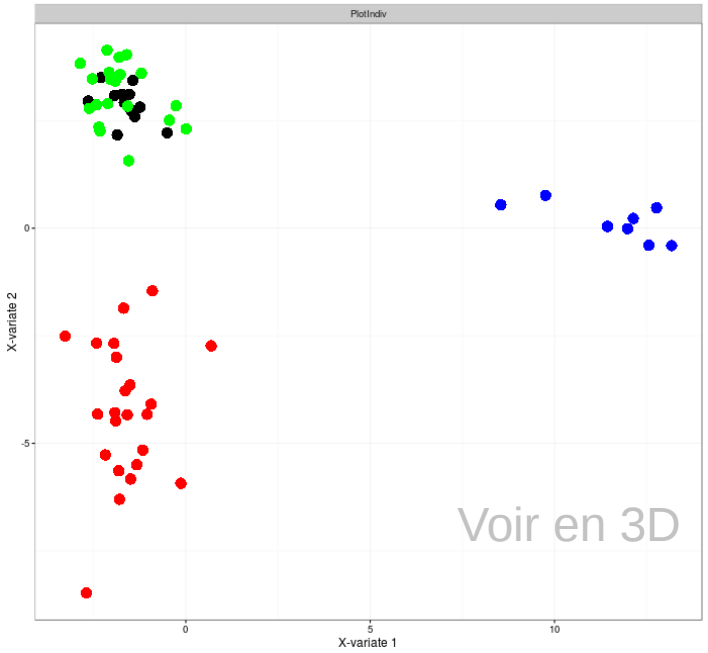


ACP



PLS-DA

The Small Round Blue Cell Tumors dataset from Khan et al., (2001) contains information of 63 samples and 2308 genes. The samples are distributed in four classes as follows: 8 Burkitt Lymphoma (BL), 23 Ewing Sarcoma (EWS), 12 neuroblastoma (NB), and 20 rhabdomyosarcoma (RMS).

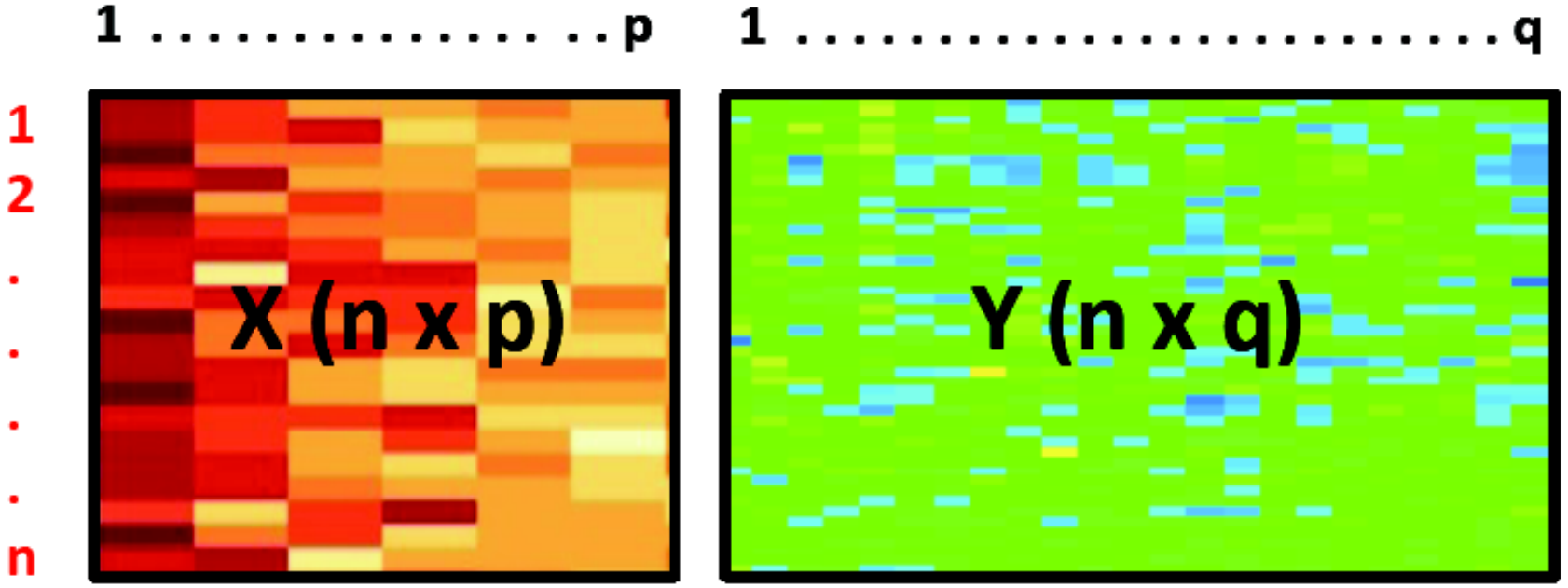


PLS-DA avec sélection de variables (voir extensions *sparse*)

Intégration de données

Objectif

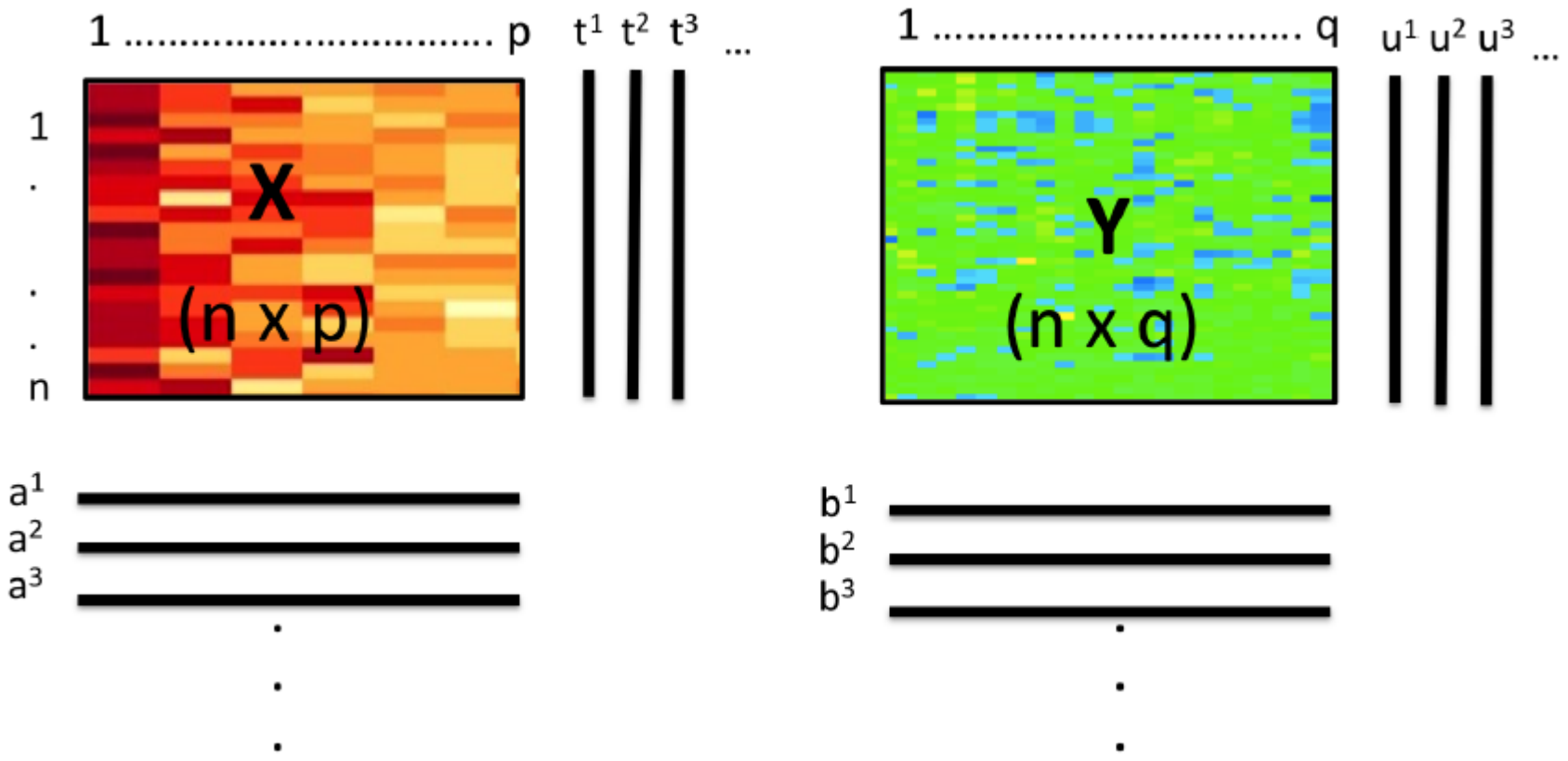
The two types of variables are measured on the same matching samples: X (n x p) and Y (n x q), n << p + q



Aims:

- Understand the correlation/covariance structure between two data sets
- Select co-regulated biological entities across samples

Principe

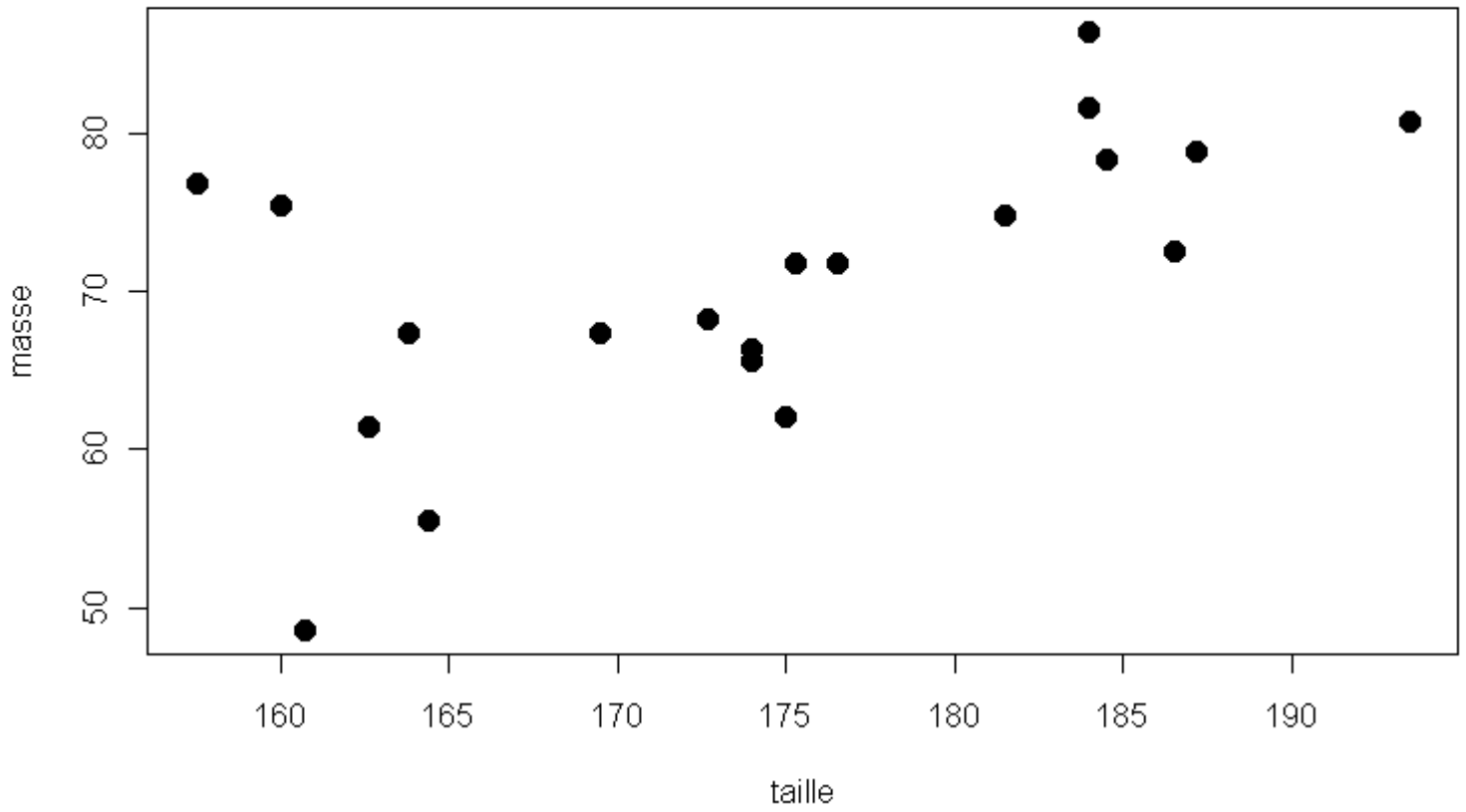


Methods generate a set of components(*) and loading(*) vectors associated to each dataset and are unsupervised.

