Principal Component Analysis
Analyse en Composantes Principales

Sébastien Déjean

math.univ-toulouse.fr/~sdejean

The Batman principle

It's not who I am underneath,
but what I do that defines me.

*From **Batman Begins***
www.youtube.com/watch?v=PmwLPU5H6_Q

What does PCA do?

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

Prerequisites

- Variance and standard deviation
- Linear combination
- Data pre-processing

Variance and Standard Deviation

$$var(X) = \frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})^2$$

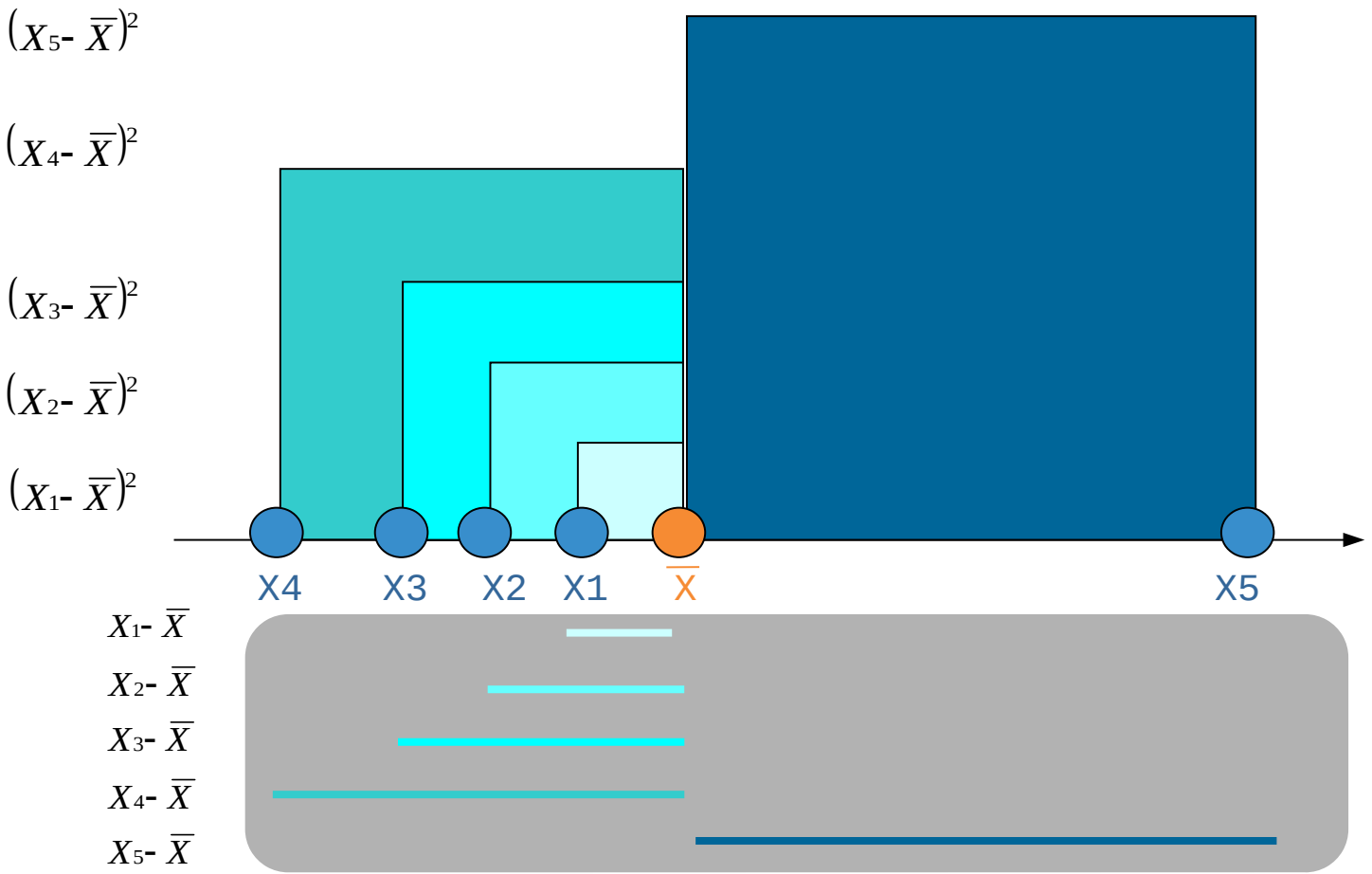
Mean of the squared deviations to the mean

$$sd(X) = \sqrt{var(X)}$$

Square root of the variance

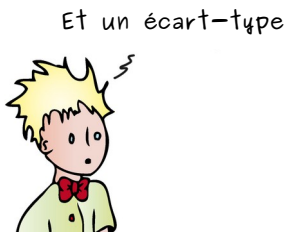
Variance and Standard Deviation

Square root of the mean of the squared deviations to the mean



Variance

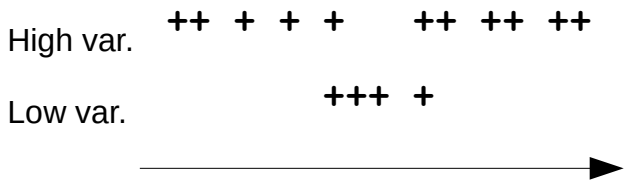
Standard deviation



Variance and Standard Deviation

To put it in a nutshell

- Variance and standard deviation are **spread** indicators
- The more scattered the valules, the higher the variance (and the standard deviation)
- Positive
- Unchanged by translation
- If the data are expressed in **m** then the standard deviation also express in **m** (as the mean) and the variance in **m²**!



Linear Combination

2 variables

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

2 coefficients

c1 = 0.5
c2 = 2

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

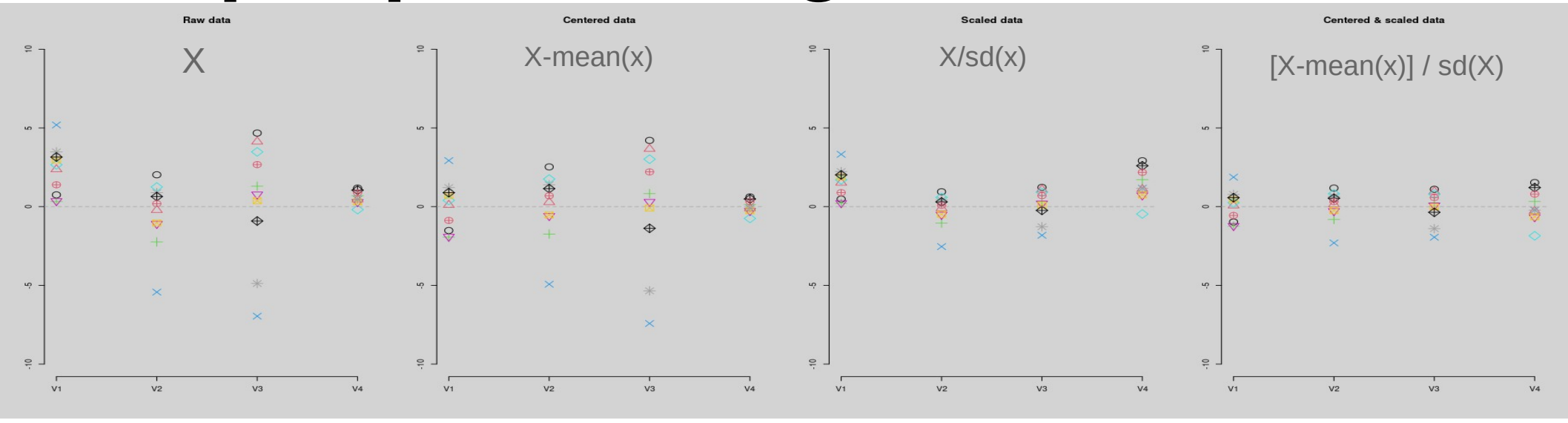
LC = 0.5	174.0	+ 2	65.6	=	218.20
	175.3		71.8		231.25
	193.5		80.7		258.15
	186.5		72.6		238.45
	187.2		78.8		251.20
	181.5		74.8		240.35
	184.0		86.4		264.80
	184.5		78.4		249.05
	175.0		62.0		211.50
	184.0		81.6		255.20

Data pre-processing: 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).
- Sometimes called « z-transformation » ou « z-score »

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

Data pre-processing: center/scale



	x1	x2	x3	x4
○	0.7	2.0	4.7	1.2
△	2.4	-0.2	4.1	0.4
+	0.3	-2.2	1.3	0.7
×	5.2	-5.4	-7.0	0.5
◇	2.7	1.2	3.5	-0.2
▽	0.4	-1.1	0.8	0.3
⊠	3.0	-1.0	0.4	0.3
✳	3.5	0.9	-4.9	0.5
⊕	3.2	0.6	-0.9	1.1
⊕	1.4	0.2	2.7	0.9
Mean	2.3	-0.5	0.5	0.6
S.D.	1.6	2.1	3.8	0.4

	x1	x2	x3	x4
○	-1.5	2.5	4.2	0.6
△	0.1	0.3	3.7	-0.1
+	-1.9	-1.7	0.8	0.1
×	2.9	-4.9	-7.4	-0.1
◇	0.4	1.8	3.0	-0.7
▽	-1.9	-0.6	0.3	-0.3
⊠	0.7	-0.5	-0.1	-0.3
✳	1.2	1.4	-5.3	-0.1
⊕	0.9	1.1	-1.4	0.5
⊕	-0.9	0.7	2.2	0.3
Mean	0	0	0	0
S.D.	1.6	2.1	3.8	0.4

	x1	x2	x3	x4
○	0.5	0.9	1.2	2.9
△	1.5	-0.1	1.1	1.1
+	0.2	-1.0	0.3	1.7
×	3.3	-2.5	-1.8	1.2
◇	1.7	0.6	0.9	-0.5
▽	0.2	-0.5	0.2	0.7
⊠	1.9	-0.5	0.1	0.7
✳	2.2	0.4	-1.3	1.2
⊕	2.0	0.3	-0.2	2.6
⊕	0.9	0.1	0.7	2.2
Mean	1.5	-0.2	0.1	1.4
S.D.	1	1	1	1

	x1	x2	x3	x4
○	-1.0	1.2	1.1	1.5
△	0.1	0.1	1.0	-0.3
+	-1.2	-0.8	0.2	0.3
×	1.9	-2.3	-1.9	-0.2
◇	0.3	0.8	0.8	-1.9
▽	-1.2	-0.3	0.1	-0.7
⊠	0.5	-0.2	0.0	-0.6
✳	0.8	0.6	-1.4	-0.2
⊕	0.6	0.5	-0.4	1.2
⊕	-0.6	0.3	0.6	0.8
Mean	0	0	0	0
S.D.	1	1	1	1

Data pre-processing: log transform

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
$4 < 5 < 8$	$2 < \sim 2.3 < 3$
$2 < 3 < 4$	$1 < \sim 1.6 < 2$
$0.1 < 0.125$	$\sim -3.3 < -3$

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

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

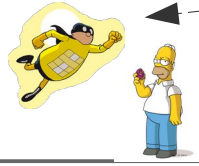
$$Y = \ln(X) \leftrightarrow X = e^Y = \exp(Y)$$

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

Teasing

Would you use a cubic box to pack a fishing rod?

