Multivariate projection methodologies for the integration of large biological data sets

Application in R using mixOmics

Exploration and

Integration of

Omics datasets

Genotoul Biostat

math.univ-toulouse.fr/biostat

Agenda

- Introduction
- Reminders (?)
- Explore one data set (PCA)
- Discriminant analysis (LDA, PLS-DA)
- Data integration (PLS, CCA, GCCA)
- Graphical outputs
- Extensions: sparse and multilevel
- Conclusion

Introduction

Research hypothesis

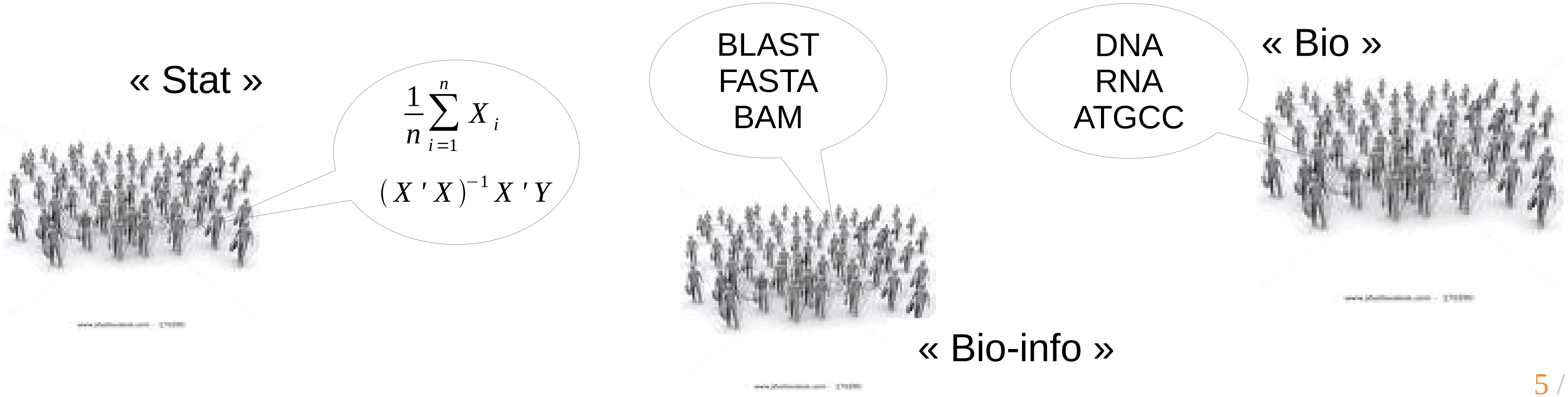
- Molecular entities act together to trigger cells' responses and need to be appropriately modelled and identified using novel statistical techniques.

- Multivariate statistical methods to shift the univariate statistics paradigm to obtain deeper insight into biological systems
  - Identify a combination of biomarkers rather than univariate biomarkers
  - Integrate multiple sources of biological data
  - Reduce the dimension of the data for a better understanding of complex biological systems

Multidisciplinarity!

- Nearly unlimited quantity of data from multiple and heterogeneous sources
- Computational issues to foresee
- Biological interpretation for validation
- Keep pace with new technologies

A close interaction between statisticians, bioinformaticians and molecular biologists is essential to provide meaningful results



« Stat »

$\frac{1}{n}\sum_{i=1}^{n} X_i$

$(X'X)^{-1}X'Y$

BLAST
FASTA
BAM

DNA
RNA
ATGCC

« Bio »

« Bio-info »

Data integration

Generally, data integration can be defined as the process of combining data residing in diverse sources to provide users with a comprehensive view of such data. There is no universal approach to data integration, and many techniques are still evolving.

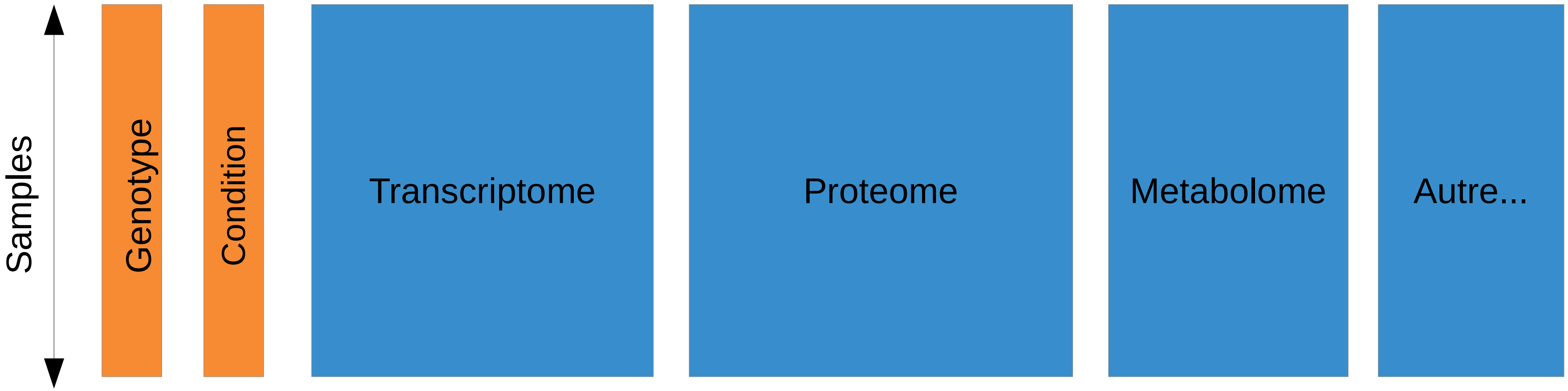
From Schneider, M. V., & Jimenez, R. C. (2012). Teaching the Fundamentals of Biological Data Integration Using Classroom Games. PLoS Computational Biology, 8(12)

mixOmics philisophy in this context:

- R toolkit for multivariate data analysis of ‘omics data
- Statistical data integration
- Data-driven approaches (≠ database or knowledge-based approaches)

Overview

🟦 Quantitative 🟧 Qualitative



- Univariate 🟦

  Mean, median, standard deviation...

- Bivariate: 2 quantitatives 🟦 🟦 or 1 quantitative + 1 qualitative 🟦 🟧 or 2 qualitatives 🟧 🟧

  Correlation, statistical test (Student, ANOVA, Chi2)

- Multivariate unsupervised 🟦

  PCA

- Multivariate supervised 🟦 🟧

  PLS-DA

- Multi-block unsupervised 🟦 🟦 🟦 🟦

  PLS (2 blocks), GCCA

- Multi-block supervised 🟦 🟦 🟦 🟦 🟧

  GCC-DA

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
  - 21,000 downloads (unique IP adress) in 2016 (4,000 in 2014, 10,000 in 2015)
  - Website: `www.mixomics.org`
  - Two web-interfaces (shiny and PHP, also Galaxy but not advertised)
  - 19 multivariate methodologies and sparse variants (13 are our own methods)
  - Team: 3 core members and 4 key contributors
  - Move to Bioconductor OPEN SOURCE SOFTWARE FOR BIOINFORMATICS in october 2018

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 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 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 Projection to Latent Sructures (PLS) canonical mode. The total number of variables is greater than the number of samples: Regularized Canonical Correlation Analysis (rCCA) or Projection to Latent Sructures (PLS) canonical mode

Practical works

- Based on the vignette of the package

  bioconductor.org/packages/release/bioc/vignettes/mixOmics/inst/doc/vignette.html

- Quick start section for every methods

- Focus on PCA, (Sparse-)PLS-DA and (Sparse-)PLS

Reminders (?)

Variance and Standard Deviation

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

Mean of the squared deviations to the mean

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

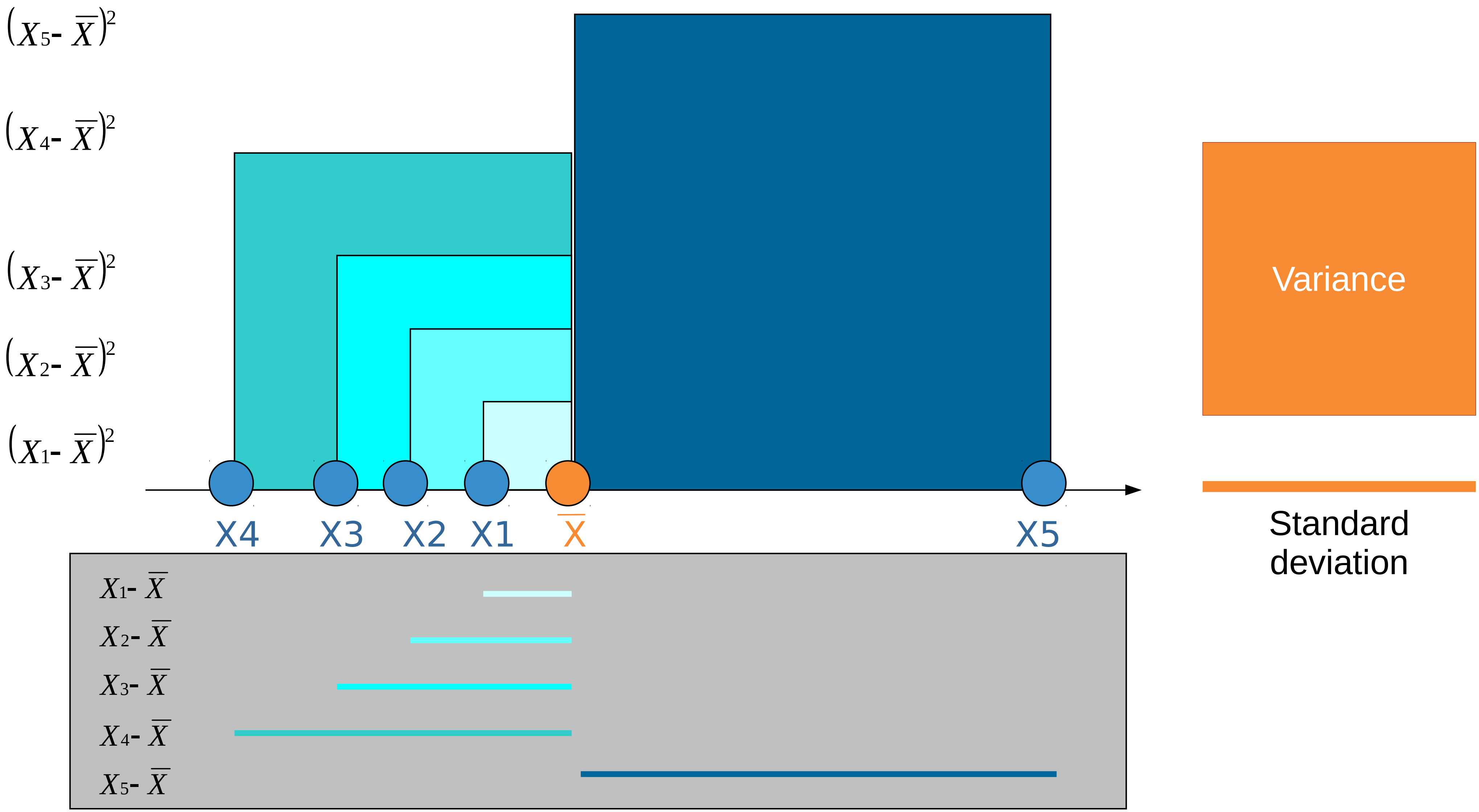
Square root of the variance

Properties of the standard deviation:

- Positif (zero if the serie is constant)
- Unchanged by translation
- Sensitive to extreme values
  - **In the same unit as the data** (as the mean but unlike the variance): *If the data are expressed in* ***m*** *then the standard deviation also express in* ***m*** *(as the mean) and the variance in* ***m²*** *!*

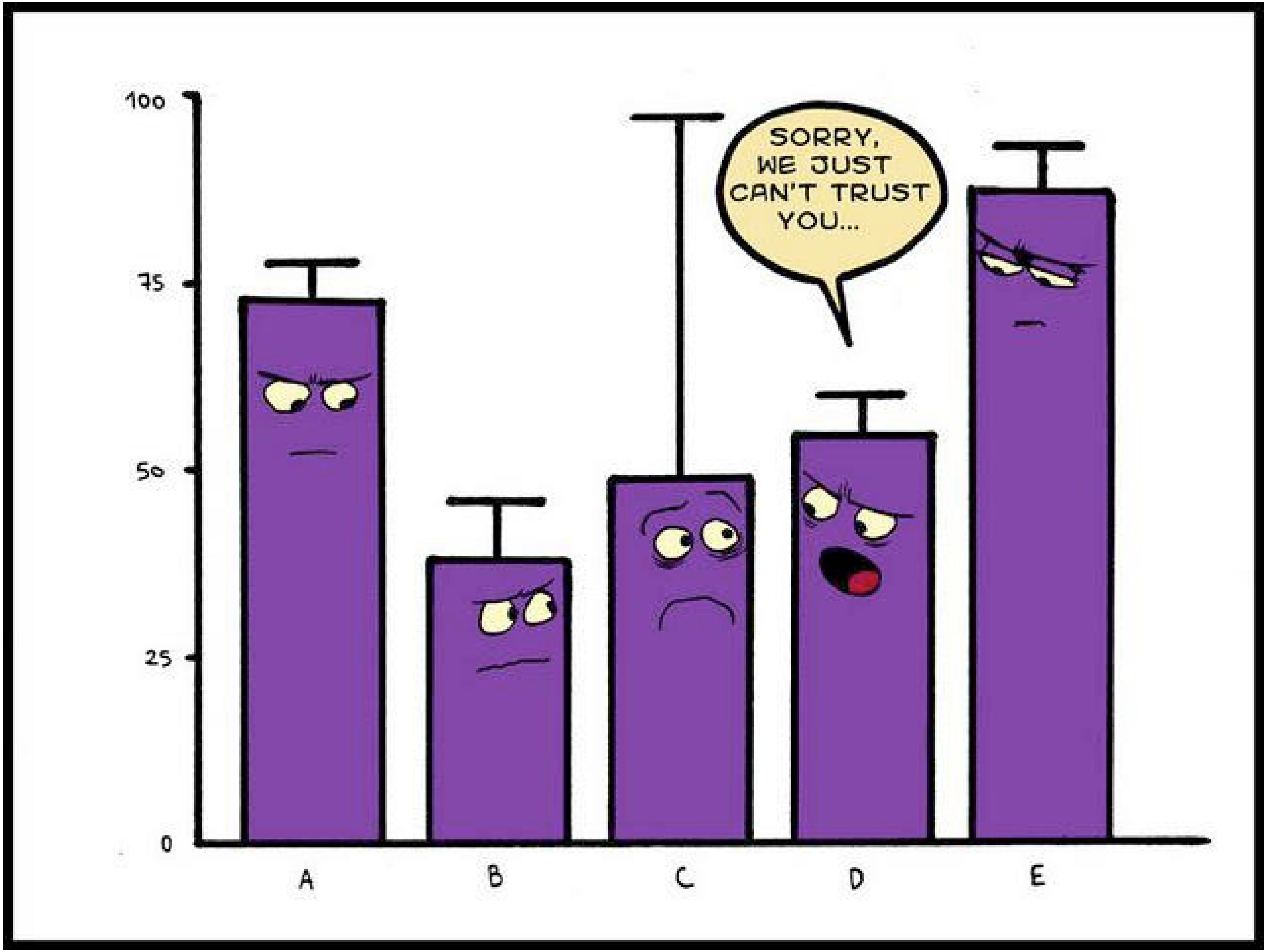
Variance and Standard Deviation

Square root of the mean of the squared deviation to the mean



Humour for statistician...

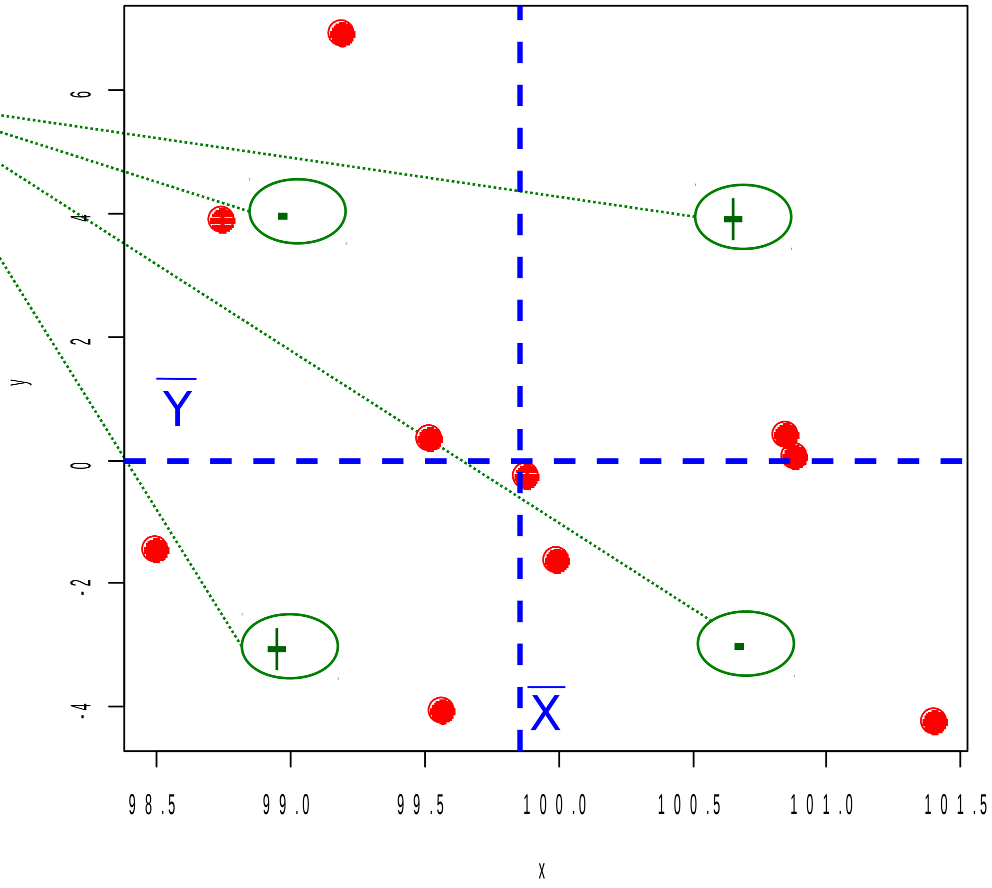
Source : xkcd.com



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)

Sign of the product (Xi-$\bar{X}$)(Yi-$\bar{Y}$)



Intuitively :

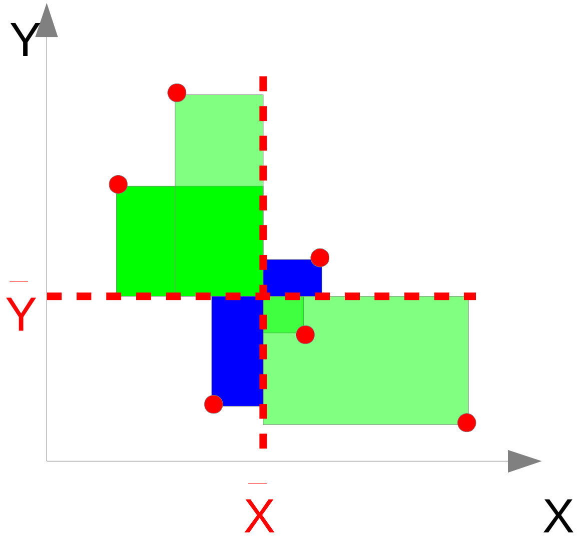
- If the **+** win

→positive linear relationship

- If the **–** win

→négative linear relationship

*On this example : cov(X,Y)=**-1.36***



The covariance depends on the physical units
➔ correlation coefficient

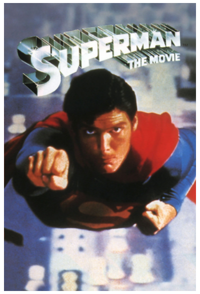
Correlation

Some properties of correlation coefficients:

- Between –1 and 1

- Pearson correlation coefficient: **linear** relationship

$$\rho(X,Y)=cov(X,Y)/(\sigma_X \sigma_Y)$$



- Spearman correlation coefficient (ranks): **monotone** relationship

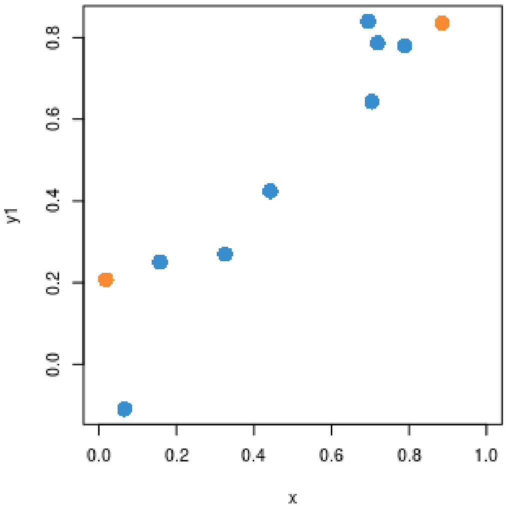
$$\rho_s(X,Y)=\rho(RX,RY)$$

X	Y	RX	RY
12	5	3	1
15	3	1	2
14	2	2	3

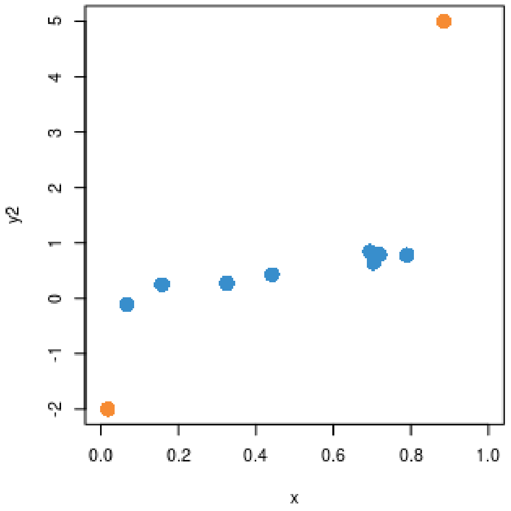
- If the coefficient is positive : when a variable is high the other is also high. Replace high with low.
- If the coefficient is negative : when a variable is high the other is low. Replace high with low and inversely.

Correlation

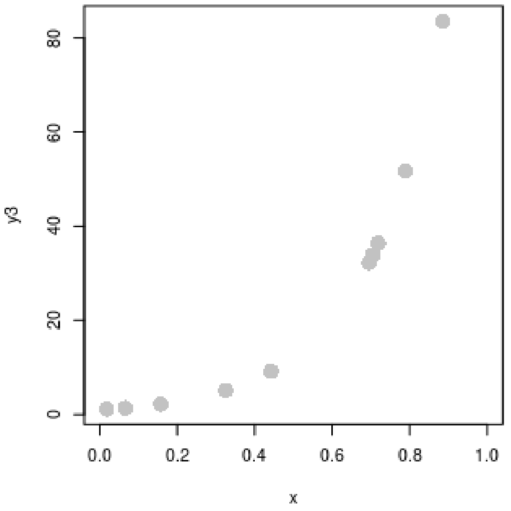
$\rho = 0.944$ - $\rho_s = 0.855$



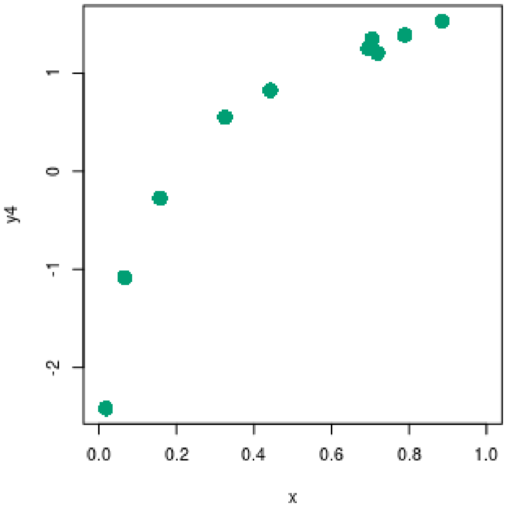
$\rho = 0.722$ - $\rho_s = 0.915$



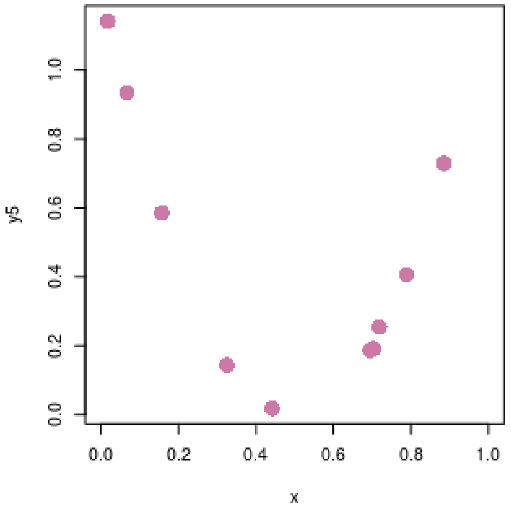
$\rho = 0.89$ - $\rho_s = 1$



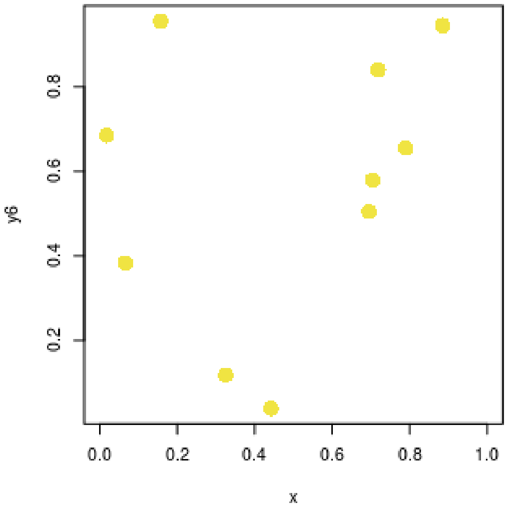
$\rho = 0.914$ - $\rho_s = 0.964$



$\rho = -0.518$ - $\rho_s = -0.248$



$\rho = 0.24$ - $\rho_s = 0.248$



Linear combination

2 variables

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

Height	Weight
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

Linear combination of the 2 variables Height and Weight with coefficients c1 and c2

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

Matrix notation: **LC** = **XW**

A principal component is a linear combination of the initial variables.

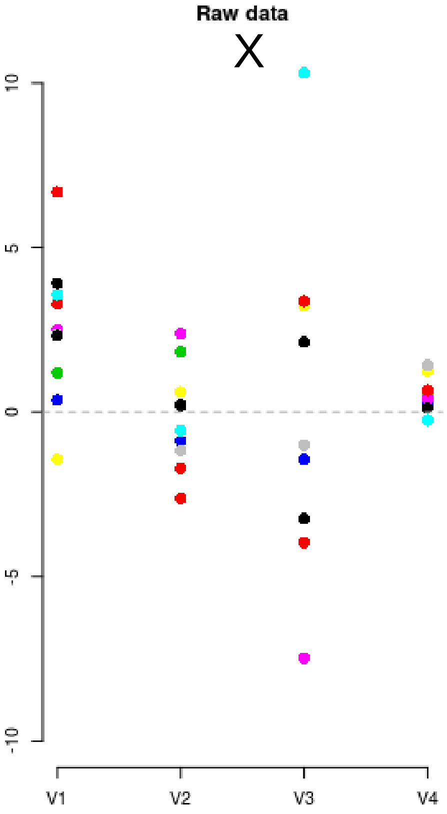
Center / scale

- **Center**: remove the mean
- **Scale**: divide by the standard deviation
- Express different variables on a common scale, without physical unit; the observations are thus expressed as **numbers of standard deviations related to the mean**.
- After centering and scaling, the mean is zero and the standard deviation is 1 (as the variance).