(*) annoyingly they have different names for different methods

Régression linéaire simple

Taille (cm) :	174.0	175.3	193.5	186.5	187.2	181.5	184.0	184.5	175.0	184.0	169.5	160.0	172.7	162.6	157.5	176.5	164.4	160.7	174.0	163.8
Masse (kg) :	65.6	71.8	80.7	72.6	78.8	74.8	86.4	78.4	62.0	81.6	67.3	75.5	68.2	61.4	76.8	71.8	55.5	48.6	66.4	67.3



Peut-on « modéliser » « correctement » par une droite la masse des individus en fonction de la taille ?

Régression linéaire simple

Équation d'une droite : $Y = aX + b + \varepsilon$

Comment déterminer a et b ?

Par exemple, critère des moindres carrés : trouver a et b qui minimisent

$$\Sigma_i (y_i - ax_i - b)^2 = \Sigma_i \varepsilon_i^2$$

On peut montrer que $\hat{a} = cov(X,Y)/var(X)$ et $b = y - \hat{a}x$

Sur l'exemple : $\hat{a} = 0.5445$; $b = -24.37$

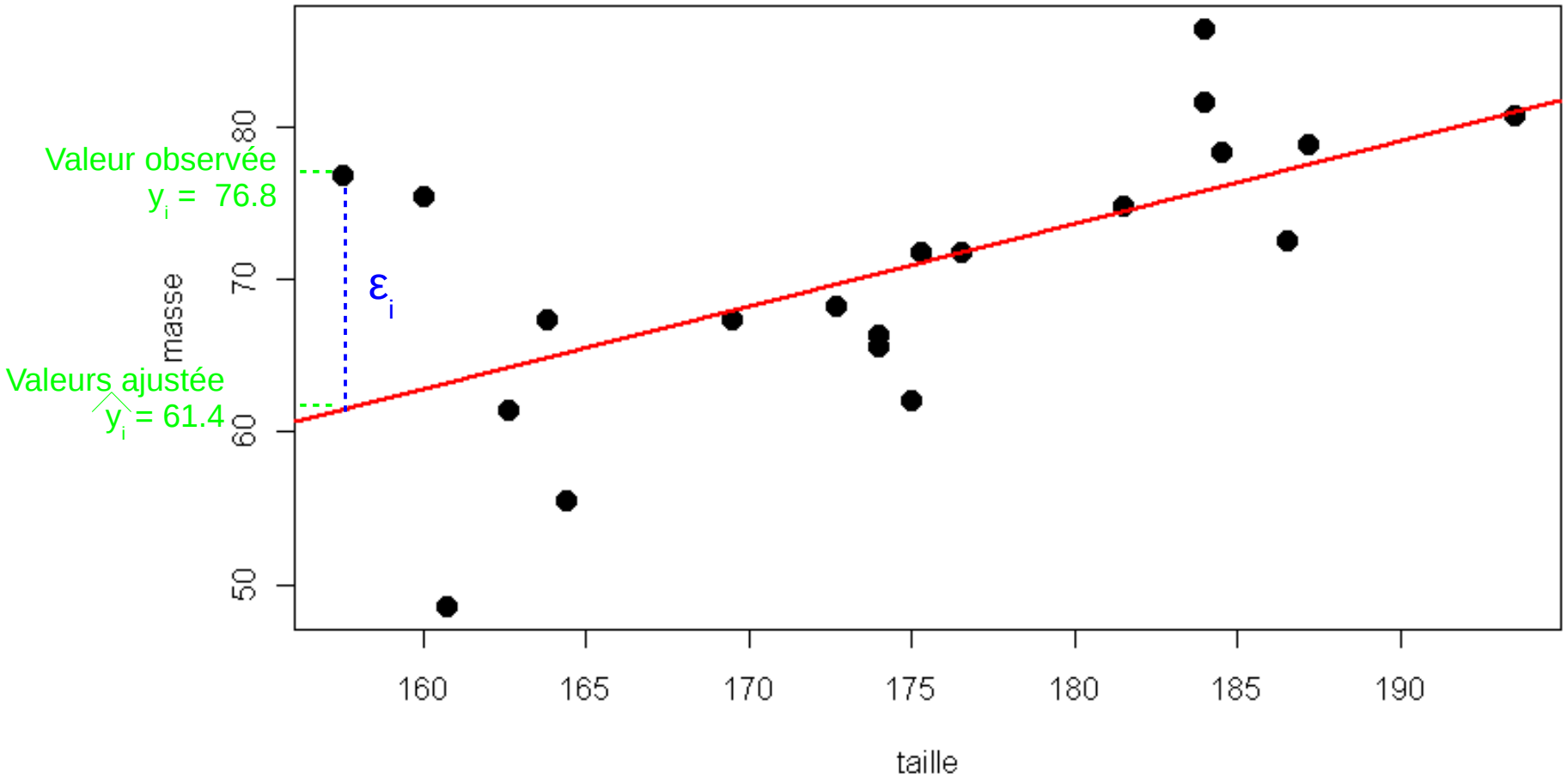
Un individu qui mesure 180cm pèse, selon le modèle, 180*0.5445-24.37 = 73.6kg

Régression linéaire simple

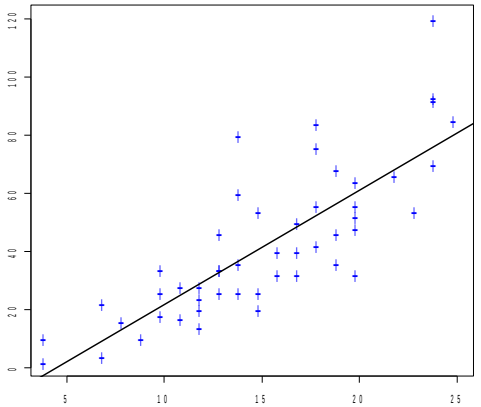
Équation de la droite :

$Y = 0.5445\ X - 24.37$

$R^2 = 0{,}3782 = var(y_i) / var(y_i)$

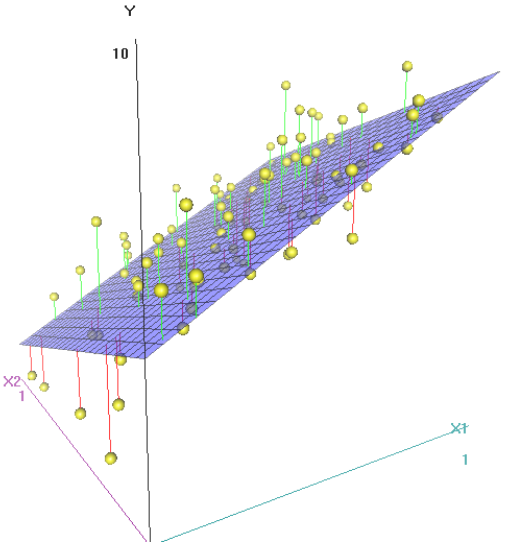


Régression linéaire multiple



Régression linéaire simple :

- **1** variable à expliquer (Y) par **1** variable explicative (X)
- trouver les paramètres a0 (ordonnée à l'origine) et a1 (pente) de la **droite** qui passe « au mieux » dans le nuage de points de Y en fonction de X
- Y = a0 + a1 X



Régression linéaire « double » :

- **1** variable à expliquer (Y) par **2** variables explicatives (X1 et X2)
- trouver les paramètres a0, a1 et a2 du **plan** qui passe « au mieux » dans le nuage de points de Y en fonction de X1 et X2
- Y = a0 + a1 X1 + a2 X2

Régression linéaire multiple :

- **1** variable à expliquer (Y) par **p** variables explicatives (X1, ... Xp)
- trouver les paramètres a0, a1, … ap de l'**hyperplan** qui passe « au mieux » dans le nuage de points de Y en fonction de X1, X2, ... et Xp
- Y = a0 + a1 X1 + a2 X2 + … + ap Xp

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

Régression linéaire multiple

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

$$\hat{Y}_i = \hat{a}_0 + \hat{a}_1 X_1 + \hat{a}_2 X_2 + \ldots + \hat{a}_p X_p$$

- Qualité du modèle

(part de variance expliquée par le modèle)

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

- Sélection de variables (choix de modèles) : R2 ajusté, Cp 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-08