To PCA

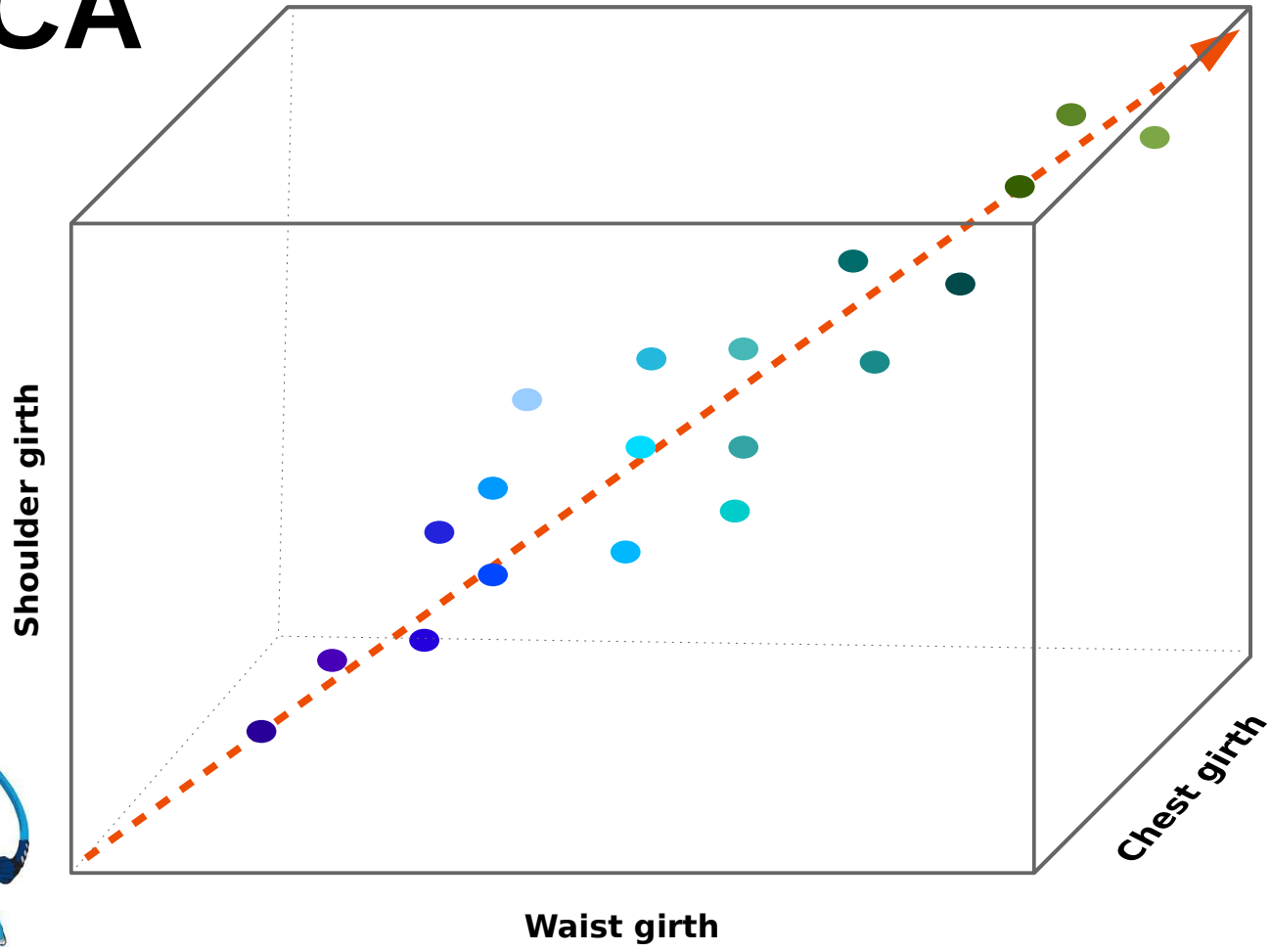


Shoulder girth

Waist girth

Chest girth

To PCA



Do we really need 3 dimensions to represent ‘standard’ individuals?

1st Principal Component: «beefyness»

PCA: (verbose) comments

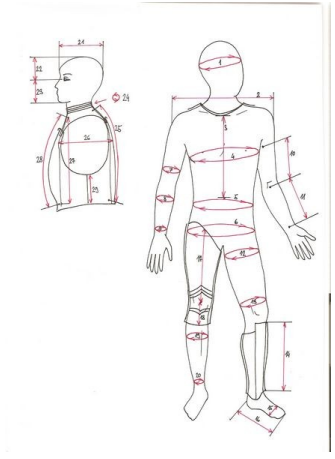
- The measurements are rather **strongly correlated.** Indeed, one can assume that a person with a high shoulder girth will also have high chest girth (even if exceptions exist...). In these conditions, the information brought by the 3 variables are **redundant**. Graphically, in the cube determined by shoulder girth, chest girth and waist girth, there are nearly empty 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.

- PCA allows to 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 retain 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.

A toy example

- 20 individuals
- 5 variables

```
s.g : shoulder girth (cm)
c.g : chest girth (cm)
w.g : waist girth (cm)
w   : weight (kg)
h   : height (cm)
```



Id	s.g	c.g	w.g	w	h
I1	106.2	89.5	71.5	65.6	174.0
I2	110.5	97.0	79.0	71.8	175.3
I3	115.1	97.5	83.2	80.7	193.5
I4	104.5	97.0	77.8	72.6	186.5
I5	107.5	97.5	80.0	78.8	187.2
I6	119.8	99.9	82.5	74.8	181.5
I7	123.5	106.9	82.0	86.4	184.0
I8	120.4	102.5	76.8	78.4	184.5
I9	111.0	91.0	68.5	62.0	175.0
I10	119.5	93.5	77.5	81.6	184.0
I11	105.0	89.0	71.2	67.3	169.5
I12	100.2	94.1	79.6	75.5	160.0
I13	99.1	90.8	77.9	68.2	172.7
I14	107.6	97.0	69.6	61.4	162.6
I15	104.0	95.4	86.0	76.8	157.5
I16	108.4	91.8	69.9	71.8	176.5
I17	99.3	87.3	63.5	55.5	164.4
I18	91.9	78.1	57.9	48.6	160.7
I19	107.1	90.9	72.2	66.4	174.0
I20	100.5	97.1	80.4	67.3	163.8

First computations

Raw data

Id	s.g	c.g	w.g	w	h
I1	106.2	89.5	71.5	65.6	174.0
I2	110.5	97.0	79.0	71.8	175.3
I3	115.1	97.5	83.2	80.7	193.5
I4	104.5	97.0	77.8	72.6	186.5
I5	107.5	97.5	80.0	78.8	187.2
I6	119.8	99.9	82.5	74.8	181.5
I7	123.5	106.9	82.0	86.4	184.0
I8	120.4	102.5	76.8	78.4	184.5
I9	111.0	91.0	68.5	62.0	175.0
I10	119.5	93.5	77.5	81.6	184.0
I11	105.0	89.0	71.2	67.3	169.5
I12	100.2	94.1	79.6	75.5	160.0
I13	99.1	90.8	77.9	68.2	172.7
I14	107.6	97.0	69.6	61.4	162.6
I15	104.0	95.4	86.0	76.8	157.5
I16	108.4	91.8	69.9	71.8	176.5
I17	99.3	87.3	63.5	55.5	164.4
I18	91.9	78.1	57.9	48.6	160.7
I19	107.1	90.9	72.2	66.4	174.0
I20	100.5	97.1	80.4	67.3	163.8

Univariate analysis

Mean	108.1	94.2	75.3	70.6	174.4
Variance	68.6	37.5	50.8	85.7	109.3

Bivariate analysis

Covariance matrix

	s.g	c.g	w.g	w	h
s.g	**68.6**	37.7	28.1	55.3	61.2
c.g	37.7	**37.5**	33.9	45.7	32.4
w.g	28.1	33.9	**50.8**	56.6	27.7
w	55.3	45.7	56.6	**85.7**	59.5
h	61.2	32.4	27.7	59.5	**109.3**

Pearson correlation matrix

	s.g	c.g	w.g	w	h
s.g	1.0	0.7	0.5	0.7	0.7
c.g	0.7	1.0	0.8	0.8	0.5
w.g	0.5	0.8	1.0	0.9	0.4
w	0.7	0.8	0.9	1.0	0.6
h	0.7	0.5	0.4	0.6	1.0

351.9 represents the quantity of information contained in the data.

68.6 + 37.5 + 50.8 + 85.7 + 109.3 = **351.9**

The core of PCA

Coefficients of linear combination (or loadings)

