Exploration and Integration of Heterogeneous Biological Data Sets with mixOmics

Sébastien Déjean

www.math.univ-toulouse.fr/~sdejean

Outline

- **Introduction**: interdisciplinarity, data integration, answer a question
- **Tools:** mixOmics R package, workflow
- **Methods**: understand PCA, extend to integration problems, sparsity, vertical integration
- **Examples**: simulated toy examples, liver toxicity data set

Interdisciplinarity

*The biological sciences are **today** in the process of changing from being primarily descriptive **to being very much quantitative**. As a result, biologists find themselves **confronted more and more with large amounts of numerical data** […]. But the mere collecting and recording of data achieve nothing; having been collected, they must be **investigated to see what information may be contained concerning the biological problem** at hand.[…]*

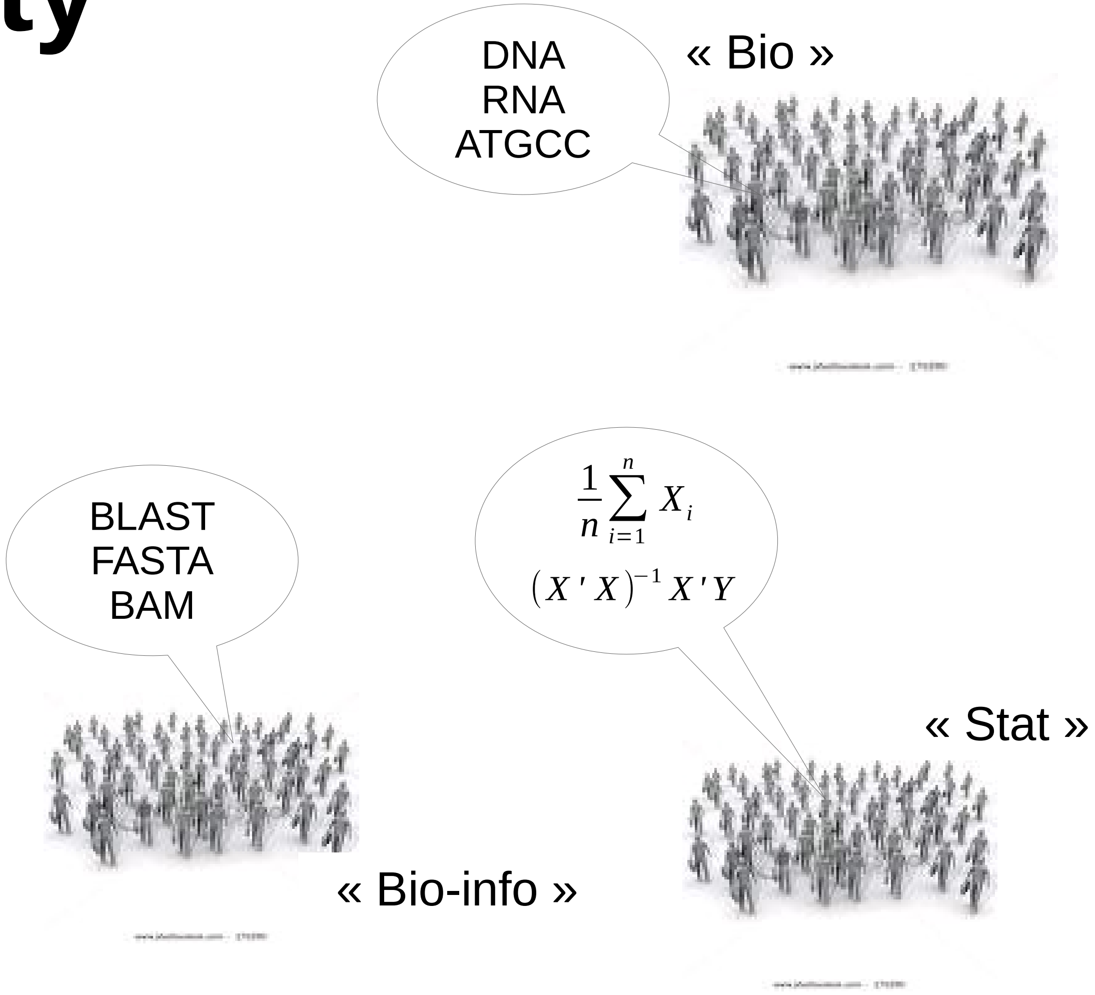
*Frequently, however, biologists have to subject their data to more complex calculations, requiring procedures that **involve mathematical details beyond their general experience**. In order to carry out the mathematics the biologist in this situation must either **learn the procedures himself**, or at least **learn something of the language of mathematics**, that he may **communicate satisfactorily with the mathematician** whose aid he enlists.*

S.R Searle (**1966**)
Matrix Algebra for the biological sciences

Interdisciplinarity

- 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



DNA
RNA
ATGCC

« Bio »

BLAST
FASTA
BAM

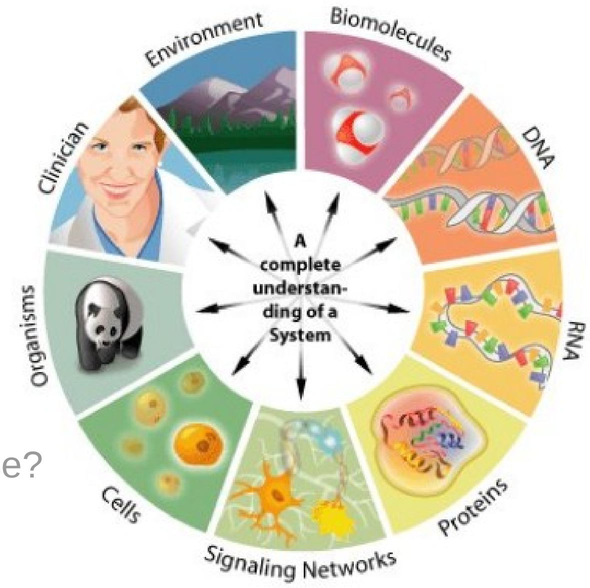
$$\frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\left(X'X\right)^{-1}X'Y$$