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 ***
```

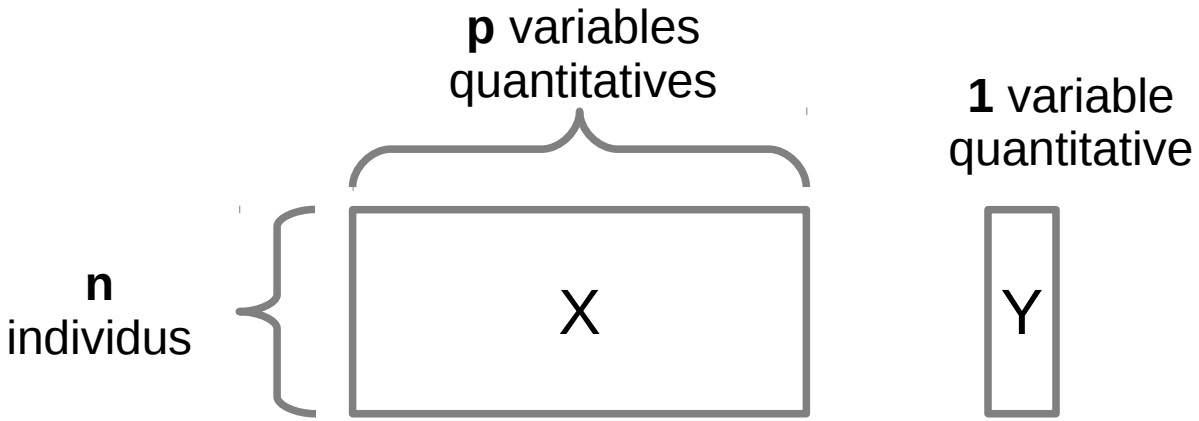
Régression linéaire multiple

- Limites
  - Nombre d'observations nécessaires supérieur au nombre de variables
  - Colinéarité des variables

- Alternatives
  - Sélection de variables type ascendant (forward)
    - limite : certaines variables explicatives ne feront plus partie du modèle
  - Régression sur composantes principales (RCP)
  - Régression PLS

RCP et PLS : principe

M. Tenenhaus. La régression PLS - Théorie et pratique. 1998, Technip



- Transformation de la matrice X en une matrice T (n*k, k<m)
  T = XW (combinaison linéaire)

- Modélisation de Y en fonction en fonction de T

RCP

- T : matrice des composantes principales
- Régression sur les CP

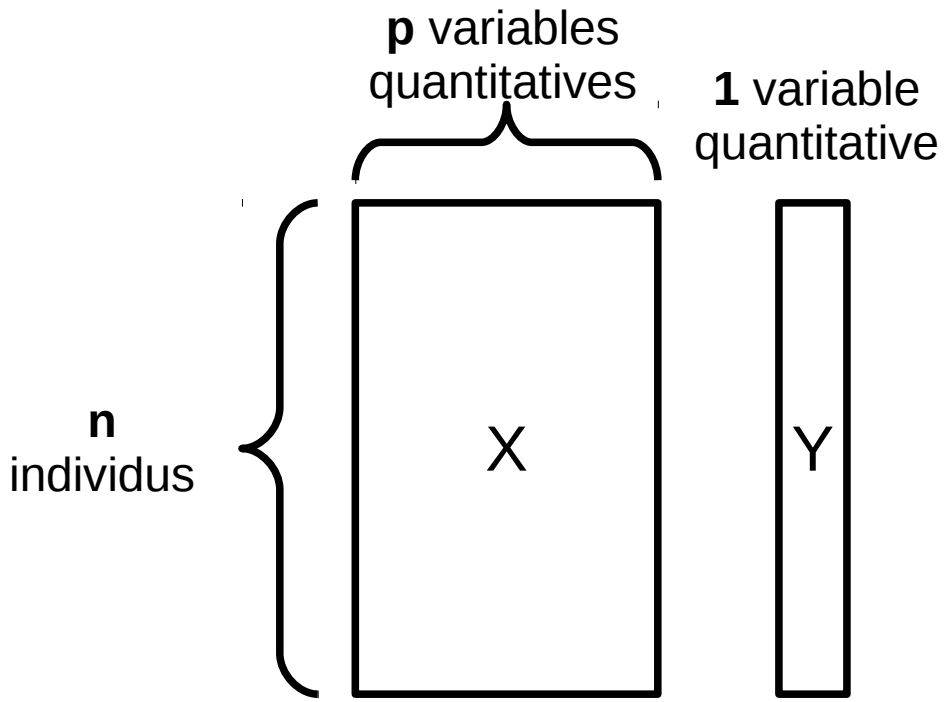
$$y = q_1 t_1 + q_2 t_2 + \ldots + q_k t_k$$

PLS

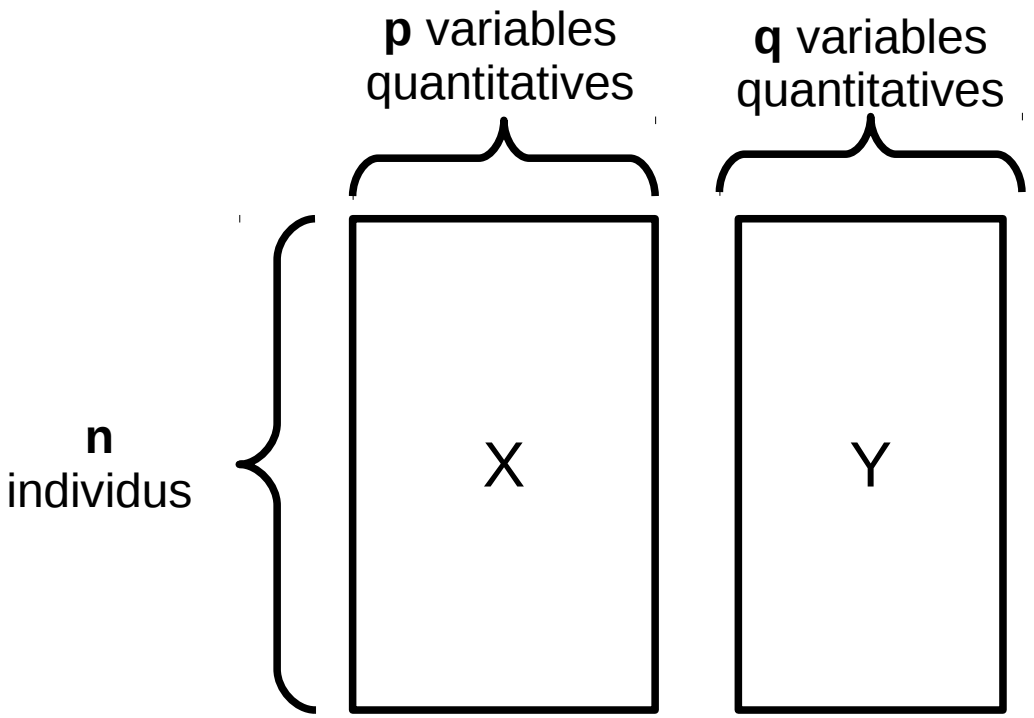
- **Le calcul de T tient compte de Y**
- Double modélisation

$$X = TP + R$$
$$Y = TQ + F$$

Méthodes PLS



PLS 1



PLS 2

Régression PLS1

- Construire une première composante t_1 :

$$t_1 = w_{11}x_1 + \ldots + w_{1p}x_p$$

- Régression simple de y sur t1

$$y = c_1 t_1 + y_1$$

- D'où : $y = c_1 w_{11} x_1 + \ldots + c_1 w_{1p} x_p + y_1$
- Pour ajouter, si nécessaire, une deuxième composante t_2 (non corrélée à t1) :

$$t_2 = w_{21}x_{11} + \ldots + w_{2p}x_{1p}$$

où les x_{1j} sont les résidus des régressions des variables x_j sur t_1.

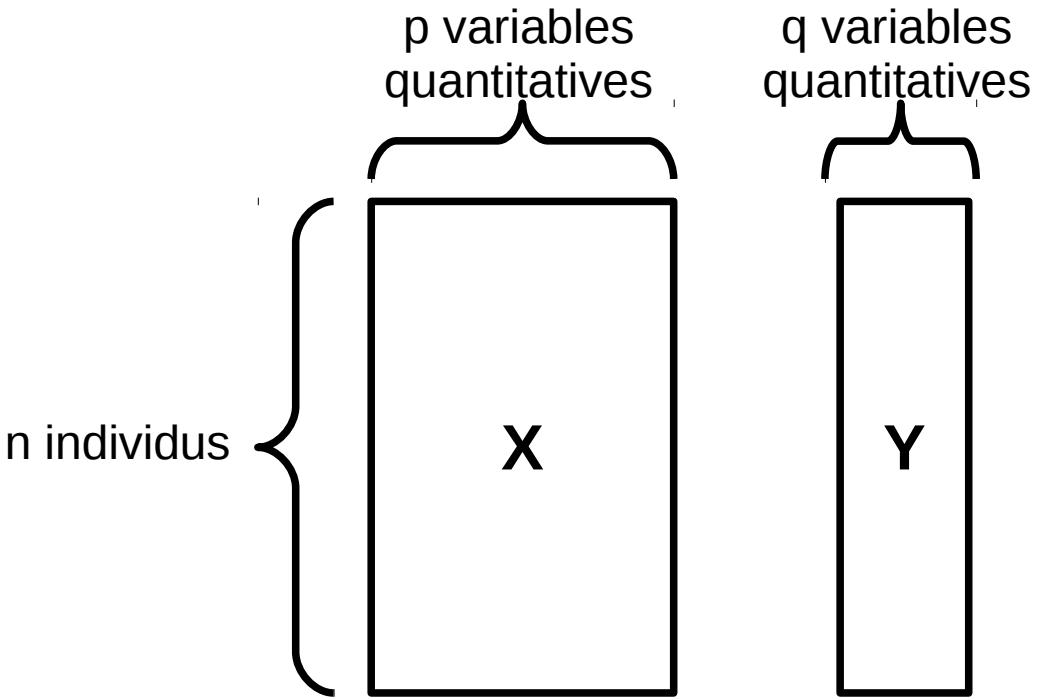
- Nouvelle régression : $y = c_1 t_1 + c_2 t_2 + y_2$
- ...

Régression PLS2

- La régression PLS s'applique aussi au cas où Y est un ensemble de variables quantitatives. On recherche dans ce cas des combinaisons linéaires de chaque paquet de variables ayant la plus grande covariance possible.
- Analogie avec l'analyse des corrélations canoniques (CCA) : recherche des combinaisons linéaires de variables de chaque paquet ayant la plus grande corrélation possible.

Analyse des corrélations Canoniques (CCA)

Objectif : décrire les relations entre deux tableaux de données constitués de variables **quantitatives**.



CCA : exemple simulé

X

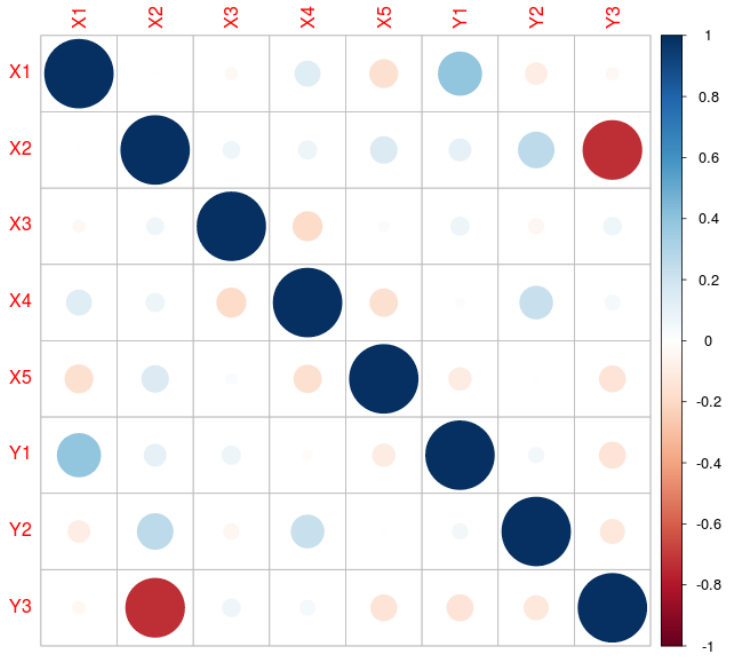
X1	X2	X3	X4	X5
0.87	0.31	0.24	0.06	0.29
0.76	0.8	0.52	0.1	0.95
0.65	0.76	0.57	0.1	0.17
0.86	0.47	0.00	0.21	0.75
0.65	0.46	0.41	0.23	0.86
0.11	0.56	0.84	0.14	0.49
0.85	0.81	0.42	0.65	0.39
0.74	0.73	0.15	0.81	0.80
0.75	0.30	0.72	0.48	0.99
0.55	0.06	0.30	0.87	0.67
0.41	0.52	0.21	0.51	0.59
0.59	0.87	0.99	0.67	0.28
0.34	0.35	0.56	0.03	0.56
0.07	0.02	0.59	0.04	0.54
0.17	0.08	0.50	0.37	0.89
0.39	0.54	0.53	0.65	0.46