	PC1	PC2	PC3	PC4	PC5
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
...
```

What is underneath? ~~Bruce Wayne~~ Eigen decomposition of the covariance matrix.

Around the core

Centered data

Ex: -6.50 = **0.45***(-1.9) + **0.32***(-4.7) + **0.34***(-3.8) + **0.54***(-5) + **0.54***(-0.4)

Id	s.g	c.g	w.g	w	h
I1	-1.9	-4.7	-3.8	-5.0	-0.4
I2	2.4	2.8	3.7	1.2	0.9
I3	7.0	3.3	7.9	10.1	19.1
I4	-3.6	2.8	2.5	2.0	12.1
I5	-0.6	3.3	4.7	8.2	12.8
I6	11.7	5.7	7.2	4.2	7.1
I7	15.4	12.7	6.7	15.8	9.6
I8	12.3	8.3	1.5	7.8	10.1
I9	2.9	-3.2	-6.8	-8.6	0.6
I10	11.4	-0.7	2.2	11.0	9.6
I11	-3.1	-5.2	-4.1	-3.3	-4.9
I12	-7.9	-0.1	4.2	4.9	-14.4
I13	-9.0	-3.4	2.6	-2.4	-1.7
I14	-0.5	2.8	-5.8	-9.2	-11.8
I15	-4.1	1.2	10.7	6.2	-16.9
I16	0.3	-2.4	-5.4	1.2	2.1
I17	-8.8	-6.9	-11.8	-15.1	-10.0
I18	-16.2	-16.1	-17.4	-22.0	-13.7
I19	-1.0	-3.3	-3.1	-4.2	-0.4
I20	-7.6	2.9	5.1	-3.3	-10.6

	PC1	PC2	PC3	PC4	PC5
s.g	0.45	-0.16	0.78	-0.18	0.36
c.g	0.32	0.25	0.26	0.72	-0.49
w.g	0.34	0.53	-0.33	0.24	0.66
w	0.54	0.36	-0.17	-0.60	-0.44
h	0.54	-0.70	-0.43	0.17	0.02

Apply loadings

	PC1	PC2	PC3	PC4	PC5
I1	-6.50	-4.48	-0.37	-1.03	1.27
I2	4.40	2.04	0.81	1.87	1.38
I3	22.66	-5.94	-6.18	0.11	1.97
I4	7.78	-5.24	-8.38	4.10	-1.74
I5	13.73	-2.67	-8.02	0.82	-2.15
I6	15.67	-0.15	4.49	2.33	4.40
I7	26.99	3.19	6.29	0.04	-3.08
I8	18.41	-3.43	5.63	1.09	-1.96
I9	-6.25	-8.48	4.97	0.79	1.86
I10	16.78	-3.67	1.99	-7.08	1.22
I11	-8.83	-0.78	0.28	-3.02	0.07
I12	-7.28	15.41	-2.31	-3.00	-2.35
I13	-6.45	2.25	-7.60	0.95	1.15
I14	-12.51	2.68	8.91	4.27	-1.53
I15	-3.65	20.76	-0.30	-2.45	1.99
I16	-0.63	-4.62	0.34	-3.46	-2.80
I17	-23.61	-5.07	2.20	1.19	-1.15
I18	-37.50	-9.07	-1.33	-1.89	-0.02
I19	-4.98	-3.61	0.33	-0.50	1.02
I20	-8.24	10.89	-1.74	4.86	0.44

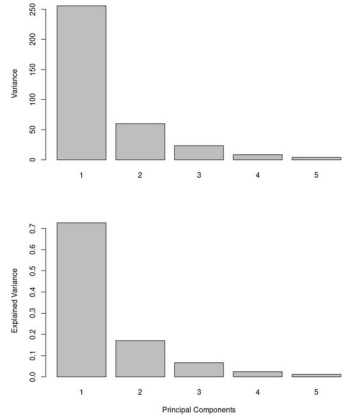
255.7 is the greatest variance we can obtain with a linear combination of the initial variables.

Mean	0	0	0	0	0	
Var.	**255.7**	**60.2**	**23.5**	**8.6**	**4.0**	= **351.9**

Graphical outputs (1/3)

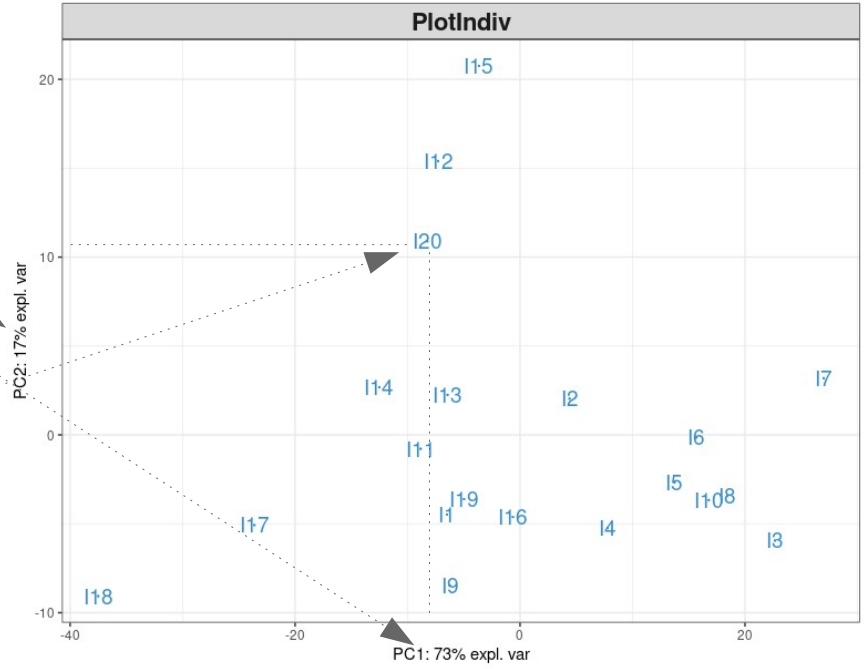
	PC1	PC2	PC3	PC4	PC5
Variance	255.7	60.2	23.5	8.6	4.0
% variance	72.6	17.1	6.7	2.4	1.1

Screeplot



	PC1	PC2	PC3	PC4	PC5
I1	-6.50	-4.48	-0.37	-1.03	1.27
I2	4.40	2.04	0.81	1.87	1.38
I3	22.66	-5.94	-6.18	0.11	1.97
I4	7.78	-5.24	-8.38	4.10	-1.74
I5	13.73	-2.67	-8.02	0.82	-2.15
I6	15.67	-0.15	4.49	2.33	4.40
I7	26.99	3.19	6.29	0.04	-3.08
I8	18.41	-3.43	5.63	1.09	-1.96
I9	-6.25	-8.48	4.97	0.79	1.86
I10	16.78	-3.67	1.99	-7.08	1.22
I11	-8.83	-0.78	0.28	-3.02	0.07
I12	-7.28	15.41	-2.31	-3.00	-2.35
I13	-6.45	2.25	-7.60	0.95	1.15
I14	-12.51	2.68	8.91	4.27	-1.53
I15	-3.65	20.76	-0.30	-2.45	1.99
I16	-0.63	-4.62	0.34	-3.46	-2.80
I17	-23.61	-5.07	2.20	1.19	-1.15
I18	-37.50	-9.07	-1.33	-1.89	-0.02
I19	-4.98	-3.61	0.33	-0.50	1.02
I20	-8.24	10.89	-1.74	4.86	0.44

Individual plot



Graphical outputs (2/3)

Id	s.g	c.g	w.g	w	h
I1	106.2	89.5	71.5	65.6	174.0
I2	110.5	97.0	79.0	71.8	175.3
I3	115.1	97.5	83.2	80.7	193.5
I4	104.5	97.0	77.8	72.6	186.5
I5	107.5	97.5	80.0	78.8	187.2
I6	119.8	99.9	82.5	74.8	181.5
I7	123.5	106.9	82.0	86.4	184.0
I8	120.4	102.5	76.8	78.4	184.5
I9	111.0	91.0	68.5	62.0	175.0
I10	119.5	93.5	77.5	81.6	184.0
I11	105.0	89.0	71.2	67.3	169.5
I12	100.2	94.1	79.6	75.5	160.0
I13	99.1	90.8	77.9	68.2	172.7
I14	107.6	97.0	69.6	61.4	162.6
I15	104.0	95.4	86.0	76.8	157.5
I16	108.4	91.8	69.9	71.8	176.5
I17	99.3	87.3	63.5	55.5	164.4
I18	91.9	78.1	57.9	48.6	160.7
I19	107.1	90.9	72.2	66.4	174.0
I20	100.5	97.1	80.4	67.3	163.8

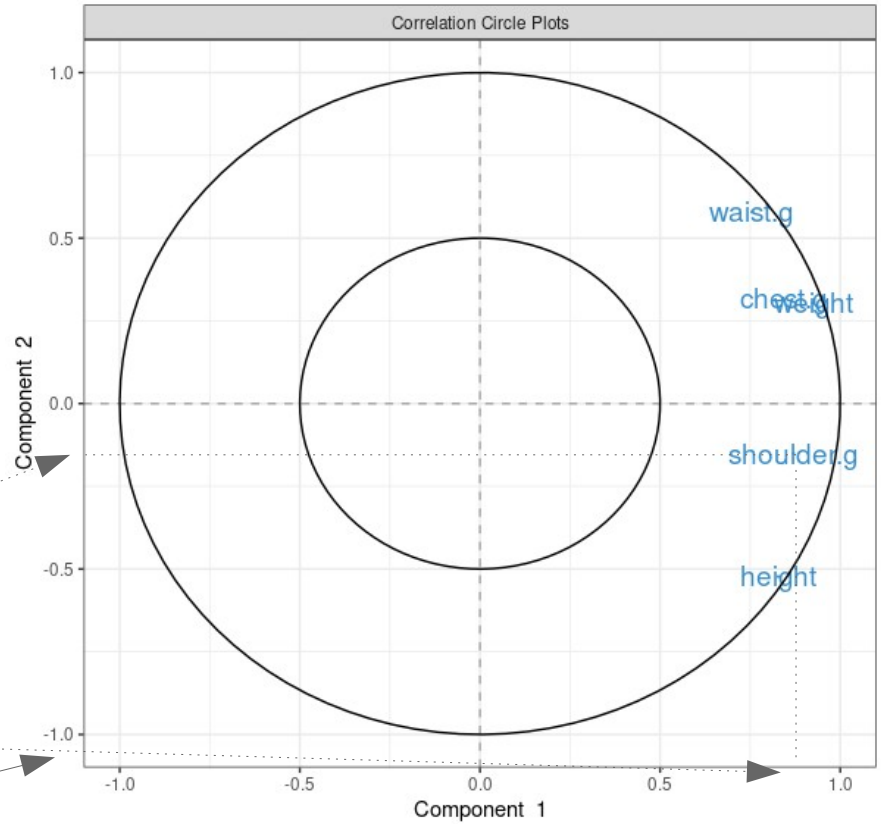
	PC1	PC2
I1	-6.50	-4.48
I2	4.40	2.04
I3	22.66	-5.94
I4	7.78	-5.24
I5	13.73	-2.67
I6	15.67	-0.15
I7	26.99	3.19
I8	18.41	-3.43
I9	-6.25	-8.48
I10	16.78	-3.67
I11	-8.83	-0.78
I12	-7.28	15.41
I13	-6.45	2.25
I14	-12.51	2.68
I15	-3.65	20.76
I16	-0.63	-4.62
I17	-23.61	-5.07
I18	-37.50	-9.07
I19	-4.98	-3.61
I20	-8.24	10.89