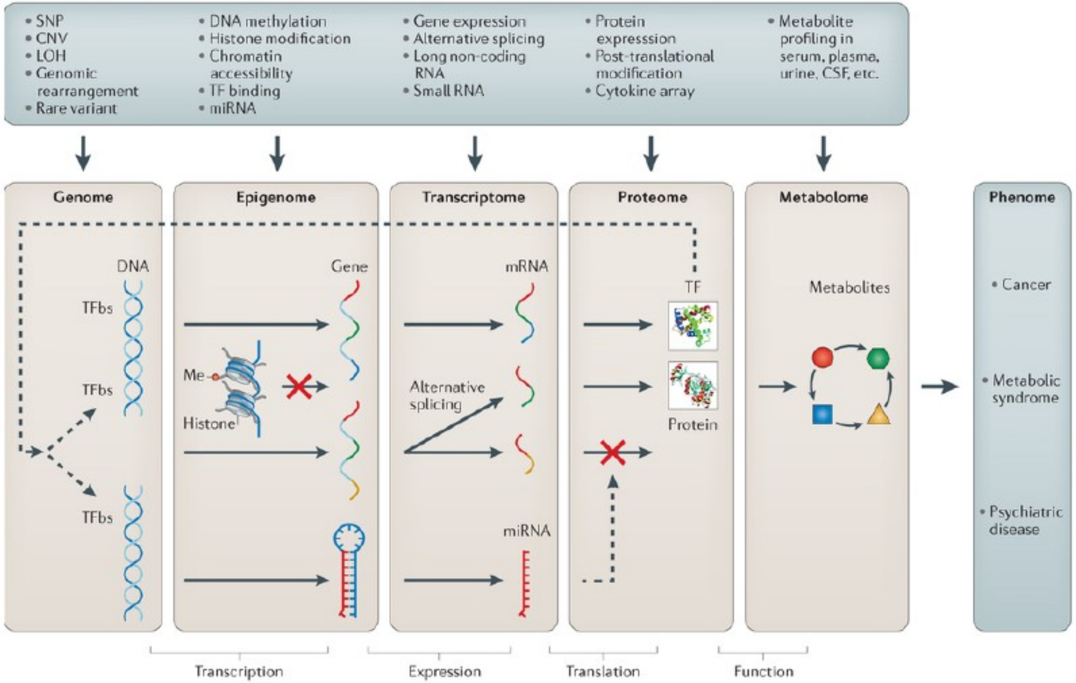
« Stat »

« Bio-info »

Data integration



Source?

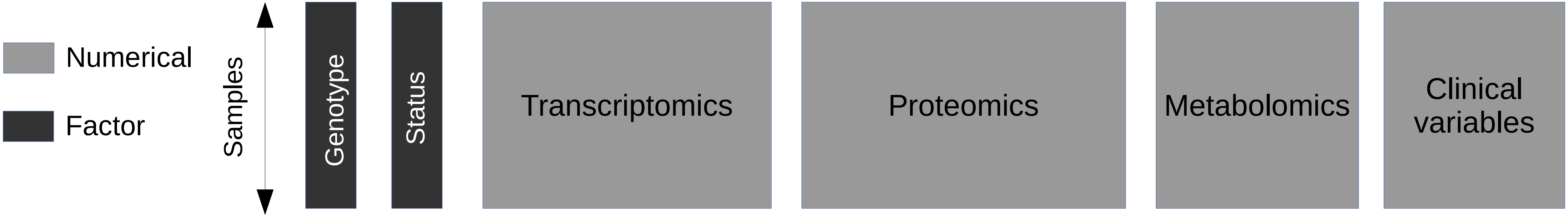


From Ritchie et al. (2015), *Nature reviews. Genetics*
Methods of integrating data to uncover genotype-phenotype interactions.

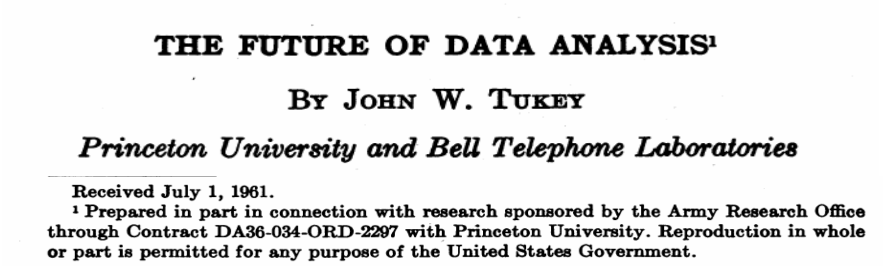
Generally, data integration can be defined as the process of combining data residing in diverse sources to provide users with a comprehensive view of such data. There is no universal approach to data integration, and many techniques are still evolving.
From Schneider, M. V., & Jimenez, R. C. (2012). Teaching the Fundamentals of Biological Data Integration Using Classroom Games. PLoS Computational Biology, 8(12)

Data integration with statistics

Goal: extract knowledge from data



Answer a question



THE FUTURE OF DATA ANALYSIS[1]

By John W. Tukey

Princeton University and Bell Telephone Laboratories

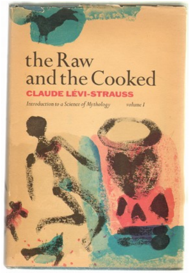
Received July 1, 1961.

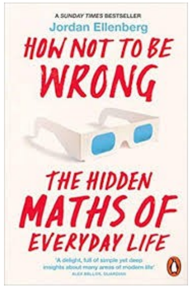
[1] Prepared in part in connection with research sponsored by the Army Research Office through Contract DA36-034-ORD-2297 with Princeton University. Reproduction in whole or part is permitted for any purpose of the United States Government.

Far better an approximate answer to ***the right question [...]****, than an exact answer to the wrong question [...].*

The scientific mind does not so much provide the right answers as ***ask the right questions****.*

C. Lévi-Strauss. The Raw and the Cooked (1964)





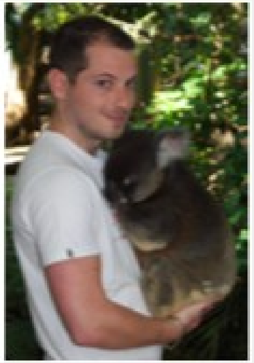
[…] in order to give a sensible answer, you need to know more than just numbers […] It's ***only after you've started to formulate these questions*** *that you take out the calculator. But* ***at that point the real mental work is already finished****. Dividing one number by another is mere computation; figuring out what you should divide by what is mathematics.*

Tool(s)

- Package for R
  `r-project.org`
- Freely available on Bioconductor
  `bioconductor.org/packages/release/bioc/html/mixOmics.html`
- Web site `mixomics.org`
- Forum
  `mixomics-users.discourse.group`

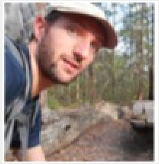
mixOmics facebook

- Core team



Sébastien Déjean,
Kim-Anh Lê Cao,
Ignacio Gonzalez,
Florian Rohart

- Key developers / contributors



Benoit Gautier

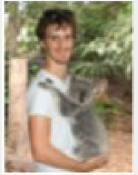
Xin-Yi Chua

Amrit Singh

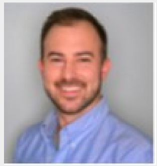
Max Bladen

Al J Abadi

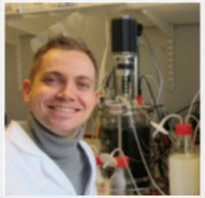
François Bartolo



Casey Shanon

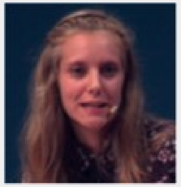


- Tutors / teachers contributors



Olivier Chapleur

Eva Yiwen Wang



Laëtitia Cardona

David Rengel

Yannick Lippi

Jerôme Mariette

- Many users and trainees

mixOmics workflow

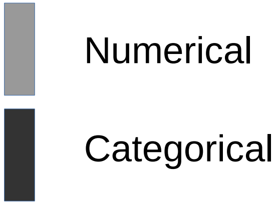
1) Run a method: `pca(DataX), spca(DataX, keepX=c(…)), pls(DataX, DataY), spls(DataX, DataY, keepX=c(…), keepY=c(…)), plsda(DataX, Categ), splsda(DataX, Categ, keepX=c(…)), block.pls(list(…)), block.spls(list(…)), block.plsda(list(…), Categ), block.splsda(list(…), Categ, keepX=list(…)), mint.pca(DataX, study=…), mint.plsda(DataX, Categ, study=…), …`