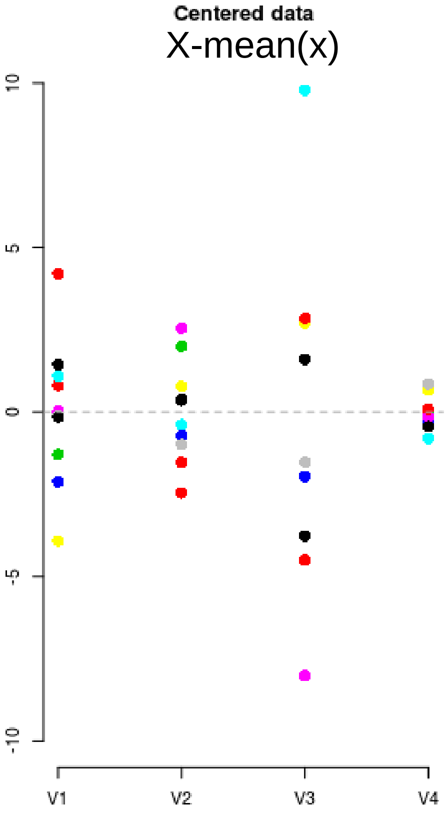
$$Z_i = \frac{X_i - \overline{X}}{\sigma_X}$$

- Sometimes called « z-transformation » ou « z-score »

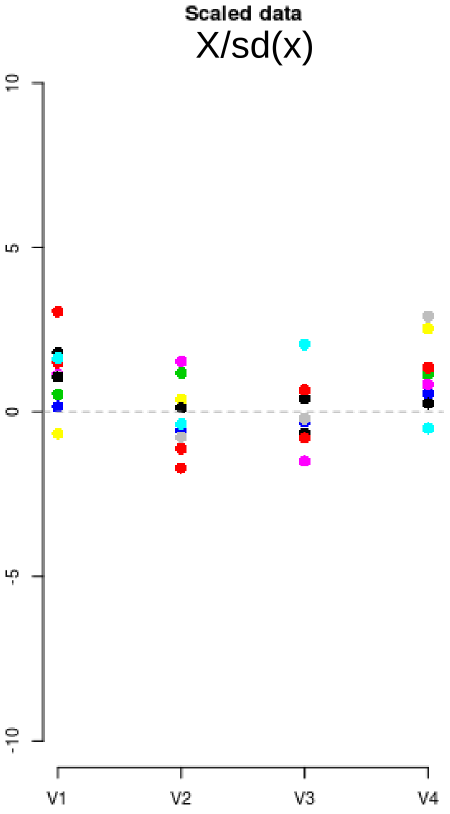
Center / scale


Raw data

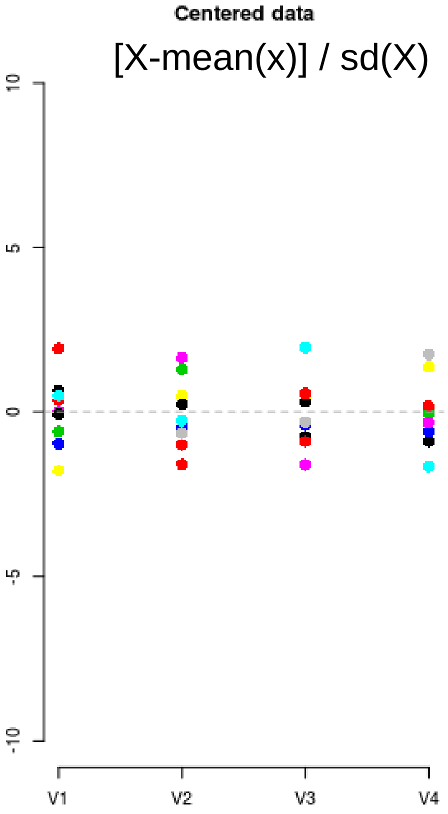
X


Centered data

X-mean(x)


Scaled data

X/sd(x)


Centered data

[X-mean(x)] / sd(X)

```
        v1   v2   v3   v4
1      3.9  0.2 -3.2  0.6
2      3.3 -1.7 -4.0  0.6
3      1.2  1.8  3.3  0.6
4      0.4 -0.9 -1.4  0.3
5      3.6 -0.6 10.3 -0.2
6      2.5  2.4 -7.5  0.4
7     -1.4  0.6  3.2  1.2
8      2.4 -1.2 -1.0  1.4
9      2.3  0.2  2.1  0.1
10     6.7 -2.6  3.4  0.7

Mean   2.5 -0.2  0.5  0.6
S.D.   2.2  1.5  5.0  0.5
```

```
        v1   v2   v3   v4
1      1.4  0.4 -3.8  0.1
2      0.8 -1.5 -4.5  0.0
3     -1.3  2.0  2.8  0.0
4     -2.1 -0.7 -2.0 -0.3
5      1.1 -0.4  9.8 -0.8
6      0.0  2.5 -8.0 -0.2
7     -3.9  0.8  2.7  0.7
8     -0.1 -1.0 -1.5  0.9
9     -0.2  0.4  1.6 -0.4
10     4.2 -2.5  2.8  0.1

Mean   0    0    0    0
S.D.   2.2  1.5  5.0  0.5
```

```
        v1   v2   v3   v4
1      1.1  0.1 -0.6  0.8
2      1.0 -1.1 -0.8  0.8
3      0.3  1.2  0.7  0.7
4      0.1 -0.6 -0.3  0.4
5      1.0 -0.4  2.0 -0.3
6      0.7  1.5 -1.5  0.5
7     -0.4  0.4  0.6  1.6
8      0.7 -0.7 -0.2  1.8
9      0.7  0.1  0.4  0.2
10     2.0 -1.7  0.7  0.9

Mean   1.1 -0.1  0.1  1.2
S.D.   1    1    1    1
```

```
        v1   v2   v3   v4
1      0.7  0.3 -0.8  0.1
2      0.4 -1.0 -0.9  0.0
3     -0.6  1.3  0.6  0.0
4     -1.0 -0.5 -0.4 -0.6
5      0.5 -0.3  2.0 -1.7
6      0.0  1.7 -1.6 -0.3
7     -1.8  0.5  0.5  1.4
8     -0.1 -0.6 -0.3  1.7
9     -0.1  0.2  0.3 -0.9
10     1.9 -1.6  0.6  0.2

Mean    0    0    0    0
S.D.    1    1    1    1
```

Log transformation

X	Log2(X)
0.125 = 2^{-3}	-3
0.25 = 2^{-2}	-2
0.5 = 2^{-1}	-1
1 = 2^{0}	0
2 = 2^{1}	1
4 = 2^{2}	2
8 = 2^{3}	3
$4 < 5 < 8$	$2 < \sim 2.3 < 3$
$2 < 3 < 4$	$1 < \sim 1.6 < 2$
$0.1 < 0.125$	$\sim -3.3 < -3$

$$Y = \log_2(X) \leftrightarrow X = 2^Y$$

$$Y = \log_{10}(X) \leftrightarrow X = 10^Y$$

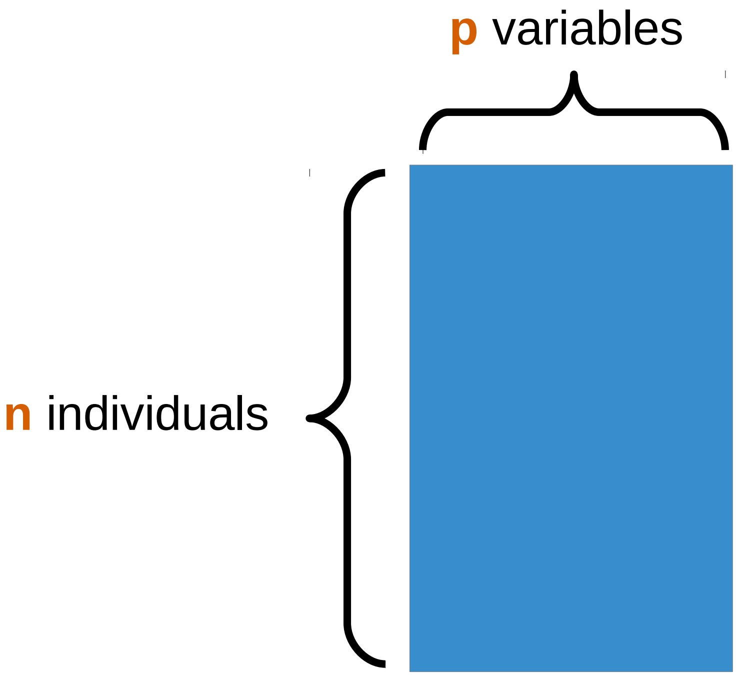
$$Y = \ln(X) \leftrightarrow X = e^Y = \exp(Y)$$

City	Population	Log10
Toulouse	441 802	5.65
Colomiers	35 186	4.55
Tournefeuille	25 340	4.40
Muret	23 864	4.38
...		
Castanet-Tolosan	11 033	4.04
Saint-Orens...	10 918	4.04
Saint-Jean	10 259	4.01
Revel	9 361	3.97
Portet-sur-Garonne	9 435	3.97
Auterive	9 107	3.96
...		
La Magdelaine-sur-T/	1 006	3.00
Grépiac	990	2.99
Landorthe	946	2.98
Vigoulet-Auzil	944	2.97
...		
Belbèze-de-Lauragais	104	2.02
Saint-Germier	103	2.01
Seyre	102	2.01
Gouzens	95	1.98
Lourde	98	1.99
Pouze	97	1.99
...		
Saccourvielle	13	1.11
Cirès	13	1.11
Bourg-d'Oueil	8	0.90
Trébons-de-Luchon	8	0.90
Caubous	6	0.78
Baren	5	0.70

Explore one data set

Principal Components Analysis

Describe with no prior a data set exclusively composed of **quantitatives** variables



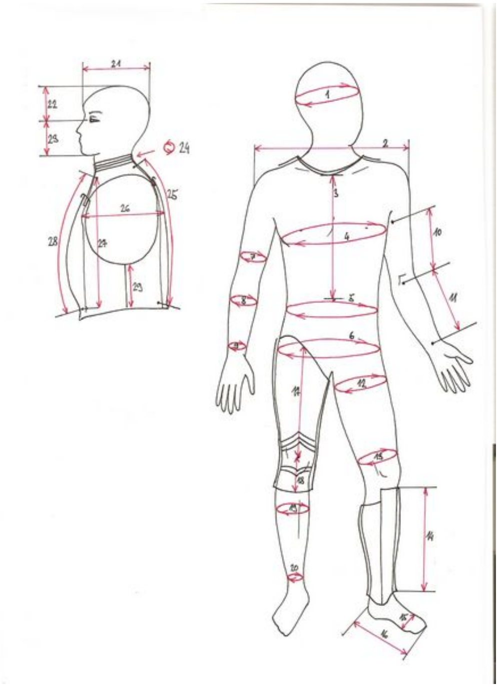
p variables

n individuals

Body data set

- 20 individuals
- 5 variables

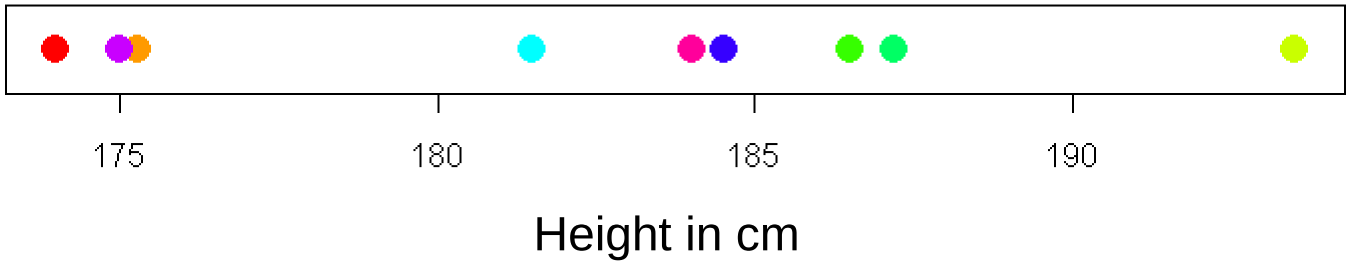
V1 : shoulder girth (cm)
V2 : chest girth (cm)
V3 : waist girth (cm)
V4 : weight (kg)
V5 : height (cm)



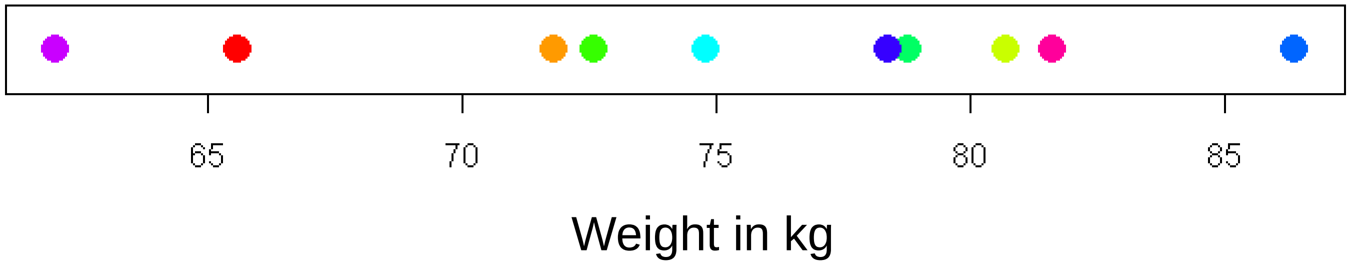
```
         V1     V2    V3    V4     V5
H 1    106.2   89.5  71.5  65.6  174.0
H 2    110.5   97.0  79.0  71.8  175.3
H 3    115.1   97.5  83.2  80.7  193.5
H 4    104.5   97.0  77.8  72.6  186.5
H 5    107.5   97.5  80.0  78.8  187.2
H 6    119.8   99.9  82.5  74.8  181.5
H 7    123.5  106.9  82.0  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
```

1D graphical output: stripchart

	●	●	●	●	●	●	●	●	●	●
Height	174.0	175.3	193.5	186.5	187.2	181.5	184.0	184.5	175.0	184.0

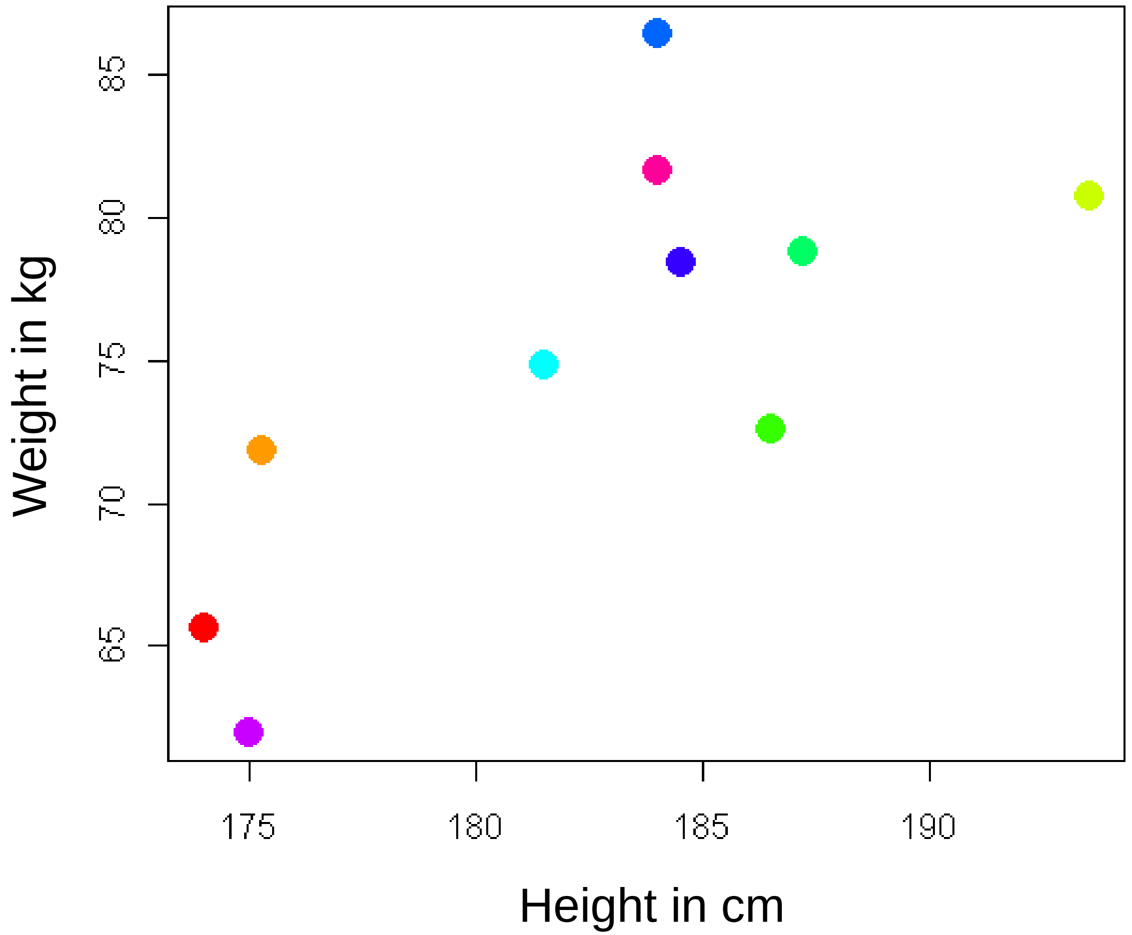


	●	●	●	●	●	●	●	●	●	●
Weight	65.6	71.8	80.7	72.6	78.8	74.8	86.4	78.4	62.0	81.6



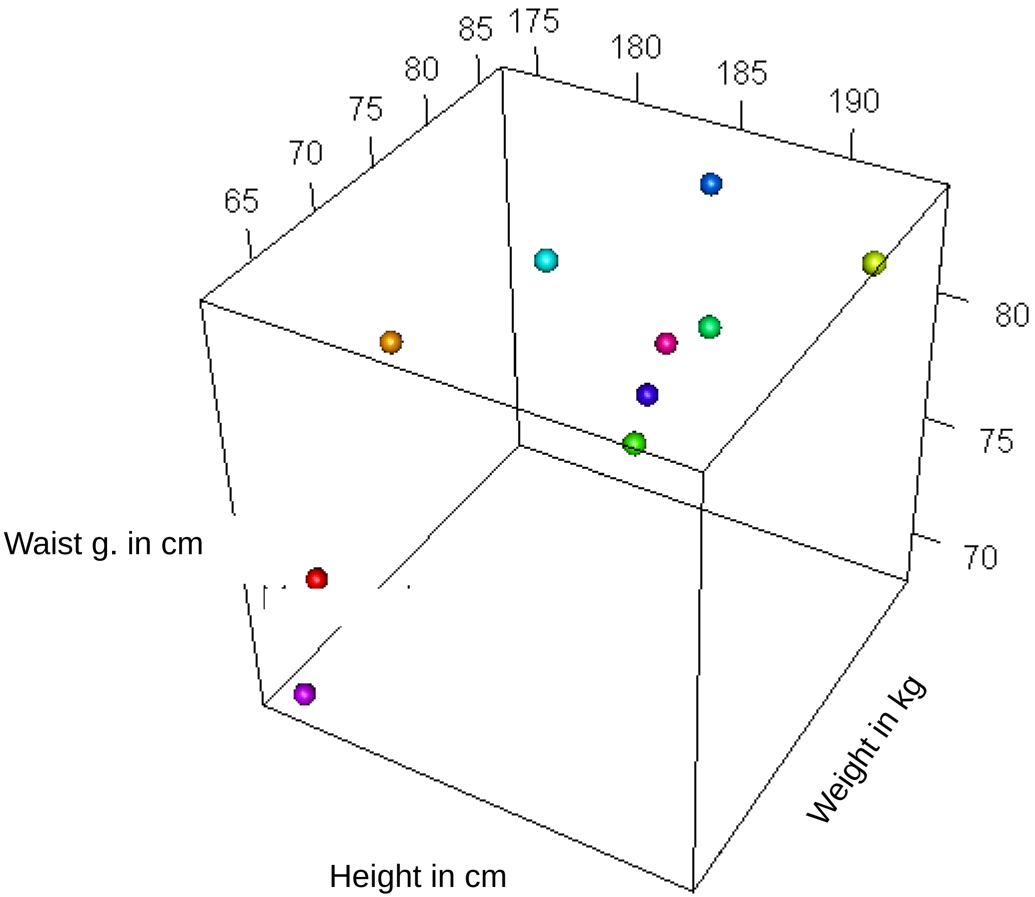
2D graphical output: scatter plot

	●	●	●	●	●	●	●	●	●	●
Height	174.0	175.3	193.5	186.5	187.2	181.5	184.0	184.5	175.0	184.0
Weight	65.6	71.8	80.7	72.6	78.8	74.8	86.4	78.4	62.0	81.6



3D graphical output: scatter plot

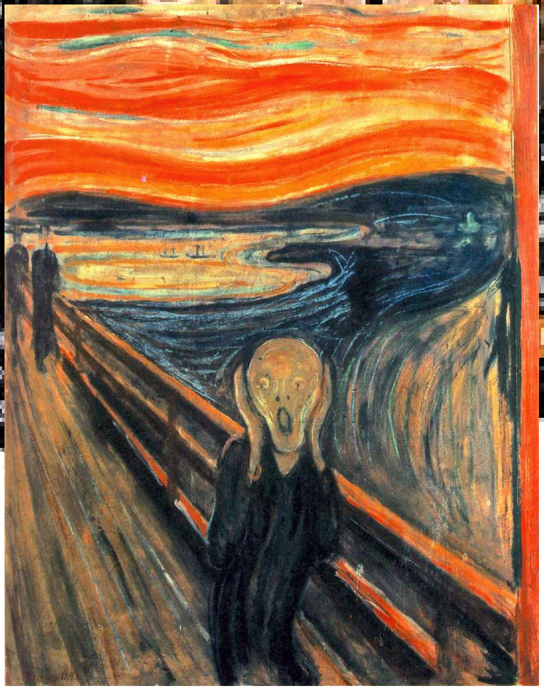
	●	●	●	●	●	●	●	●	●	●
Height	174.0	175.3	193.5	186.5	187.2	181.5	184.0	184.5	175.0	184.0
Weight	65.6	71.8	80.7	72.6	78.8	74.8	86.4	78.4	62.0	81.6
Waist g.	71.5	79.0	83.2	77.8	80.0	82.5	82.0	76.8	68.5	77.5



4D ?

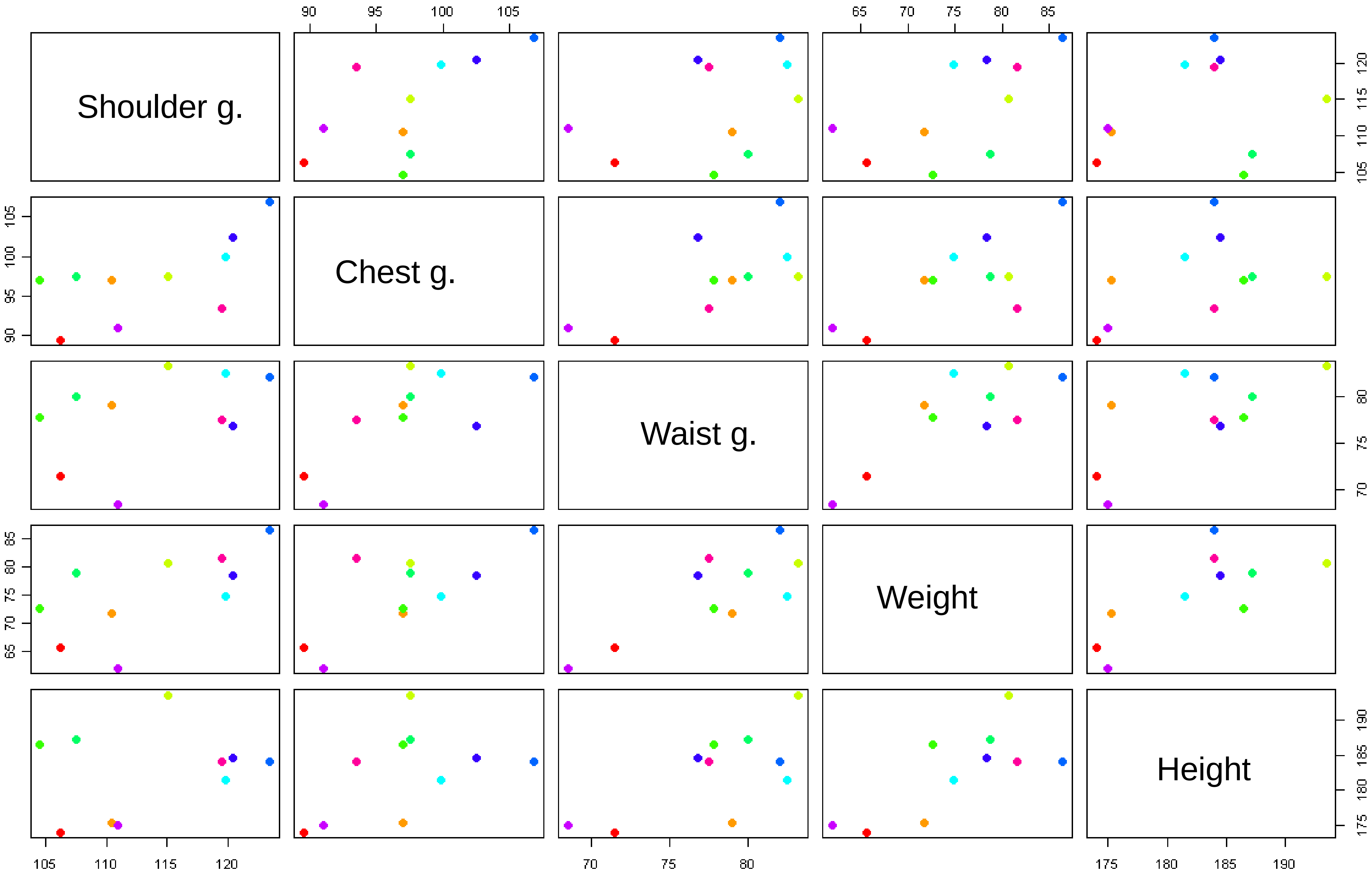
THE TWILIGHT ZONE
LA QUATRIÈME DIMENSION
LA SÉRIE ORIGINALE
E=mc²

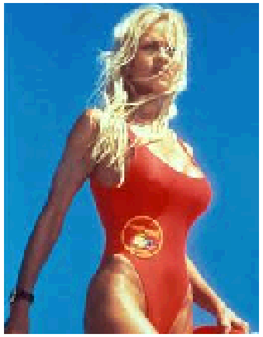
DAY!