```
cor(s.g, PC1) =  0.87
cor(s.g, PC2) = -0,15

cor(c.g, PC1) =  0.84
cor(c.g, PC2) =  0.32
...
```

	PC1	PC2
shoulder.g	0.87	-0.15
chest.g	0.84	0.32
waist.g	0.75	0.58
weight	0.92	0.30
height	0.83	-0.52

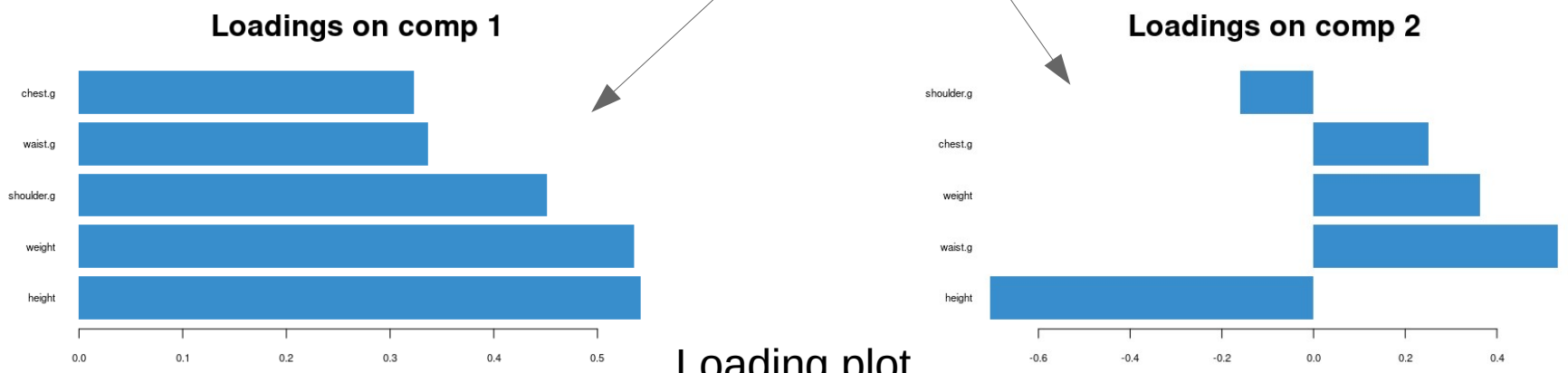
Variable plot



Graphical outputs (3/3)

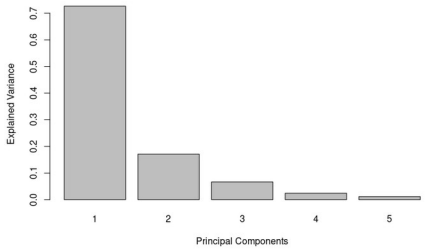
Loadings

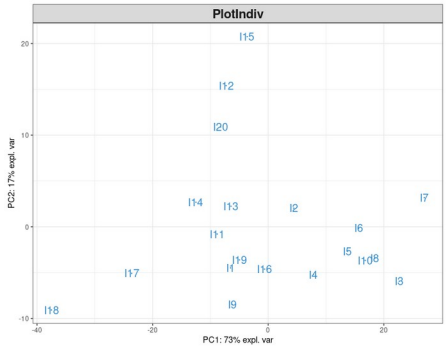
	PC1	PC2
shoulder.g	0.45	-0.16
chest.g	0.32	0.25
waist.g	0.34	0.53
weight	0.54	0.36
height	0.54	-0.70

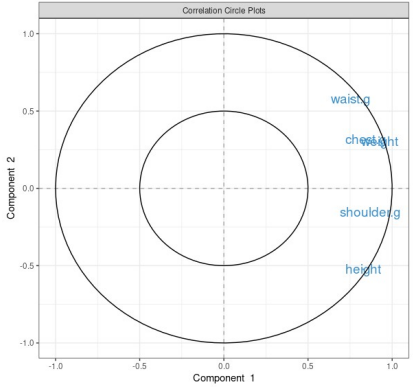


Loading plot

How to intepret plots?







- How many components?
- Here is the deal:
  - 5 PCs ↔ 100 %
  - 2 PCs ↔ 90 %
  - 3 PCs ↔ 97 %

- Clusters, outliers…
- Caution: visual proximities

- Correlation between variables
- Interpret components:
  - 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.

Focus on the individual plot?

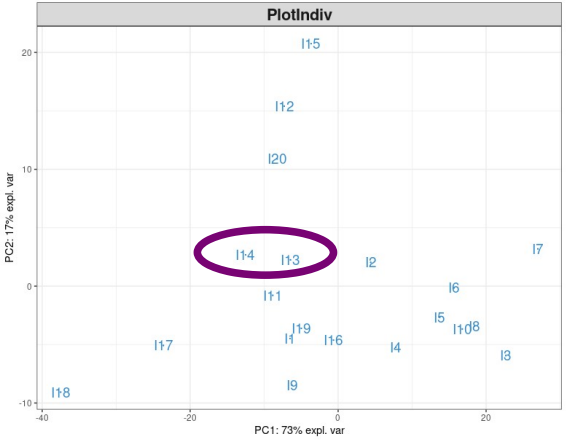
- To interpret the graphical results of PCA must be done keeping in mind that one is looking at a projection on a plane (2D space).
- Be careful when interpreting visual proximities
- Illustration in comics with the only true super-heros ...

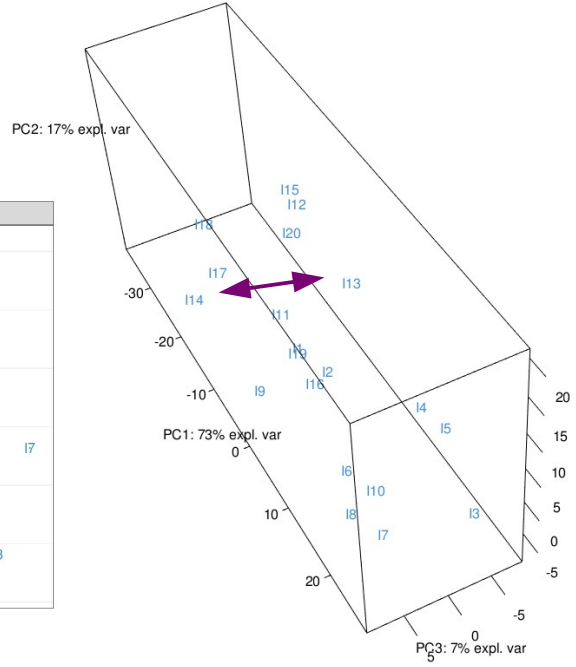
Scenario & illustration
Pascal Jousselin
pjousselin.free.fr
Colour
Laurence Croix

Focus on the individual plot?

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

I13	99.1	90.8	77.9	68.2	172.7
I14	107.6	97.0	69.6	61.4	162.6



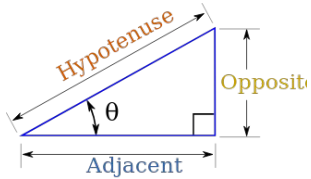


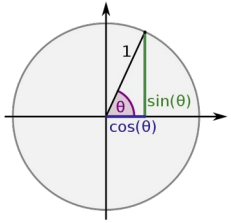
- I13 and I14 are close on PC1-PC2 but they do not have very close values
- They are separated on PC3!

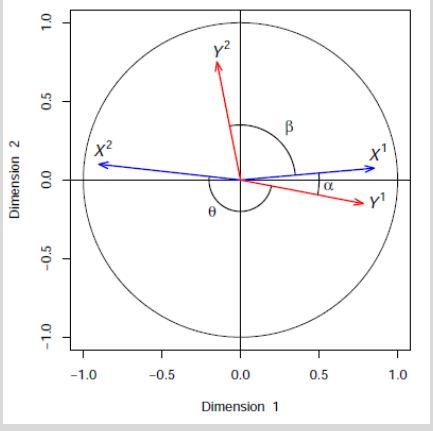
Focus on the variable plot?

Correlation ↔ cosine

Remember trigonometry and right triangles:





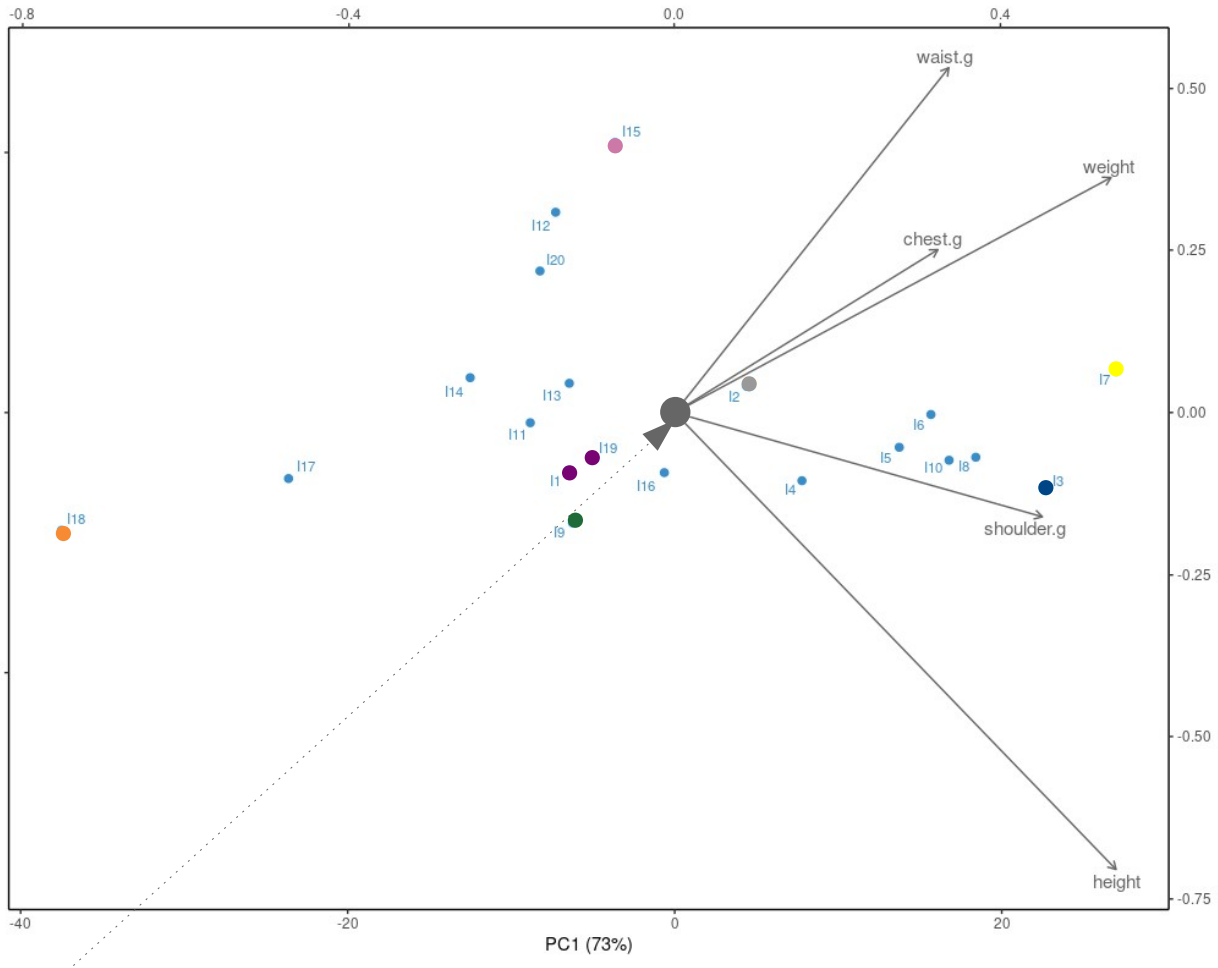


The correlation between two variables is represented as:

- An acute angle ($\cos(\alpha) > 0$) if it is positive
- An obtuse angle ($\cos(\theta) < 0$) if it is negative
- A right angle ($\cos(\beta) \approx 0$) if it is near zero

And so what?

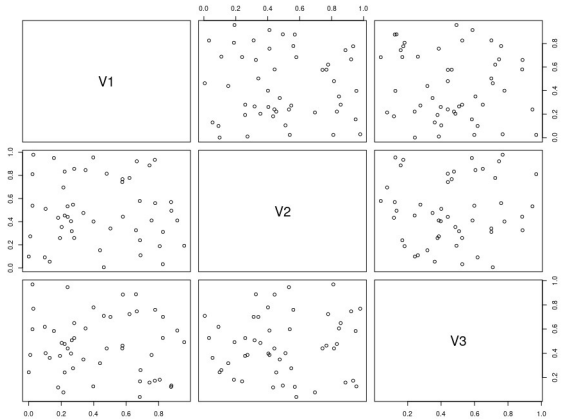
Id	s.g	c.g	w.g	w	h
I1	106.2	89.5	71.5	65.6	174.0
I2	110.5	97.0	79.0	71.8	175.3
I3	115.1	97.5	83.2	80.7	193.5
I4	104.5	97.0	77.8	72.6	186.5
I5	107.5	97.5	80.0	78.8	187.2
I6	119.8	99.9	82.5	74.8	181.5
I7	123.5	106.9	82.0	86.4	184.0
I8	120.4	102.5	76.8	78.4	184.5
I9	111.0	91.0	68.5	62.0	175.0
I10	119.5	93.5	77.5	81.6	184.0
I11	105.0	89.0	71.2	67.3	169.5
I12	100.2	94.1	79.6	75.5	160.0
I13	99.1	90.8	77.9	68.2	172.7
I14	107.6	97.0	69.6	61.4	162.6
I15	104.0	95.4	86.0	76.8	157.5
I16	108.4	91.8	69.9	71.8	176.5
I17	99.3	87.3	63.5	55.5	164.4
I18	91.9	78.1	57.9	48.6	160.7
I19	107.1	90.9	72.2	66.4	174.0
I20	100.5	97.1	80.4	67.3	163.8
Mean	108.1	94.2	75.3	70.6	174.4



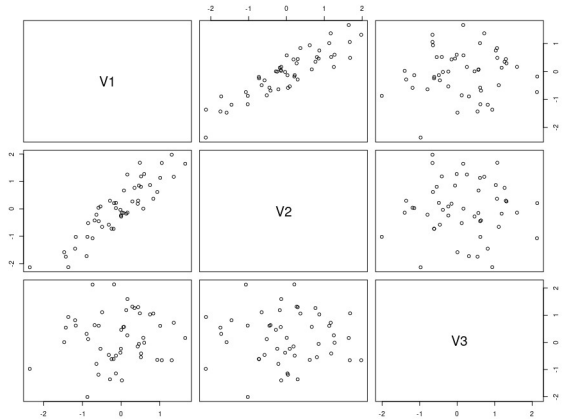
Simulated examples

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

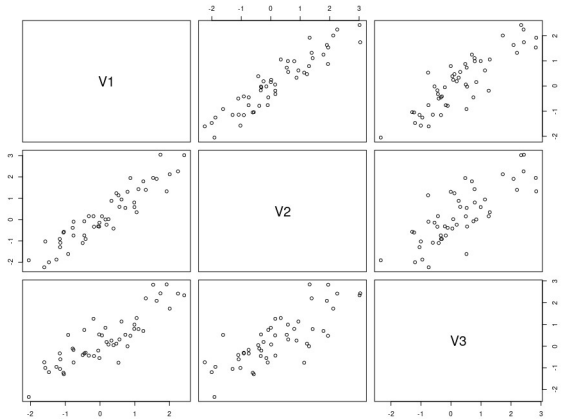
Case 1)
{V1} - {V2} - {V3}



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



Case 3)
{V1 - V2 - V3}



Pearson Correlation matrices

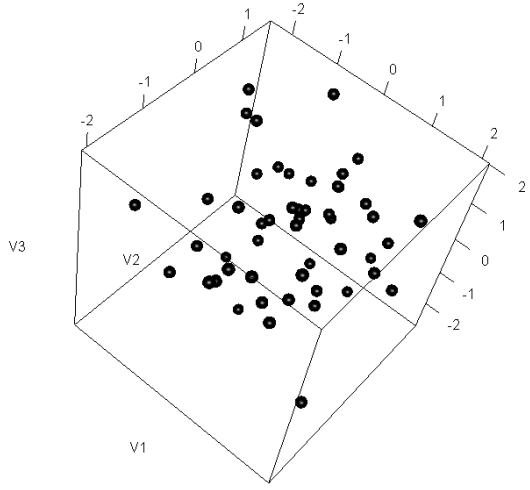
1)	V1	V2	V3
V1	1.00	-0.05	-0.12
V2	-0.05	1.00	0.06
V3	-0.12	0.06	1.00

2)	V1	V2	V3
V1	1.00	0.90	0.08
V2	0.90	1.00	-0.01
V3	0.08	-0.01	1.00

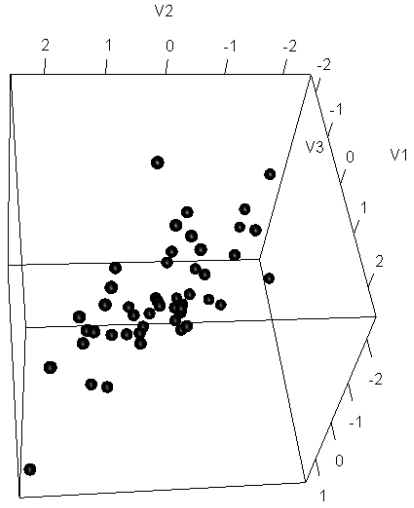
3)	V1	V2	V3
V1	1.00	0.93	0.87
V2	0.93	1.00	0.79
V3	0.87	0.79	1.00

Simulated examples

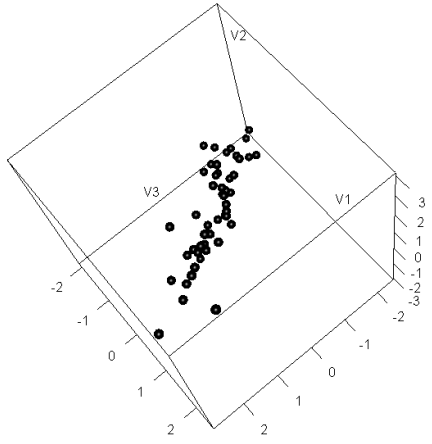
Case 1)



Case 2)



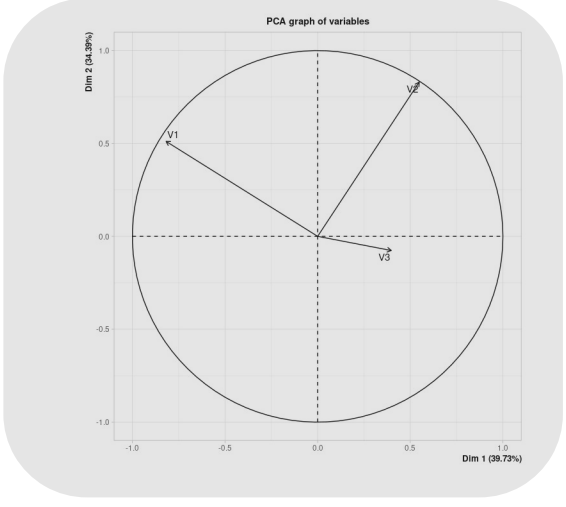
Case 3)

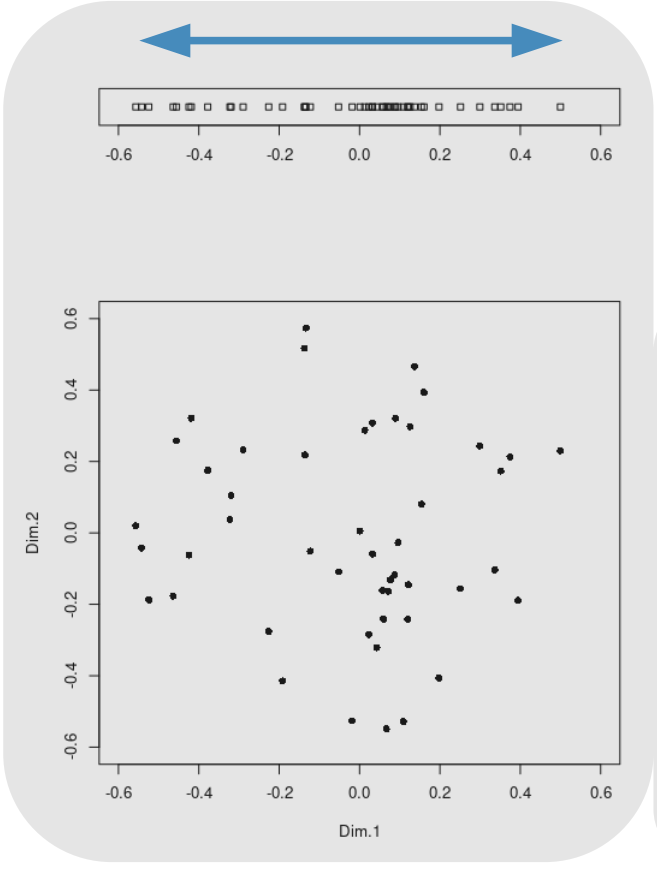


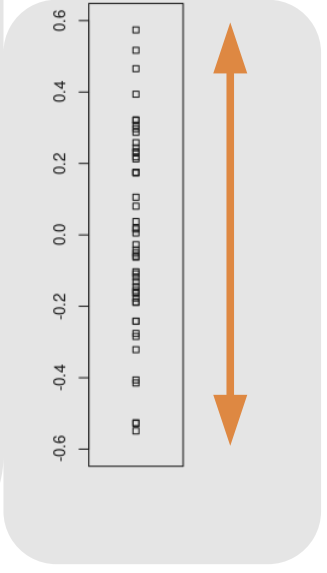
Simulated examples

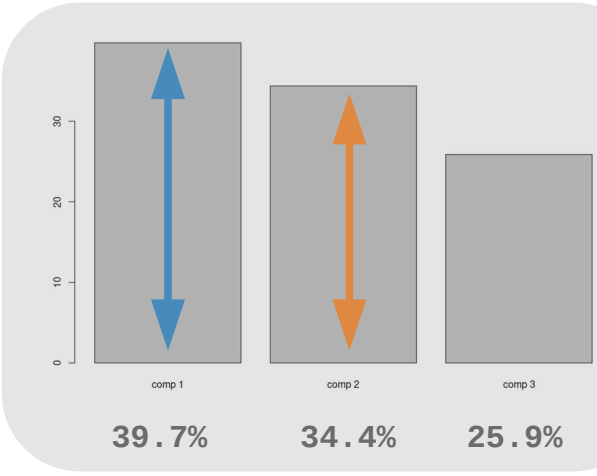
Loadings

	Dim.1	Dim.2	Dim.3
V1	-0.23	0.14	0.07
V2	0.15	0.23	-0.03
V3	0.10	-0.02	0.22