`Optional argument for compositional data (microbial dataset):` **`logratio = ‘CLR’`**

2) Represent individuals: `plotIndiv()`

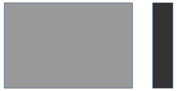
3) Represent variables: `plotVar(), plotLoadings(), cim(), network()`

Overview of statistical methods

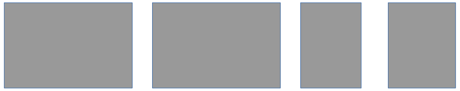


- **Multivariate unsupervised**
One numerical dataset `pca(), spca()`

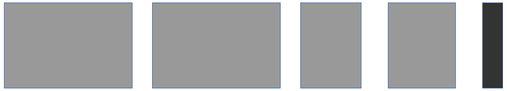
- **Multivariate supervised**
One numerical dataset and one categorical variable `plsda(), splsda()`



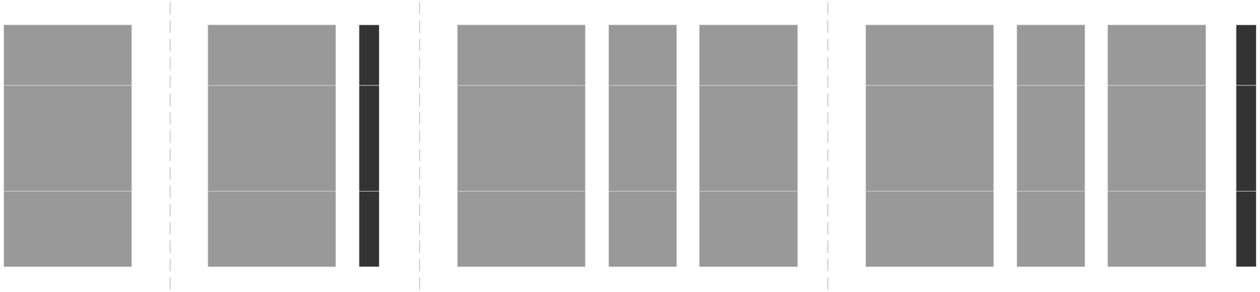
- **Multi-block unsupervised**
Several numerical datasets, same samples
`pls(), spls(), block.pls(), block.spls()`



- **Multi-block supervised**
Several numerical datasets and one categorical variable, same samples
`block.plsda(), block.splsda()`



- **Multi-group analyses**
Same as above with samples divided pre-defined in groups (batch, study…)
`mint.pca(), mint.plsda(), mint.splsda(), mint.block.pls(), mint.block.spls(), mint.block.plsda(), mint.block.splda()`



Methods

- Understand Principal Component Analysis
- Extend to integration problems
- Sparsity

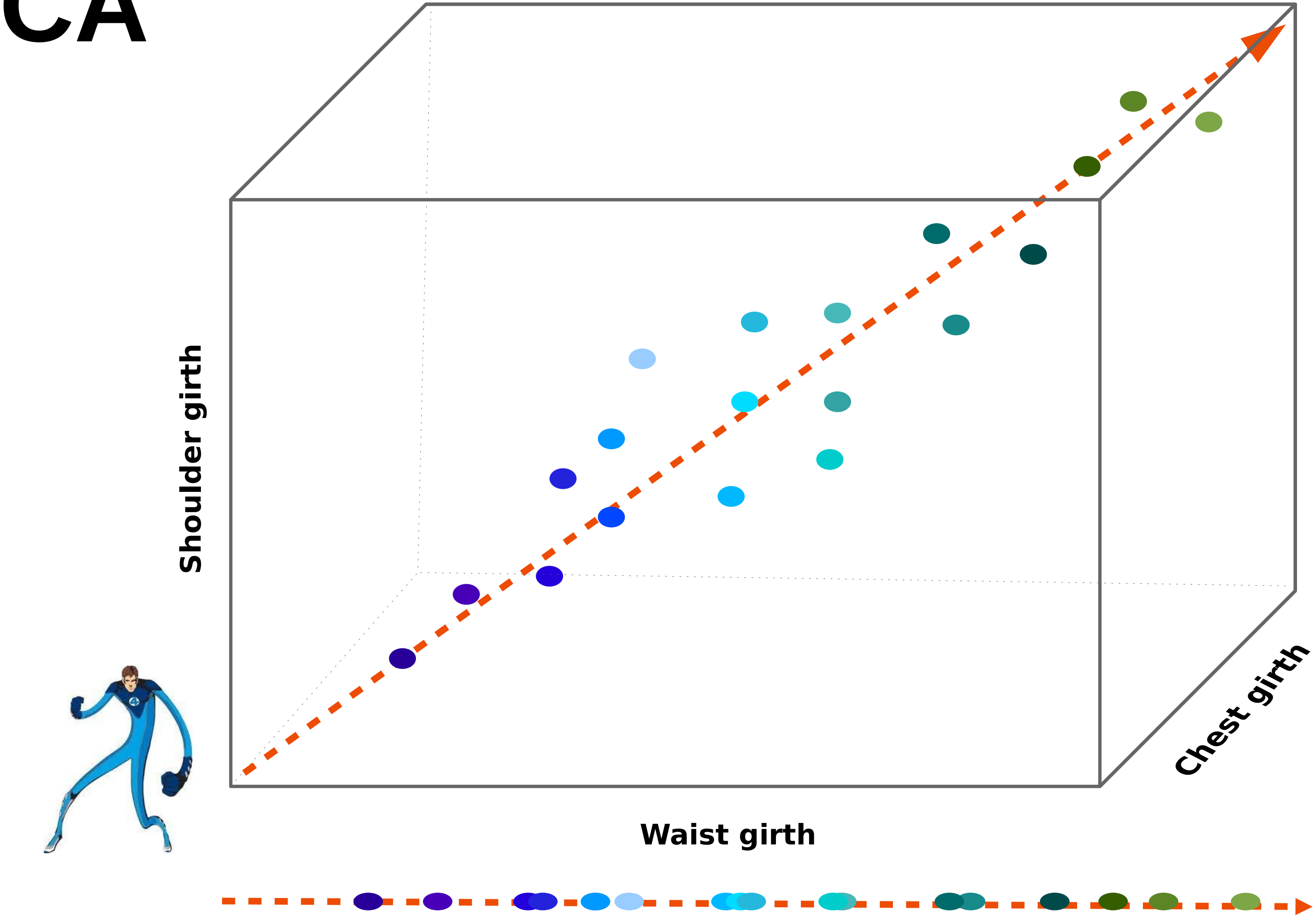
The Batman principle: *It's not who I am underneath, but what I do that defines me.*

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

Understand PCA

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

PCA



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

=

Do we need a cubic box to pack a fishing rod?