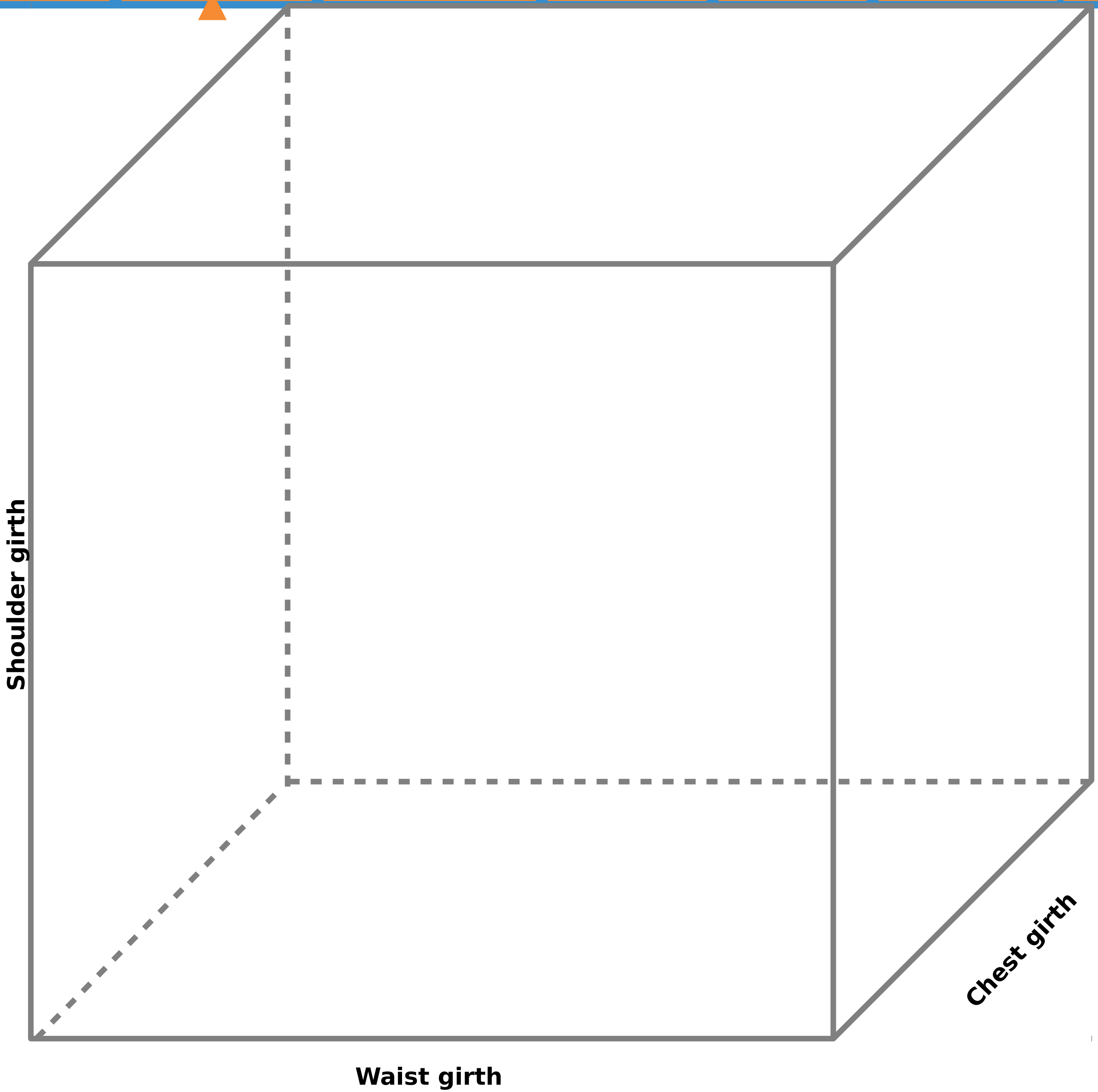


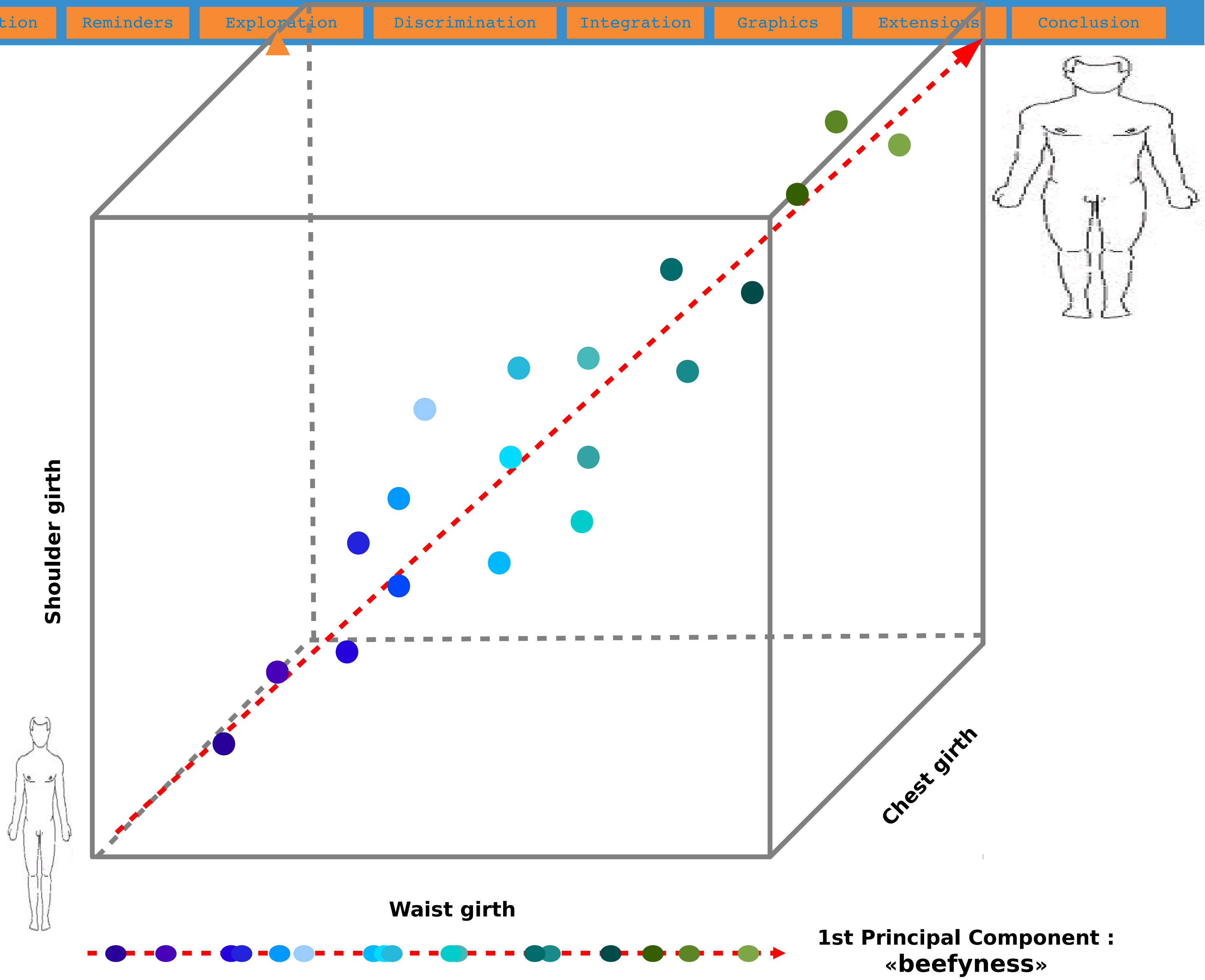
Ship Stolen
Need Money
To Get Home

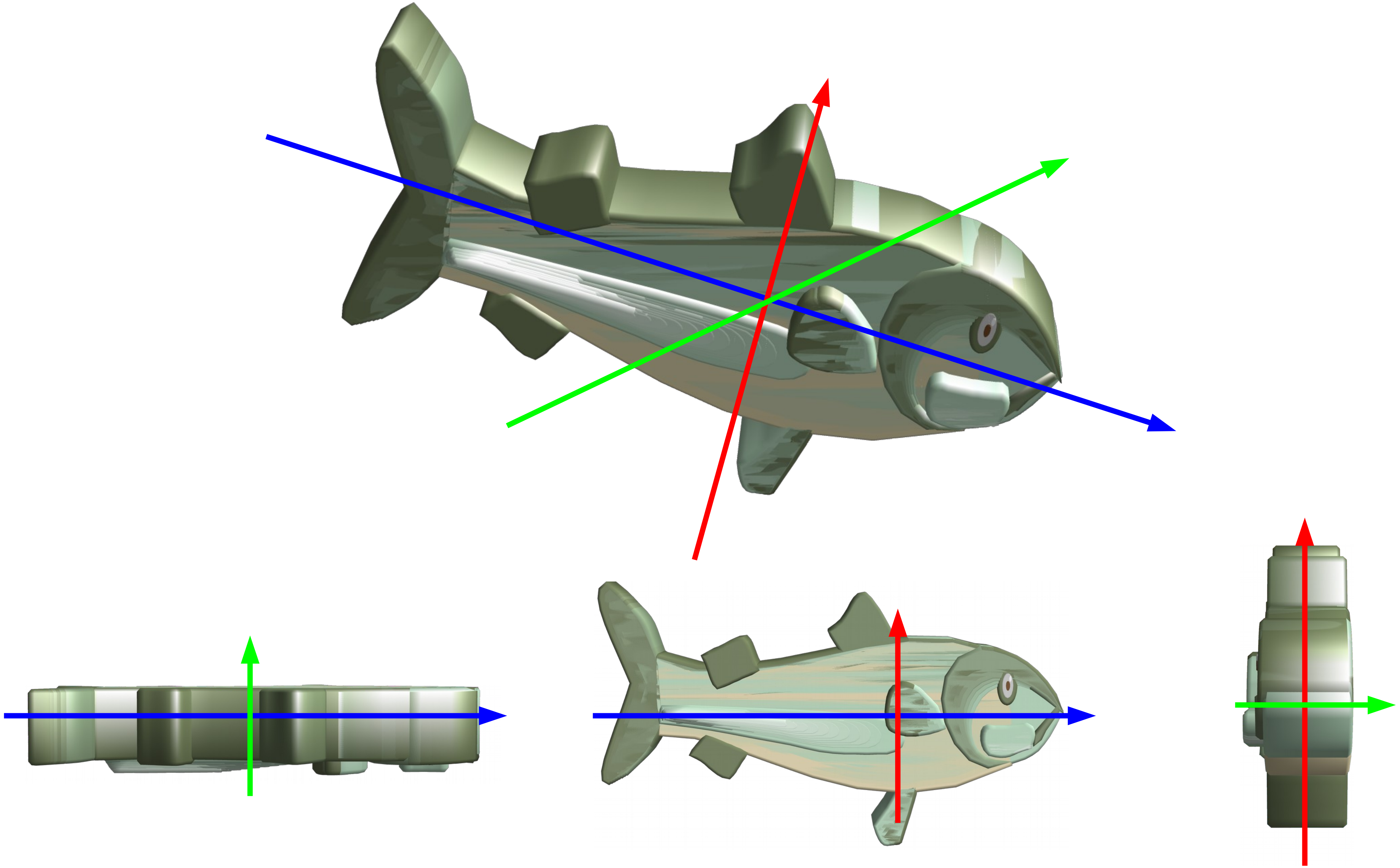
Alternative to 4D (or more)



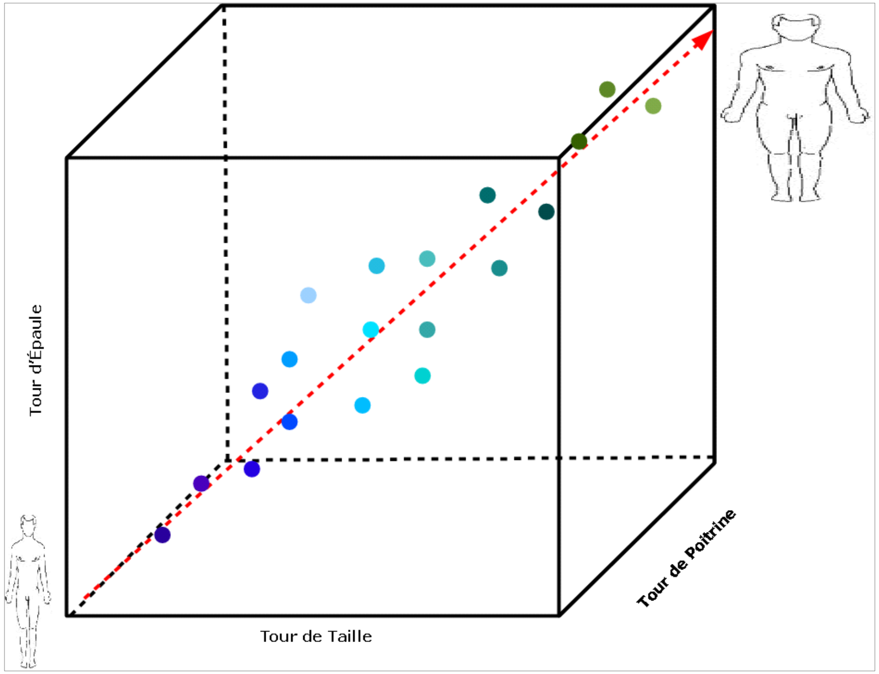



Shoulder girth
Chest girth
Waist girth

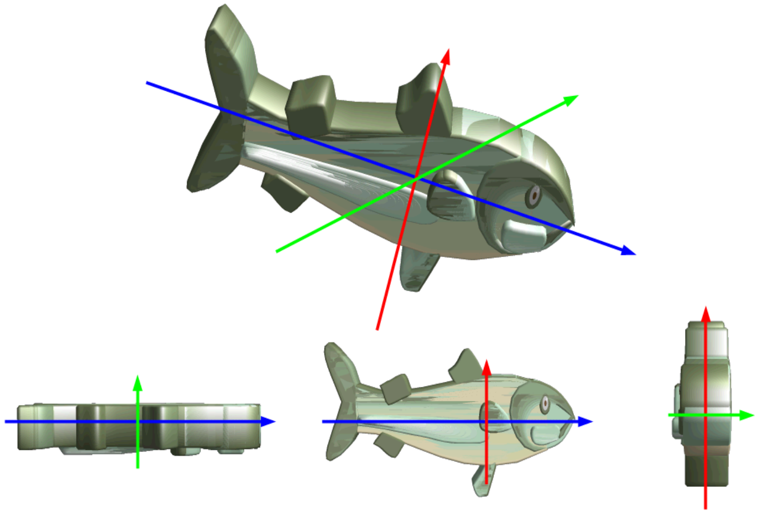

Shoulder girth
Waist girth
Chest girth
1st Principal Component :
«beefyness»



Comments



The measurements are rather **strongly correlated**. Indeed, one can assume that a person with a high shoulder girth will also have high chest girth (even if exceptions exist...). In these conditions, the information brought by the 5 variables are **redundant**. Graphically, in the cube determined by shoulder girth, chest girth and waist girth, there are nearly emplty areas. One variable calculated as a **combination** of these 3 variables (represented as the dotted arrow) would be enough to represent the individuals with a **minimal loss in information** because all the points are located along these direction that is the first principal component.



Among the potential projections in 2D spaces, all does not enable to identify easily the object. Among the 3 proposed projections, the image at the center is the **nearest from the original**. One can easily recongnize the initial object because the projection was made on the plane formed by the 2 directions along which the object **spreads out the most** (high variability). The information brought by the 3rd dimension is minimal and its loss is not a problem to recongnize the fish.

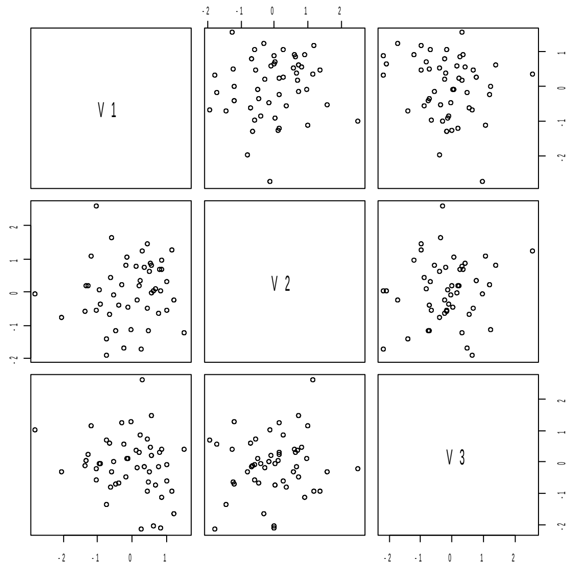
In other words

- PCA allow determine the sub-spaces of lower dimension than the initial space on which the projection of the individuals is the **least modified**, that is to say, the sub-spaces that keep the **greatest part of the information** (i.e. **variability**).
- The principle of PCA consists in finding a direction (the first PC), calculated as a **linear combination of the initial variables**, such that the **variance** of the points around this direction is **maximal**. Iterate this process in orthogonal directions to determine the following principal components. The number of PC that can be calculated is equal to the number of initial variables.
- Concerning the variables, the PCA keeps at best the **correlation structure** between the initial variables.

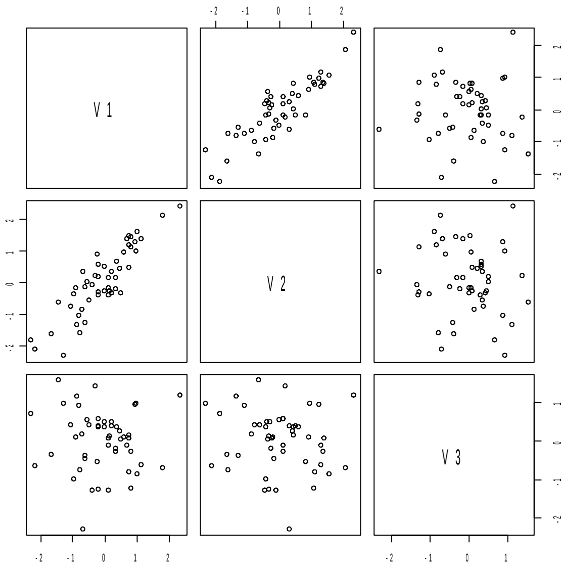
PCA: simulated examples

Data set : 50 observations, 3 variables (V1 – V2 - V3)

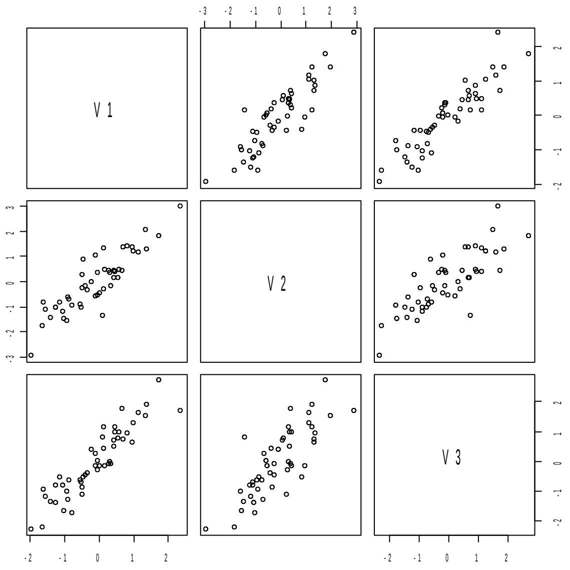
Case 1)
{V1} - {V2} - {V3}



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



Case 3)
{V1 - V2 - V3}



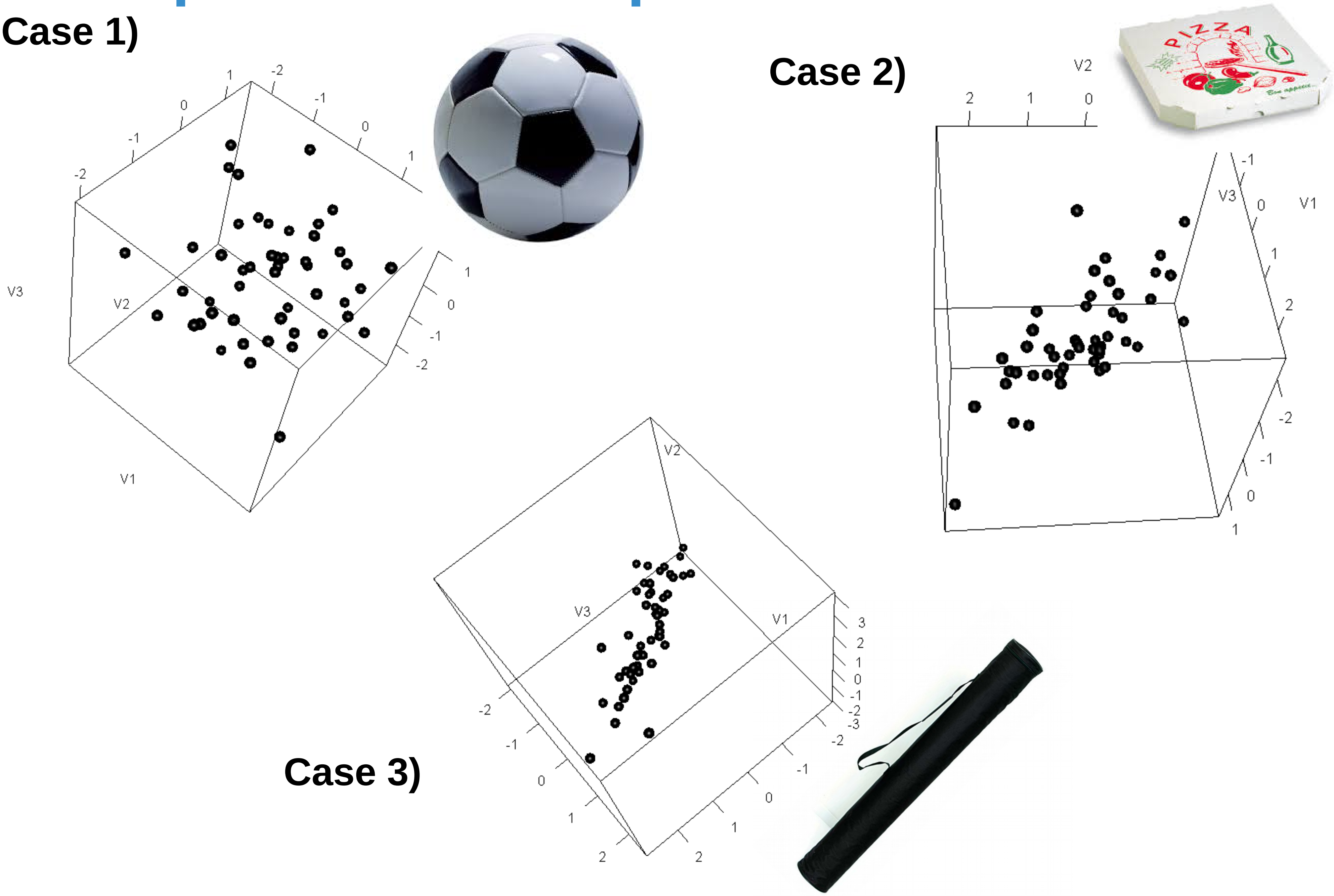
Pearson Correlation matrices

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

Example: 3D scatter plots



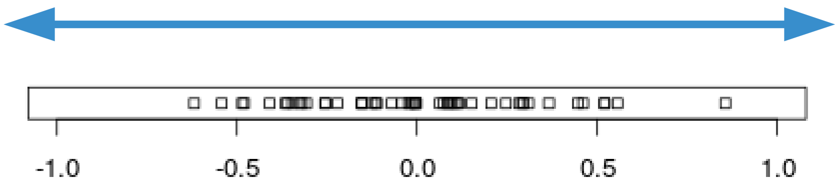
Case 1)

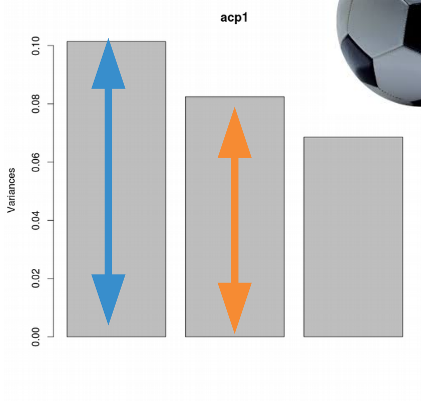
Case 2)

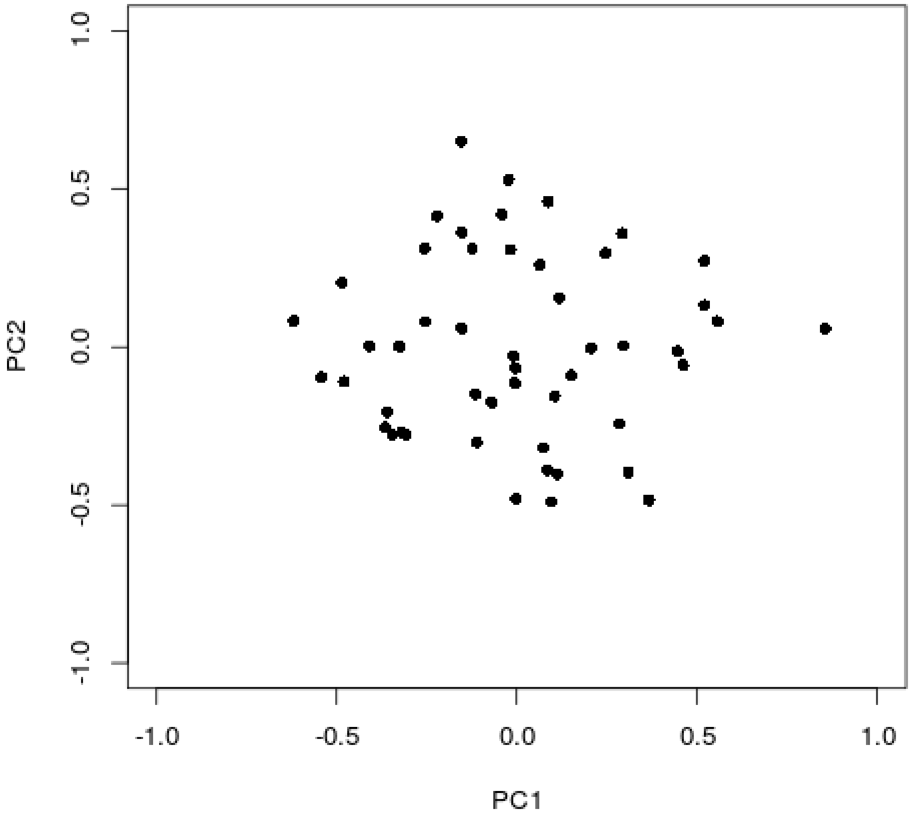
Case 3)

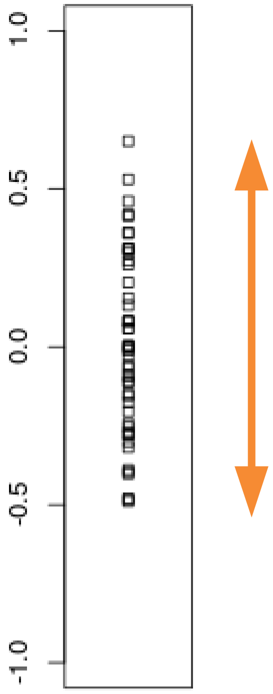
Example: individuals plot

1)



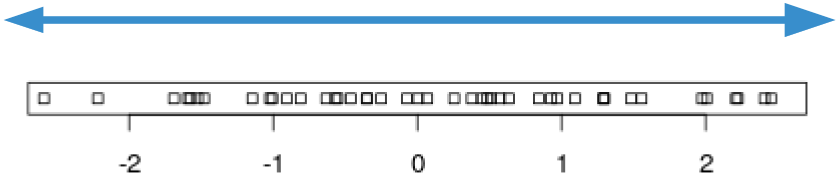


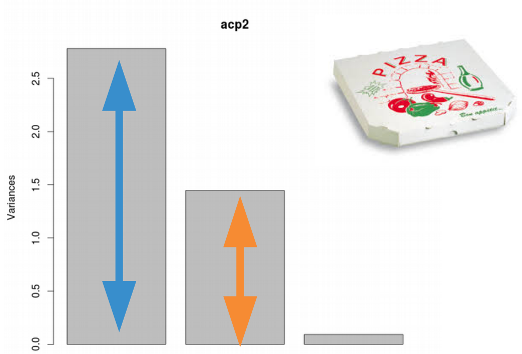


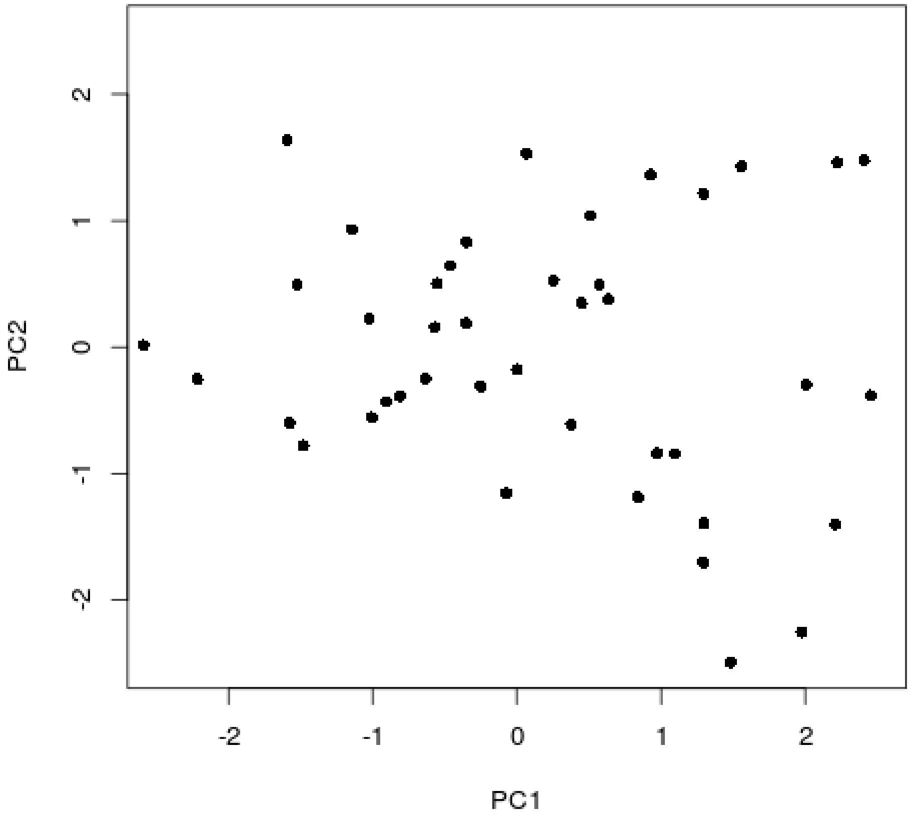


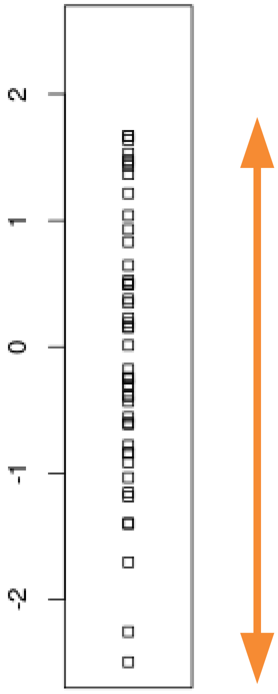
Example: individuals plot

2)



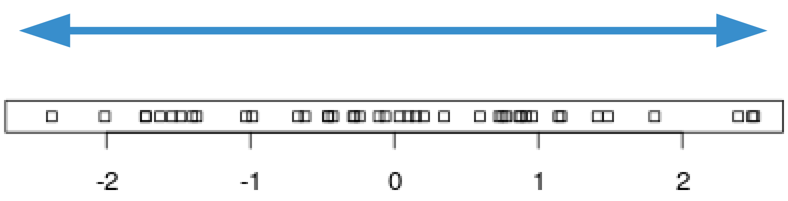


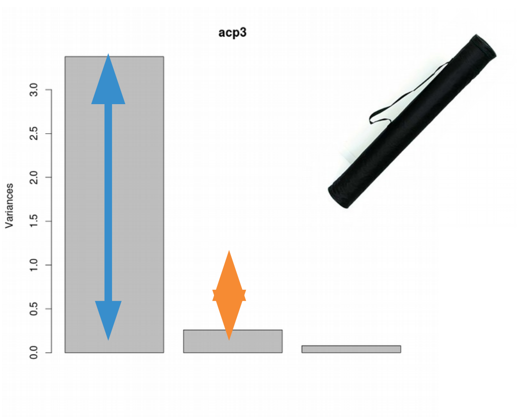


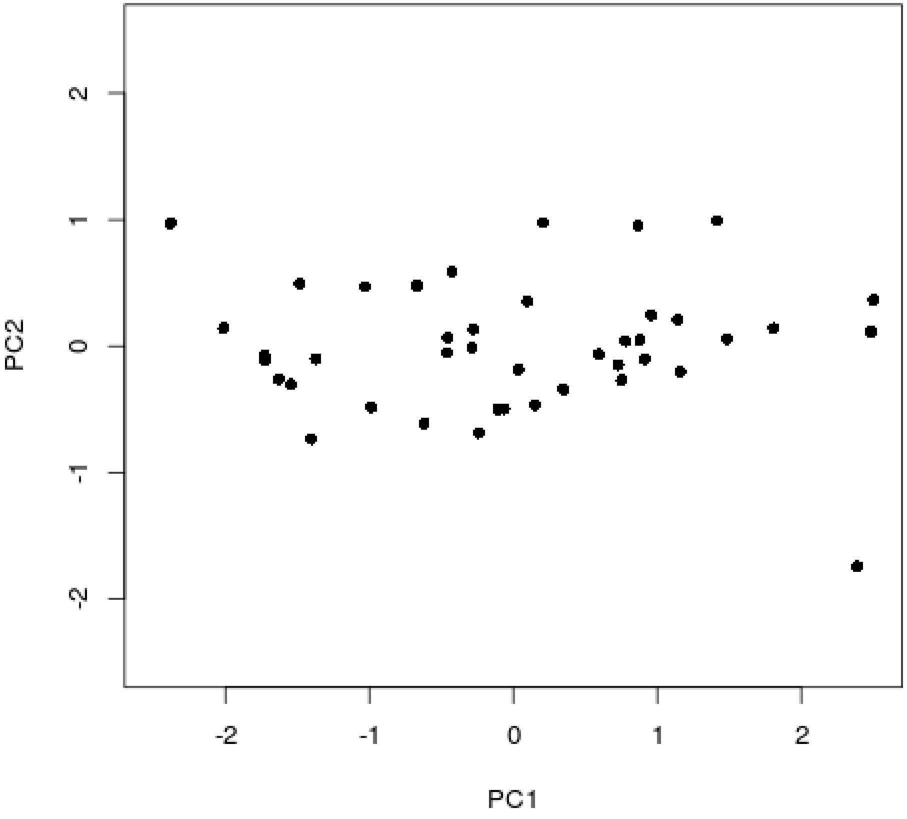


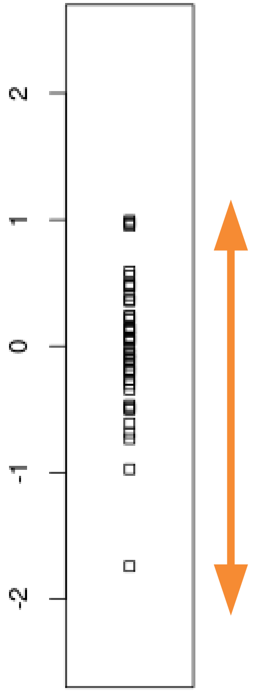
Example: individuals plot

3)