0.06	0.17	0.28	0.82	0.46
0.22	0.83	0.90	0.17	0.49
0.83	0.27	0.51	0.38	0.55
0.02	0.51	0.56	0.34	0.99
0.04	0.46	0.81	0.47	0.46
0.32	0.95	0.65	0.10	0.43
0.42	0.27	0.17	0.36	0.37
0.39	0.68	0.94	0.79	0.87
0.48	0.30	0.83	0.60	0.22
0.84	0.25	0.54	0.00	0.52
0.31	0.14	0.33	0.48	0.38
0.15	0.80	0.09	0.87	0.29
0.99	0.07	0.81	0.96	0.01
0.26	0.21	0.20	0.24	0.66
0.99	0.07	0.86	0.84	0.36
0.91	0.19	0.82	0.04	0.25
0.46	0.17	0.48	0.38	0.02
0.95	0.94	0.41	0.83	0.48
0.80	0.34	0.54	0.72	0.58
0.09	0.01	0.81	0.02	0.63
0.93	0.75	0.54	0.79	0.90
0.78	0.99	0.67	0.08	0.84
0.83	0.05	0.04	0.70	0.41
0.97	0.68	0.37	0.88	0.34
0.13	0.35	0.16	0.95	0.81
0.5	0.04	0.17	0.49	0.15
0.37	0.64	0.55	0.96	0.14
0.01	0.98	0.48	0.94	0.76
0.40	0.44	0.80	0.40	0.94
0.44	0.67	0.67	0.42	0.20
0.92	0.07	0.48	0.92	0.06
0.30	0.39	0.54	0.23	0.92
0.60	0.75	0.22	0.60	0.50
0.25	0.77	0.02	0.51	0.18

Y

Y 1	Y2	Y3
0.71	0.33	-0.53
0.62	0.07	-0.78
0.77	0.10	-0.52
0.49	0.57	-1.09
0.76	0.67	-0.30
0.53	0.84	-0.55
0.71	0.57	-0.75
0.24	0.89	-0.50
1.62	0.18	-0.80
-0.51	0.16	0.25
0.29	0.72	-0.61
1.11	0.80	-0.95
0.49	0.27	-0.06
0.51	0.02	-0.46
0.20	0.48	0.36
0.27	0.88	-0.48
0.61	0.98	-0.51
0.02	0.82	-0.74
0.40	0.08	-0.39
-0.53	0.46	-0.69
0.49	0.59	-0.28
0.07	0.61	-1.19
0.06	0.51	-0.31
0.05	0.76	-0.18
-0.25	0.25	-0.13
0.96	0.11	-1.58
0.24	0.74	0.41
0.23	0.89	-1.57
0.06	0.76	-0.29
0.42	0.61	-0.22
0.64	0.09	0.12
1.44	0.08	0.12
1.12	0.70	0.18
1.29	0.58	-1.37
1.60	0.51	-0.38
-0.02	0.23	0.05
-0.01	0.65	-1.20
1.12	0.81	-1.12
1.53	0.87	0.09
1.15	0.71	-0.52
0.28	0.23	-0.07
-0.89	0.20	0.25
1.15	0.73	-0.48
0.60	0.01	-1.49
0.28	0.64	0.23
0.71	0.61	-1.18
0.98	0.24	0.71
1.01	0.83	-0.51
0.09	0.56	-1.04
0.67	0.15	-0.87

Matrice de corrélation (X,Y)

	X1	X2	X3	X4	X5	Y1	Y2	Y3
X1	1.00	0.00	-0.03	0.13	-0.17	0.40	-0.10	-0.03
X2	0.00	1.00	0.06	0.07	0.15	0.10	0.27	-0.74
X3	-0.03	0.06	1.00	-0.18	0.02	0.07	-0.05	0.07
X4	0.13	0.07	-0.18	1.00	-0.16	-0.02	0.23	0.05
X5	-0.17	0.15	0.02	-0.16	1.00	-0.11	0.01	-0.14
Y1	0.40	0.10	0.07	-0.02	-0.11	1.00	0.05	-0.15
Y2	-0.10	0.27	-0.05	0.23	0.01	0.05	1.00	-0.12
Y3	-0.03	-0.74	0.07	0.05	-0.14	-0.15	-0.12	1.00

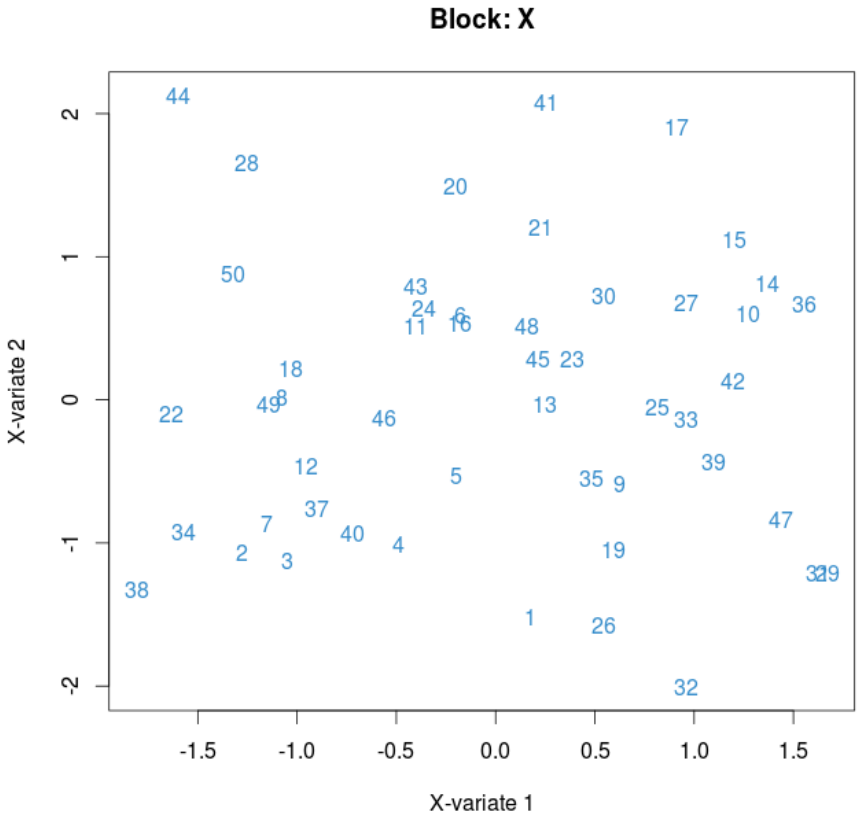


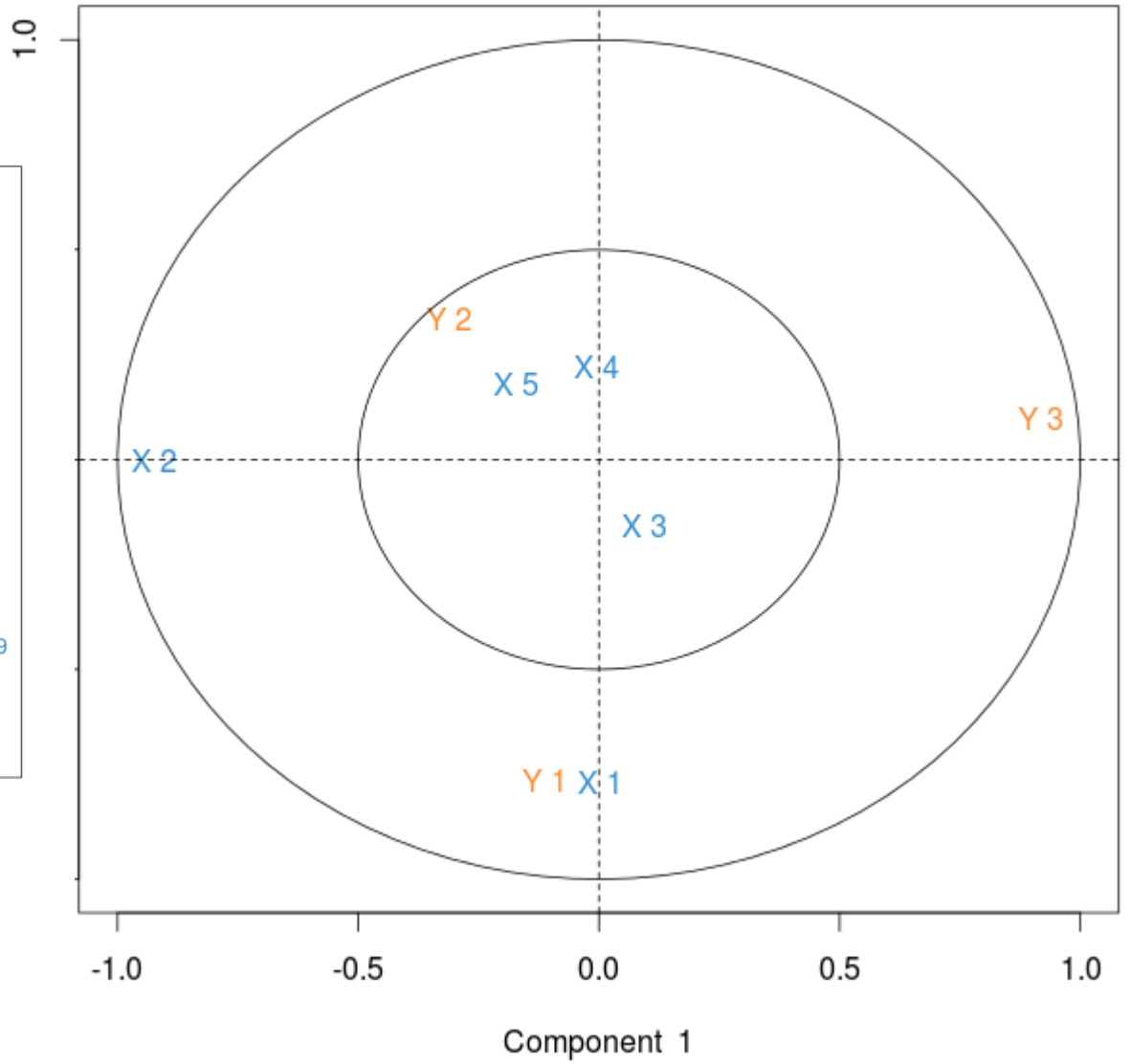
Package R
`corrplot`

CCA : représentations graphiques

Représentation des variables

Représentation des individus





CCA : principe

- Le principe de l'ACC peut-être vu comme un algorithme itératif
  - Maximiser la corrélation (ρ_1) entre des combinaisons linéaires des variables de X (t_1) d'une part et des variables de Y (u_1) d'autre part.

$$t_1 = a_{11}X_1 + a_{12}X_2 + \ldots + a_{1p}X_p$$