1st Principal Component: «beefyness»

PCA: (verbose) comments

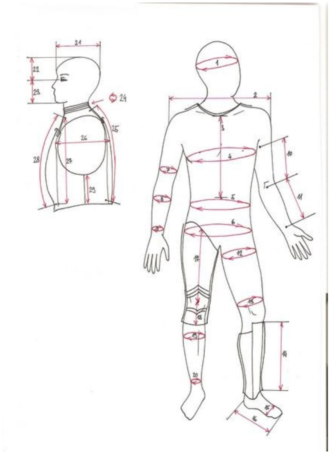
- The measurements are rather **strongly correlated.** Indeed, one can assume that a person with a high shoulder girth will also have high chest girth. In these conditions, the information brought by the 3 variables are **redundant**. Graphically, in the cube determined by shoulder girth, chest girth 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

Bivariate analysis

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

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

Univariate analysis

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

351.9 represents the quantity of information contained in the data.

$$68.6 + 37.5 + 50.8 + 85.7 + 109.3 = \mathbf{351.9}$$

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.

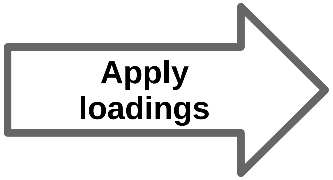
Around the core

Centered data

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

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

	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

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

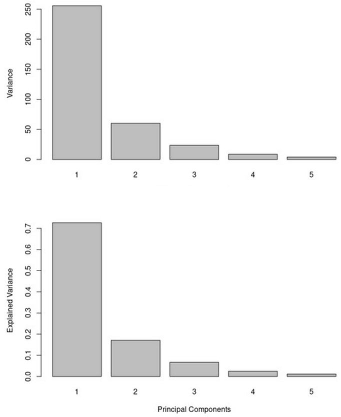
255.7 is the greatest variance we can obtain with a linear combination of the initial variables.

Graphical outputs (1/3)

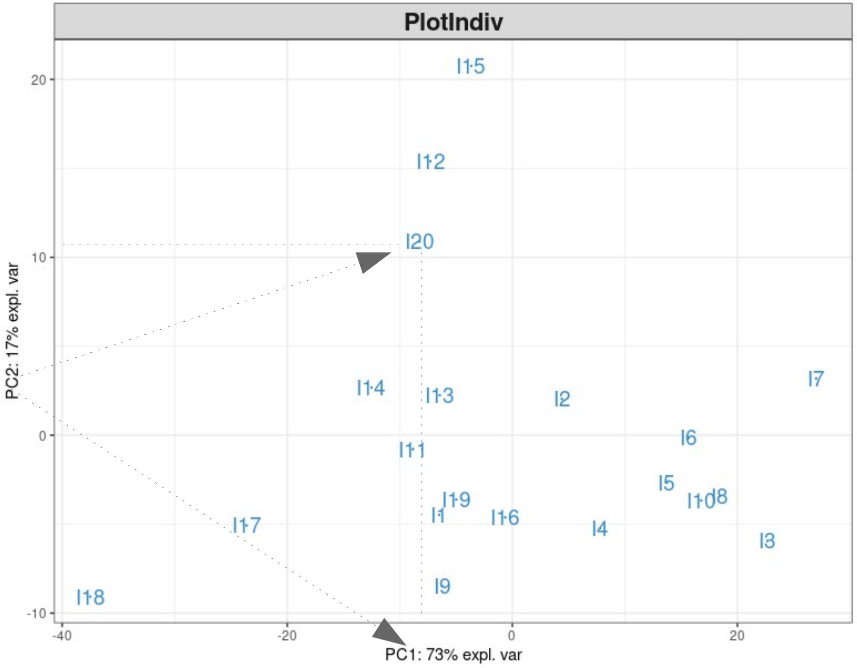
	PC1	PC2	PC3	PC4	PC5
Variance	255.7	60.2	23.5	8.6	4.0
% variance	72.6	17.1	6.7	2.4	1.1

	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

Screeplot



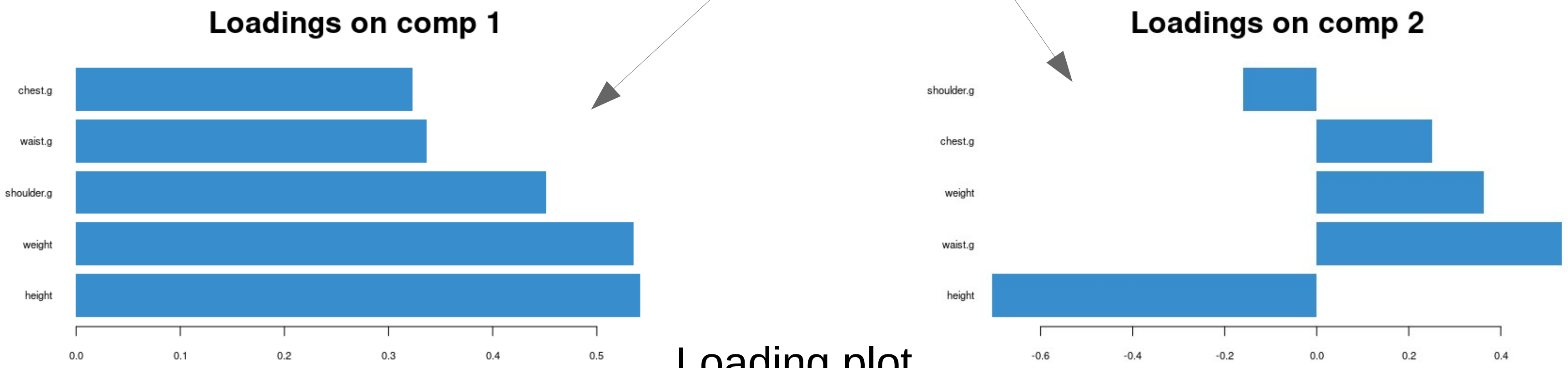
Individual plot



Graphical outputs (2/3)