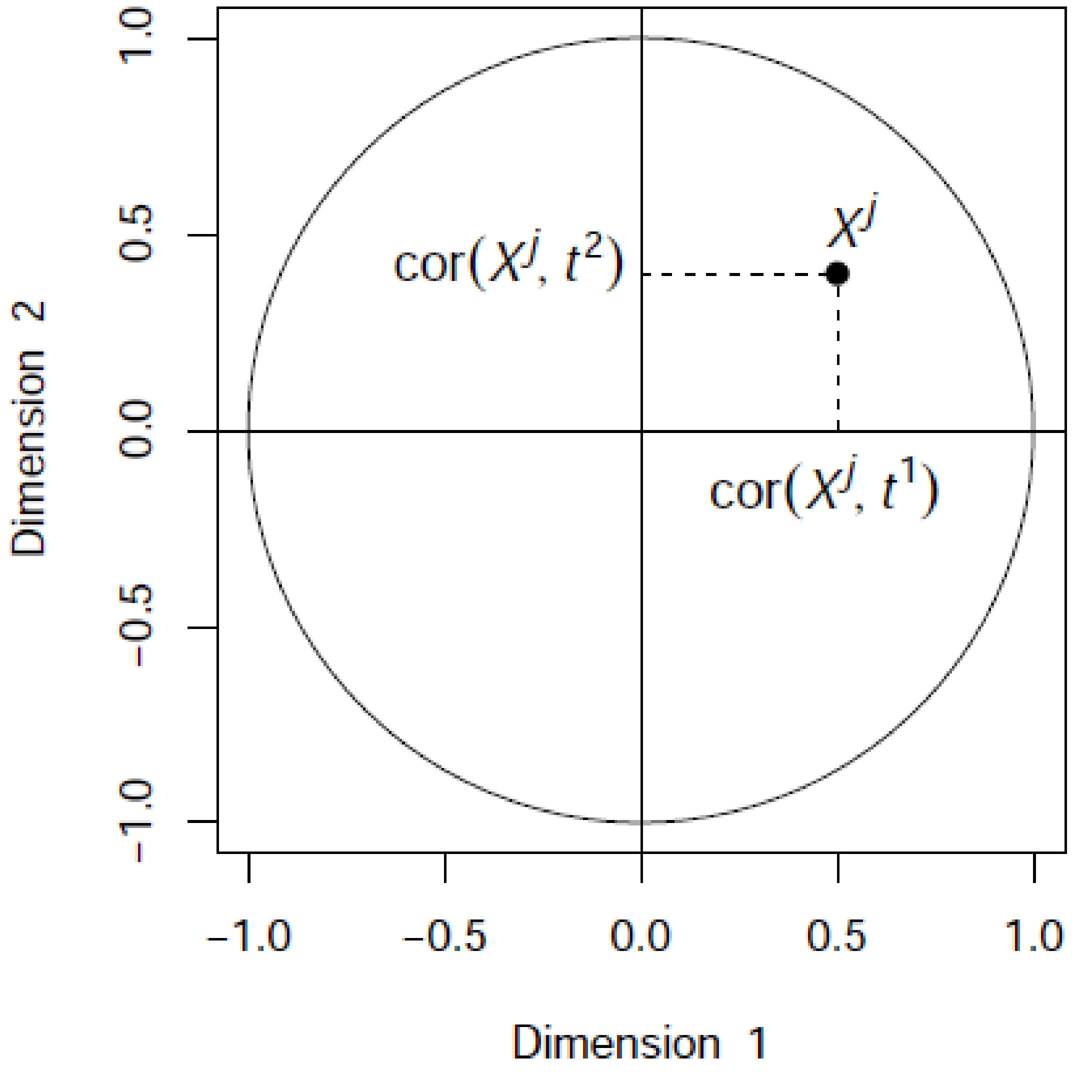








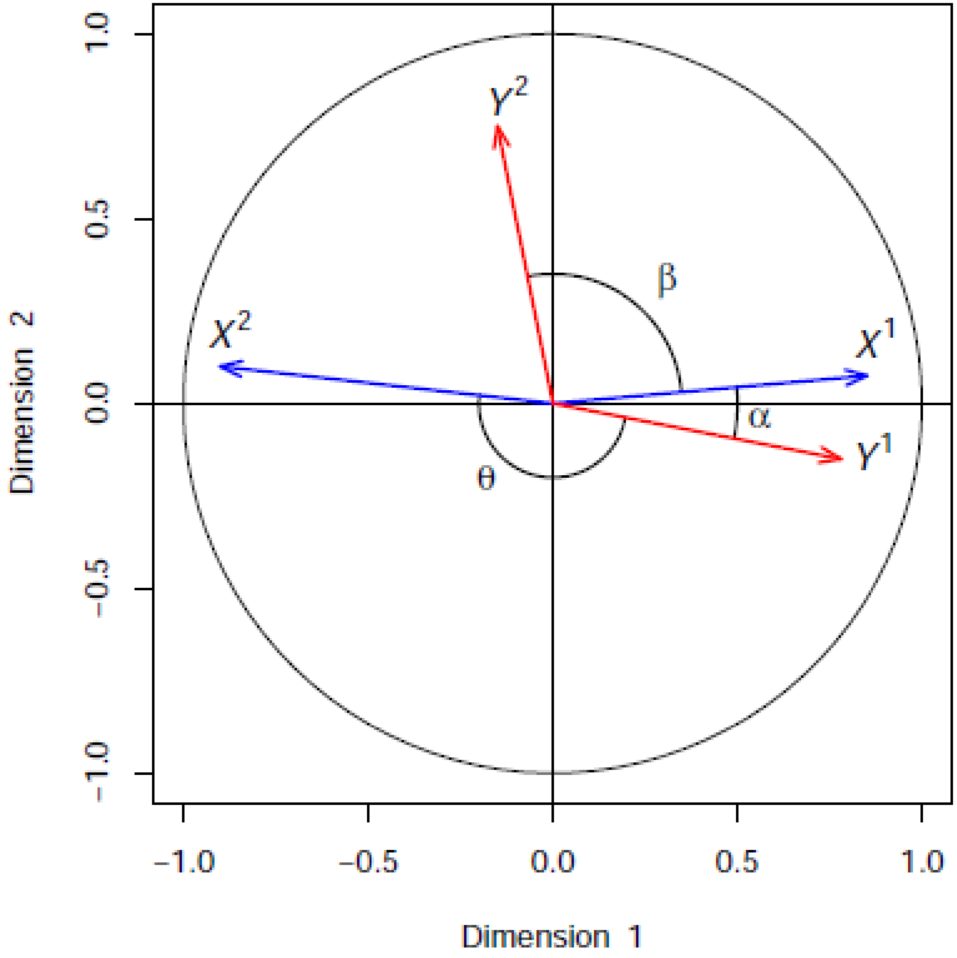
Variables plot



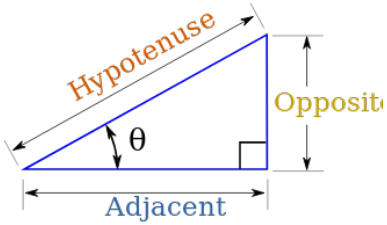
The coordinates of a variable X^j on a principal component PC^i is given by the correlation between this variable and the component PC^i.

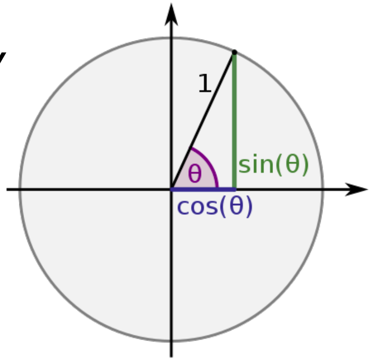
Variables plot

Correlation ≈ cosine



Remember trigonometry and right triangles:





The correlation between two variables is represented as:

- An acute angle (cos(α) > 0) if it is positive
- An obtuse angle (cos(θ) < 0) if it is negative
- A right angle (cos(β)≈0) if it is near zero

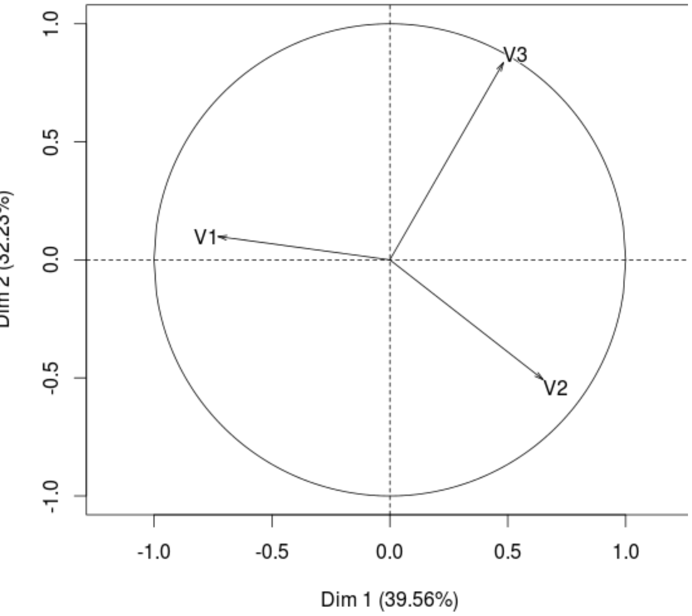
Variables plot

1)

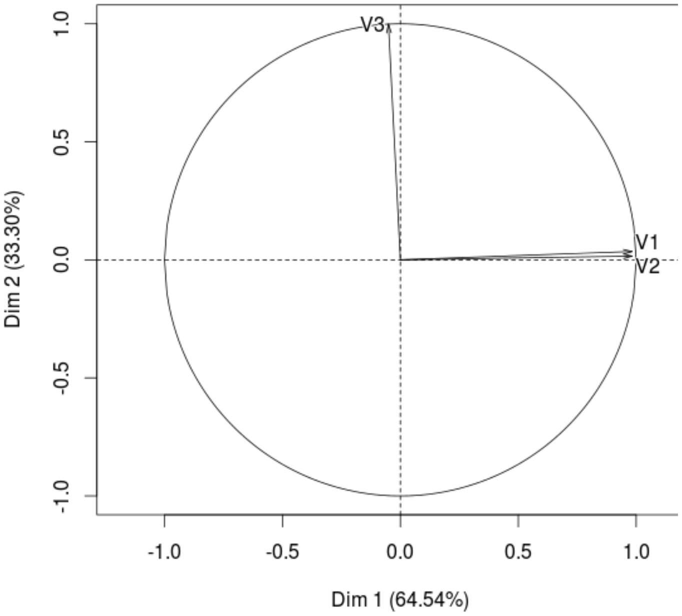
2)

3)

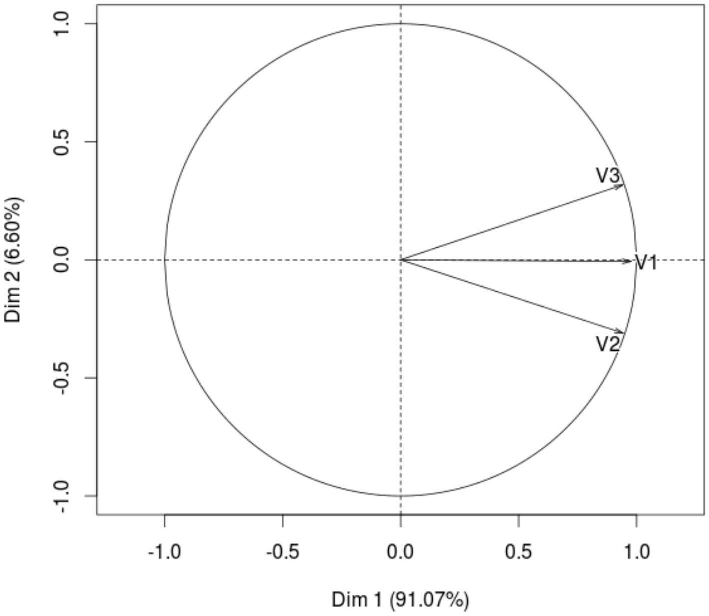
Variables factor map (PCA)



Variables factor map (PCA)



Variables factor map (PCA)



Correlation matrices

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

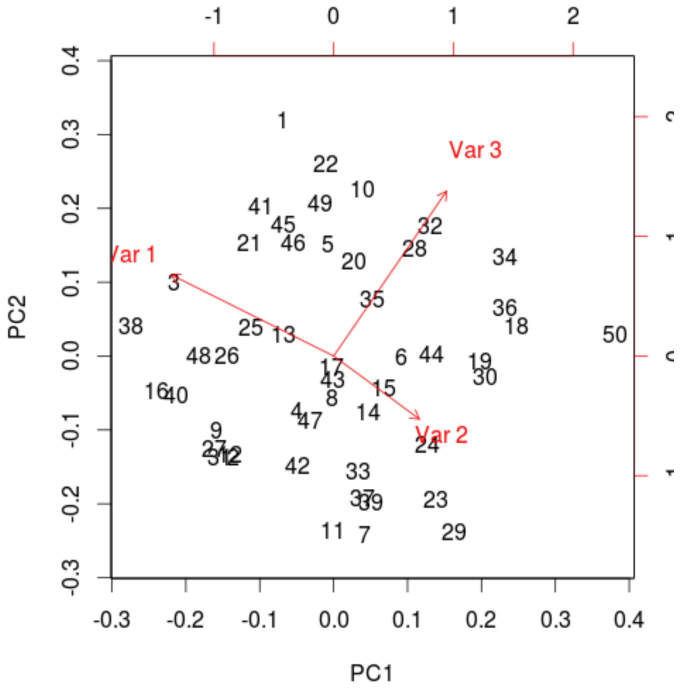
Example: biplot representation

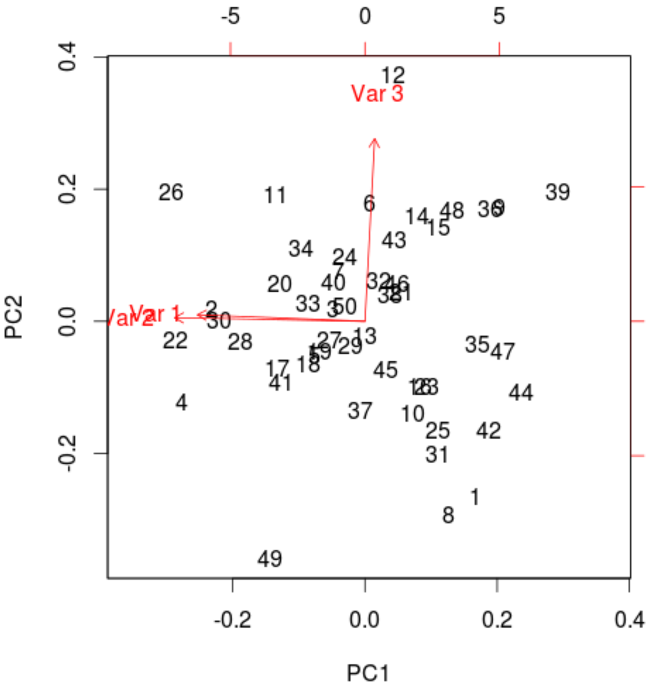
Individuals and variables are plotted on the same graph

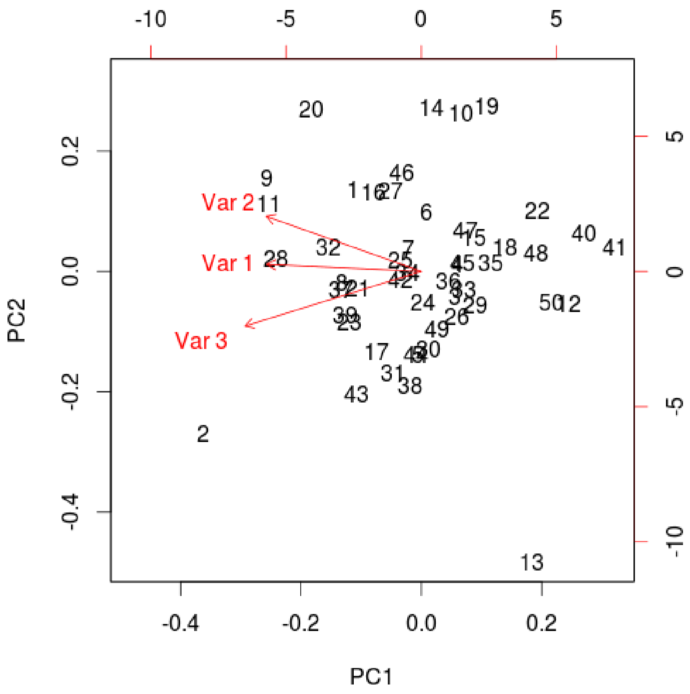
1)

2)

3)







And what about PCA?

- Mathematically, to perform a PCA consists in diagonalising the covariance (or the correlation for scaled PCA) matrix.
- Indeed, it can be shown that the sub-space in which the projected points have a maximal variance is given by the first eigen vectors of the covariance (or correlation) matrix ; the variance are given by the corresponding eigen values.
- The first eigen vector provides the direction (via the coefficients of the linear combination to apply to initial variables) that explains the greatest part of variability. The second explains the greatest past of the remaining variance and so on...

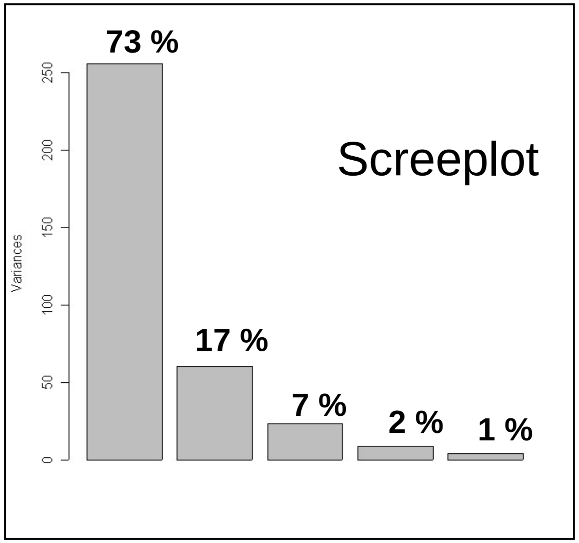
PCA: practical aspects

- Should I scale my data before performing PCA?
  - Without scaling: one variable with high variance will structure nearly alone the first principal component
  - With scaling: one noisy variable with low variability will be given the same variance as others meaningful variables
- Can I perform PCA with missing values?
  - Specific algorithms to deal with missing values exist (for instance, NIPALS - implemented in mixOmics). It can be used to impute missing values but it requires « many » components.

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

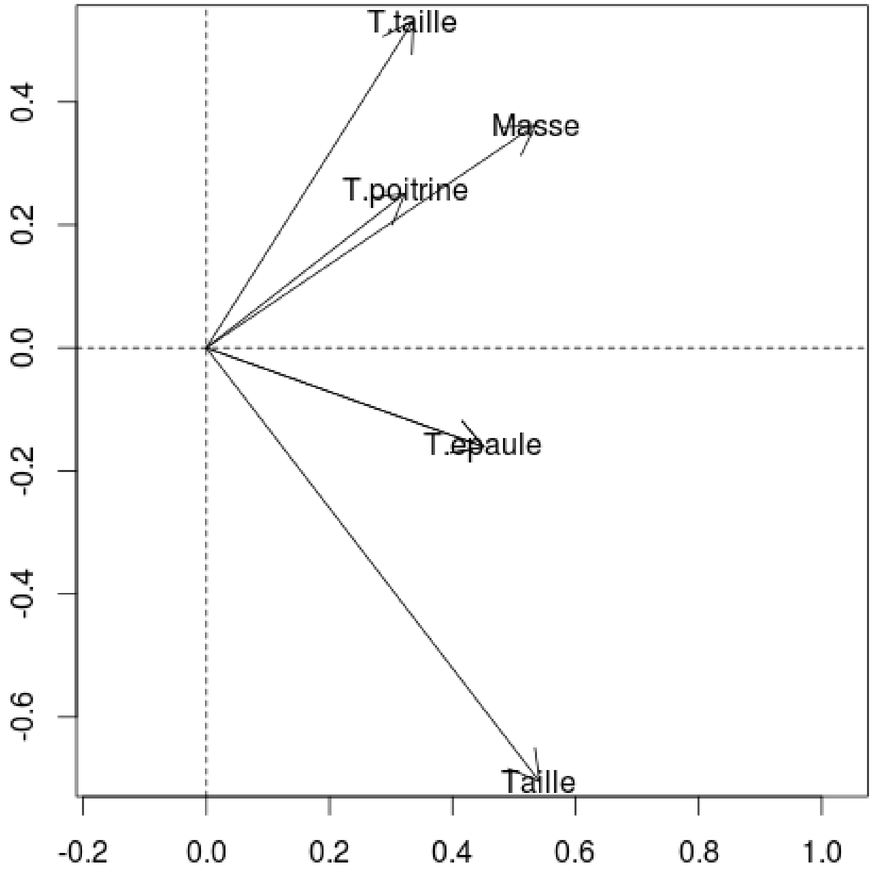
Gertrude Cox, 1900-1978, American statistician

PCA: *body* data set



- 90% of the variability is explained by the first two PCs
- 10% of the information is lost when projecting from 5 to 2 dimensions.
- PC 1 «beefyness»: separation of beefy people on the right (high values for the 5 variables) and weakling ones on the left.
- PC 2 «fatness, rotundity»: bottom, variables linked to height and shoulders; top, weight, waist and chest girth.

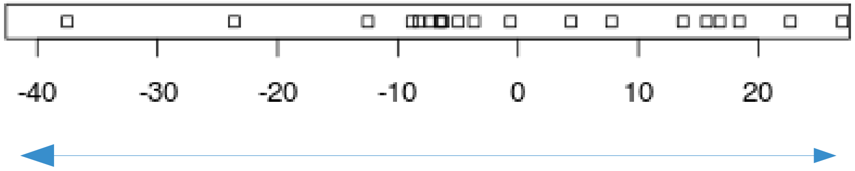
Variables plot

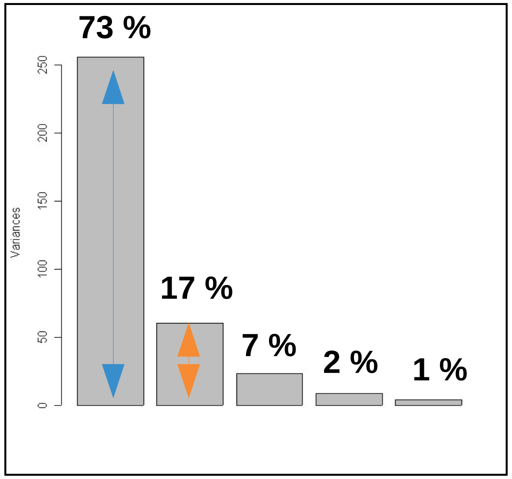


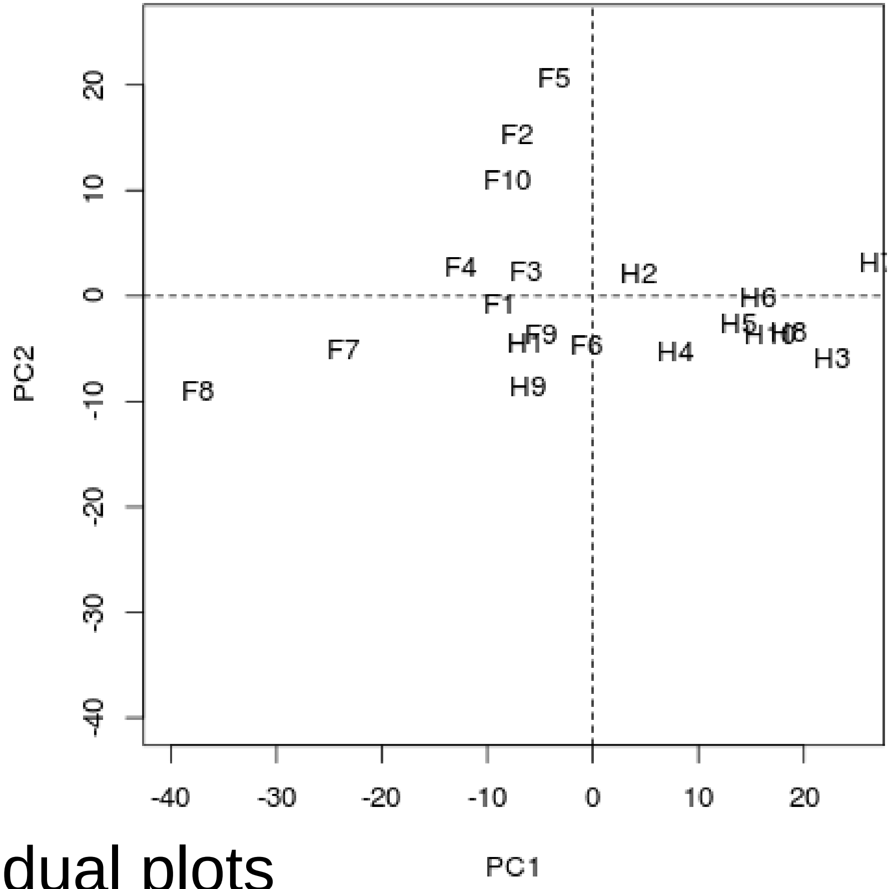
Correlation matrix

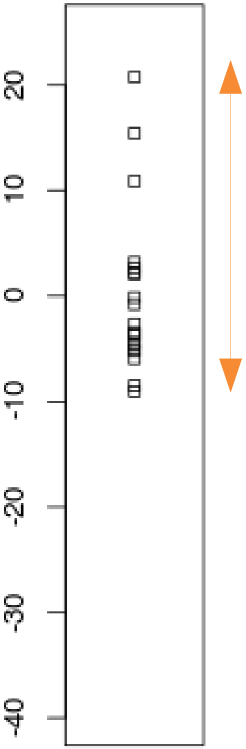
	T.ep	T.p	T.t	M	T
T.ep	1.00	0.74	0.48	0.72	0.71
T.p	0.74	1.00	0.78	0.81	0.51
T.t	0.48	0.78	1.00	0.86	0.37
M	0.72	0.81	0.86	1.00	0.61
T	0.71	0.51	0.37	0.61	1.00

PCA: *body* data set





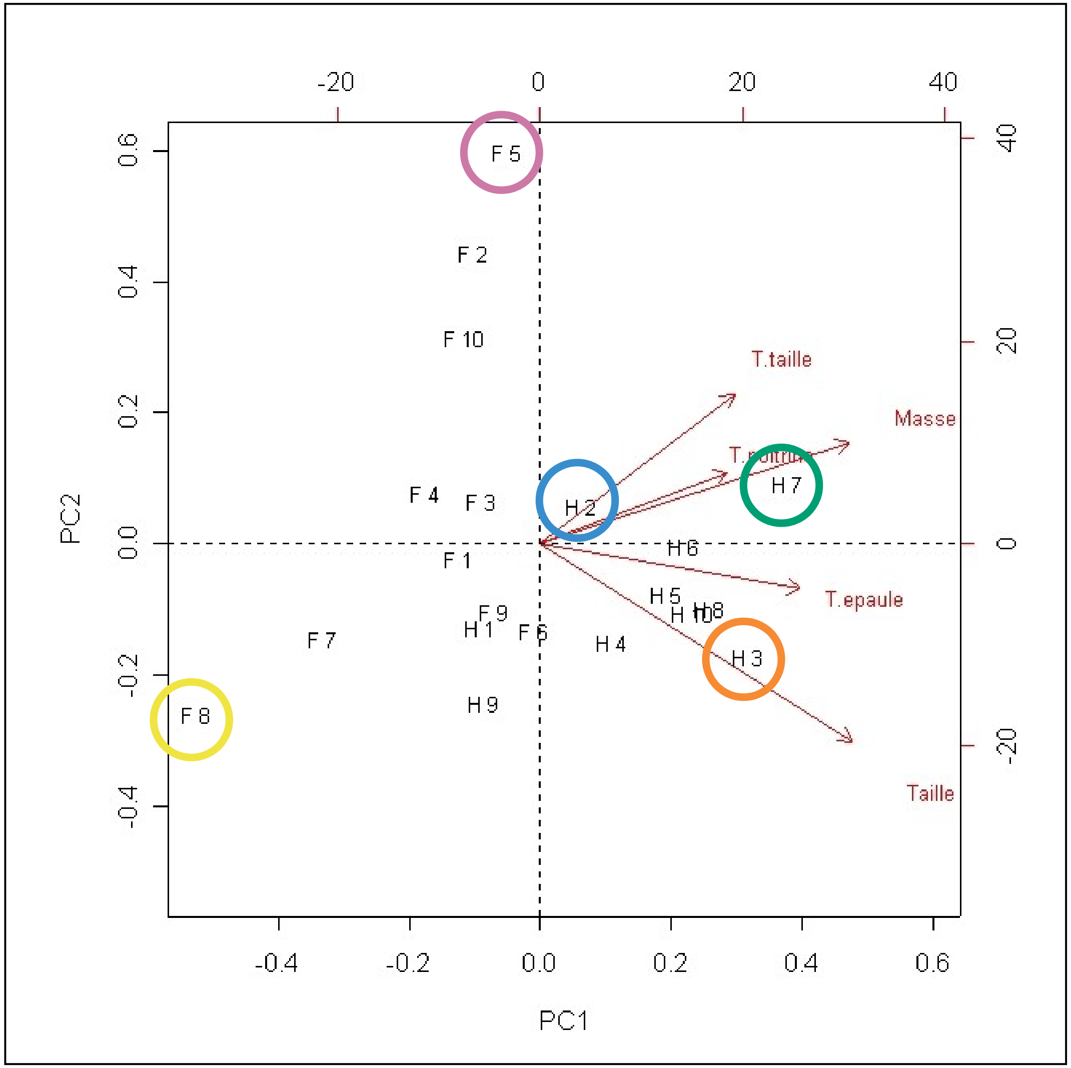




Individual plots

PCA: *body* data set

	s.g	c.g	w.g	w	h
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



Origin (coordinate (0,0)): average individual

s.g	c.g	w.g	w	h
108.1	94.2	75.4	70.6	174.4

PCA: *body* data set

Data

	s.g	c.g	w.g	w	h
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
Mean	108.1	94.2	75.3	70.6	174.4
Var.	68.6	37.5	50.8	85.7	109.3

Covariance matrix

	s.g	c.g	w.g	w	h
Shoulder.g	**68.64**	37.74	28.08	55.32	61.19
Chest.g	37.74	**37.51**	33.90	45.70	32.40
Waist.g	28.08	33.90	**50.77**	56.58	27.70
Weight	55.32	45.70	56.58	**85.71**	59.52
Height	61.19	32.40	27.70	59.52	**109.31**

68.64 + 37.51 + 50.77 + 85.71 + 109.31 = 351.94

351.94 represents (somehow) the quantity of information contained in the data.