$$u_1 = b_{11}Y_1 + b_{12}Y_2 + \ldots + b_{1q}Y_q$$

$$\rho_1 = cor(t_1,u_1) = \max_{t,u} cor(t,u)$$

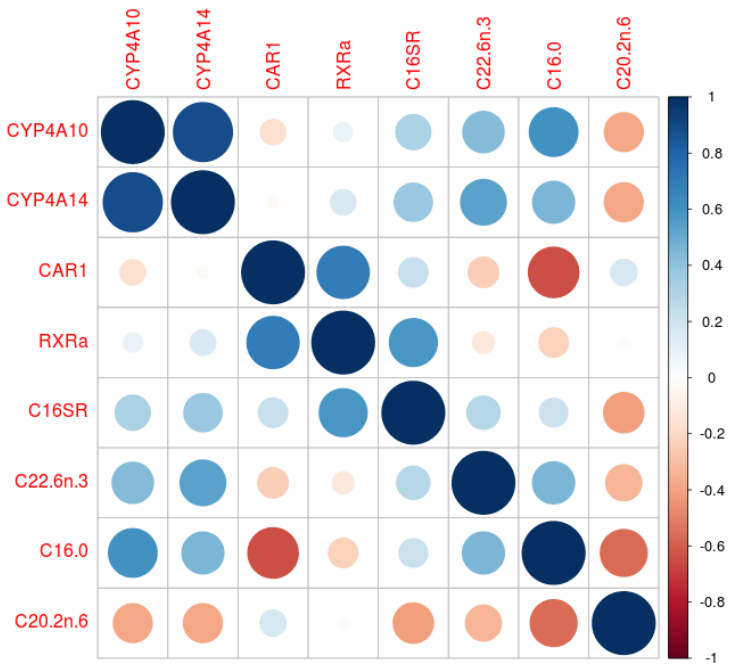
t_1 et u_1 sont les premières variables canoniques et p_1 est la première corrélation canonique

  - Pour les ordres suivants, itérer le processus sous des contraintes d'orthogonalité avec les ordres précédents
- L'ACC est similaire à l'ACP pour la construction et l'interprétation des sorties graphiques
- Les calculs se font par une décomposition en éléments propres de matrices particulières

CCA : exemple nutrimouse

- 40 souris (2 génotypes)
- Expression de 5 gènes
- Concentration de 3 lipides

Question : quelles sont les relations entre les gènes et les lipides ?

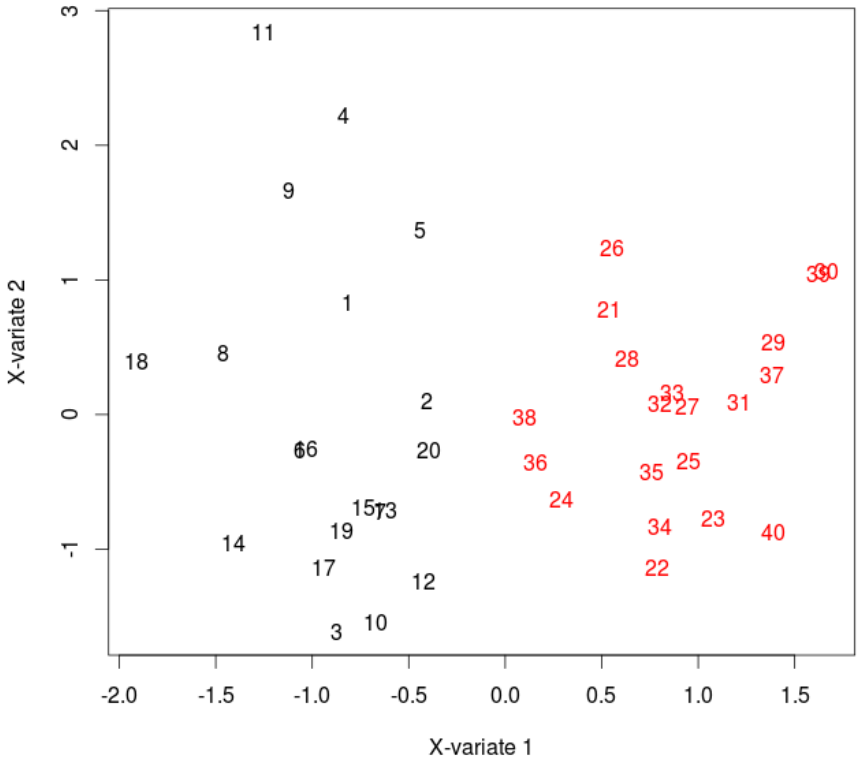


Matrice de corrélation

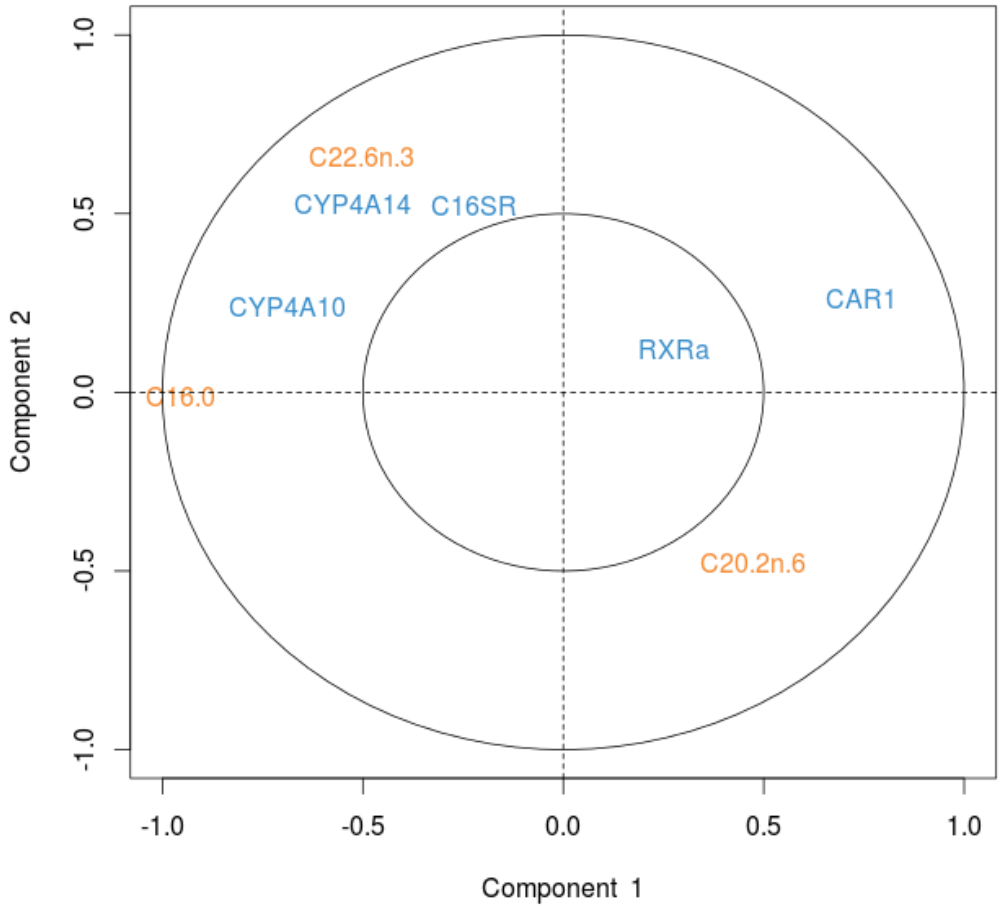
CYP4A10	CYP4A14	CAR1	RXRa	C16SR	C22.6n.3	C16.0	C20.2n.6
-0.81	-0.81	-0.97	-0.67	1.66	10.39	26.45	0.00
-0.88	-0.84	-0.92	-0.59	1.65	2.61	24.04	0.30
-0.71	-0.98	-0.98	-0.68	1.57	2.51	23.70	0.33
-0.65	-0.41	-0.97	-0.72	1.61	14.99	25.48	0.00
-1.16	-1.16	-1.06	-0.78	1.66	6.69	24.80	0.23
-0.99	-1.09	-1.03	-0.62	1.70	2.56	26.04	0.00
-0.62	-0.76	-0.91	-0.65	1.58	9.84	25.94	0.00
-0.82	-0.87	-1.11	-0.76	1.62	10.40	28.63	0.00
-0.48	-0.37	-0.85	-0.55	1.72	16.36	25.34	0.00
-0.79	-0.95	-0.99	-0.67	1.55	1.86	28.49	0.00
-0.51	-0.15	-0.92	-0.60	1.69	16.21	25.73	0.00
-1.00	-1.13	-1.02	-0.69	1.57	6.61	24.28	0.21
-0.88	-0.99	-0.99	-0.67	1.60	3.27	24.63	0.36
-1.05	-1.15	-1.19	-0.75	1.59	7.04	26.04	0.19
-0.72	-0.73	-0.93	-0.58	1.61	2.71	24.76	0.35
-0.67	-0.85	-0.99	-0.72	1.60	10.96	26.46	0.00
-1.19	-1.22	-1.15	-0.69	1.60	1.99	23.45	0.00
-0.56	-0.73	-0.95	-0.55	1.78	17.35	29.72	0.00
-1.03	-1.10	-1.02	-0.59	1.67	2.44	27.00	0.00
-1.01	-1.06	-1.01	-0.70	1.60	5.97	24.09	0.23
-1.21	-1.17	-0.91	-0.67	1.65	0.64	23.59	0.05
-1.15	-1.29	-0.90	-0.69	1.55	2.16	19.95	0.31
-1.22	-1.25	-0.88	-0.67	1.55	1.70	17.64	0.61
-1.15	-1.19	-0.90	-0.58	1.65	11.56	22.73	0.27
-1.16	-1.18	-0.87	-0.67	1.57	0.91	14.65	0.83
-0.93	-0.90	-0.73	-0.52	1.74	1.22	20.49	0.32
-1.13	-1.10	-0.83	-0.62	1.61	3.44	18.44	0.09
-1.09	-1.08	-0.85	-0.63	1.64	4.02	17.72	0.12
-1.33	-1.22	-0.85	-0.66	1.60	13.26	21.70	0.24
-1.18	-1.08	-0.74	-0.63	1.62	4.45	16.25	0.10
-1.18	-1.14	-0.84	-0.67	1.57	1.16	22.91	0.00
-0.96	-1.05	-0.70	-0.49	1.72	0.28	23.27	0.00
-1.07	-1.03	-0.83	-0.63	1.60	1.41	20.25	0.33
-1.12	-1.11	-0.84	-0.57	1.60	1.11	20.18	0.54
-1.22	-1.15	-0.90	-0.62	1.59	11.57	20.71	0.24
-1.05	-0.96	-0.88	-0.53	1.65	0.64	21.79	0.07
-1.07	-1.03	-0.73	-0.58	1.62	2.29	21.57	0.11
-1.23	-1.18	-0.98	-0.64	1.64	16.28	25.23	0.26