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

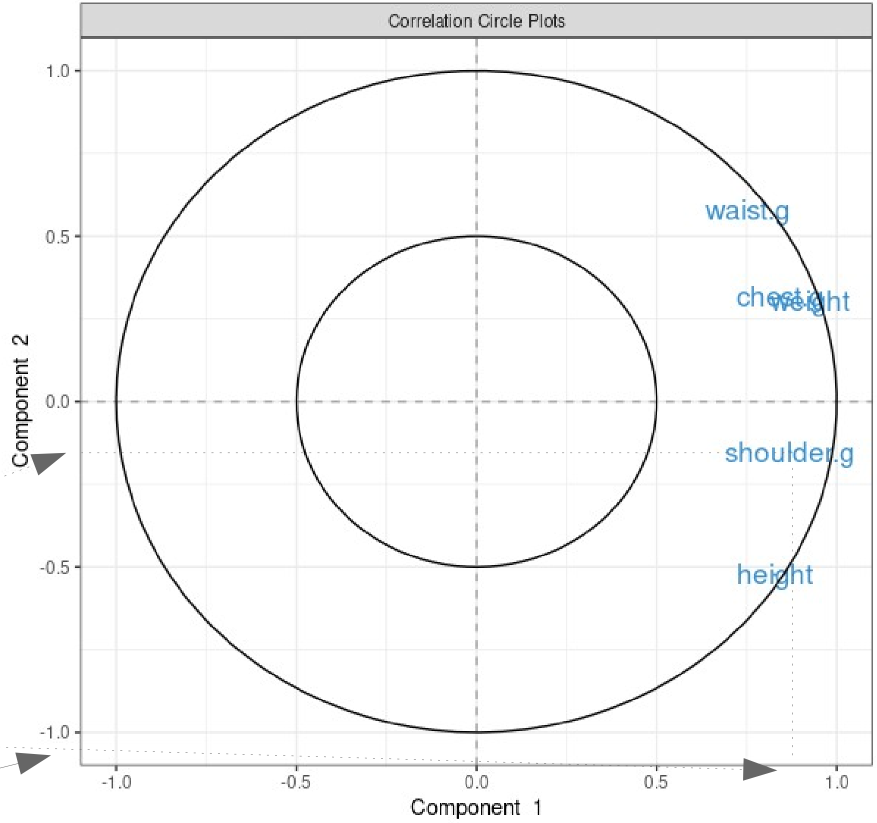
Graphical outputs (3/3)

```
Id      s.g     c.g    w.g    w      h                   PC1     PC2
I1     106.2    89.5   71.5   65.6   174.0        I1     -6.50   -4.48
I2     110.5    97.0   79.0   71.8   175.3        I2      4.40    2.04
I3     115.1    97.5   83.2   80.7   193.5        I3     22.66   -5.94
I4     104.5    97.0   77.8   72.6   186.5        I4      7.78   -5.24
I5     107.5    97.5   80.0   78.8   187.2        I5     13.73   -2.67
I6     119.8    99.9   82.5   74.8   181.5        I6     15.67   -0.15
I7     123.5   106.9   82.0   86.4   184.0        I7     26.99    3.19
I8     120.4   102.5   76.8   78.4   184.5        I8     18.41   -3.43
I9     111.0    91.0   68.5   62.0   175.0        I9     -6.25   -8.48
I10    119.5    93.5   77.5   81.6   184.0        I10    16.78   -3.67
I11    105.0    89.0   71.2   67.3   169.5        I11    -8.83   -0.78
I12    100.2    94.1   79.6   75.5   160.0        I12    -7.28   15.41
I13     99.1    90.8   77.9   68.2   172.7        I13    -6.45    2.25
I14    107.6    97.0   69.6   61.4   162.6        I14   -12.51    2.68
I15    104.0    95.4   86.0   76.8   157.5        I15    -3.65   20.76
I16    108.4    91.8   69.9   71.8   176.5        I16    -0.63   -4.62
I17     99.3    87.3   63.5   55.5   164.4        I17   -23.61   -5.07
I18     91.9    78.1   57.9   48.6   160.7        I18   -37.50   -9.07
I19    107.1    90.9   72.2   66.4   174.0        I19    -4.98   -3.61
I20    100.5    97.1   80.4   67.3   163.8        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



Focus on individual plot

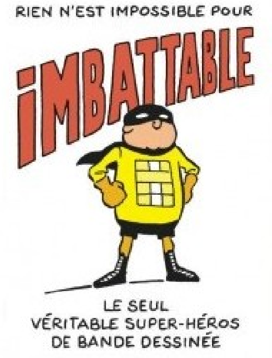
- To interpret the graphical results of PCA must be done keeping in mind that one is looking at a projection on a plane (or in a volume for 3D representation)

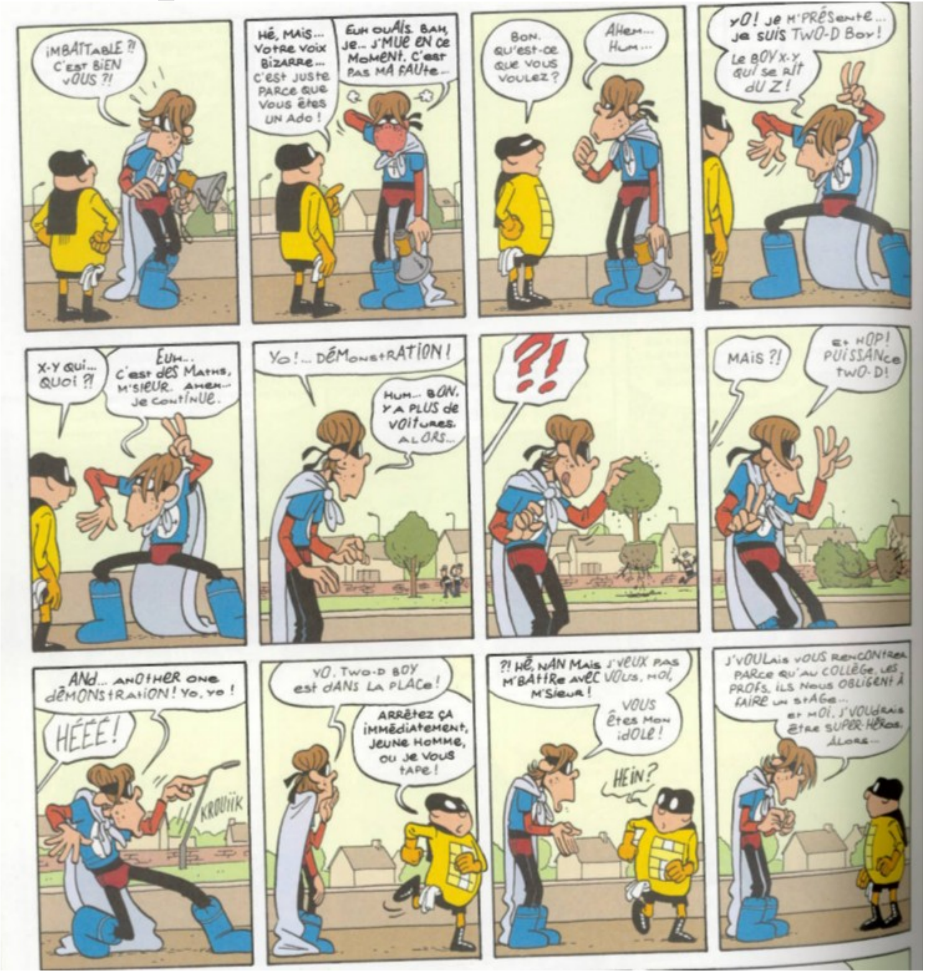
- Be careful when interpreting visual proximities

- Illustration in comics with the only true super-heros ...

Scenario & illustration: Pascal Jousselin
Colour: Laurence Croix

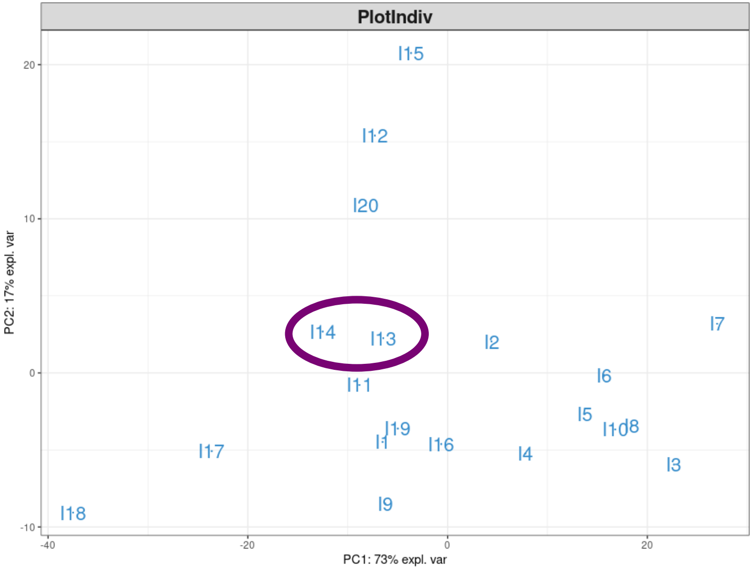
pjousselin.free.fr

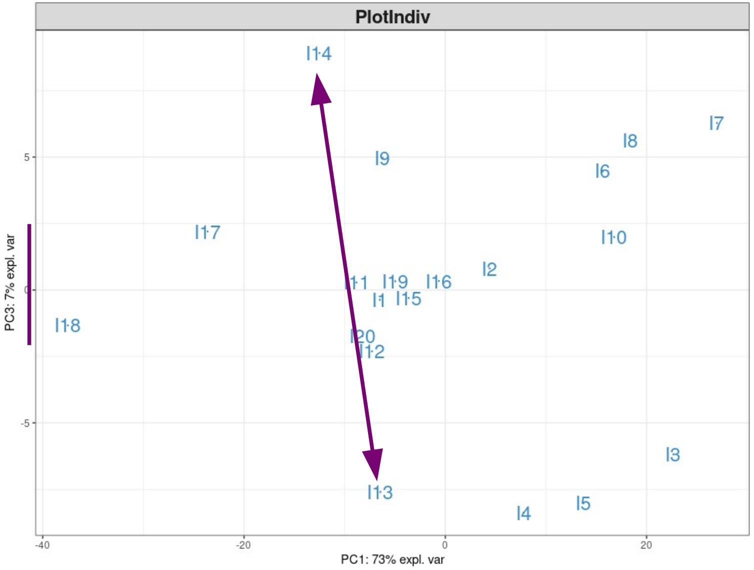

RIEN N'EST IMPOSSIBLE POUR
iMBATTABLE