PCA: *body* data set

Coefficients (optimally calculated) to build principal components

	Dim1	Dim2	Dim3	Dim4	Dim5
shoulder.g	0.45	-0.16	0.78	-0.18	0.36
chest.g	0.32	0.25	0.26	0.72	-0.49
waist.g	0.34	0.53	-0.33	0.24	0.66
weight	0.54	0.36	-0.17	-0.60	-0.44
height	0.54	-0.70	-0.43	0.17	0.02

```
PC1 = 0.45*shoulder.g + 0.32*chest.g
    + 0.34*waist.g + 0.54*weight + 0.54*height

PC2 = -0.16*shoulder.g + 0.25*chest.g
    + 0.53*waist.g + 0.36*weight — 0.70*height

PC3 = ...
```

Covariance matrix between PCs

	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 is the greatest value of variance that we can obtain on the individuals with a linear combination of the initial variables.

Coordinates of the individuals on the PCs

	Dim1	Dim2	Dim3	Dim4	Dim5
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
Mean	0	0	0	0	0
Var.	255.7	60.2	23.5	8.61	4.0

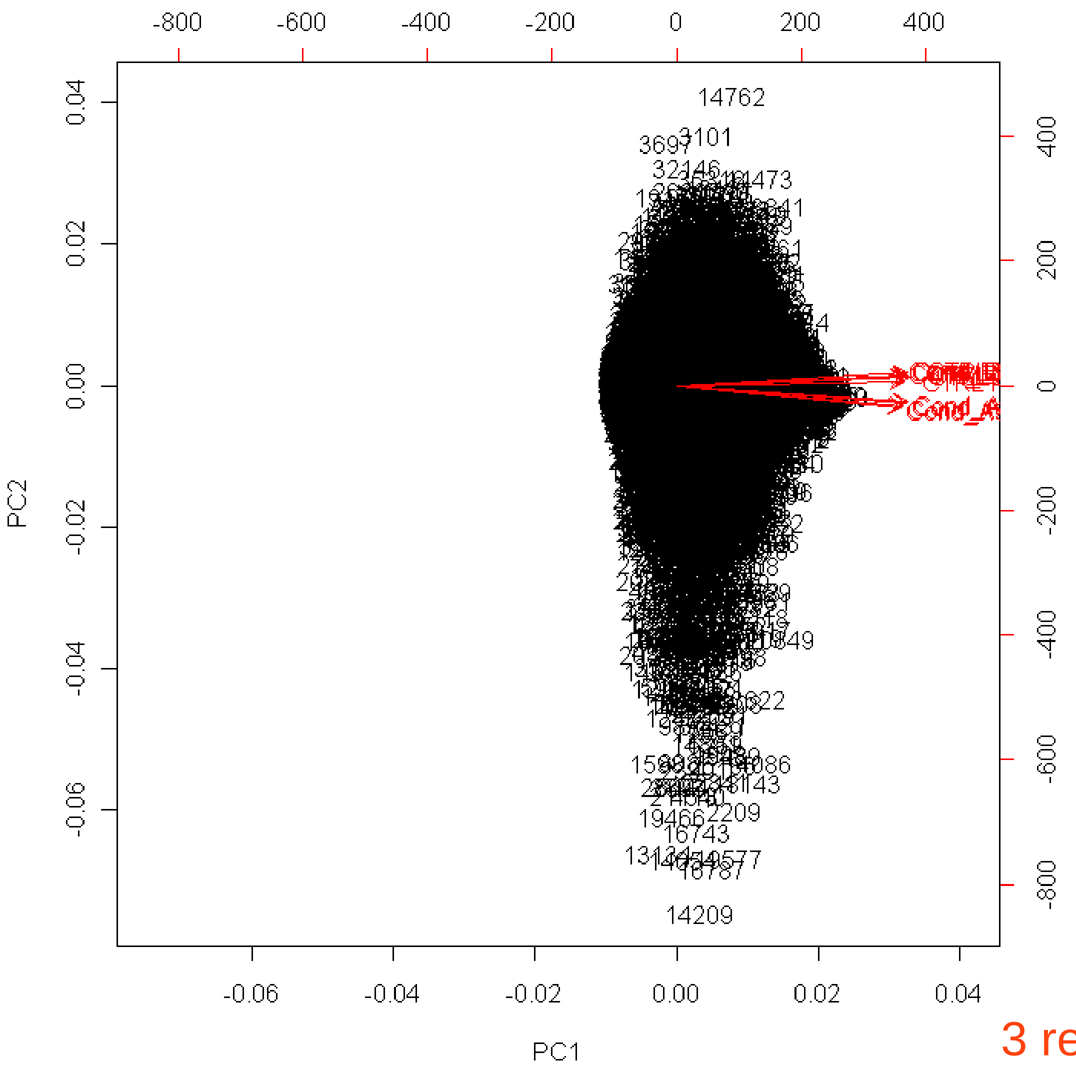
255.66 + 60.18 + 23.48 + 8.61 + 4.01
= 351.94

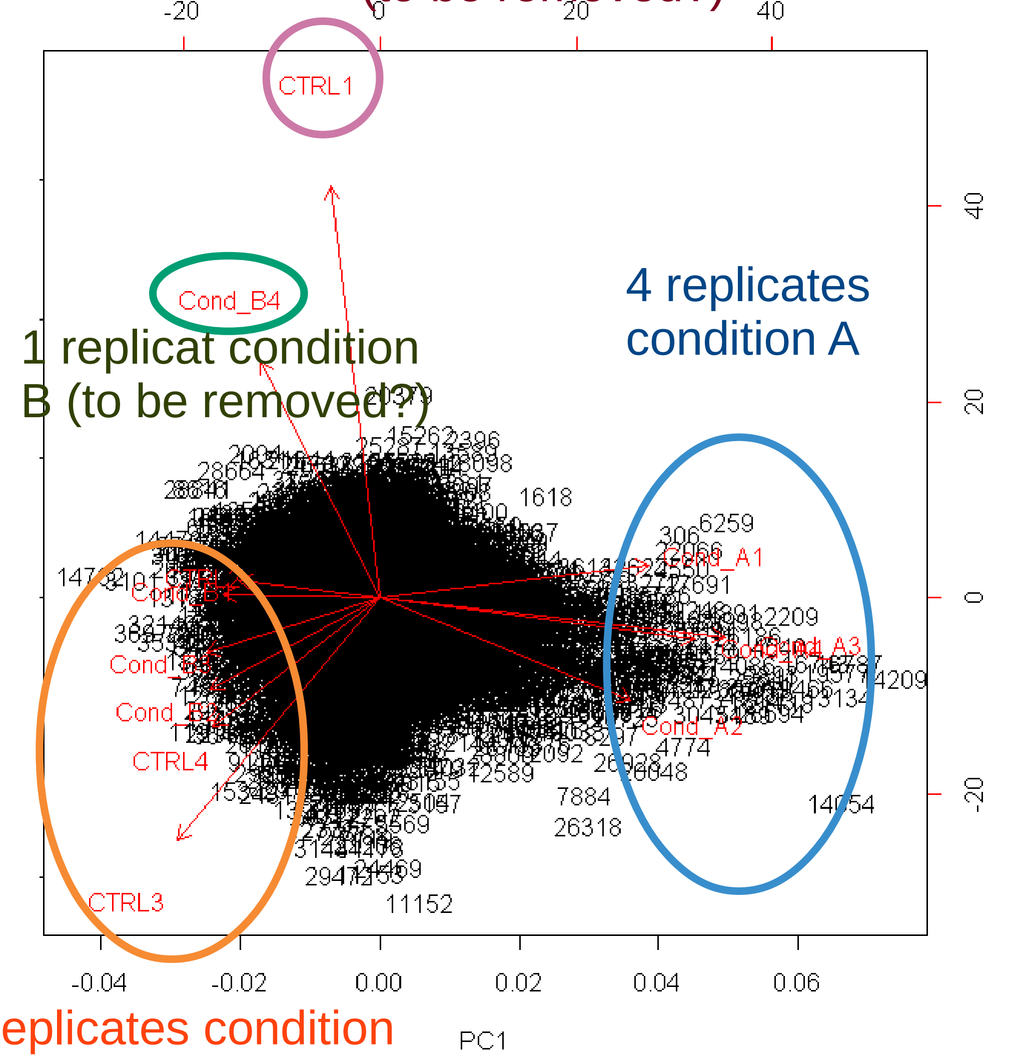
The same quantity of information (**351.94**) is kept but it is ``optimally'' allocated.

Biological data set (1)

PCA for quality control!

3 conditions, 4 replicates, 38000 genes, chip Affymetrix





1 replicate contrôle
(to be removed?)

1 replicat condition
B (to be removed?)

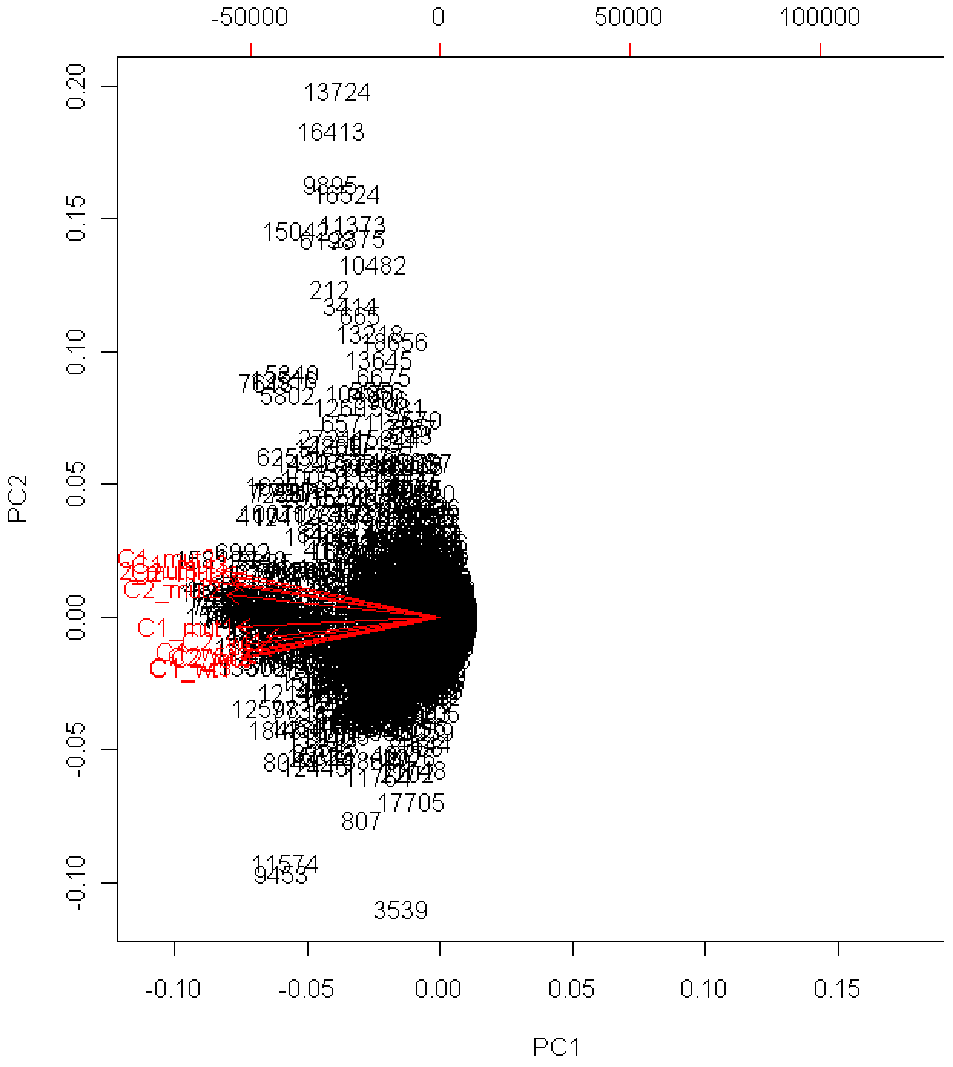
4 replicates
condition A

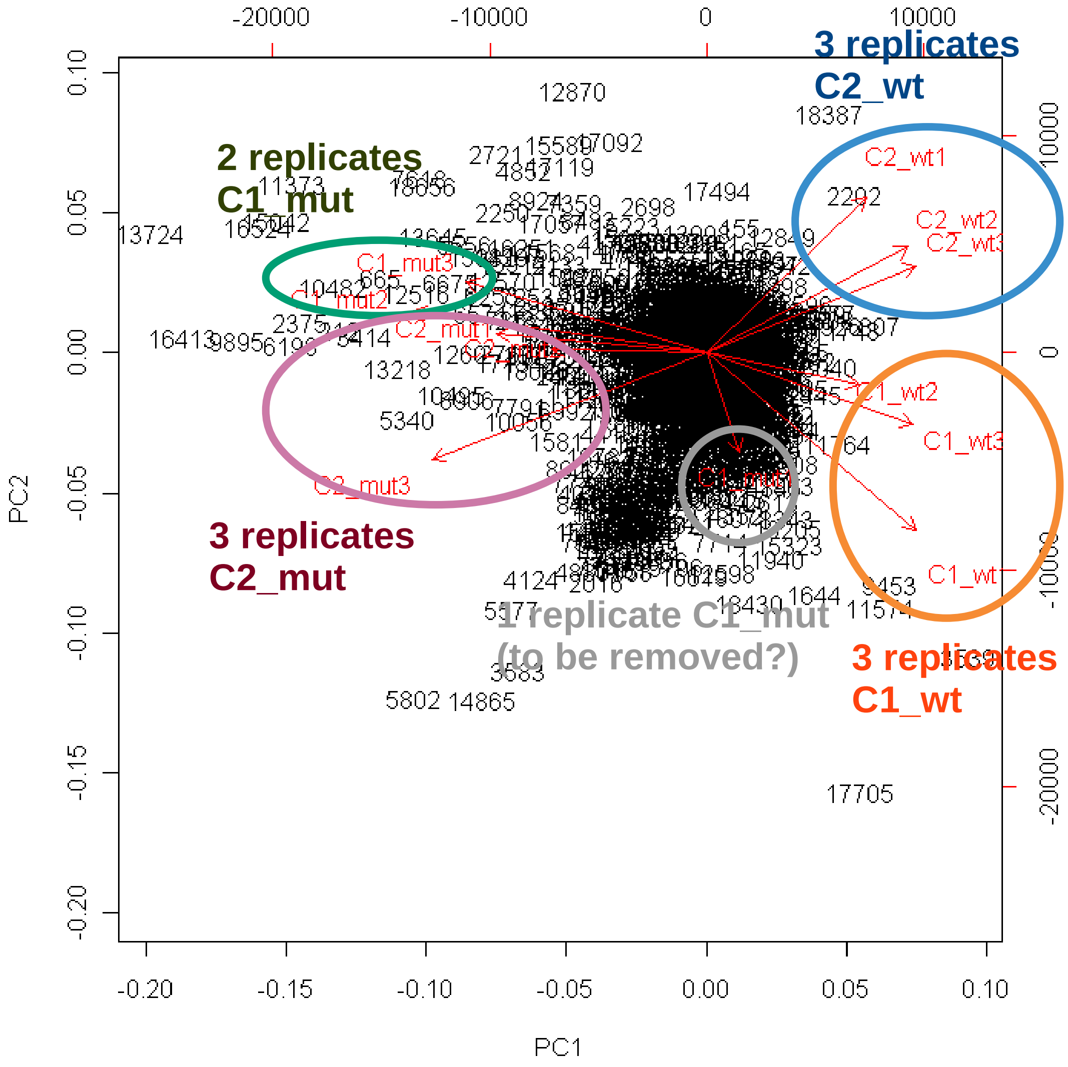
3 replicates condition
B et 3 replicats control

Biological data set (2)

PCA for quality control!

4 conditions (2 treatments * 2 genotypes), 3 replicates, 20000 genes (Affymetrix)





3 replicates C2_wt

2 replicates C1_mut

3 replicates C2_mut

1 replicate C1_mut (to be removed?)

3 replicates C1_wt

PCA = projection

- To interpret the graphical results of PCA must be done keeping in mind that one is looking at a projection on a plane (or in a volume for 3D representation).
- Be careful when interpreting visual proximities
- Illustration in comics with the *only true super-heros* ...

PCA = projection

★ *I'm TWO-D boy. The boy X-Y who doesn't care about the Z !*

¡MBATTABLE ?! C'EST BIEN VOUS ?!
HÉ, MAIS... VOTRE VOIX BIZARRE... C'EST JUSTE PARCE QUE VOUS ÊTES UN ADO !
EUH OUAIS BAH, JE... J'MUE EN CE MOMENT. C'EST PAS MA FAUTE...
BON. QU'EST-CE QUE VOUS VOULEZ ?
AHEM... HUM...
YO ! JE M'PRÉSENTE... JE SUIS TWO-D BOY !
LE BOY X-Y QUI SE RIT DU Z !
X-Y QUI... QUOI ?!
EUH... C'EST DES MATHS, M'SIEUR. AHEM... JE CONTINUE.
YO !... DÉMONSTRATION !
HUM... BON. Y A PLUS DE VOITURES. ALORS...
?!
MAIS ?!
ET HOP ! PUISSANCE TWO-D !
AND... ANOTHER ONE DÉMONSTRATION ! YO, YO !
HÉÉÉ !
KROUIIK
YO, TWO-D BOY EST DANS LA PLACE !
ARRÊTEZ ÇA IMMÉDIATEMENT, JEUNE HOMME, OU JE VOUS TAPE !
?! HÉ, NAN MAIS J'VEUX PAS M'BATTRE AVEC VOUS, MOI, M'SIEUR !
VOUS ÊTES MON IDOLE !
HEIN ?
J'VOULAIS VOUS RENCONTRER PARCE QU'AU COLLÈGE, LES PROFS, ILS NOUS OBLIGENT À FAIRE UN STAGE...
ET MOI, J'VOUDRAIS ÊTRE SUPER-HÉROS. ALORS...

Scenario & illustration
Pascal Jousselin

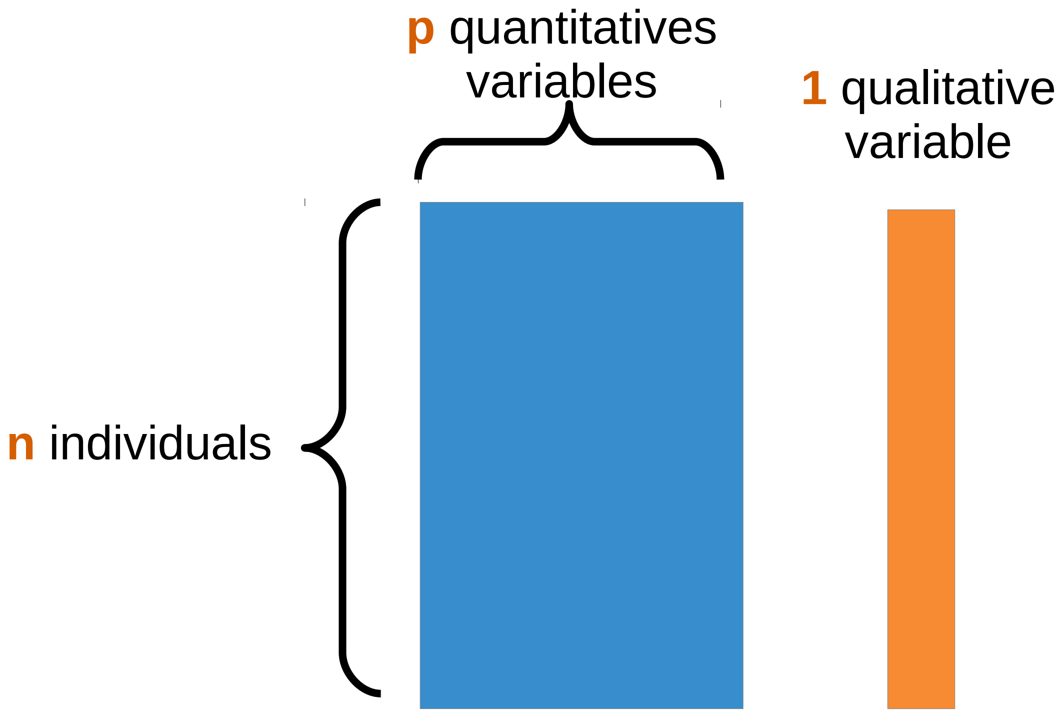
Colour
Laurence Croix

Web
`pjousselin.free.fr`

Discriminant analysis

Linear Discriminant Analysis (LDA)

Explore a data set composed of **quantitative** variables and **one qualitative** variable in order to separate the individuals based on their membership to the categories of the qualitative variable.



Body data set

Can we found a space where the projection of the individuals will separate men and women (qualitative variable S) according to the 5 body measurements (V1 to V5)?

	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	W
I 12	100.2	94.1	79.6	75.5	160.0	W
I 13	99.1	90.8	77.9	68.2	172.7	W
I 14	107.6	97.0	69.6	61.4	162.6	W
I 15	104.0	95.4	86.0	76.8	157.5	W
I 16	108.4	91.8	69.9	71.8	176.5	W
I 17	99.3	87.3	63.5	55.5	164.4	W
I 18	91.9	78.1	57.9	48.6	160.7	W
I 19	107.1	90.9	72.2	66.4	174.0	W
I 20	100.5	97.1	80.4	67.3	163.8	W

LDA: simulated example

Data set

- 50 individuals, 4 variables
- 3 quantitatives V1 – V2 – V3
- 1 qualitative Group with 2 categories A and B

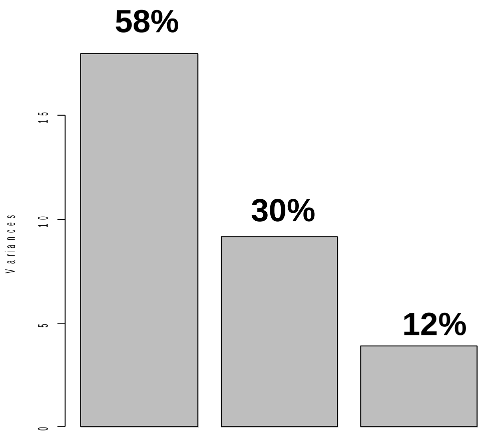
Can we find a space where the projections of the individuals from groups A and B are well separated?

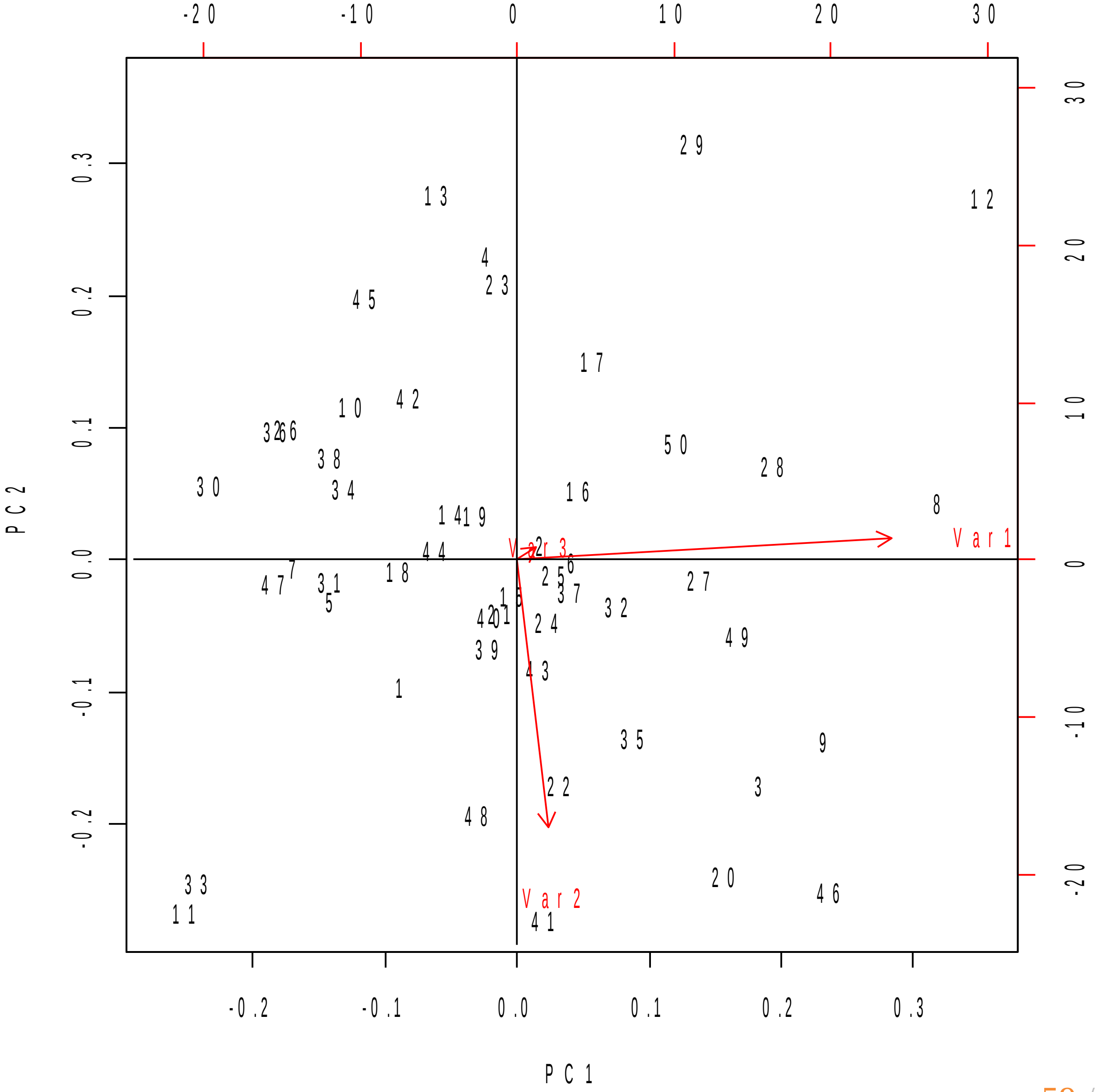
```
             V1      V2      V3
Mean          0       0       0
Variance     20      10       2
```

```
      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
```

LDA: simulated example

Results of a **PCA** applied only on the quantitative variables (**without considering the qualitative variable**).