-1.08	-1.12	-0.63	-0.53	1.72	3.87	16.20	0.13
-1.13	-1.14	-0.79	-0.61	1.55	1.83	20.70	0.59

CCA : exemple nutrimouse

Représentation des individus
Couleur selon le génotype



Représentation des variables



Corrélations canoniques : `0.853 0.627 0.253`

CCA : une méthode fondamentale...

- Si un des groupes n'a qu'une seule variable quantitative, l'ACC est équivalente à la **régression linéaire multiple**.
- Si un des groupes est constitué de variables indicatrices d'une variable qualitative et l'autre de variables quantitatives, l'ACC est équivalente à une **analyse discriminante**.
- Si les deux groupes de variables sont composées d'indicatrices de variables qualitatives, l'ACC est équivalente à l'**analyse des correspondances**.

… qui a ses limites

- La CCA ne peut fonctionner ne peut fonctionner qu'avec un nombre « suffisant » d'observations : n >> p+q
- Les variables de X et Y ne doivent pas être « trop » corrélées (X de rang p et Y de rang q)
- Alternative : version régularisée de la CCA

CCA : exemple simulé

- les variables X^1 and Y^1 sont fortement corrélées
- les variables X^2 and Y^2 sont moins fortement corrélées
- les corrélations canoniques pour X et Y sont

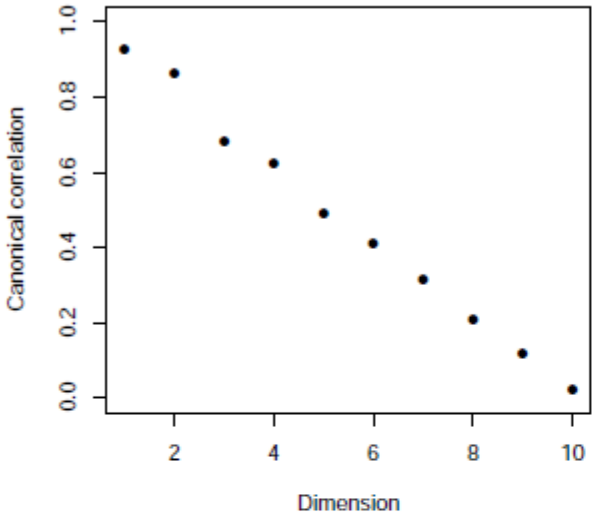
$$\rho_1 = 0.9,\ \rho_2 = 0.7 \text{ et } \rho_3 = \ldots = \rho_p = 0$$

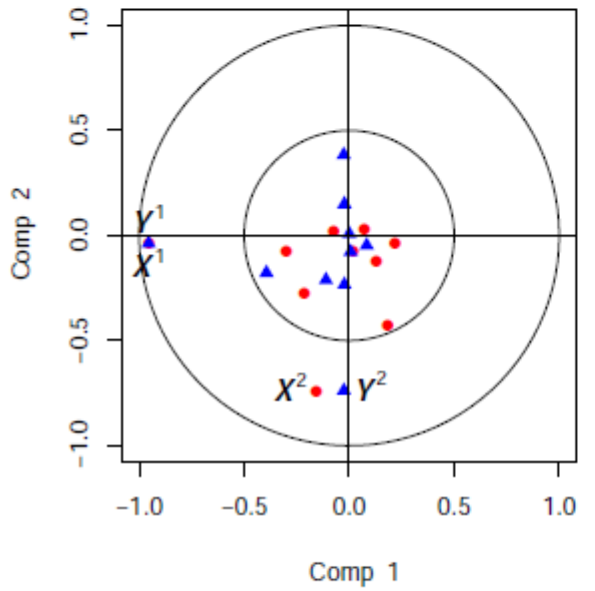
- simulations ont été réalisées pour

$$n = 50,\ p = 10 \text{ et } q = 10;\ 25 \text{ et } 39$$

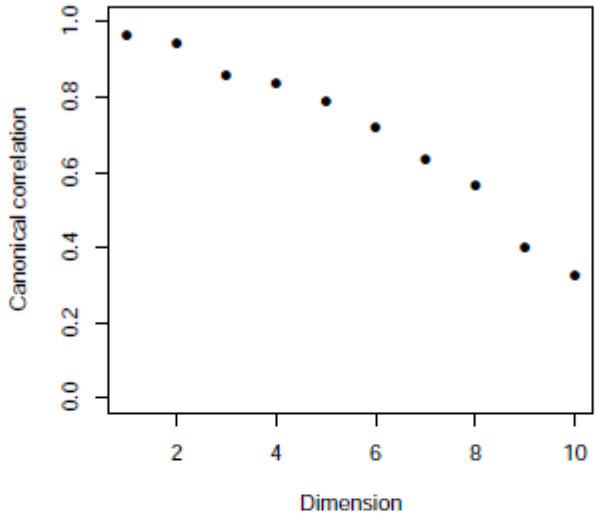
CCA : exemple simulé

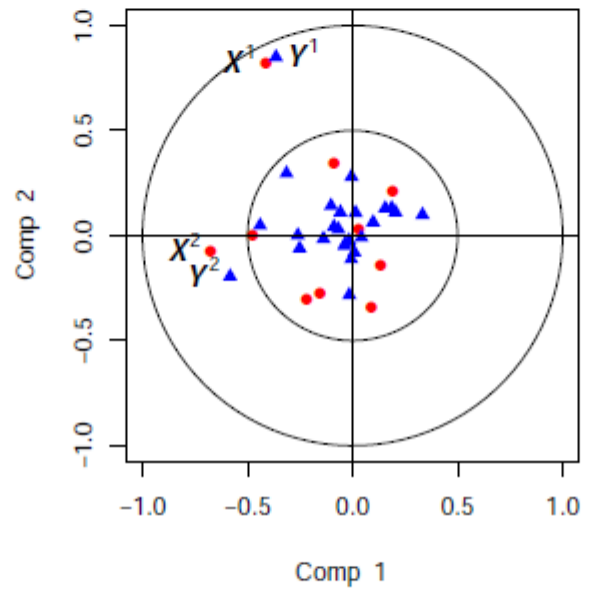
$n = 50;\ p + q = 20$



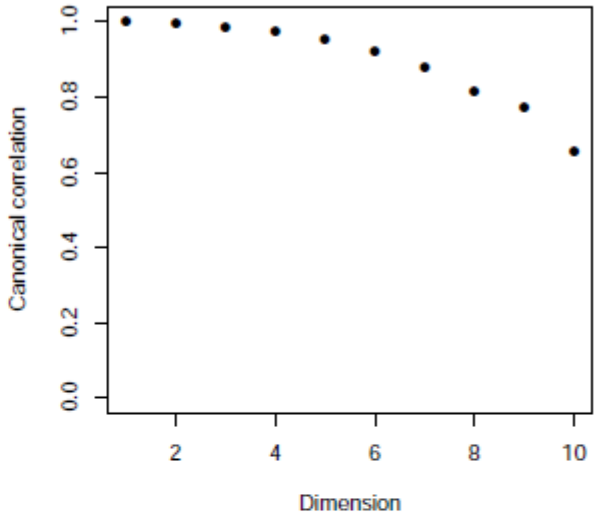


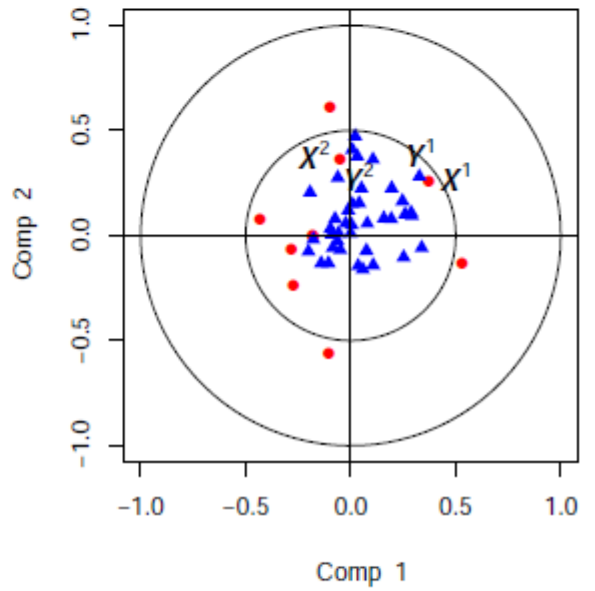
$n = 50;\ p + q = 35$



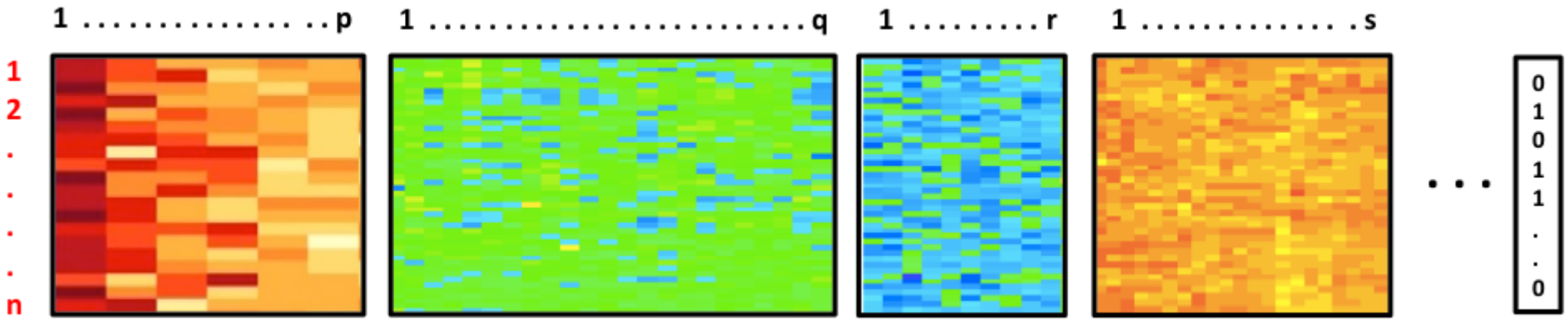


$n = 50;\ p + q = 49$





Généralisation

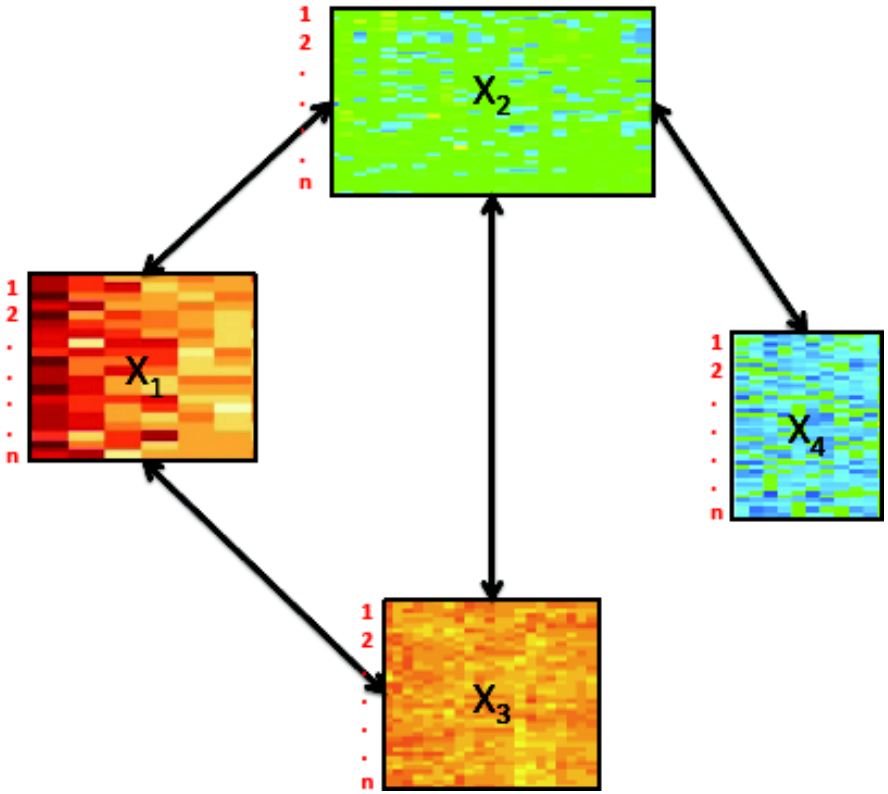


- **Generalized** CCA (GCCA): integration of more than 2 data sets ; maximizes the sum of pairwise covariances between two components at a time.
- Sparse GCCA (SGCCA): variable selection is performed on each data set

Tenenhaus, A., Philippe, C., Guillemot, V., Lê Cao K-A., Grill, J., Frouin, V. 2014, Variable selection for generalized canonical correlation analysis, Biostatistics

Définir les liens entre les jeux de données

Un lien entre 2 jeux de données indique que l'on souhaite maximiser la covariance entre ces 2 paquets de variables



Matrice de *design* exprimant ces relations

	x1	x2	x3	x4
x1	0	1	1	0
x2	1	0	1	1
x3	1	1	0	0
x4	0	1	0	0

Extensions *sparse*

Le fléau de la dimension

https://fr.wikipedia.org/wiki/Fléau_de_la_dimension