LE SEUL VÉRITABLE SUPER-HÉROS DE BANDE DESSINÉE

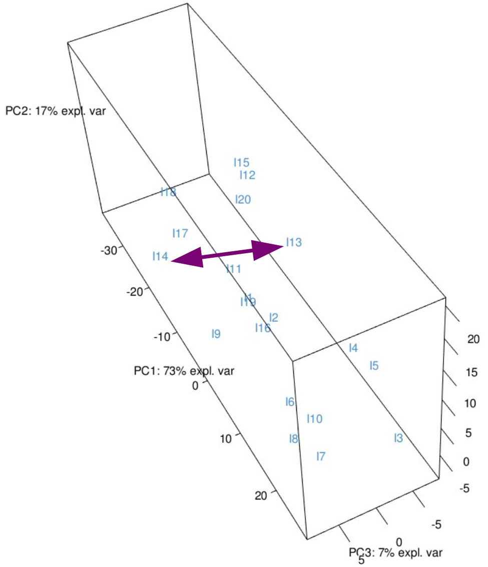

iMBATTABLE ?! C'EST BIEN VOUS ?!
HÉ, MAIS... VOTRE VOIX BIZARRE... C'EST JUSTE PARCE QUE VOUS ÊTES UN ADO !
EUH OUAIS. BAH, JE... J'MUE EN CE MOMENT. C'EST PAS MA FAUTE...
BON. QU'EST-CE QUE VOUS VOULEZ ?
AHEM... HUM...
YO ! JE M'PRÉSENTE... JE SUIS TWO-D BOY !
LE BOY X-Y QUI SE RIT DU Z !
X-Y QUI... 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...

Focus on individual plot

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



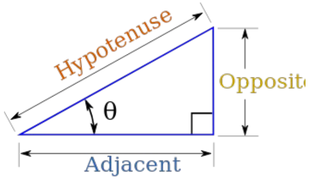


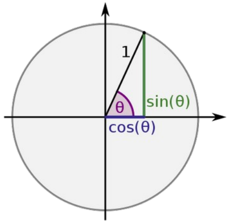


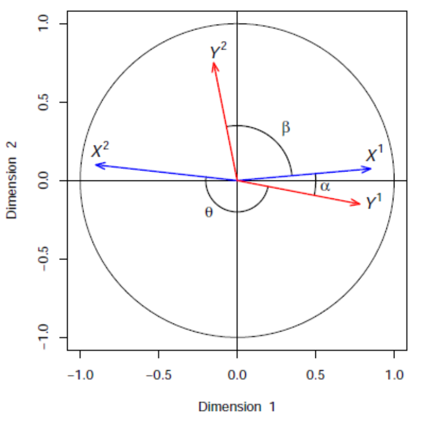
Focus on variable plot

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

Graphical outputs: summary

Screeplot

- How many components?
- 90% with 2 Pcs, 97% with 3PCs, 100% with 5PCs

Individual plot

- ‘Natural’ clusters, outliers...
- Caution: visual proximities

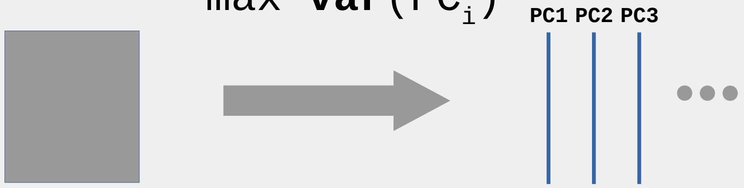
Variable plot, loading plot

- Correlation between variables
- Interpret components: PC1 « beefyness », PC2 « fatness, rotundity »

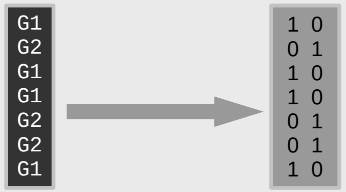
Extension to integration problems

PCA

$\max\ \mathbf{var}(PC_i)$



The trick for discriminant analyses: convert a factor into a numeric (dummy) matrix

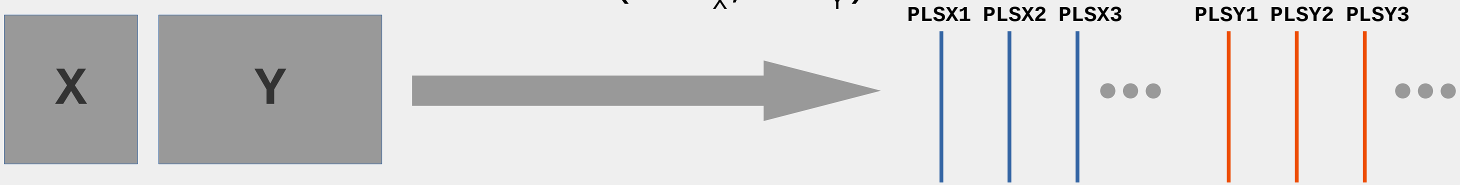


G1	1 0
G2	0 1
G1	1 0
G1	1 0
G2	0 1
G2	0 1
G1	1 0

PLS-DA → PLS

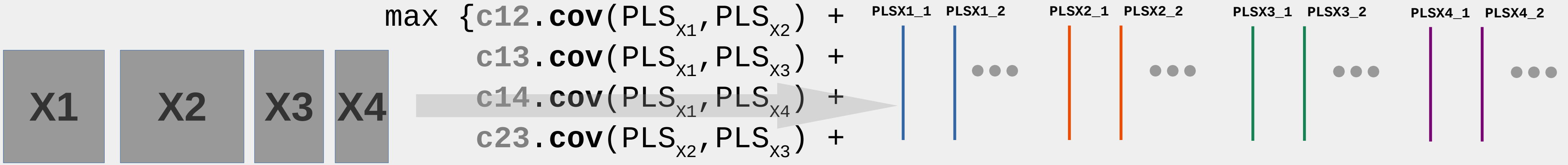
PLS, PLS-DA

$\max\ \mathbf{cov}(PLS_X, PLS_Y)$



Generalized PLS, PLS-DA

$$\begin{aligned}\max \{ & c12.\mathbf{cov}(PLS_{X1}, PLS_{X2}) + \\ & c13.\mathbf{cov}(PLS_{X1}, PLS_{X3}) + \\ & c14.\mathbf{cov}(PLS_{X1}, PLS_{X4}) + \\ & c23.\mathbf{cov}(PLS_{X2}, PLS_{X3}) + \\ & \ldots\}\end{aligned}$$



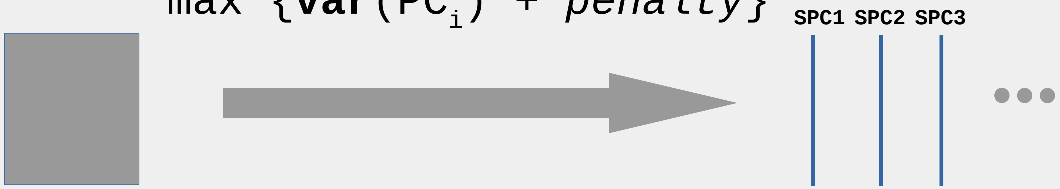
cij can be set by the user through a design matrix

Sparsity

- High throughput experiments: too many variables, noisy or irrelevant depending on the goal aimed
- Some of the variable loadings, among the smallests, are set to 0 thanks to a LASSO (L_1) penalty
- Associated variables are not taken into account when calculating the PCs

Sparse PCA

$\max \{\mathbf{var}(PC_i) + \mathit{penalty}\}$

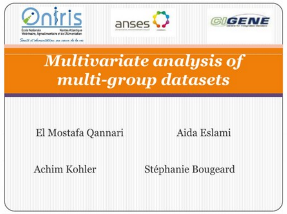