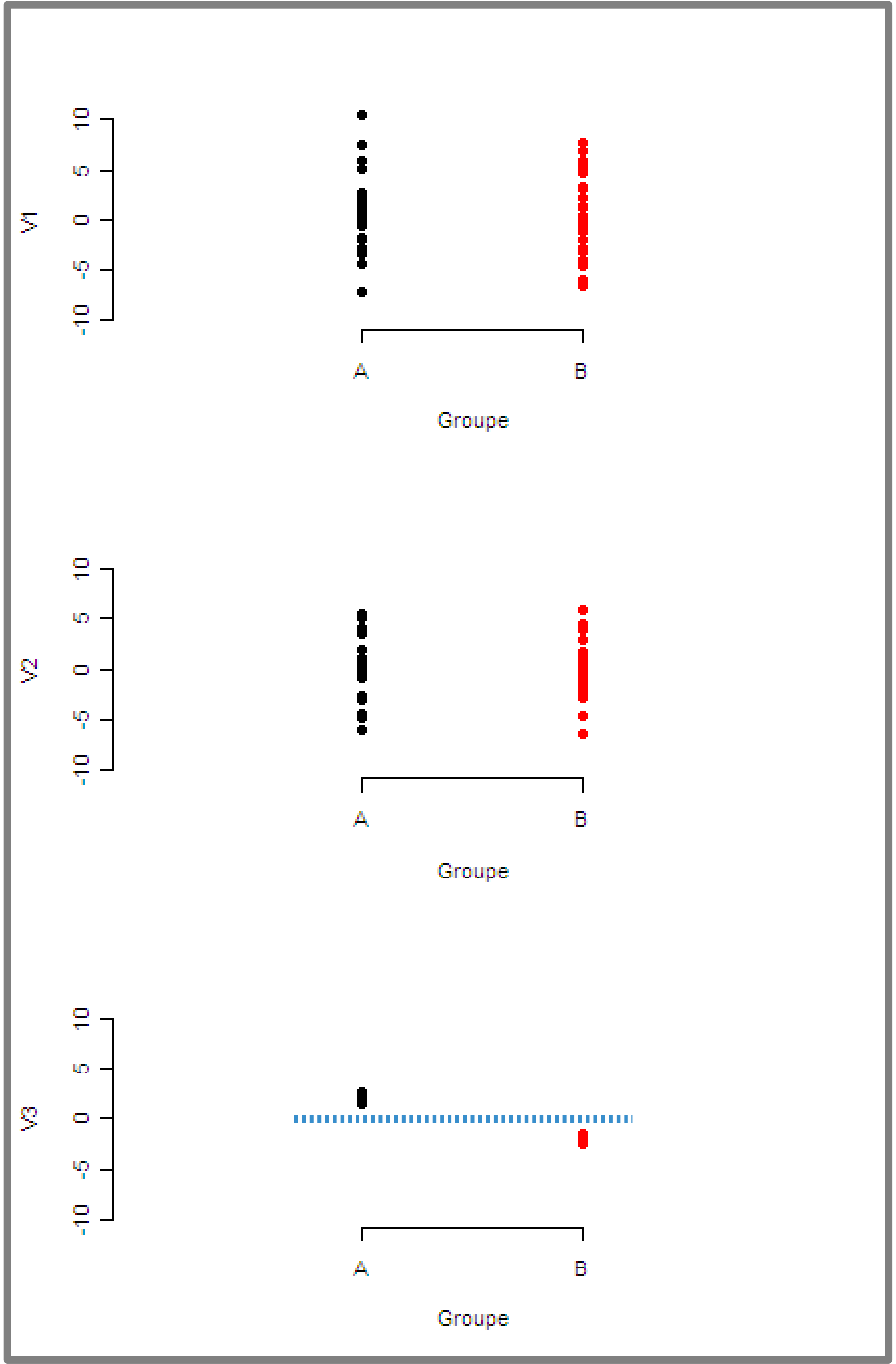




- The 3 PC are clearly identified to the 3 initial variables.
- Most part of the variability in the data is explained by V1, then by V2, then by V3.

LDA: simulated example

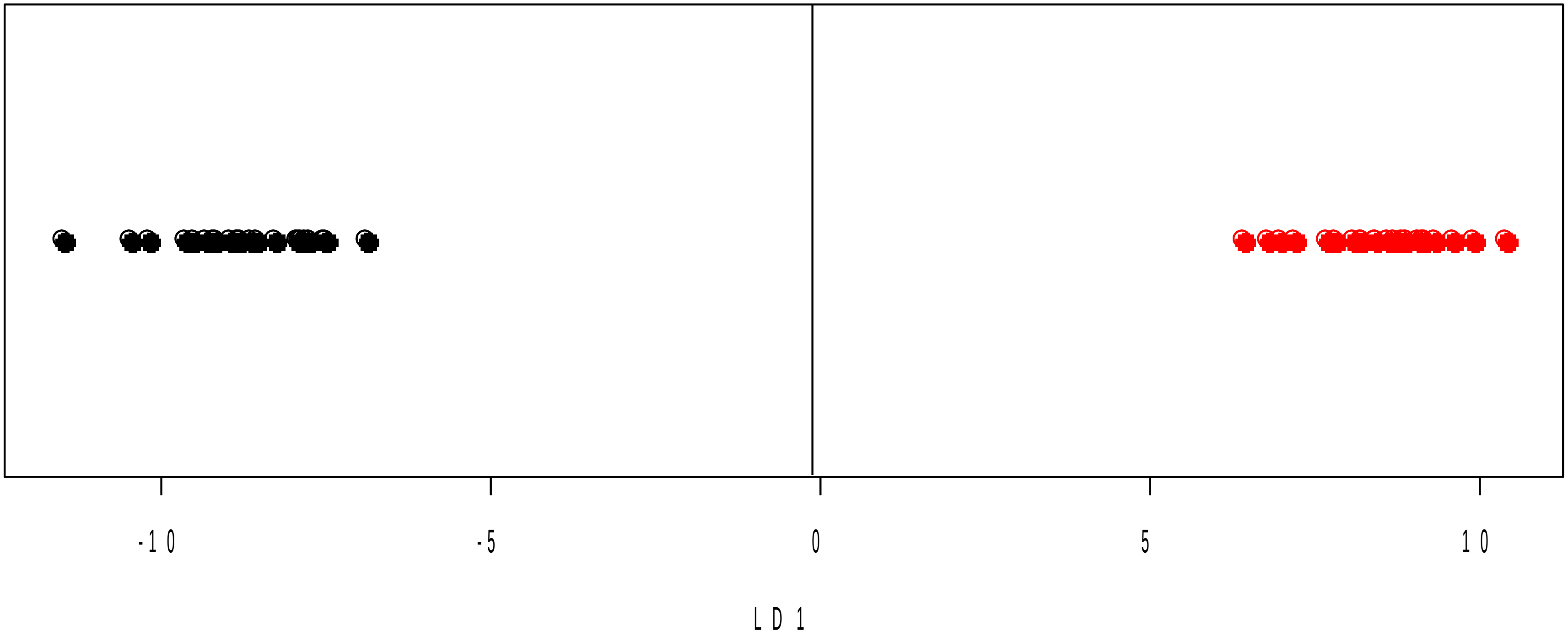
Representation of the 50 individuals according the 3 variables. Color depends on the categorie (A-black or B-red)



Although displaying the smallest variability, V3 is relevant when addressing a discrimination purpose.

LDA: simulated example

LDA result

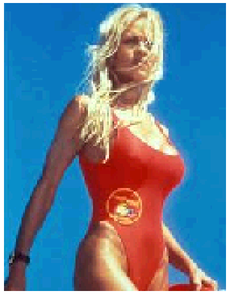


- 2 categories → 1 discriminant variable (graphical representation in)
- Linear combination of the initial variables:

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

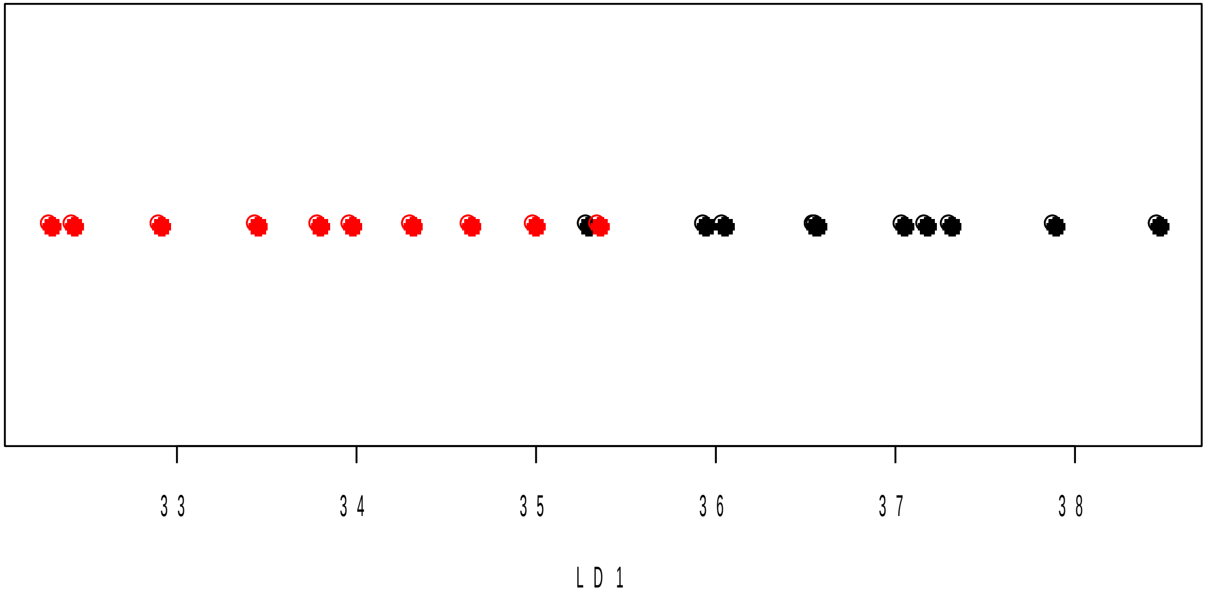
- LD1 roughly corresponds to V3 (with negative coefficient, but the sign doesn't matter).

LDA: body data set



```
Centroïd of the 2 groups
        s.g          c.g      w.g   w       h         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
shoulder g.    0.12
chest g.      -0.02
waist g.       0.11
weight        -0.11
height         0.14
```



The coefficients indicates that chest girth is the less discriminant variable (loading -0.02)... The other variables participate nearly in the same way (loadings around 0.1 in absolute value).

LDA: principle

- LDA is similar to a PCA performed on the centroid of the groups determined by the categories of the qualitative variable.
- Thus, we are looking for a sub-space of small dimension in which the centroids are the furthest possible (having a maximal variability)
- If the number of categories is 2, then the dimension the sub-space is 1; so LDA will provide only LD1.

Decision-making with LDA

- For a supplementary individual, when knowing the quantitative variables, the decision-making problem relies on the affectation of this individual to a categorie of the qualitative variable.
- Naive (and not so bad) rule: affect the new individual to the categorie whose centroid is the closest (many others more sophisticated rules exist).
- Application: credit scoring, quality control, diagnostic...

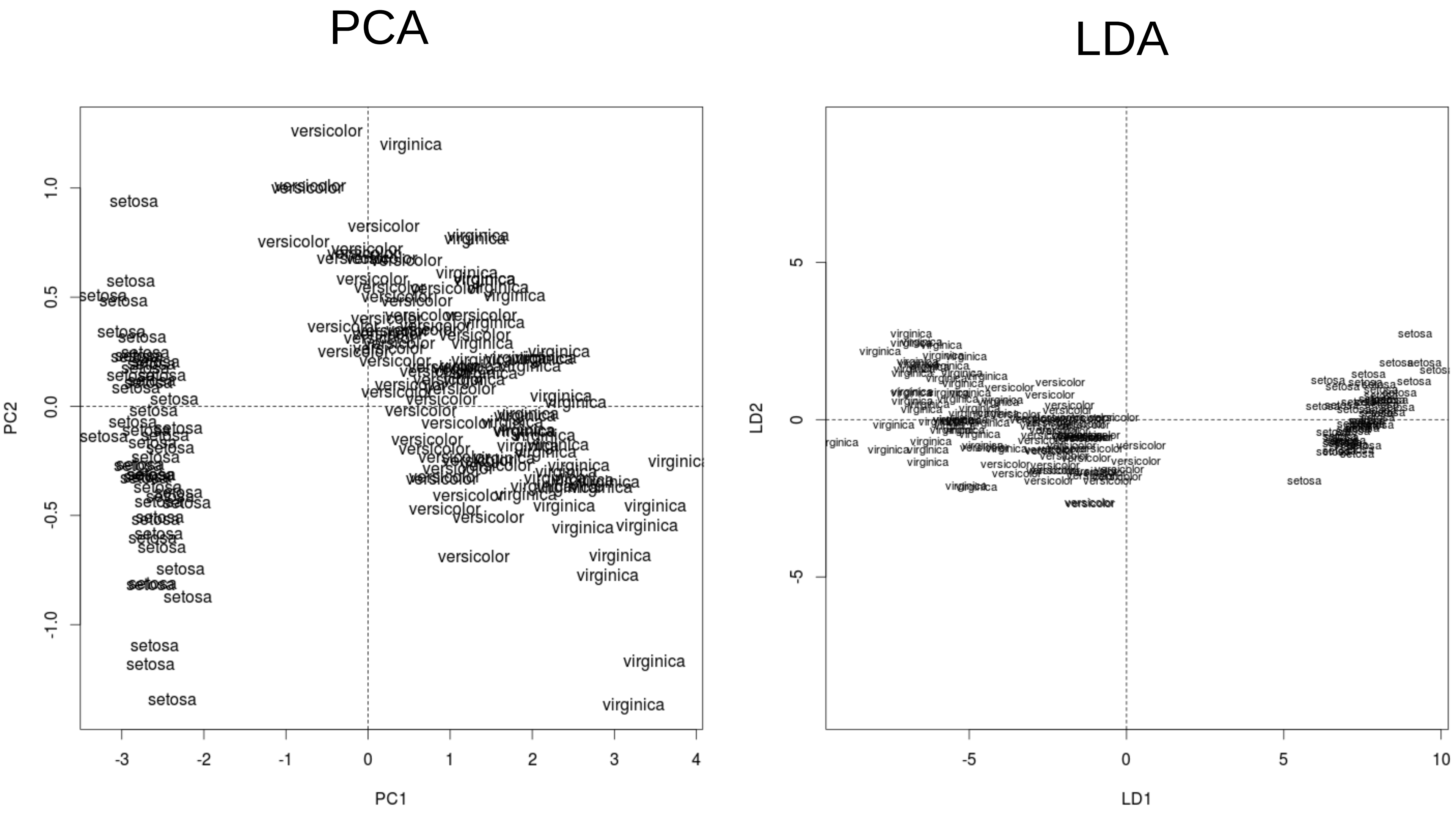
Iris data set

```
    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

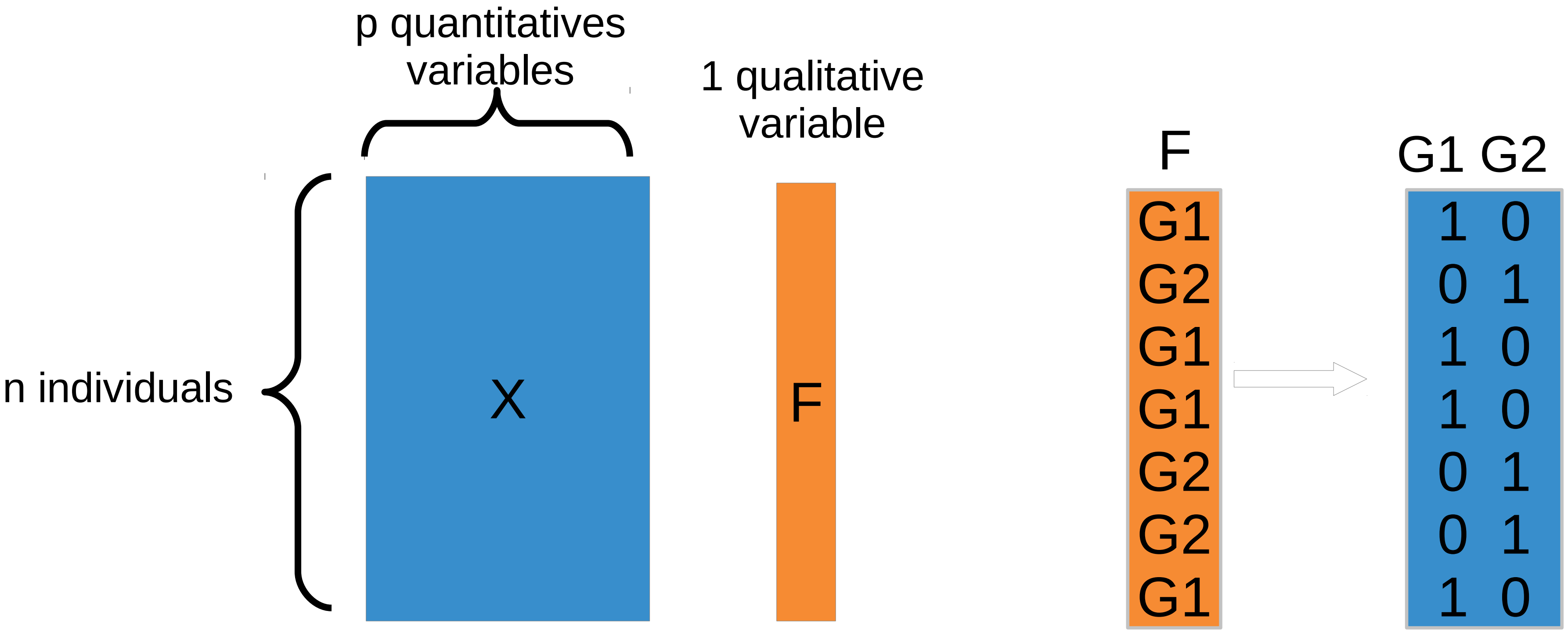
Iris data set



PCA

LDA

Projection to Latent Structure Discriminant Analysis (PLS-DA)



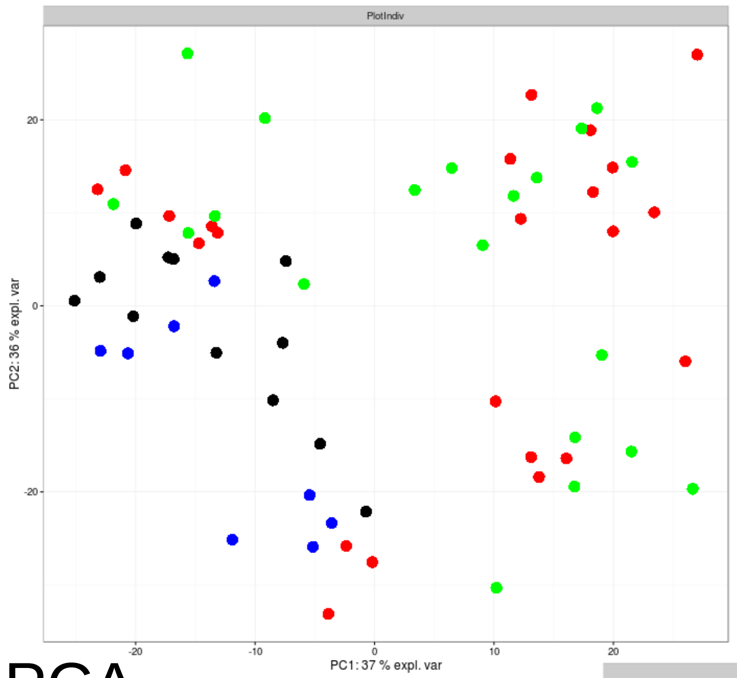
F
G1
G2
G1
G1
G2
G2
G1

G1	G2
1	0
0	1
1	0
1	0
0	1
0	1
1	0

The PLS regression(*) has been extended to deal with discrimination issues. To do that, the qualitative variable is converted into a dummy matrix (composed of 0 and 1) with as many rows as individuals and as many columns as catagories of the qualitative variable.

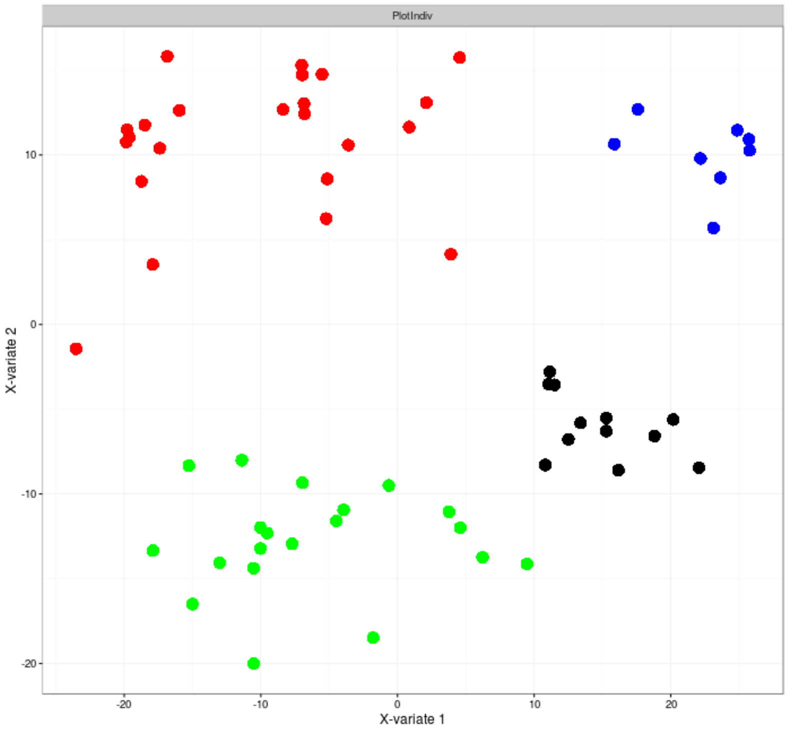
Please wait for few slides (section integration) to know more about PLS!

Comparison ACP-PLSDA

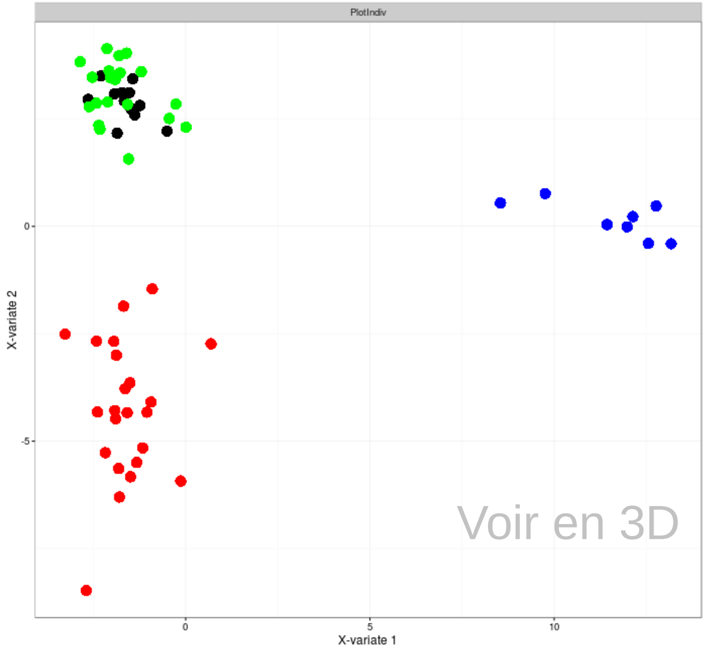


PCA

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



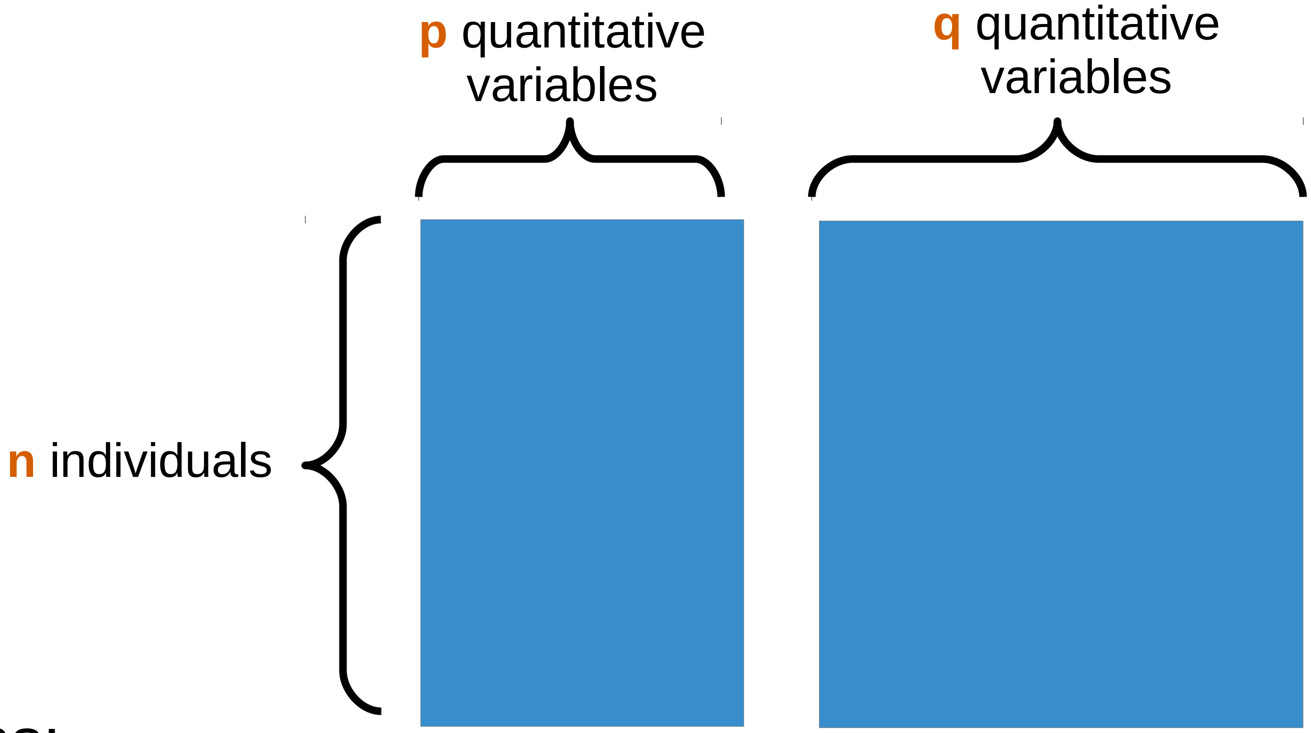
Voir en 3D

PLS-DA with variables selection (see extensions *sparse*)

Data integration

Data integration

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

CCA: simulated example

X

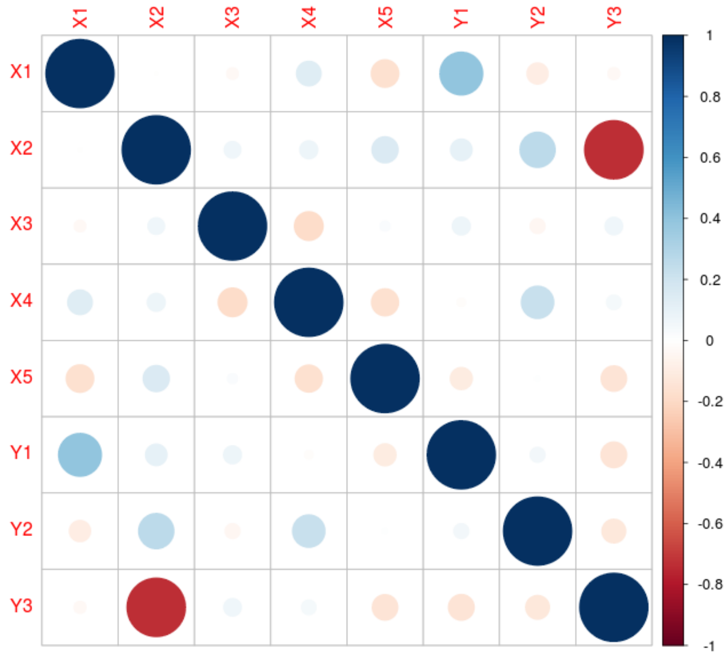
X1	X2	X3	X4	X5
0.87	0.31	0.24	0.06	0.29
0.76	0.8	0.52	0.1	0.95
0.65	0.76	0.57	0.1	0.17
0.86	0.47	0.00	0.21	0.75
0.65	0.46	0.41	0.23	0.86
0.11	0.56	0.84	0.14	0.49
0.85	0.81	0.42	0.65	0.39
0.74	0.73	0.15	0.81	0.80
0.75	0.30	0.72	0.48	0.99
0.55	0.06	0.30	0.87	0.67
0.41	0.52	0.21	0.51	0.59
0.59	0.87	0.99	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

Correlation matrix (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

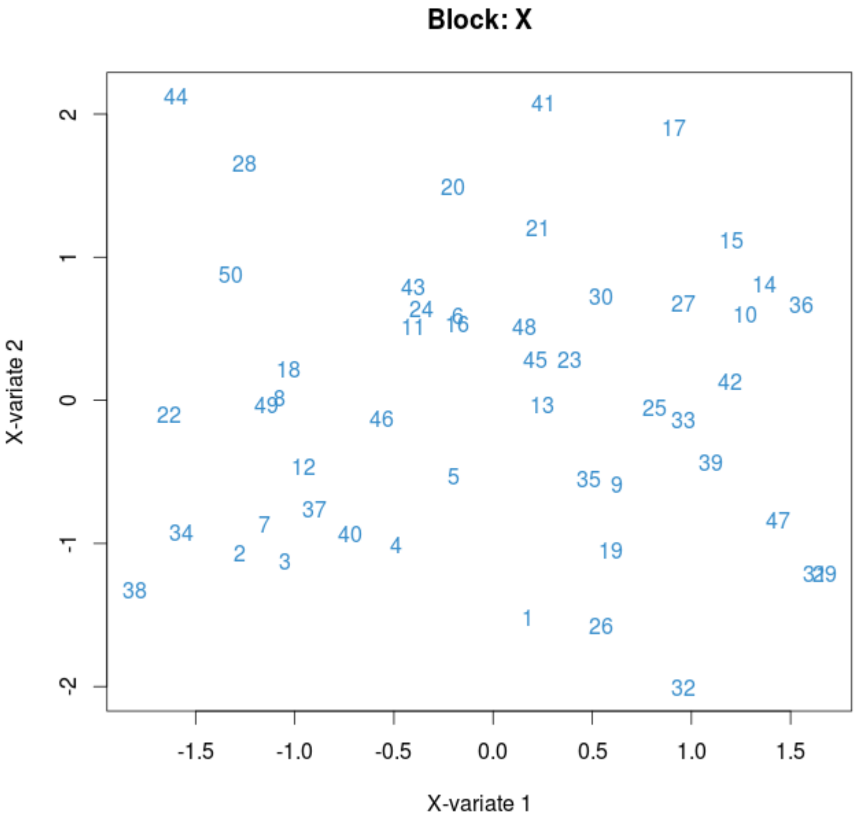


Package R corrplot

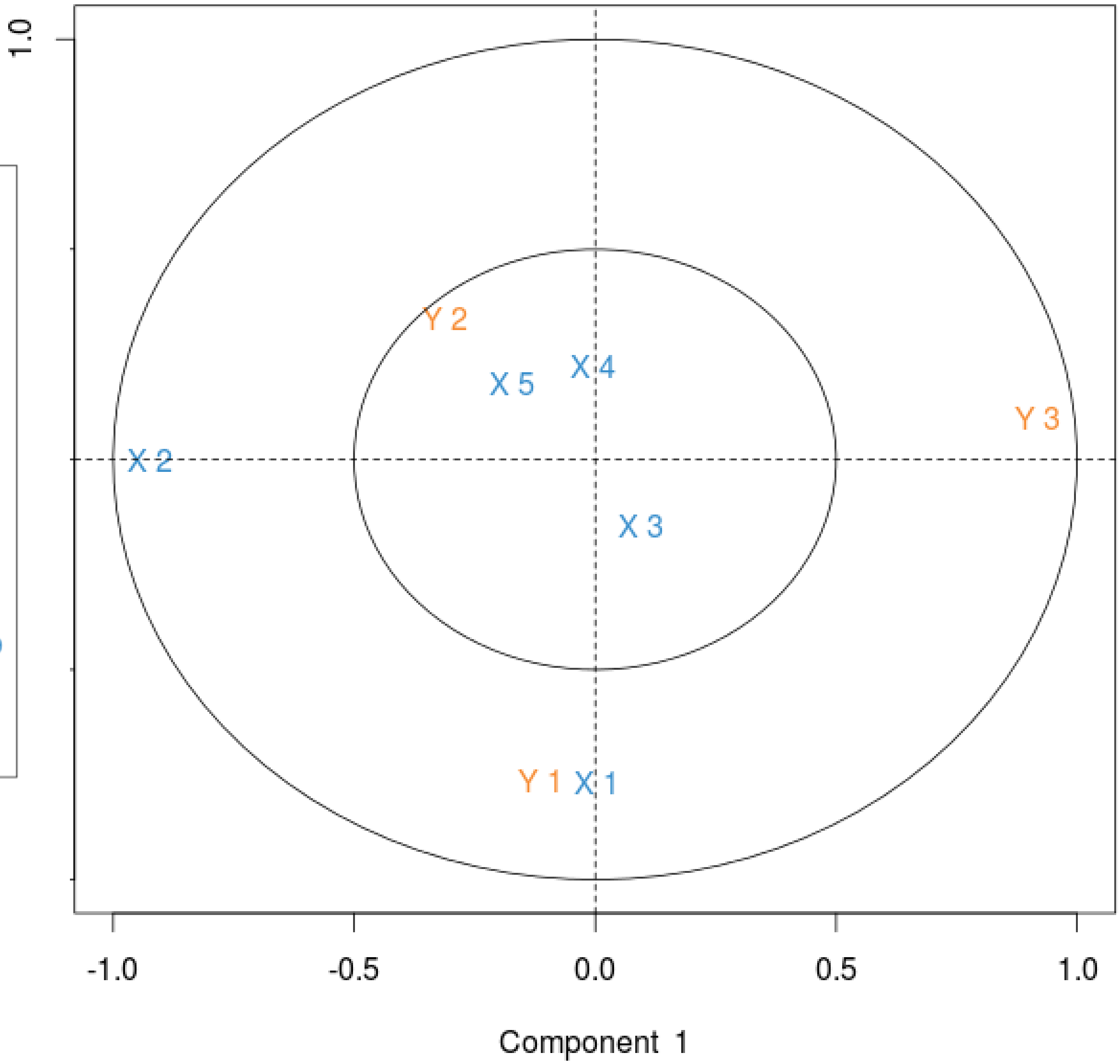
CCA: simulated example

Graphical outputs

Individual plot



Variable plot



CCA: principle

- The CCA can be viewed as an interative algorithm (like PCA)
  - Maximize the correlation (ρ_1) between two linear combinations: one from variables X (t_1), the other from variables Y (u_1).

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

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

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

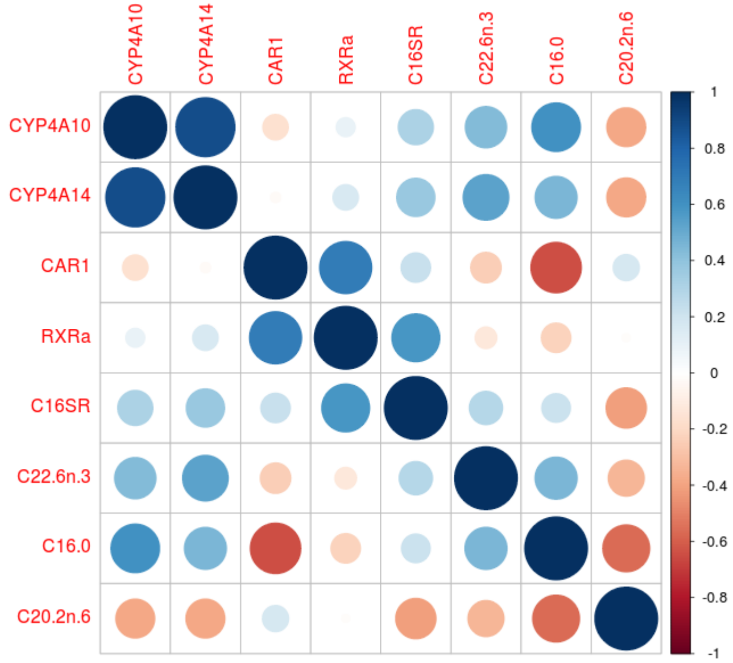
t_1 and u_1 are the first canonical variates and p_1 is the first canonical correlation.

  - For next levels, iterate the process under orthogonality constraints
- CCA is analog to PCA for the production and the interpretation of graphical outputs.
- Mathematical aspects are in the same vein as PCA (eigen decomposition of matrices)

CCA: nutrimouse data set

- 40 mice (2 genotypes)
- Expression of 5 genes
- Concentration of 3 lipids

Question: are there any genes related to lipids?



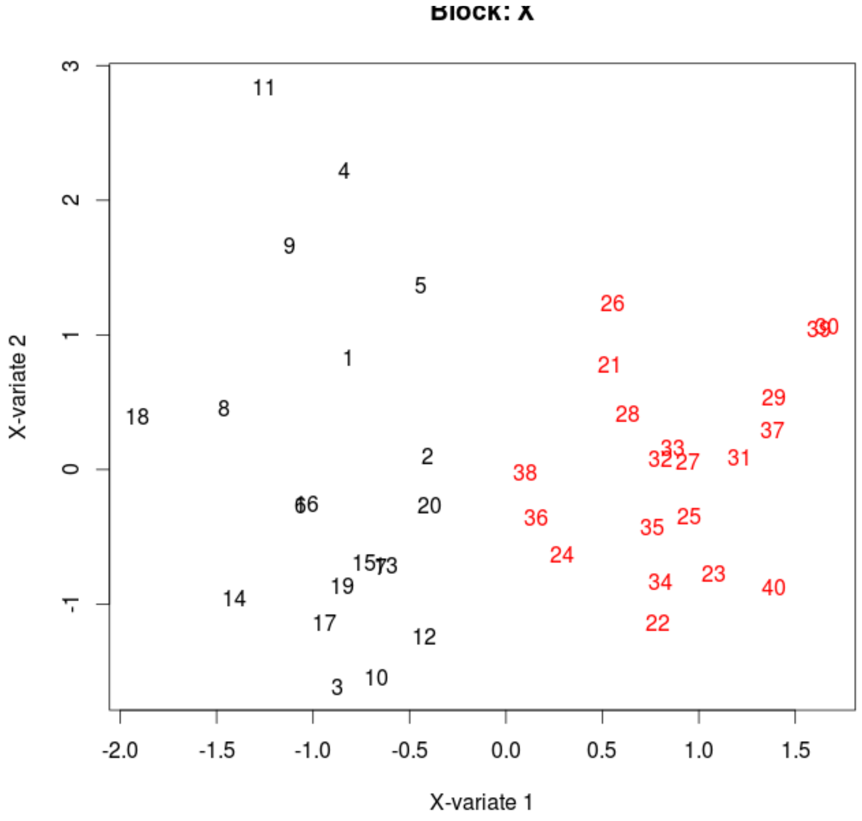
Correlation matrix