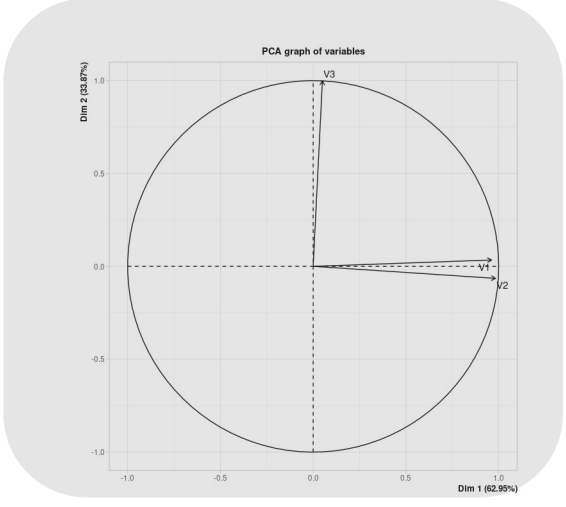


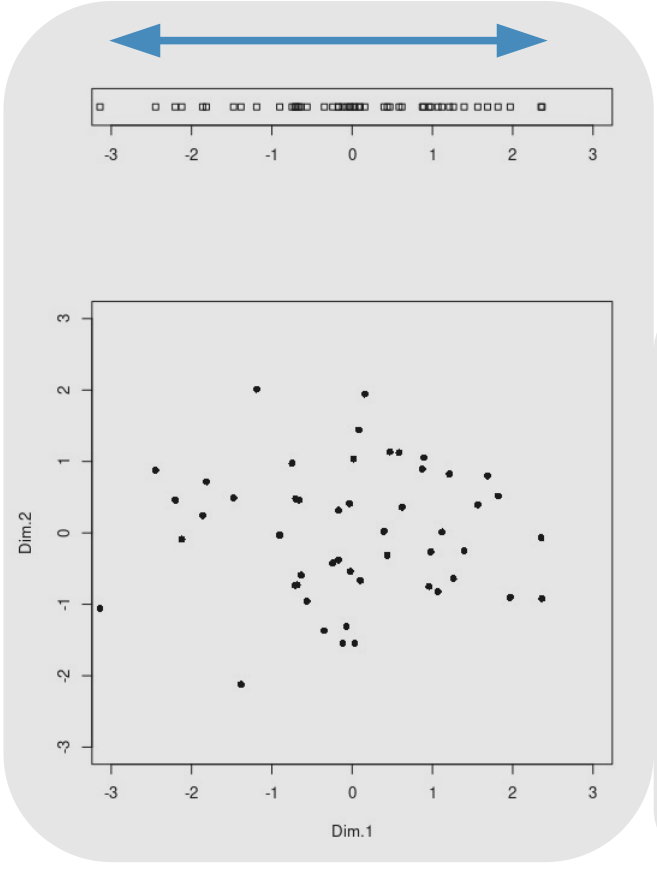


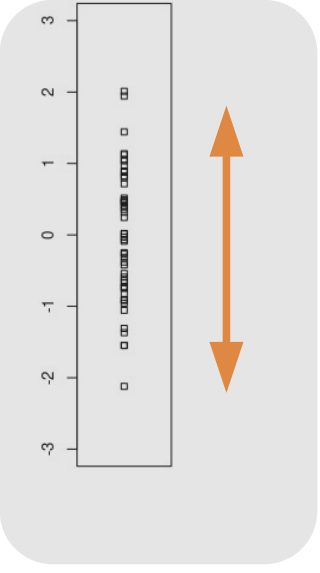
Simulated examples

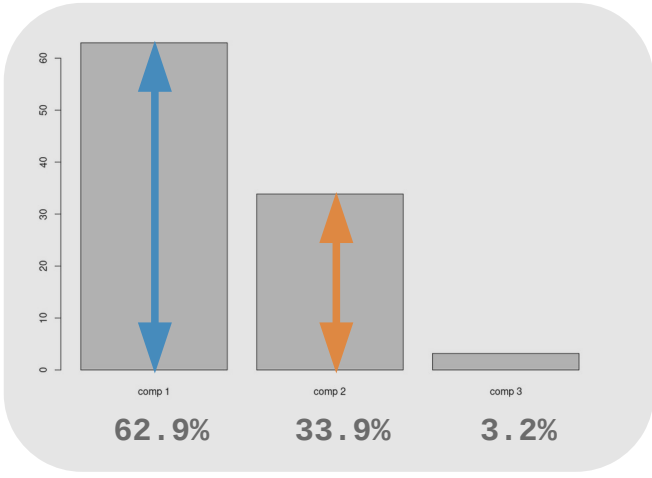
Loadings

	Dim.1	Dim.2	Dim.3
V1	0.77	0.03	0.22
V2	0.97	-0.06	-0.17
V3	0.05	0.91	-0.02





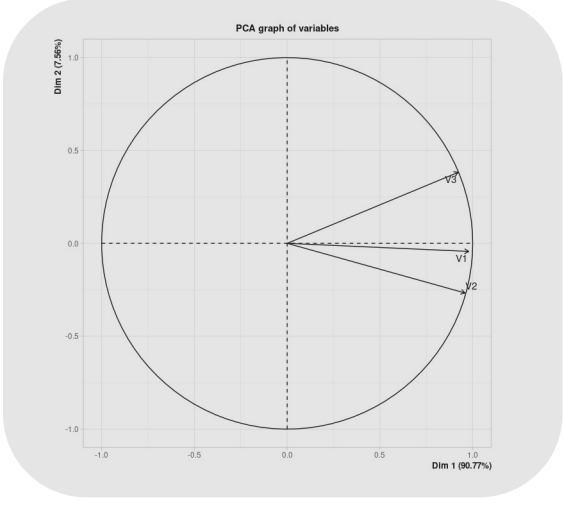


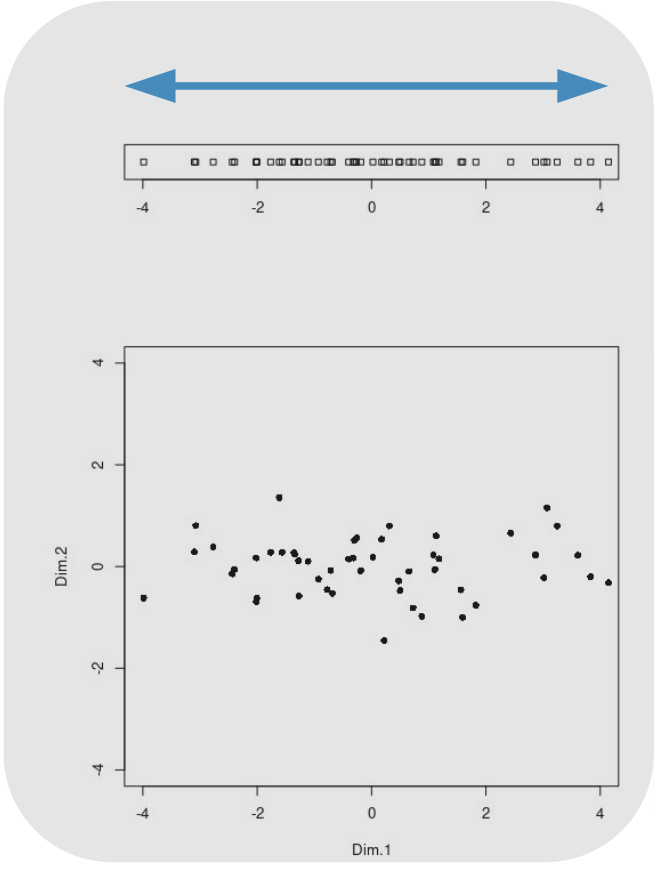


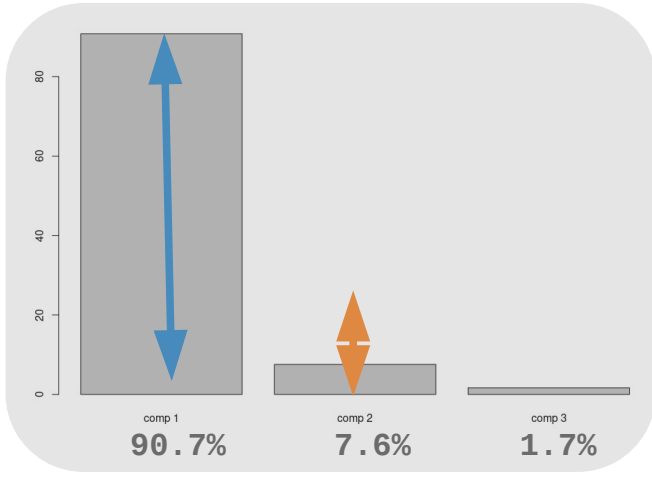
Simulated examples

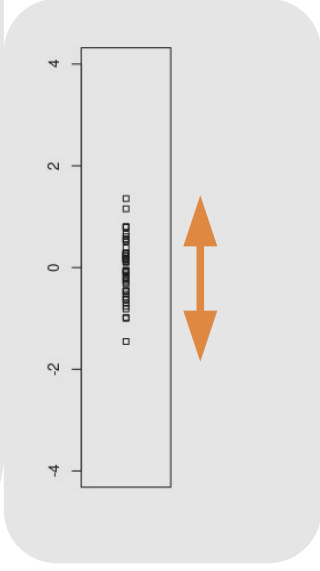
Loadings

	Dim.1	Dim.2	Dim.3
V1	1.07	-0.05	0.22
V2	1.23	-0.34	-0.13
V3	1.07	0.44	-0.07









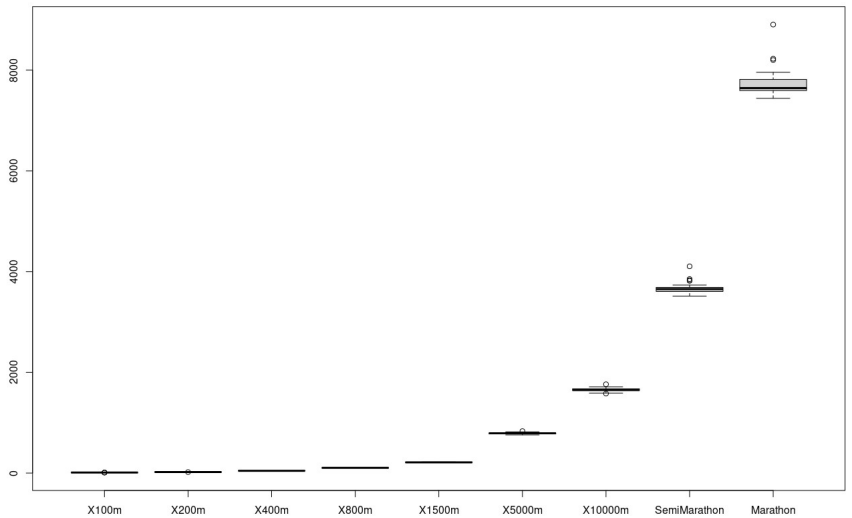
Athletics data set

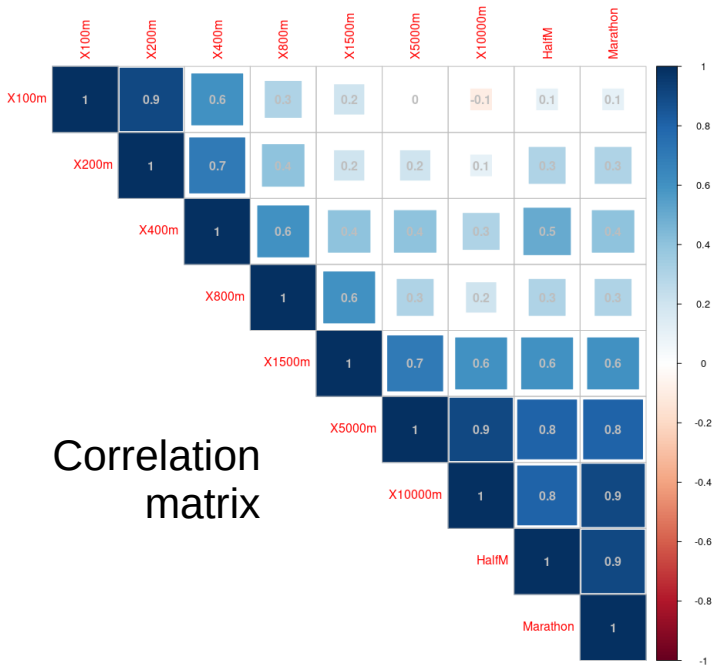
National records in seconds in athletics (needs an update...)

	100m	200m	400m	800m	1500m	5000m	10000m	HalfMarathon	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

Athletics data set

	100m	200m	400m	800m	1500m	5000m	10000m	HalfMarathon	Marathon
Mean	10.07	20.25	44.93	104.10	213.10	789.98	1655.54	3673	7754
Variance	0.04	0.2	0.5	2.2	13.3	297	1393	14326	92319
Std. Dev	0.20	0.46	0.7	1.5	3.7	17	37	119	303

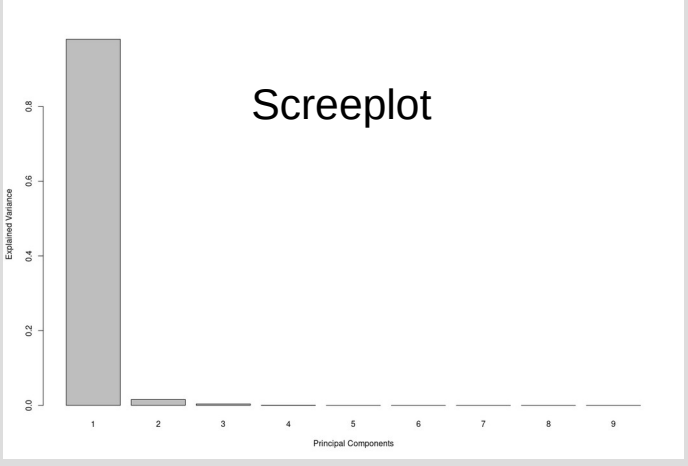




Correlation matrix

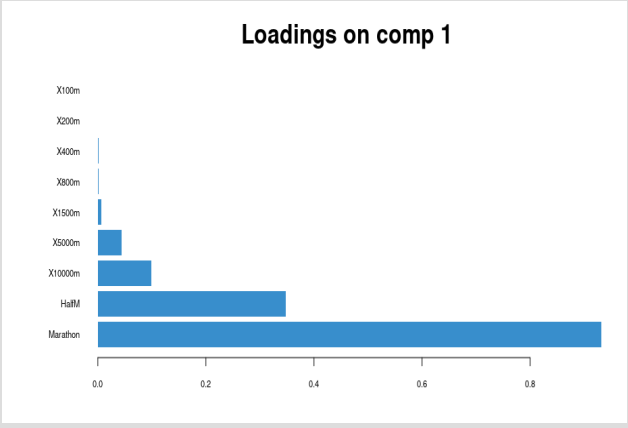
Athletics data set

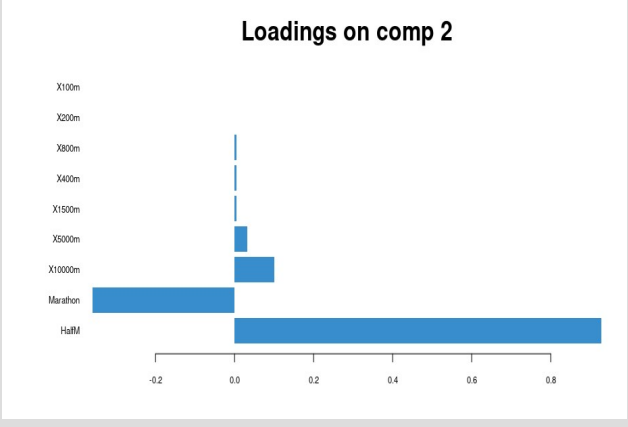
```
Importance of components:
                          PC1     PC2      PC3     PC4     PC5     PC6     PC7    PC8     PC9
Standard deviation       325.9   41.3     20.2    6.2     2.5     1.1     0.5    0.3     0.1
Proportion of Variance   0.9801  0.01575  0.00376 0.00035 0.00006 0.00001 0.0000 0.0000 0.00000
Cumulative Proportion    0.9801  0.99582  0.99958 0.99993 0.99999 1.00000 1.0000 1.0000 1.00000
```



Loadings

	PC1	PC2	PC3
100m	0.000	0.000	-0.003
200m	0.000	0.000	-0.008
400m	0.001	0.005	-0.009
800m	0.001	0.004	-0.002
1500m	0.007	0.006	0.028
5000m	0.044	0.033	0.396
10000m	0.099	0.101	0.905