```
SPC1 =    0.X1 + a12.X2 + a13.X3 + … +    0.Xp
SCP2 = a21.X1 +    0.X2 +    0.X3 + … + a2p.Xp
...
```

Vertical integration

A. Eslami, E.M Qannari, A. Kohler, S. Bougeard (2013). Analyses factorielles de données structurées en groupes d'individus. Journal de la SfdS, vol. 154(3). journal-sfds.fr/article/view/208

www.rocq.inria.fr/axis/modulad///sda11/HCSDA11-Qannari.PDF



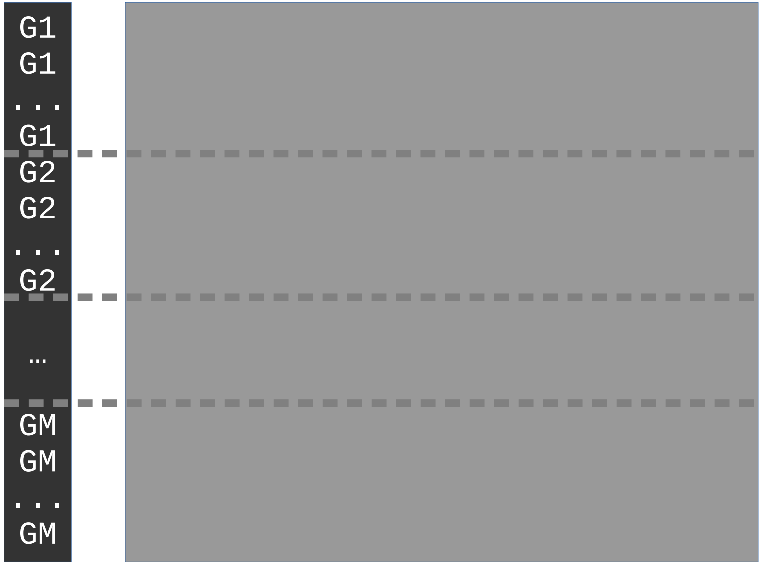
Multivariate analysis of multi-group datasets

El Mostafa Qannari Aida Eslami

Achim Kohler Stéphanie Bougeard

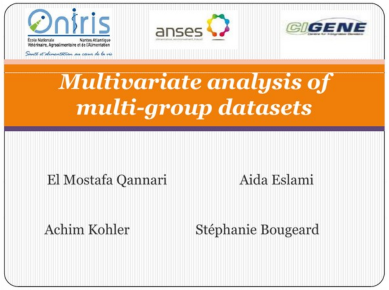
- *Setting: the same variables measured on individuals portioned into several groups*

- *The same setting as in discriminant analysis* ***but*** *the main aim herein is to investigate the relationships among individuals within the various groups*



Ask the right question!

Vertical integration



Multivariate analysis of multi-group datasets

El Mostafa Qannari Aida Eslami

Achim Kohler Stéphanie Bougeard

How to investigate the relationships among individuals within the various groups?

- **Perform PCA on each group separately**

→ Too many parameters (stability and interpretation problems)