Le fléau de la dimension ou malédiction de la dimension (*curse of dimensionality*) est un terme inventé par Richard Bellman en 1961 pour désigner divers phénomènes qui ont lieu lorsque l'on cherche à analyser ou organiser des données dans des espaces de grande dimension alors qu'ils n'ont pas lieu dans des espaces de dimension moindre.

→ Les méthodes ***sparse*** (parcimonieuse) visent à gérer les problèmes liés à la grande dimension.

PARCIMONIE n. f. XVIe siècle. Emprunté du latin parsimonia, de même sens, lui-même dérivé de parcere, « épargner ». Épargne minutieuse, qui porte sur les plus petites dépenses ; mesquinerie. Il est d'une parcimonie proche de l'avarice. Loc. adv. Avec parcimonie, en mesurant de façon stricte, chichement. Accorder des subsides avec parcimonie. Fig. Décerner des louanges avec parcimonie.

Dictionnaire de l'Académie Française, http://atilf.atilf.fr

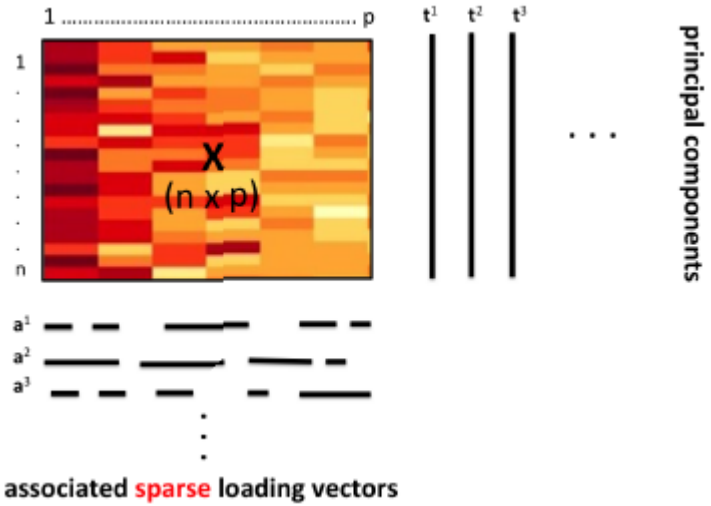
En science et en philosophie, la parcimonie est un principe consistant à n'utiliser que le minimum de causes élémentaires pour expliquer un phénomène.

https://fr.wikipedia.org/wiki/Parcimonie

Sparse PCA

High throughput experiments: too many variables, noisy or irrelevant. PCA is difficult to visualise and understand.
→ clearer signal if some of the variable weights $\{a_1, \ldots, a_p\}$ were set to 0 for the 'irrelevant' variables (small weights):

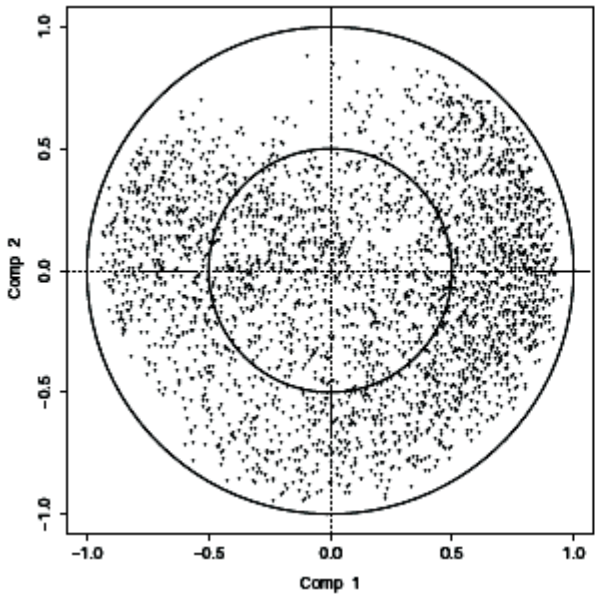
$$t = 0.x1 + a_2.x_2 + \ldots + 0.x_p$$



- Important weights : important contribution to define the Pcs.
- Null weights : those variables are not taken into account when calculating the PCs

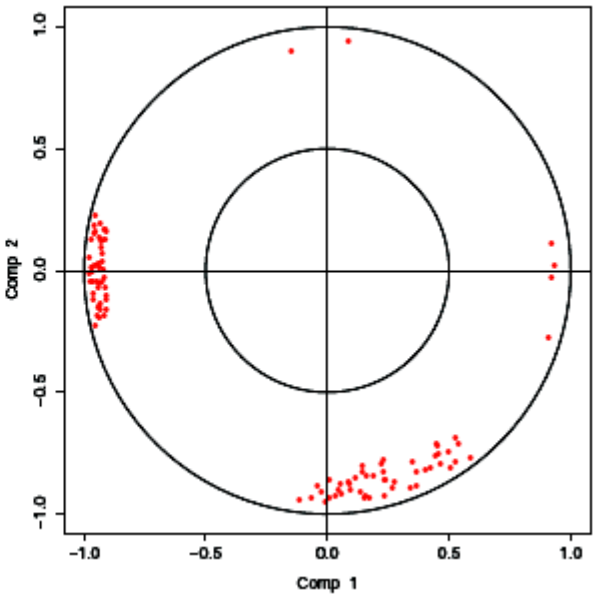
Représentations graphiques

PCA

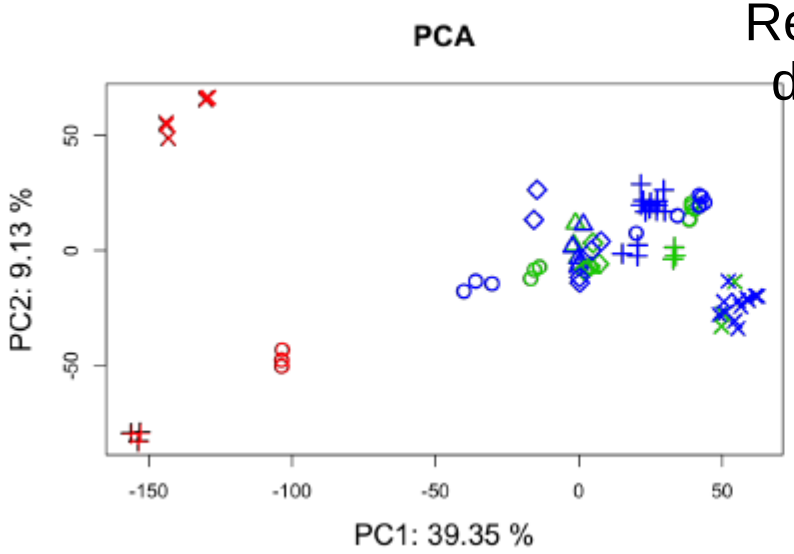


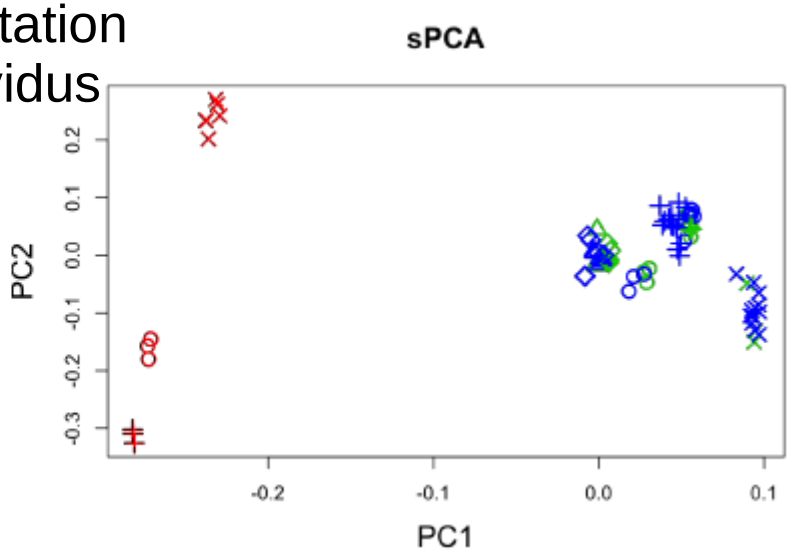
Représentation des variables

Sparse PCA



Représentation des individus





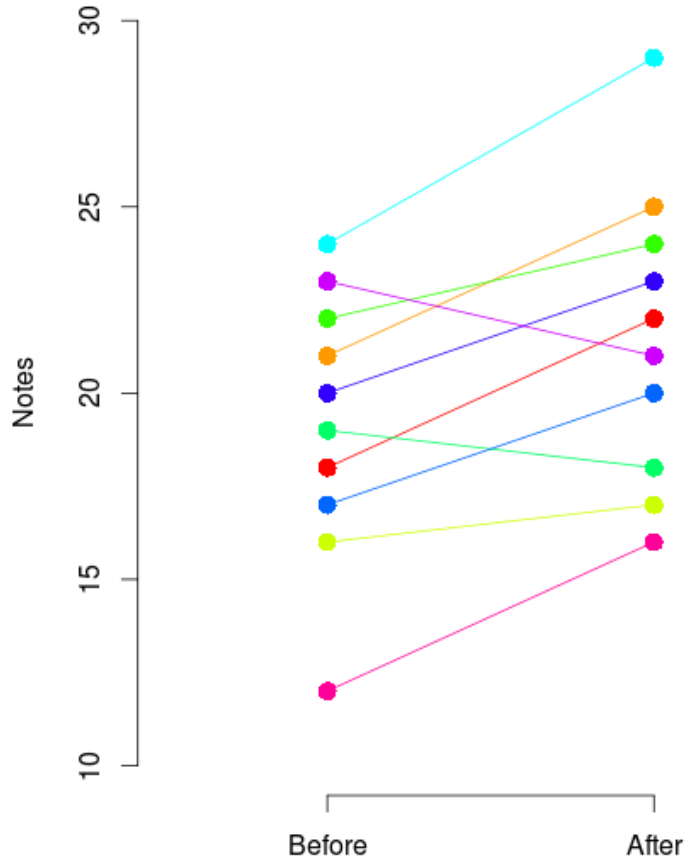
Extensions *multilevel*

Principe

- In repeated measures experiments, the subject variation can be larger than the time/treatment variation
- Multivariate projection based methodes make the assumption that samples are independent of each other
- In univariate analysis we use a paired t-test rather than a t-test
- In multivariate analysis we use a multilevel approach:
- Different sources of variation can be separated (treatment effect within subjects and differences between subjects)
- Gain in power

Données appariées

	Before	After
Louise	18	22
Léo	21	25
Emma	16	17
Gabriel	22	24
Chloé	19	18
Adam	24	29
Lola	17	20
Timéo	20	23
Inès	23	21
Raphaël	12	16

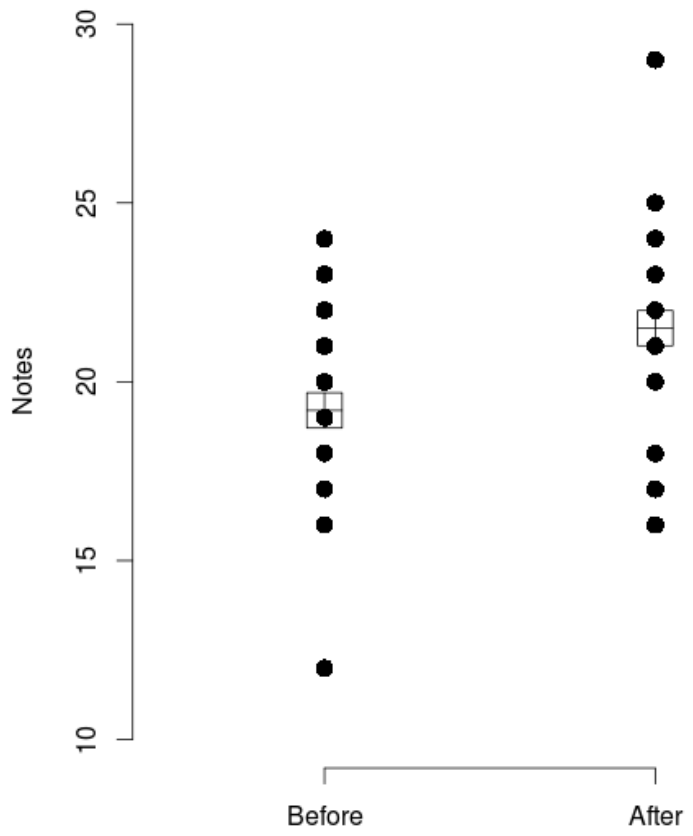