```
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
```

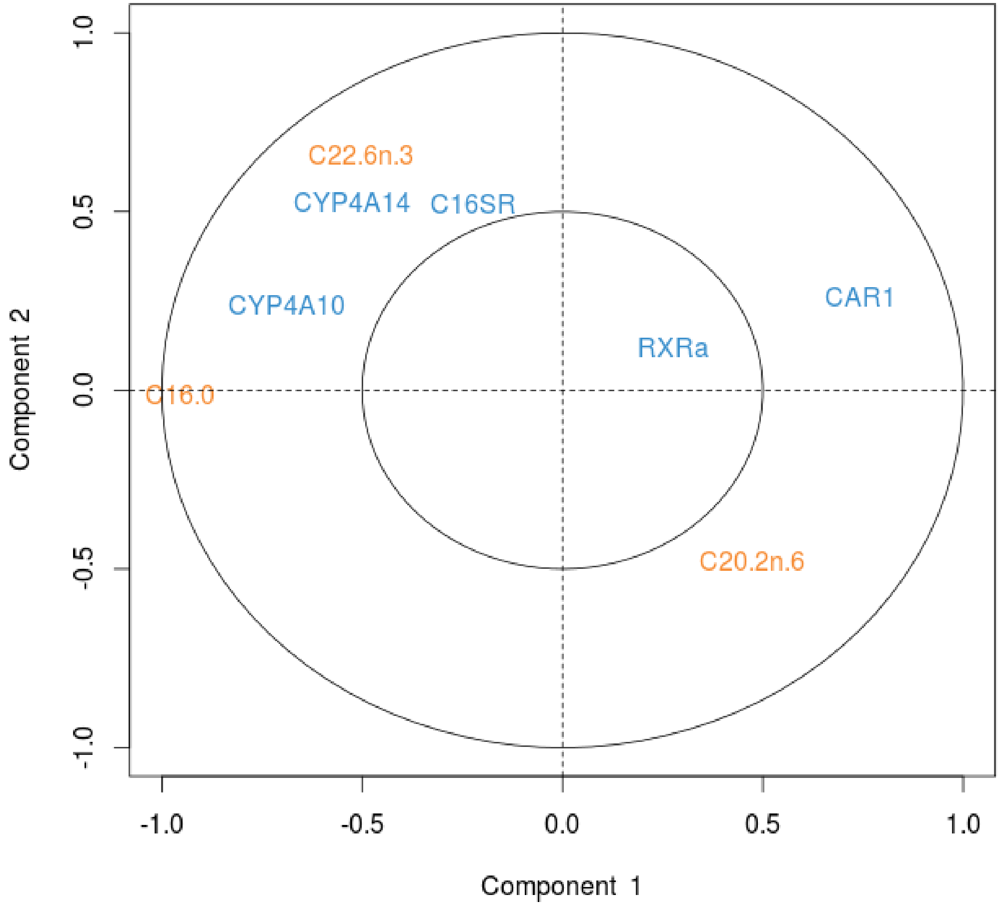
CCA: nutrimouse data set

Individual plot

color depending on the genotype added a posteriori



Variable plot



Canonical correlations : 0.853 0.627 0.253

CCA: a fundamental method...

- If one data set has only one quantitative variable, CCA is equivalent to **multiple linear regression**.
- If one data set is a dummy matrix corresponding to the categories of a qualitative variable, CCA is equivalent to **Linear Discriminant Analysis**.
- If the two data sets are dummy matrices corresponding to the categories of two qualitative variables, CCA is equivalent to **Correspondance Analysis**.

... with limits

- CCA can only be perfomed with « enough » observations: n >> p+q (sounds like a joke regarding ‘omics data...)
- Variables X and Y must not be « too » correlated
- Alternative: regularised CCA

Alternatives

- PLS related methods. In PLS, the algorithm is equivalent to find linear combinations from X and Y variables that have the greatest covariance.

- Regularized CCA. Apply a « ridge » penalty using regularization parameters to make the computation possible.

CCA: simulated example

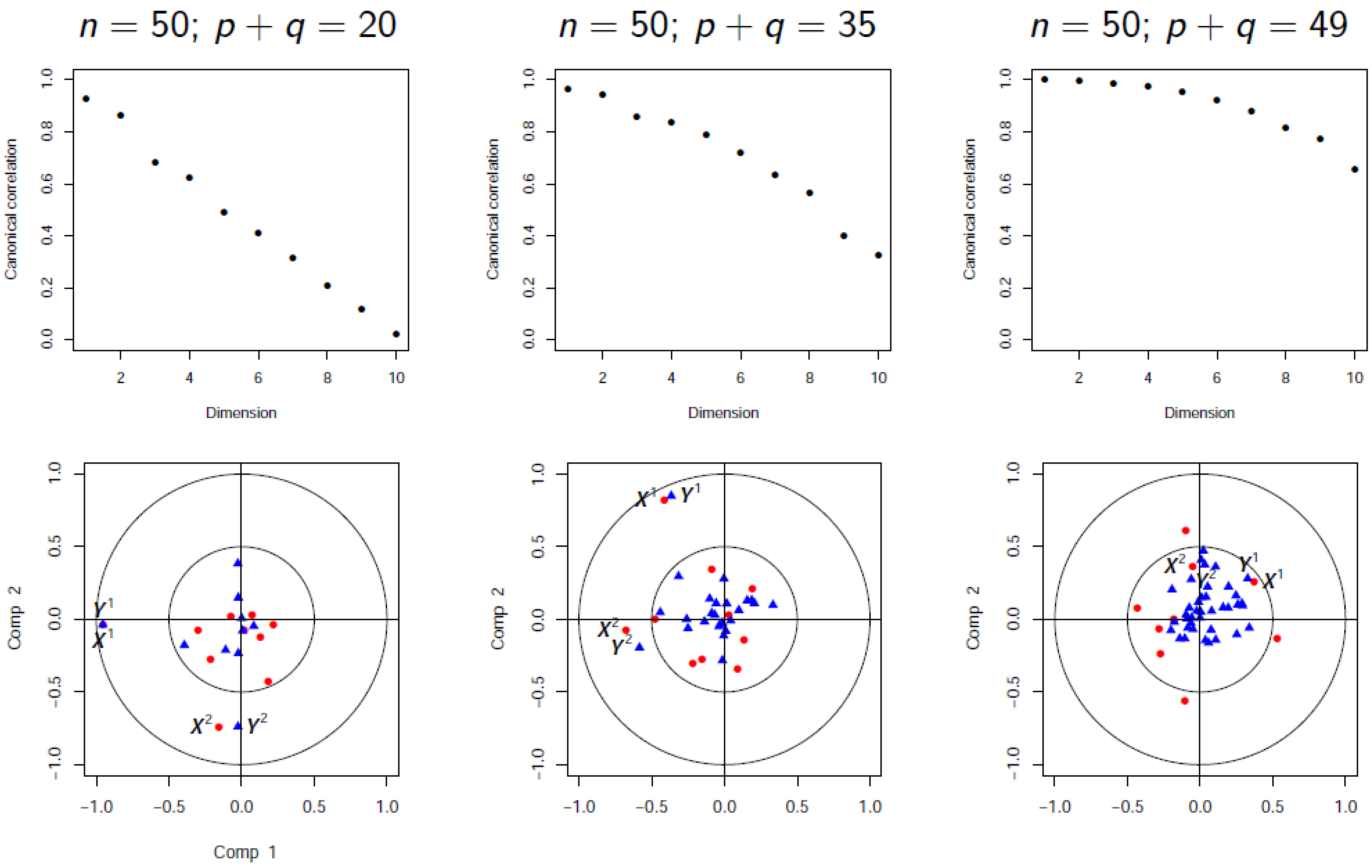
- variables X^1 and Y^1 are strongly correlated (0.9)
- variables X^2 and Y^2 are less strongly correlated (0.7)
- Canonical correlations for X et Y are approximately

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

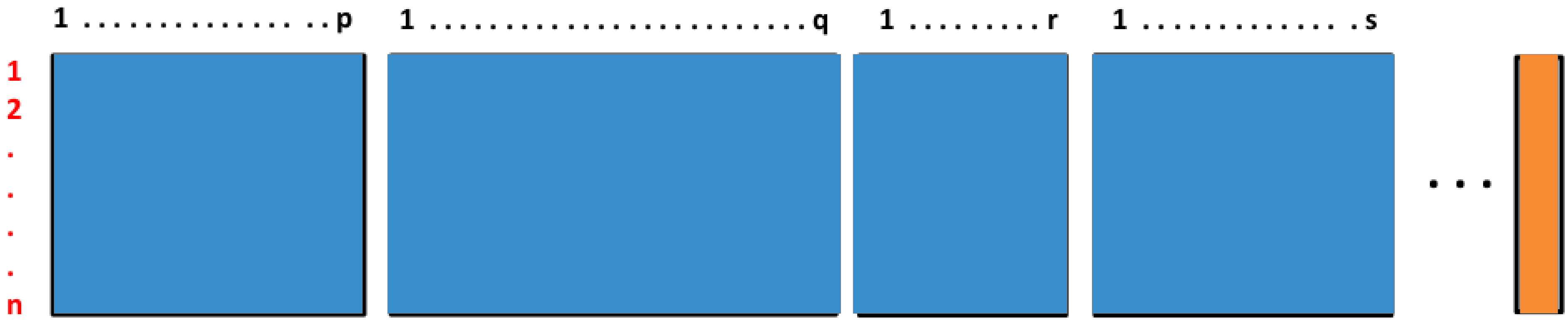
- simulations are run for

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

CCA: simulated example



Generalisation


1 . . . . p
1 . . . . q
1 . . . . r
1 . . . . s
1
2
.
.
.
.
n
. . .

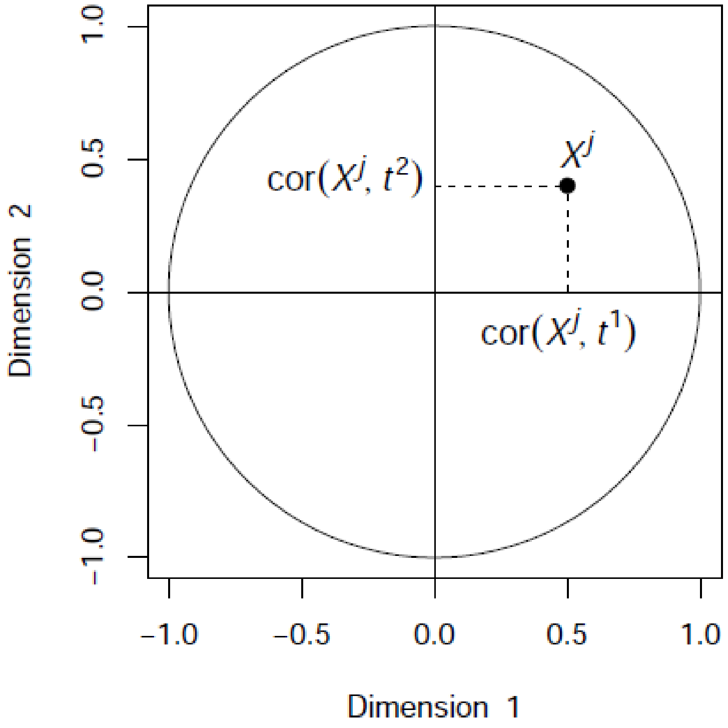
- **Generalized** CCA (GCCA): integration of more than 2 data sets ; maximizes the sum of every pairwise covariances between two components.
- Sparse (see *extensions*) GCCA (SGCCA): variable selection is performed on each data set

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

Graphical outputs

The same principles as those for PCA are still true for other multivariate methods mentioned here:

- Individuals plots: the coordinates of the individuals are given by the components calculated with the method (PCA, PLS-DA, PLS, CCA...)
- Variables plot: the variables are usually represented using their correlation with the components defining the axes. In other words, the coordinate of one variable X^j on the component t^i is given by $cor(X^j,t^i)$



Alternative graphical outputs

Motivations: usual plots are difficult to read and interpret when

- The numer of variables is too high
- The number of relevant components is greater than 2 inducing a « more than 2D » representation space.

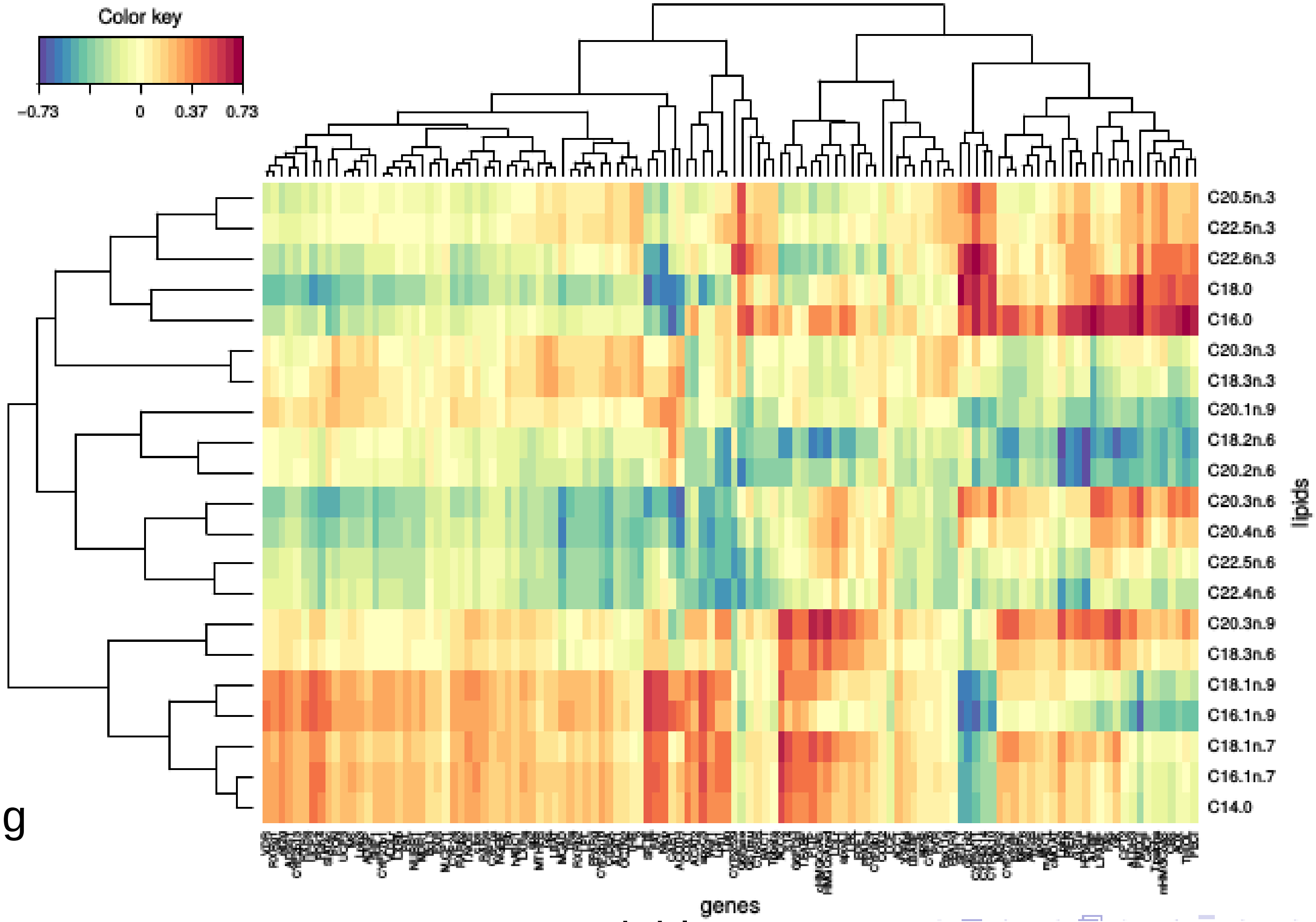
Propositions:

- Identify the pairs of highly related variables
- Produce graphical display making easy the interpretation

I. González, K-A. Lê Cao, M. Davis, S. Déjean (2012) – *Visualising associations between paired 'omics' data sets*. BioData Mining

Alternative graphical outputs

- Clustered Image Maps (CIM), Weinstein et al. (1997)
- Heatmaps, Eisent et al. (1998)



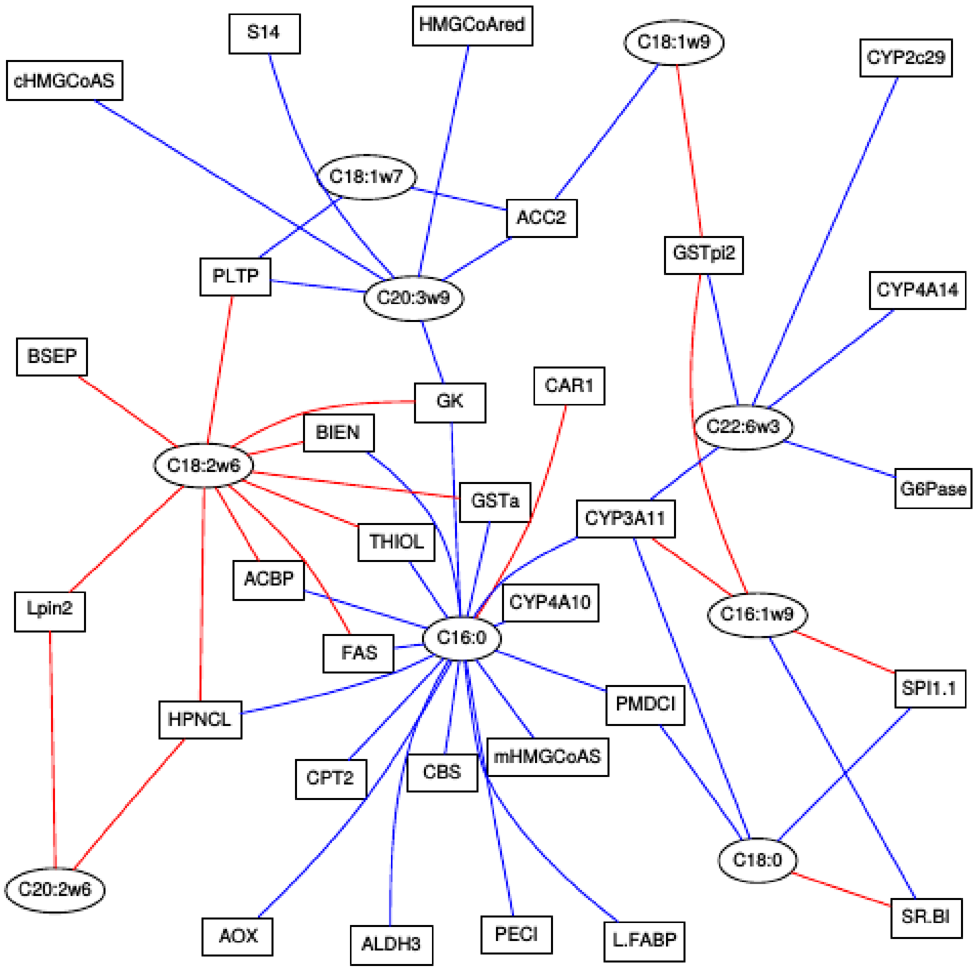
Variables set 2

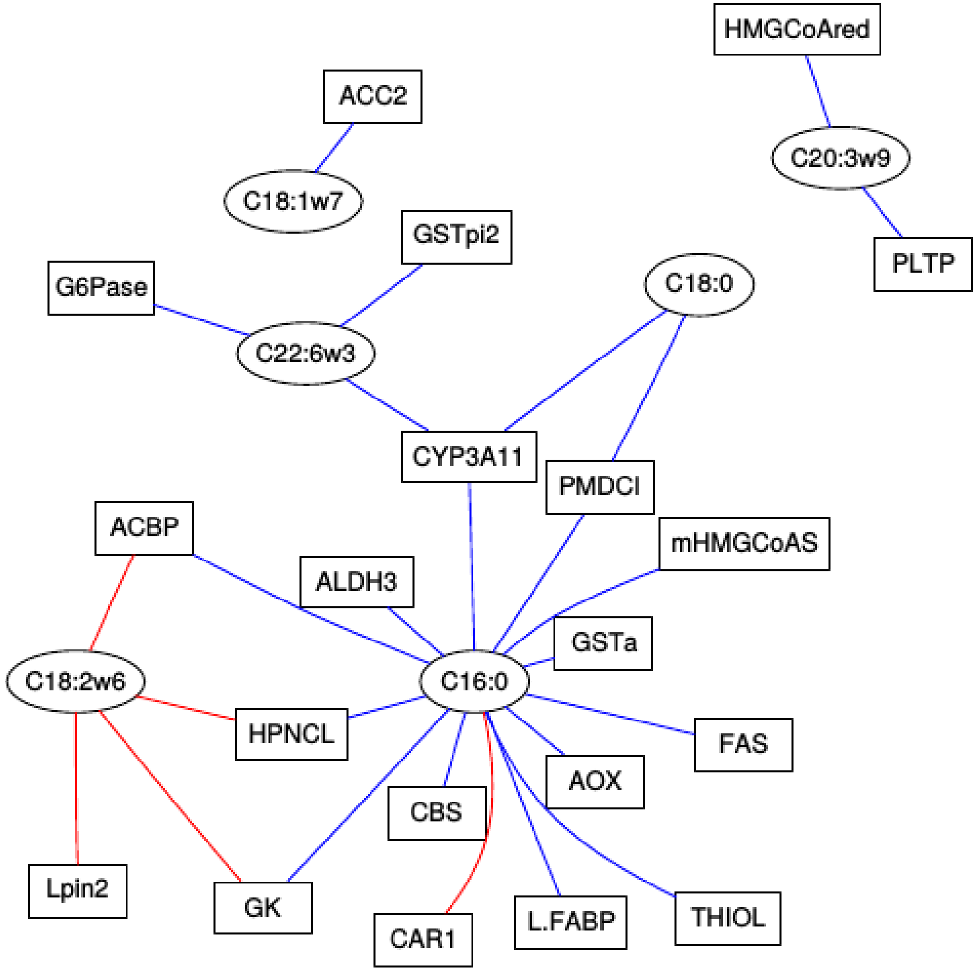
Differ from usual heatmaps crossing individuals and variables

Variables set 1

Alternative graphical outputs

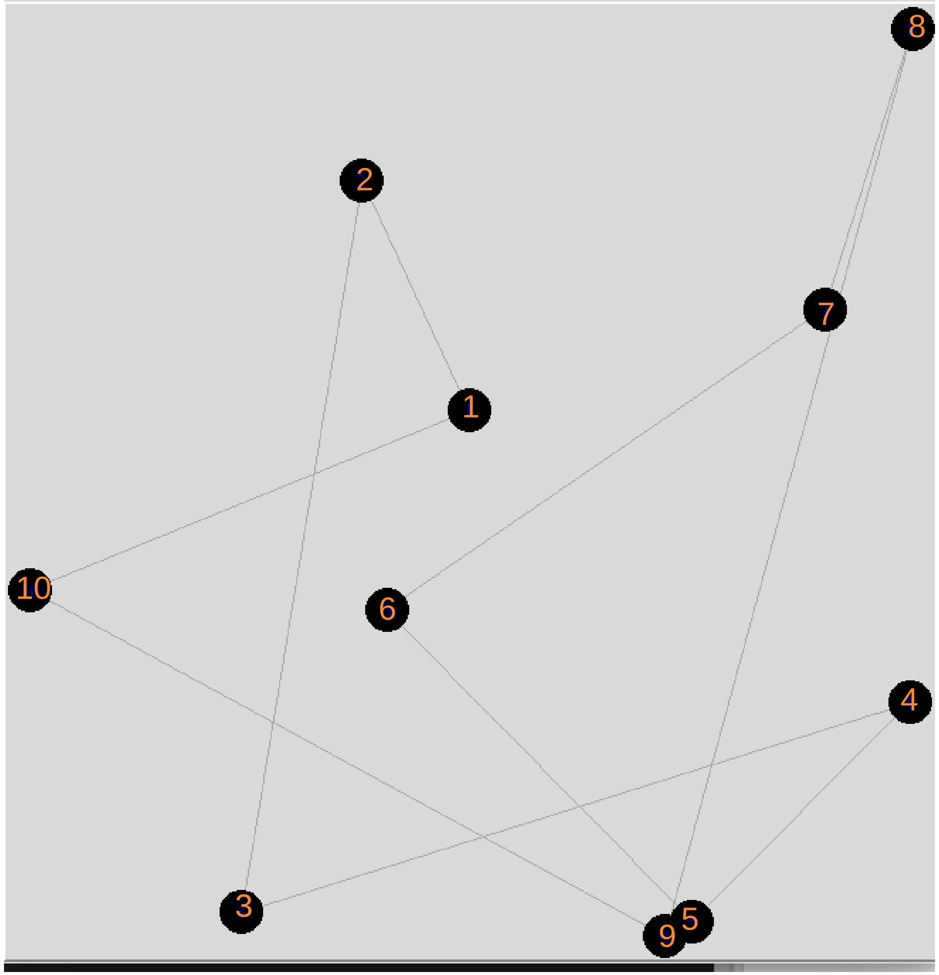
- Covariance Graph, Cox et Wermuth (1993)
- Relevance Network, Butte et al. (2000)

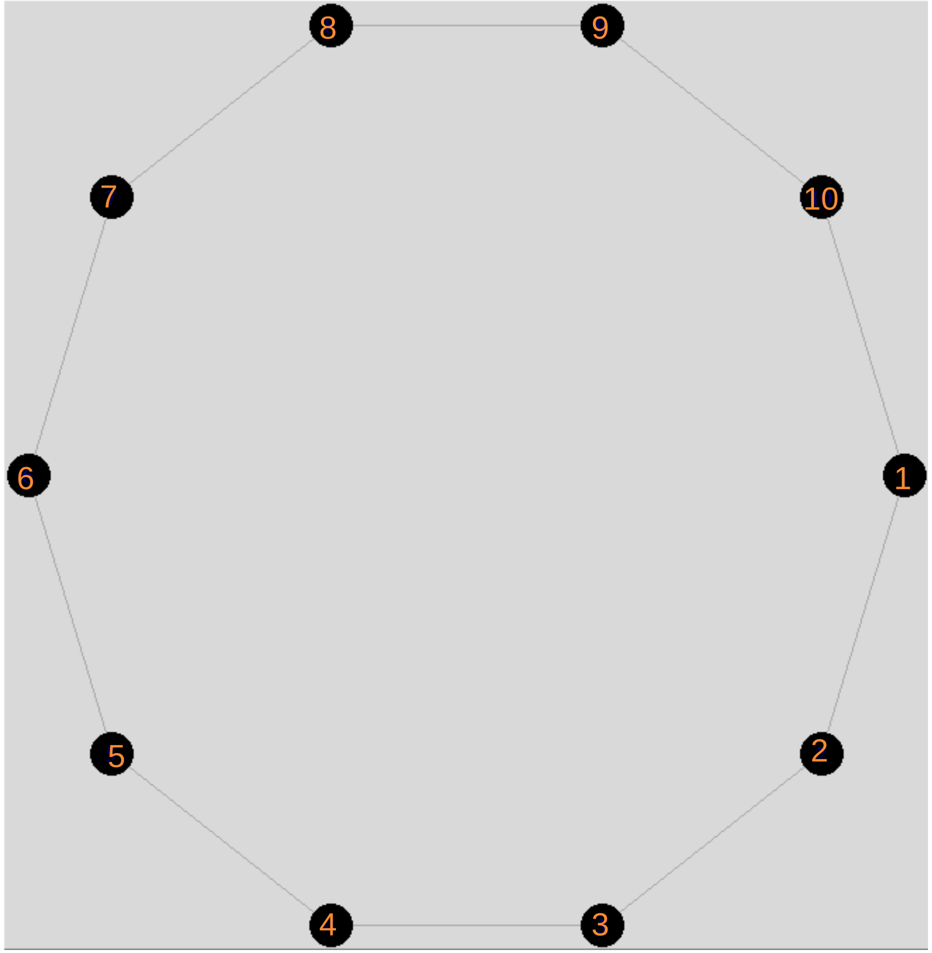




Alternative graphical outputs

- Be careful when interpreting network-based visualisation!
- The same network (same links between same edges) can be represented in very different ways.





Extensions *sparse*

Curse of dimensionality

https://en.wikipedia.org/wiki/Curse_of_dimensionality

The curse of dimensionality refers to various phenomena that arise when analyzing and organizing data in high-dimensional spaces (often with hundreds or thousands of dimensions) that do not occur in low-dimensional settings such as the three-dimensional physical space of everyday experience. The expression was coined by Richard E. Bellman when considering problems in dynamic optimization.

→ **Sparse** methods aim at dealing with problems related to the high dimension of the data.

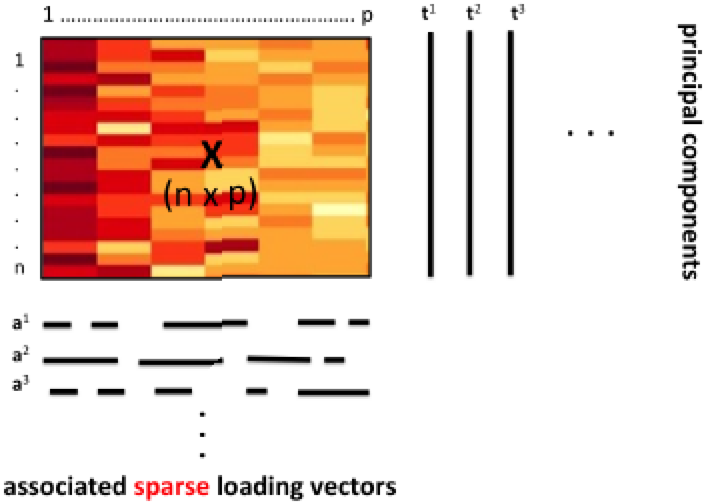