HalfM	0.348	0.927	-0.137
Marathon	0.931	-0.359	-0.064

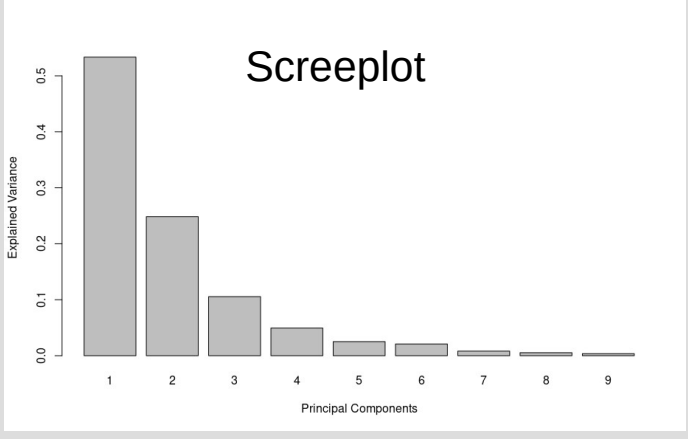




→ Marathon hides other distances. What to do?

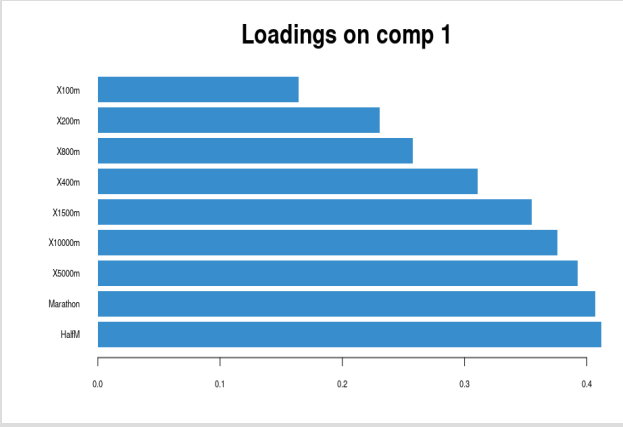
Athletics data set, scaled

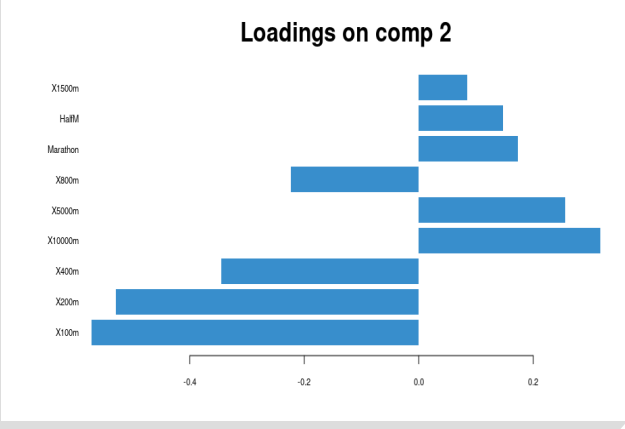
```
Importance of components:
                          PC1    PC2    PC3    PC4     PC5     PC6     PC7    PC8     PC9
Standard deviation      2.1909 1.4954 0.9744 0.6668 0.47542 0.43444 0.27107 0.21995 0.18124
Proportion of Variance  0.5334 0.2485 0.1055 0.0494 0.02511 0.02097 0.00816 0.00538 0.00365
Cumulative Proportion   0.5334 0.7818 0.8873 0.9367 0.96184 0.98281 0.99097 0.99635 1.00000
```



Loadings

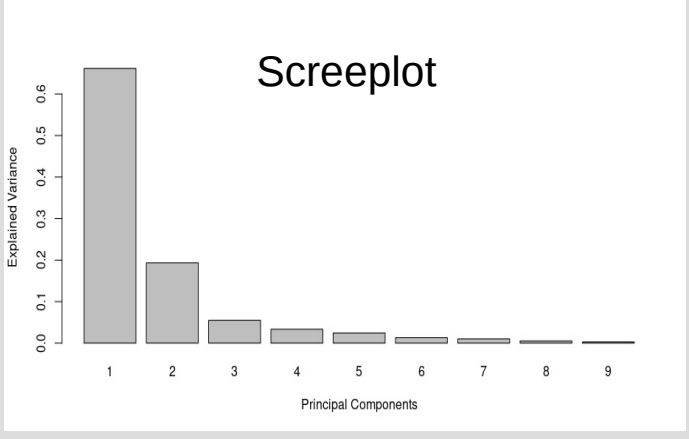
	PC1	PC2	PC3
100m	0.164	-0.571	-0.204
200m	0.231	-0.529	-0.255
400m	0.311	-0.345	-0.014
800m	0.258	-0.223	0.726
1500m	0.355	0.085	0.488
5000m	0.393	0.255	-0.054
10000m	0.376	0.317	-0.156
HalfM	0.412	0.147	-0.215
Marathon	0.407	0.174	-0.234





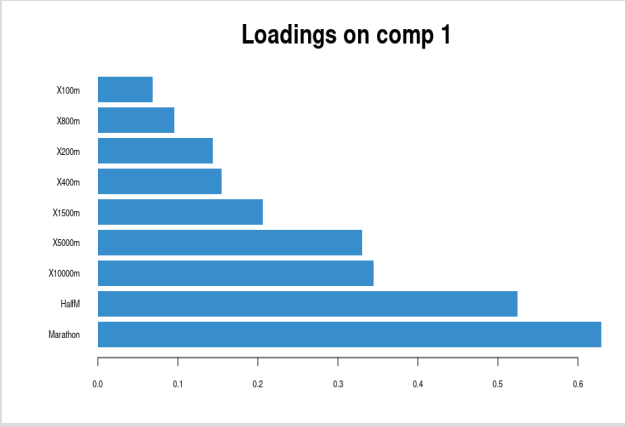
Athletics data set, log transform

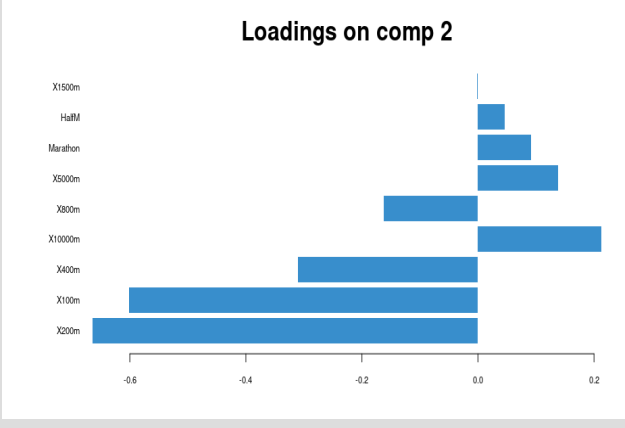
```
Importance of components:
                          PC1    PC2    PC3    PC4     PC5    PC6    PC7     PC8     PC9
Standard deviation     0.0580 0.0313 0.0167 0.0130 0.0111 0.00819 0.00724 0.00521 0.00385
Proportion of Variance 0.6616 0.1934 0.0550 0.0335 0.0245 0.01318 0.01031 0.00534 0.00292
Cumulative Proportion  0.6616 0.8550 0.9101 0.9436 0.9682 0.98143 0.99174 0.99708 1.00000
```



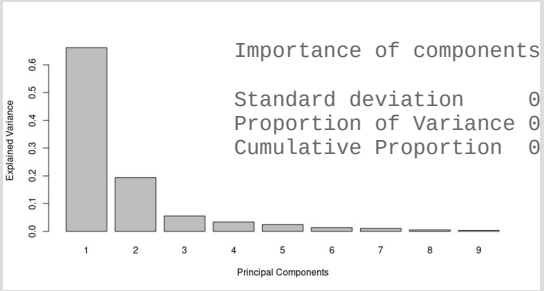
Loadings

	PC1	PC2	PC3
100m	0.068	-0.602	0.048
200m	0.144	-0.664	0.153
400m	0.155	-0.310	-0.121
800m	0.095	-0.163	-0.618
1500m	0.206	-0.002	-0.632
5000m	0.331	0.138	-0.245
10000m	0.345	0.212	-0.043
HalfM	0.525	0.046	0.186
Marathon	0.629	0.091	0.287





Athletics data set, - log transform

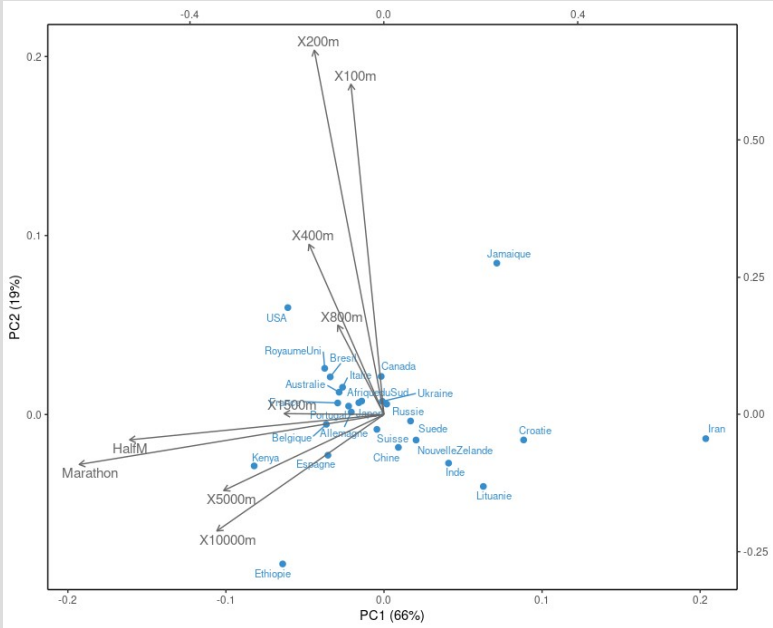