```
> wilcox.test(x,y, paired=TRUE)
Wilcoxon signed rank test with
continuity correction

V = 5, p-value = 0.02428
alternative hypothesis: true location
shift is not equal to 0
```

```
> t.test(x,y, paired=TRUE)
Paired t-test

t = -3.1461, df = 9, p-value = 0.01181
alternative hypothesis: true difference in
means is not equal to 0
```

Données indépendantes

	Before	After	
Louise	18	22	Lucas
Léo	21	25	Alice
Emma	16	17	Hugo
Gabriel	22	24	Jade
Chloé	19	18	Jules
Adam	24	29	Léa
Lola	17	20	Louis
Timéo	20	23	Manon
Inès	23	21	Arthur
Raphaël	12	16	Anna




```
> wilcox.test(x,y, paired=FALSE)
Wilcoxon rank sum test with continuity correction

W = 35, p-value = 0.2716
alternative hypothesis: true location shift is
not equal to 0
```

```
> t.test(x,y, paired=FALSE)
Two Sample t-test

t = -1.3529, df = 18, p-value = 0.1928
alternative hypothesis: true
difference in means is not equal to 0
```

Décomposition des données

Variance decomposition of the data into within and between variances

$$X = X_m + X_b + X_w$$

offset term — between-sample variation — within-sample variation

- The multilevel approach extracts the within variation matrix
- Classical multivariate tools can then be applied on the within matrix

→ We take into account the repeated measures design of the experiment

Liquet, B. Lê Cao, K-A., et al. (2012). A novel approach for biomarker selection and the integration of repeated measures experiments from two platforms, BMC Bioinformatics, 13:325.

Exemple

Westerhuis et al. (2009). Multivariate paired data analysis: multilevel PLSDA versus OPLSDA. Metabolomics 6(1).

Plan d'expérience : **20** individus, **3** variables mesurées (A, B, C), **2** conditions (control, treatment). Chaque individu est son propre contrôle.

condition	subject	A	B	C
control	1	20	10	20
control	2	18	12	17
control	3	16	15	14
control	4	14	16	11
control	5	10	2	8
control	6	9	3	5
control	7	7	7	2
control	8	7	7	8
control	9	3	9	14
control	10	2	9	17
treatment	1	21	12	20
treatment	2	21	14	17
treatment	3	17	17	14
treatment	4	17	18	11
treatment	5	11	4	8
treatment	6	12	5	5
treatment	7	8	9	2
treatment	8	10	9	8
treatment	9	4	11	14
treatment	10	5	11	17

X

Subject	$\bar{A}$	$\bar{B}$	$\bar{C}$
1	20.5	11	20
2	19.5	13	17
3	16.5	16	14
4	15.5	17	11
5	10.5	3	8
6	10.5	4	5
7	7.5	8	2
8	8.5	8	8
9	3.5	10	14
10	3.5	10	17

La matrice X_b contient l'effet « sujet » dont on souhaite s'affranchir.

X_b

DA	DB	DC
-1	-2	0
-3	-2	0
-1	-2	0
-3	-2	0
-1	-2	0
-3	-2	0
-1	-2	0
-3	-2	0
-1	-2	0
-3	-2	0
1	2	0
3	2	0
1	2	0
3	2	0
1	2	0
3	2	0
1	2	0
3	2	0
1	2	0
3	2	0

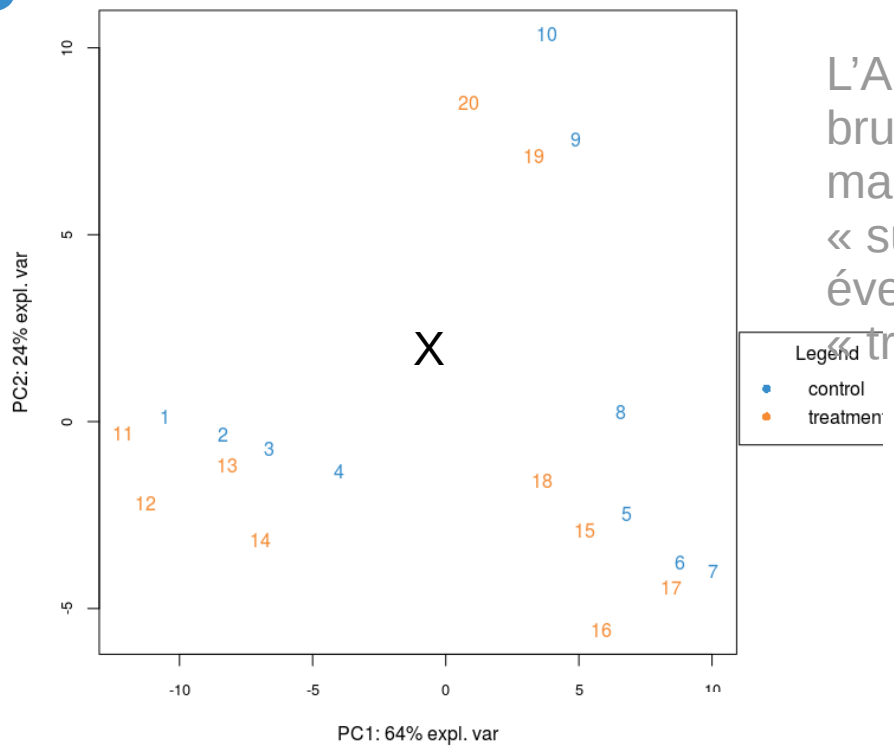
La matrice X_w contient l'information des données initiales X débarrassées de l'effet « sujet ». Elle est la matrice d'intérêt dans une approche *multilevel*.

X_w

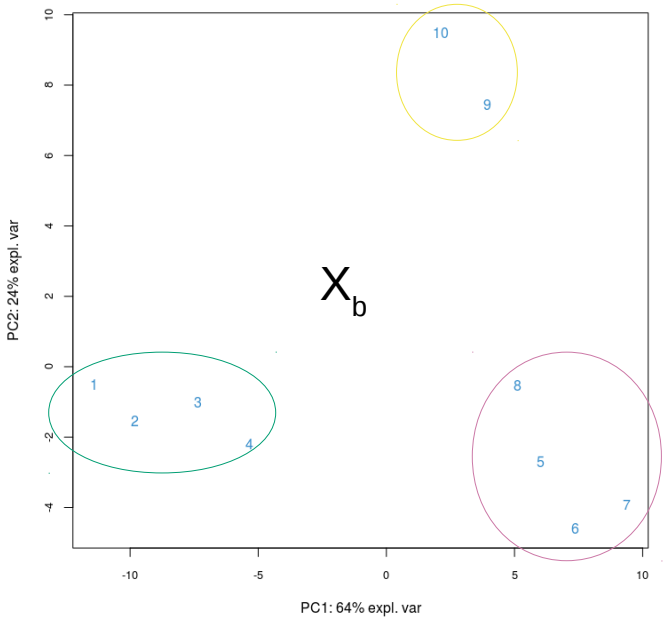
Exemple : ACP

L'ACP de la matrice X_b montre la répartition des individus indépendamment de l'effet « traitement ».

Subject	A	B	C
1	20.5	11	20
2	19.5	13	17
3	16.5	16	14
4	15.5	17	11
5	10.5	3	8
6	10.5	4	5
7	7.5	8	2
8	8.5	8	8
9	3.5	10	14
10	3.5	10	17

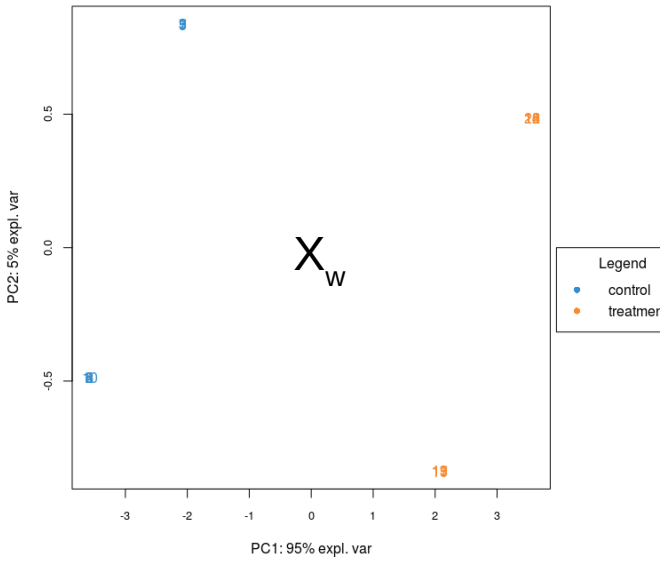


L'ACP des données brutes (matrice X) est marquée par un fort effet « sujet » qui masque un éventuel effet « traitement ».



L'ACP de la matrice Xw montre un net effet « traitement » (control à gauche, treatment à droite).

Dans l'article de Westerhuis et al., les données sont bruitées pour distinguer les individus.



To put it in a nustshell

- Multivariate linear methods enables to answer a wide range of biological questions
  - data exploration
  - classification
  - integration of multiple data sets
- Variable selection (*sparse*)
- Cross-over design (*multilevel*)

- Principles

PCA : max var(aX) → a ?
PLS1 : max cov(aX, by) → a, b ?
PLS2 : max cov(aX, bY) → a, b ?
CCA : max cor(aX,bY) → a, b ?
PLSDA → PLS2
GCCA : max Σ cov(a_iX_i,b_jX_j) → a_i, b_i ?

- Future of mixOmics
  - Time course modelling
  - Other workshops coming up! (on demand !)

Questions, *feedback*

Site web avec tutoriel :

www.mixomics.org

Contact : mixomics@math.univ-toulouse.fr

Register to our newsletter for the latest updates :

http://mixomics.org/a-propos/contact-us/

mixOmics n'existerait pas sans...

mixOmics development

Kim-Anh Lê Cao, UQ: Univ QLD
Ignacio González, INRA Toulouse
Benoît Gautier, UQDI
Florian Rohart, TRI, UQ
Sébastien Déjean, Univ. Toulouse
François Bartolo, Methodomics
Xin Yi Chua, QFAB

Methods development

Amrit Singh, UBC, Vancouver
Benoît Liquet, Univ. Pau
Jasmin Straube, QFAB
Philippe Besse, INSA Toulouse
Christèle Robert, INRA Toulouse

Data providers and biological point of view

Pascal Martin, INRA Toulouse

ANR

Australian Government
Australian Research Council

Australian Government
National Health and Medical Research Council

And many many mixOmics users and attendees!