Occam's razor (law of parsimony): this principle states that among competing hypotheses, the one with the fewest assumptions should be selected.

https://en.wikipedia.org/wiki/Occam's_razor

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):

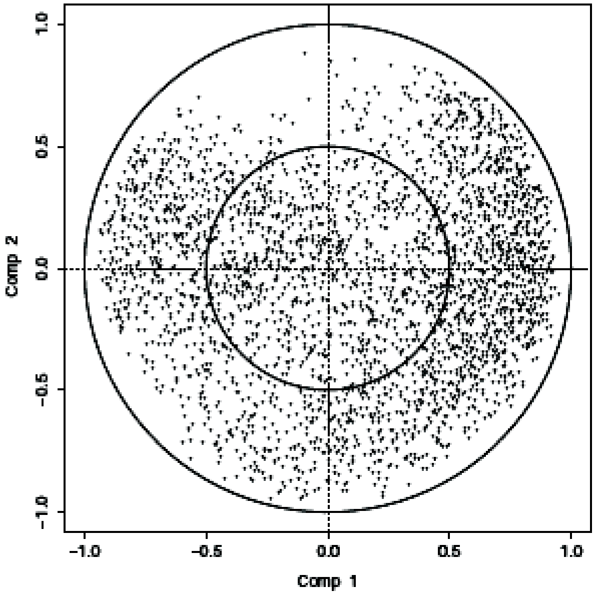
$$t = 0.x1 + a_2.x_2 + \ldots + 0.x_p$$



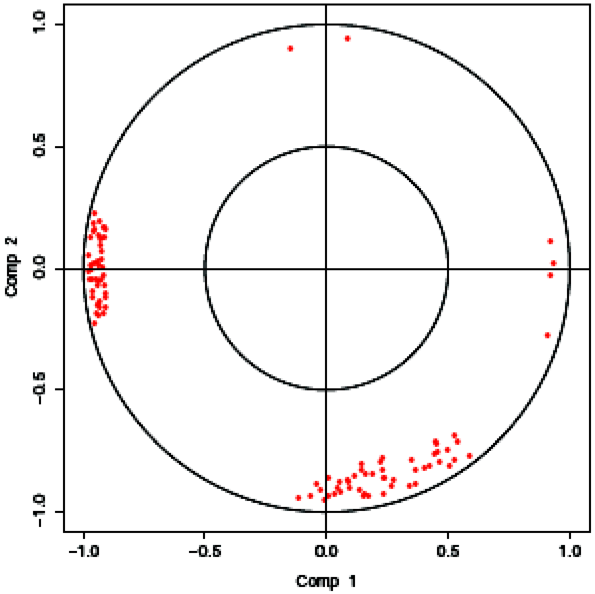
- Important weights : important contribution to define the Pcs.
- Null weights : those variables are not taken into account when calculating the PCs

Graphical outputs

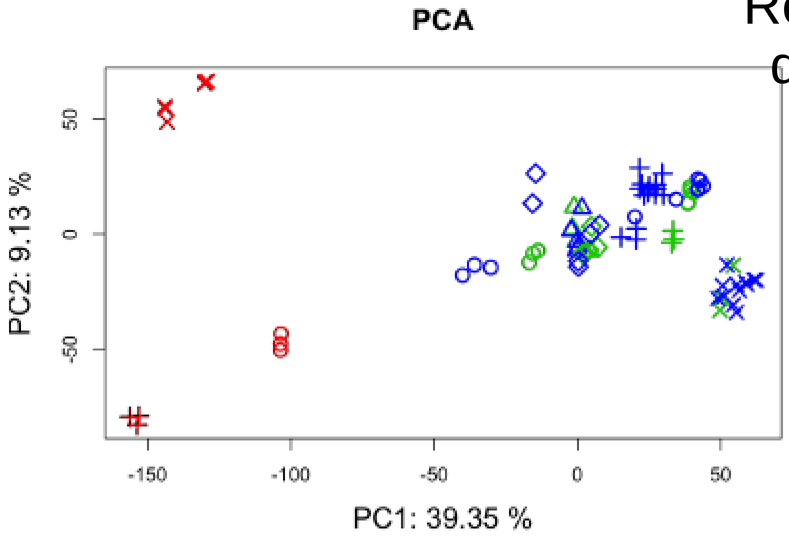
PCA



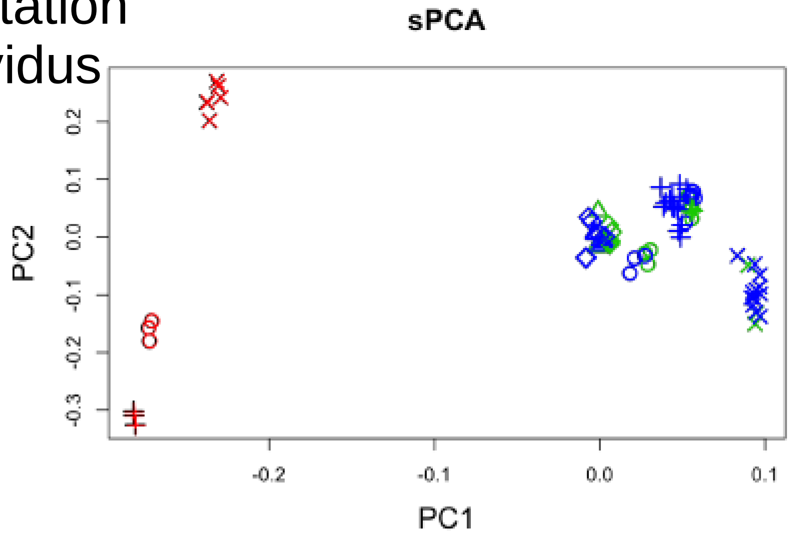
Sparse PCA



Représentation des variables



Représentation des individus

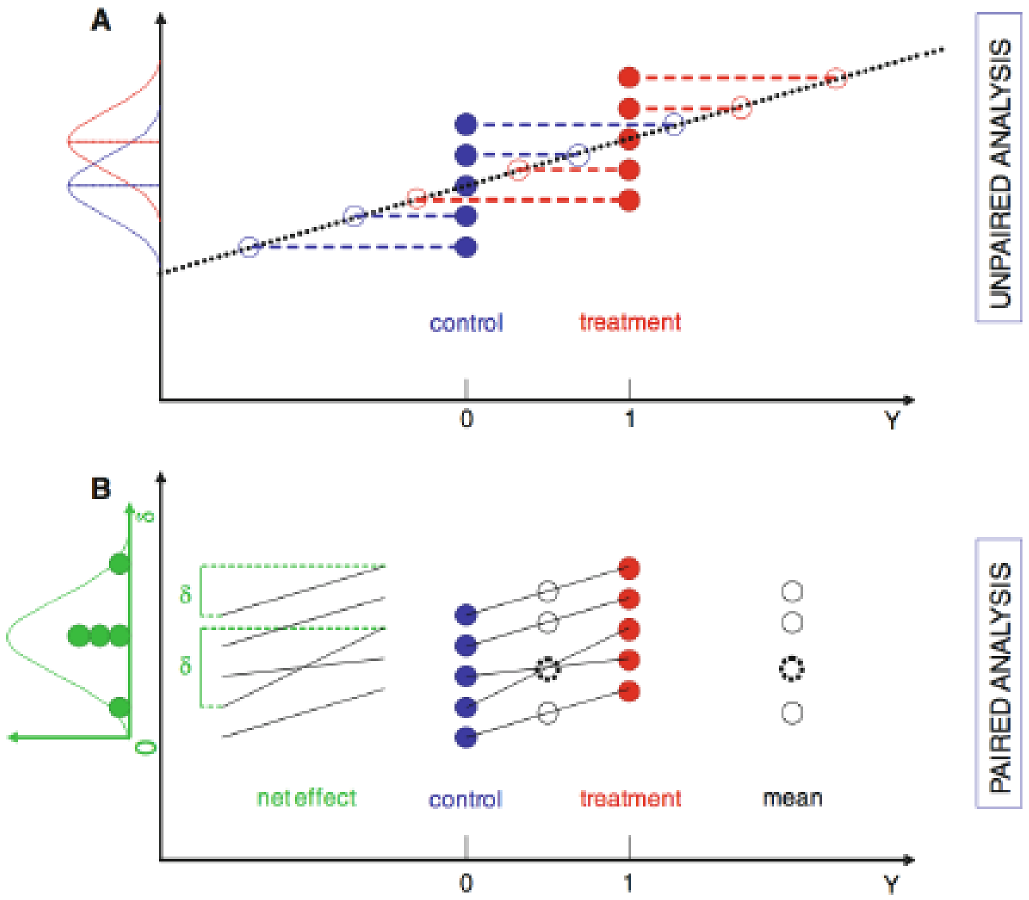


Extensions *multilevel*

Principle

- 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)

Paired data



- No paired structure (A): no significant difference between ctrl and trt

- Paired analysis (B): the data is decomposed into a mean (black circles) and a difference (d) per subject. The differences (net treatment effect) are projected on the Y-axis per subject, and are all different from 0.

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

Data decomposition

Decomposition of the data into within and between variations

$$X = \underset{\text{offset term}}{X_m} + \underset{\text{between-sample}}{X_b} + \underset{\text{within-sample}}{X_w}$$

- 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 of the design of experiments

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.

Multilevel: simulated example

3 variables (A, B, C) measured for 10 sujets (1...10) in 2 conditions *control* ou *treatment*.

Raw data set

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

Between-subject matrix

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
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

Within-subject matrix

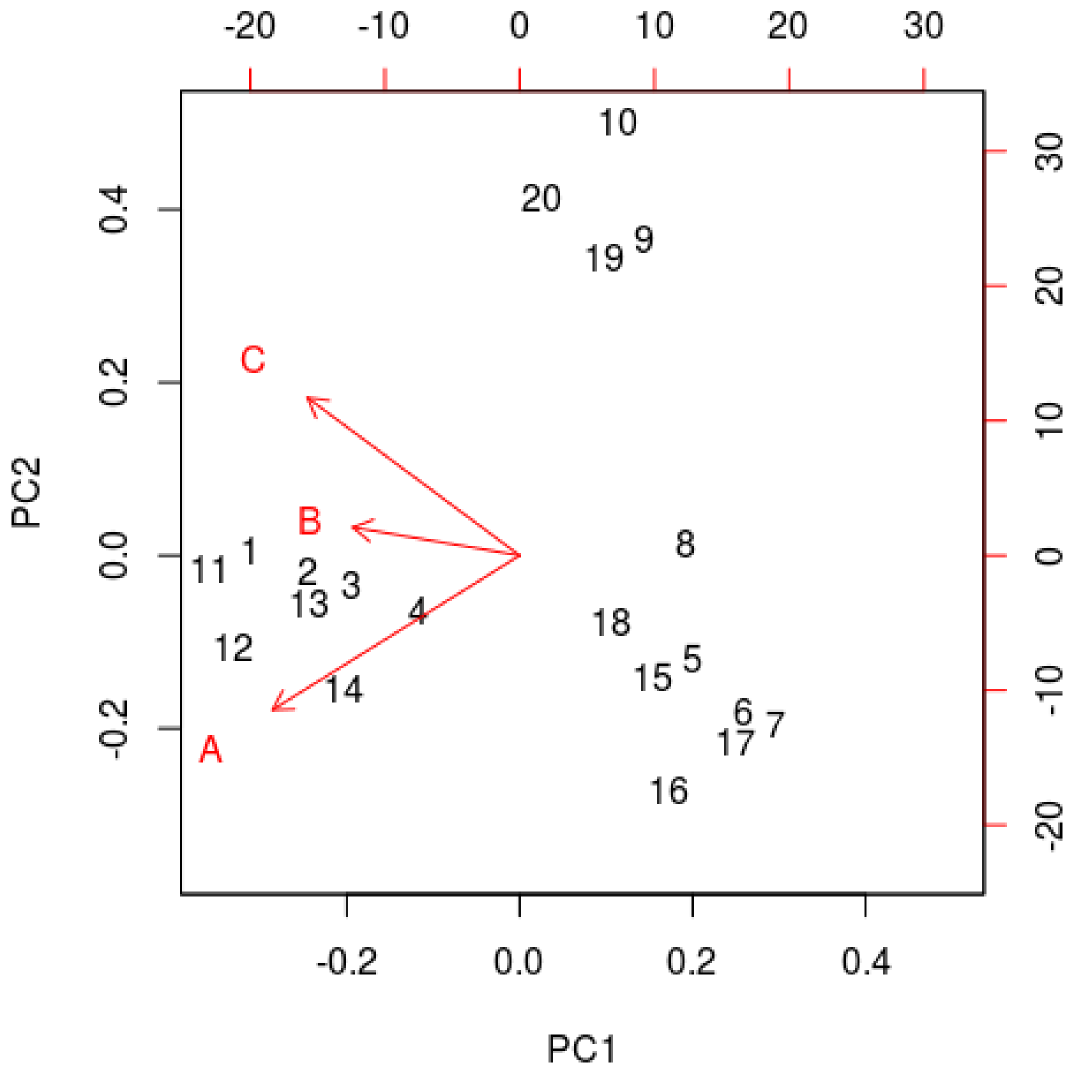
DA	DB	DC
-1	-2	0
-3	-2	0
-1	-2	0
-3	-2	0
-1	-2	0
-3	-2	0
-1	-2	0
-3	-2	0
-1	-2	0
-3	-2	0
1	2	0
3	2	0
1	2	0
3	2	0
1	2	0
3	2	0
1	2	0
3	2	0
1	2	0
3	2	0

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

Multilevel: simulated example

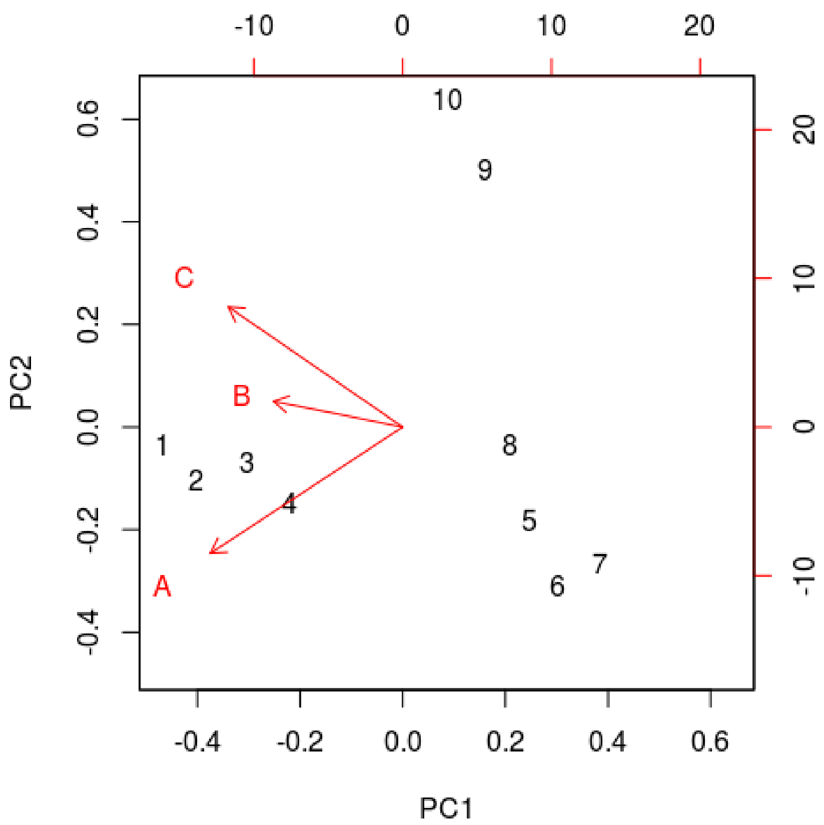
PCA on raw data

- The main information relies on the close locations of the two measurements made on each subject (1-11, 2-12, …, 9-19, 10-20)
- No treatment effect can be observed

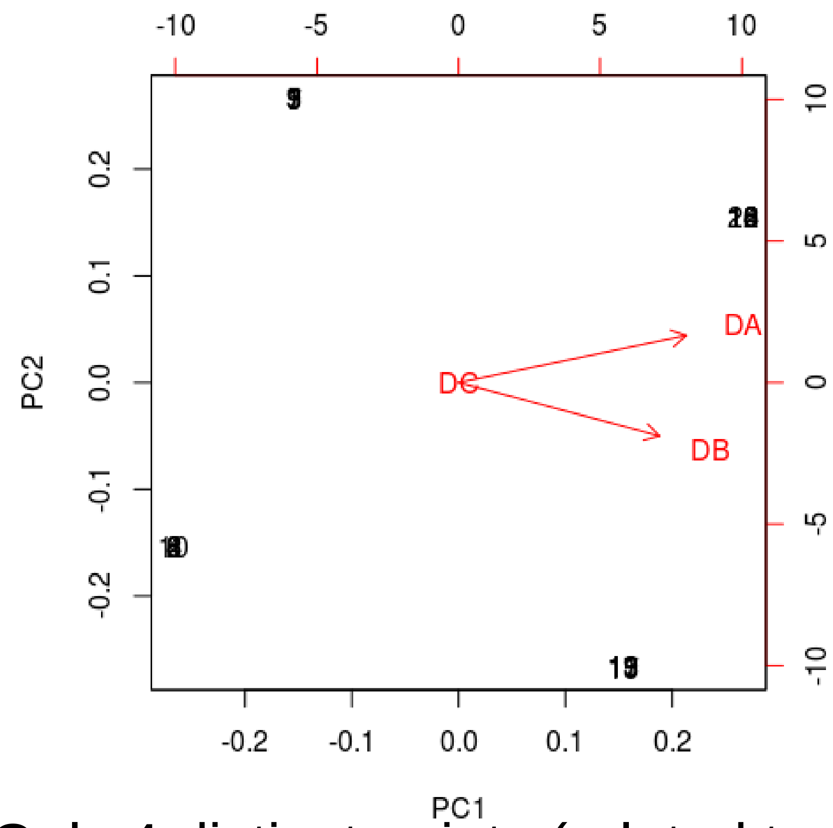


Multilevel: simulated example

PCA on between matrix



PCA on within matrix



- Nearly the same information as obtained on the raw data
- Because variability between subjects is greater than the variability due to the treatment

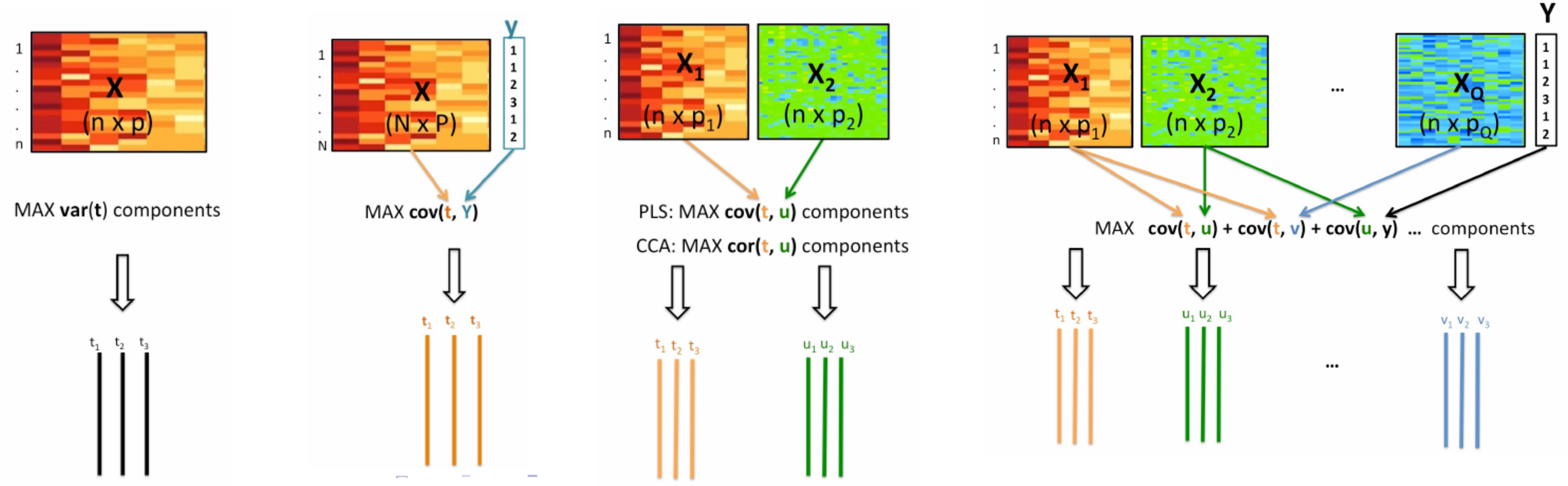
- Only 4 distinct points (related to the 4 unique rows in the within matrix)
- Treatment effect clearly appears

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 ?



Questions, *feedback*

mixOmics

mixOmics is collaborative project developed by the mixOmics team and several key collaborators. The project started at the Institut de Mathématiques de Toulouse, Université Paul Sabatier, Toulouse, France and was then further extended at the University of Queensland, Brisbane, Australia.

Why multivariate methods?

It is now generally admitted that the single `omics analysis does not provide enough information to give a deep understanding and insight of a biological system. However, we can get a more precise picture of a system by combining multiple omics analyses. In the mixOmics R package we propose a whole range of multivariate methods that we developed and validated on many biological studies.

mixOmics offers a wide range of multivariate methods for
the ***exploration*** *and* ***integration*** *of biological datasets with a particular focus on*

Recent Posts

- mixOmics 6_0_0
- Workshop 11-13 July 2016, Clermont-Ferrand
- Workshop 12-14 Sept 2016, Toulouse, travel grants available
- Workshop Ecole Nationale Vétérinaire, Nantes, 24-25 March 2016
- mixOmics for 2016

Recent Comments

Archives

- June 2016
- May 2016
- March 2016
- November 2015
- September 2015
- August 2015
- July 2015
- May 2015

Web site with tutorial :

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 would not exist without...

mixOmics development

Kim-Anh Lê Cao, Univ. Melbourne
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!