```
Importance of components:
                            PC1     PC2     PC3
Standard deviation     0.05804 0.03138 0.01675
Proportion of Variance 0.66162 0.19343 0.05508
Cumulative Proportion  0.66162 0.85505 0.91013
```

Loadings

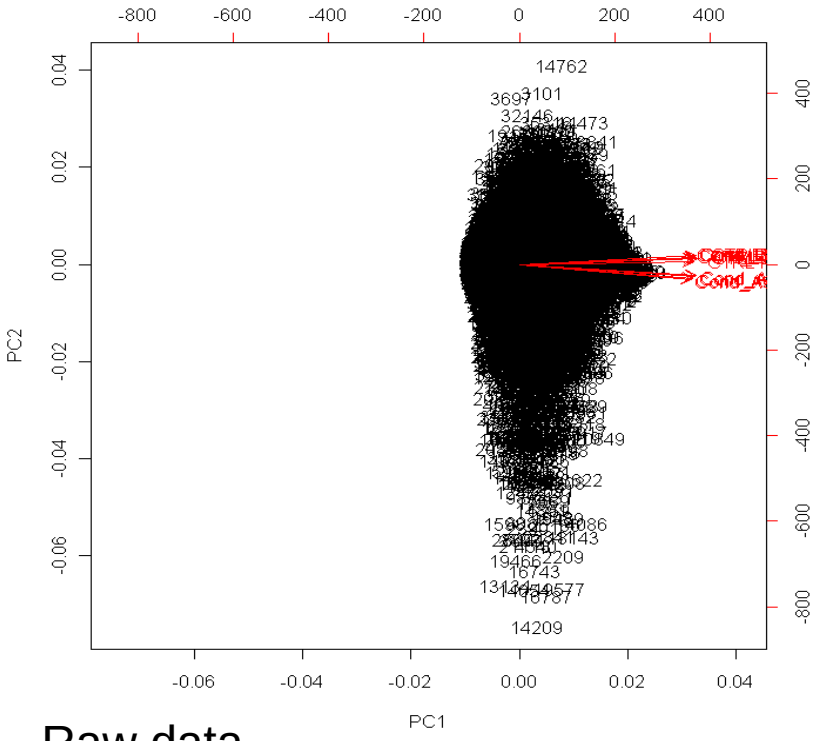
	PC1	PC2	PC3
100m	**-0.068**	**0.602**	-0.048
200m	**-0.144**	**0.664**	-0.153
400m	**-0.155**	0.310	0.121
800m	**-0.095**	0.163	**0.618**
1500m	**-0.206**	0.002	**0.632**
5000m	**-0.331**	-0.138	0.245
10000m	**-0.345**	-0.212	0.043
HalfM	**-0.525**	-0.046	-0.186
Marathon	**-0.629**	-0.091	-0.287

- **The trick : -**log (**minus** log) transform to locate countries in the direction of the distance, they have good performance (otherwise the better the country, the smaller the value)
- **PC1**: every variable hase the same sign to define PCA → highlights a global performance,from right to left, coutries are globally better
- **PC2**: emphasizes short distances (100m and 200m)
- **PC3**: provides information on intermediate distances (800m and 1500m)



Transcriptomics data set (1)

- Design of experiment: 3 conditions (CTRL, A, B) x 4 replicates
- Expression of 38000 genes, Affymetrix chip



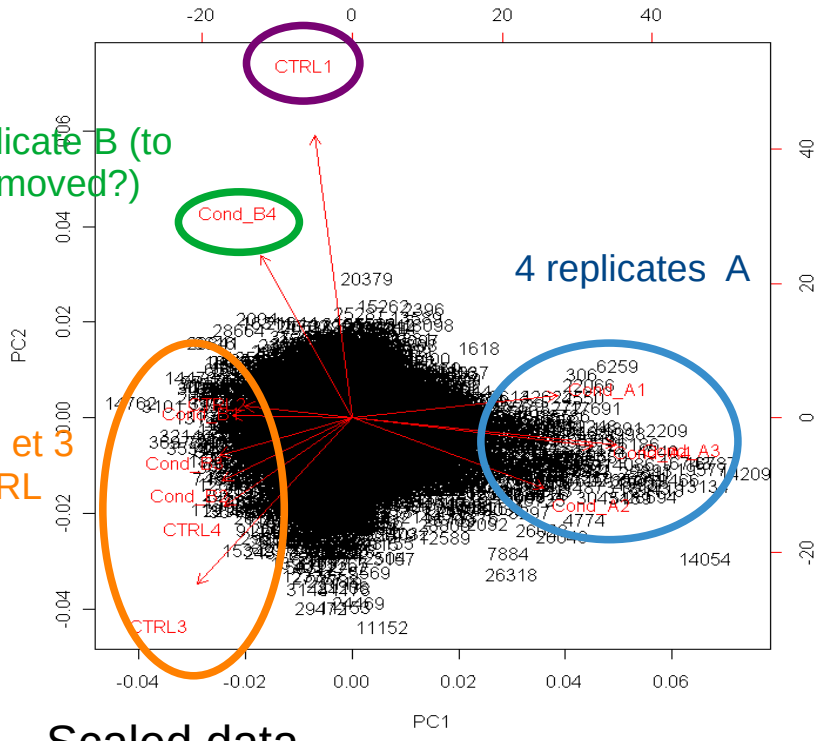
Raw data

1 replicate CTRL
(to be removed?)

1 replicate B (to
be removed?)

3 replicates B et 3
replicates CTRL

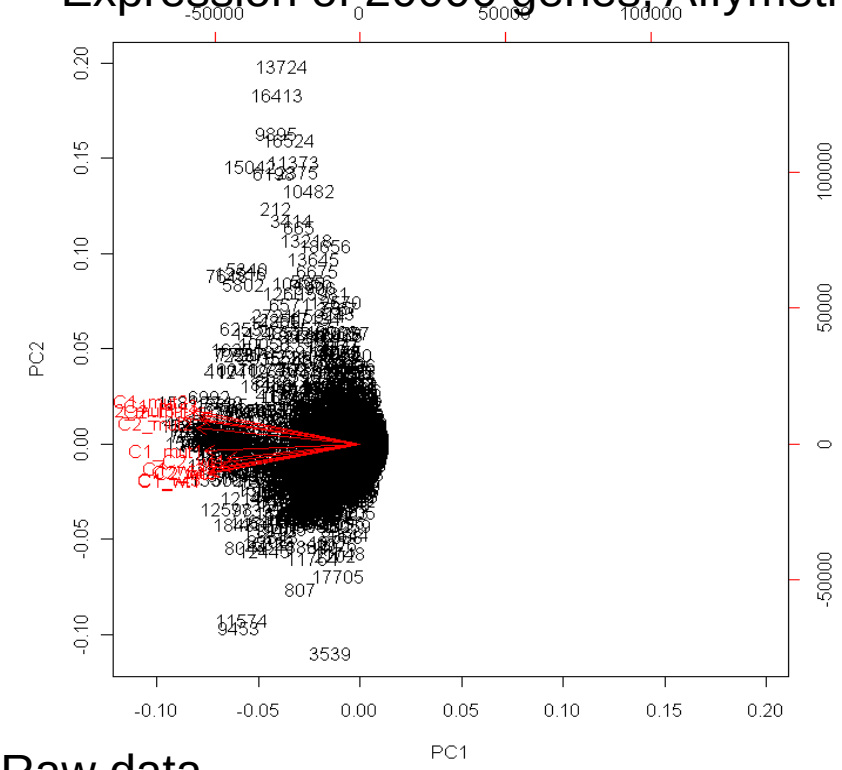
4 replicates A



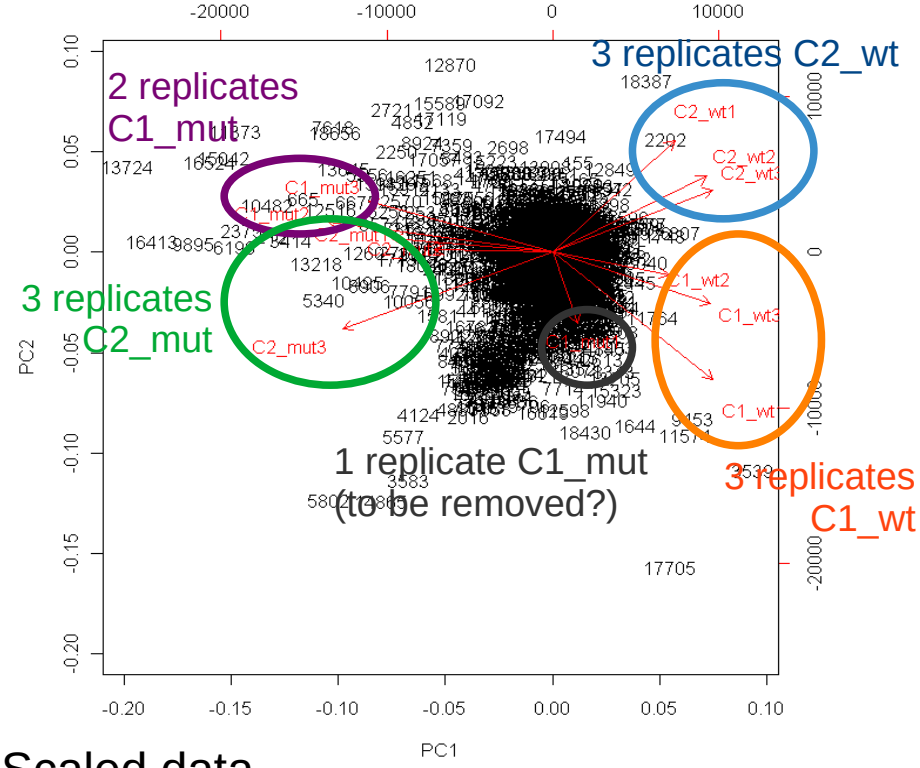
Scaled data

Transcriptomics data set (2)

- Design of experiment: 2 crossed-factor Genotype (WT, Mut) x Treatment (C1, C2) x 3 replicates
- Expression of 20000 genes, Affymetrix chip



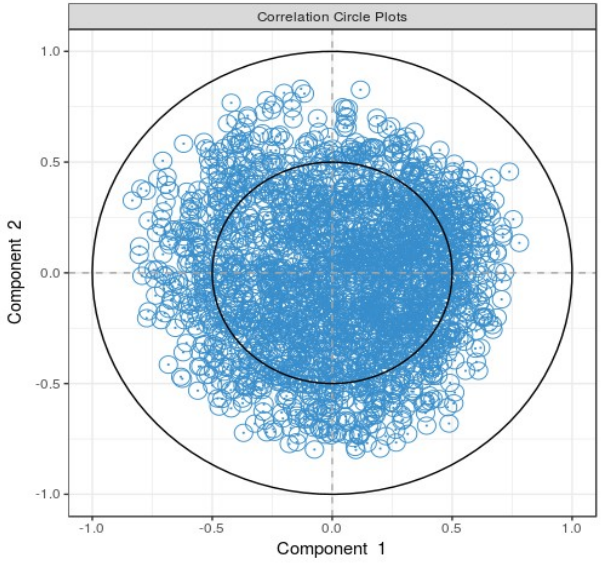
Raw data



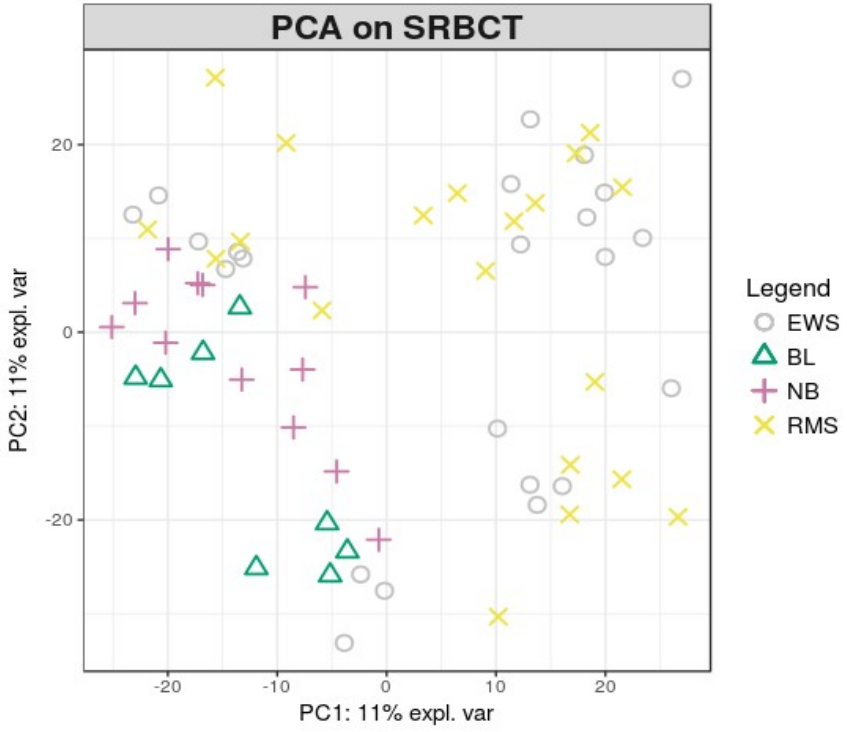
Scaled data

Transcriptomics data set (3)

- Expression of 2308 genes
- 63 samples divided in 4 classes : 23 EWS, 8 BL, 12 NB, 20 RMS



Variable plot: not very meaningful



Individual plot: even if PCA is an unsupervised method (i.e. the class is not taken into account for the computation), the class can be used to color the individuals to see if they naturally cluster according to their class membership.

Take home message

To put it in a nutshell

- Practice on your own data! The best way to understand what the method has to tell you.
- Do not bypass elementary analyses (univariate, bivariate)
- PCA is an **unsupervised method**, it is not dedicated to identify clusters, so please avoid this kind of sentences “PCA is not a good method, I can’t see my clusters...”