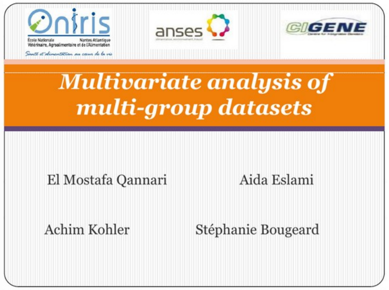
- **Perform PCA on the concatenated dataset**

→ The total variance recovered by the principal components mix up both the between and within groups variances

- **Multi-group PCA**

→ Perform PCA on the concatenated dataset **after centering by group**

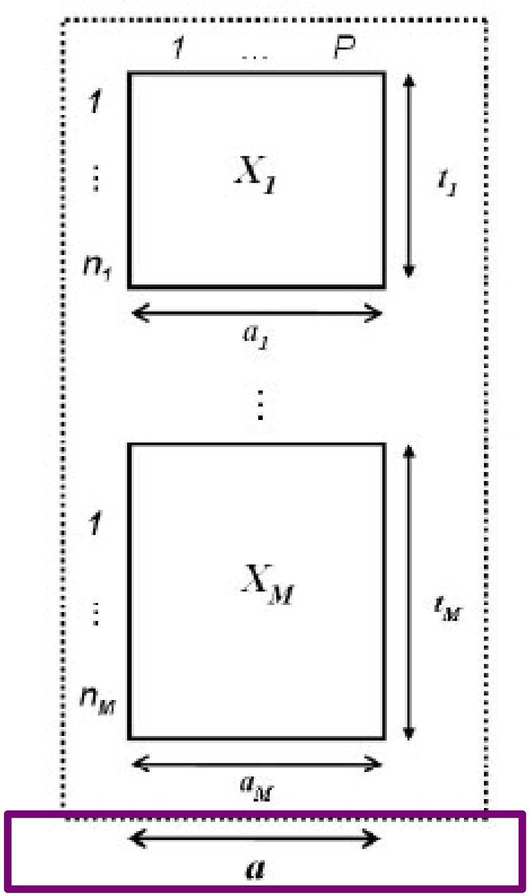
Vertical integration: mgPCA



Multivariate analysis of multi-group datasets

El Mostafa Qannari Aida Eslami

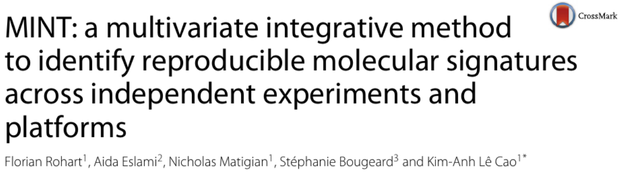
Achim Kohler Stéphanie Bougeard



a : vector of common loadings

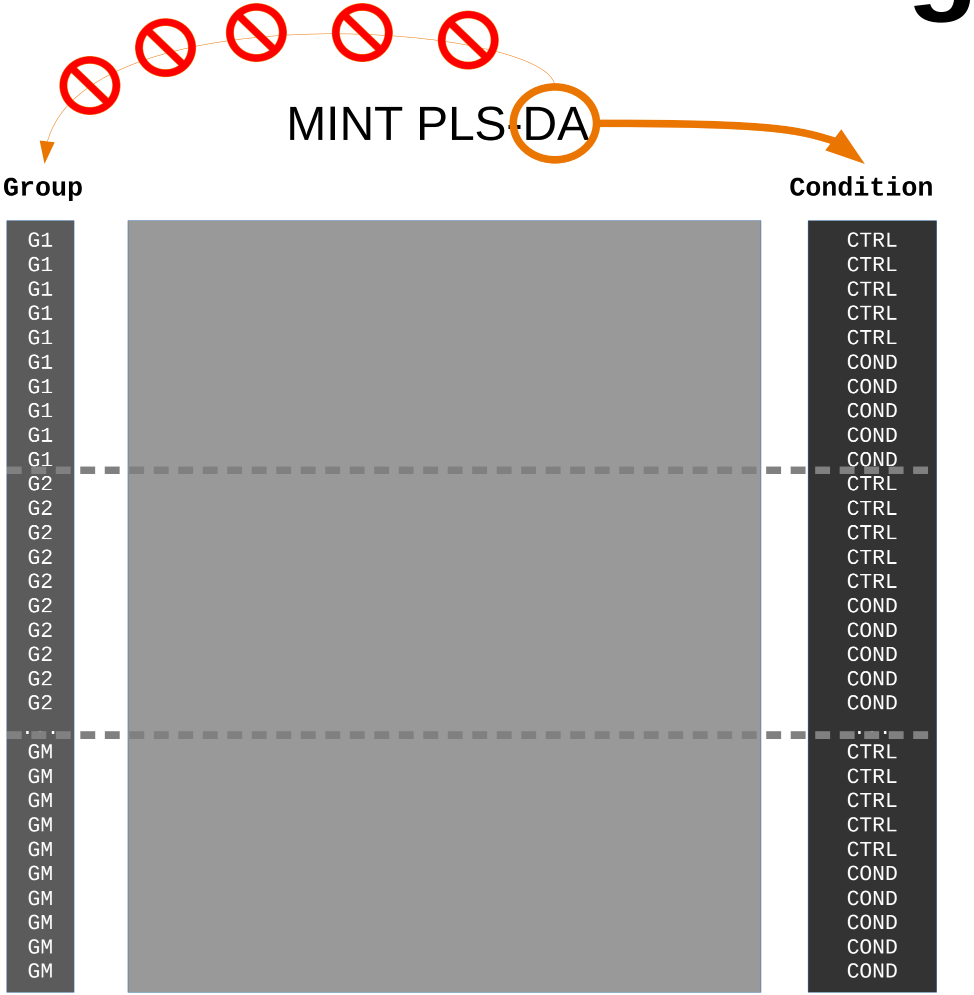
→ the same variable plot for every group

Vertical data integration



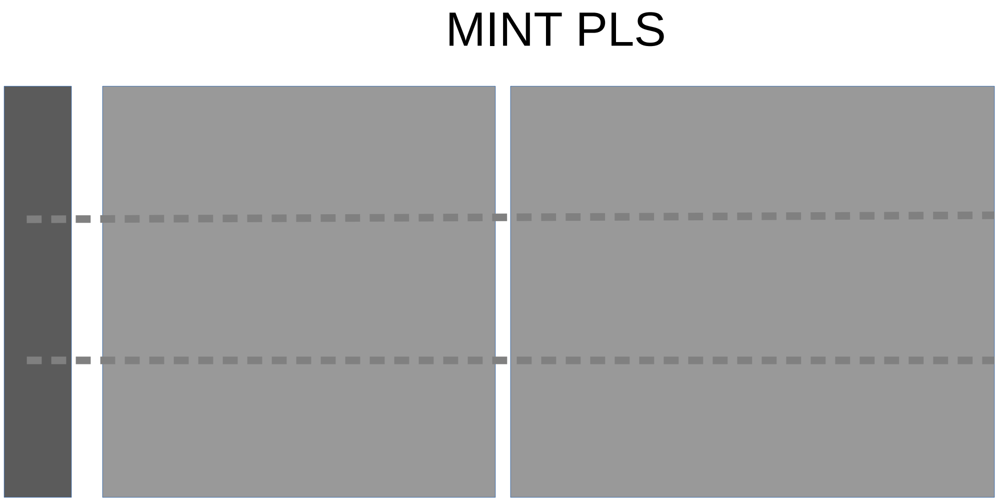
MINT: a multivariate integrative method to identify reproducible molecular signatures across independent experiments and platforms

Florian Rohart[1], Aida Eslami[2], Nicholas Matigian[1], Stéphanie Bougeard[3] and Kim-Anh Lê Cao[1*]

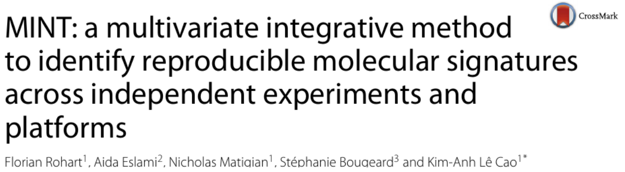


Group	Condition
G1	CTRL
G1	CTRL
G1	CTRL
G1	CTRL
G1	CTRL
G1	COND
G1	COND
G1	COND
G1	COND
G1	COND
G2	CTRL
G2	CTRL
G2	CTRL
G2	CTRL
G2	CTRL
G2	COND
G2	COND
G2	COND
G2	COND
G2	COND
...	...
GM	CTRL
GM	CTRL
GM	CTRL
GM	CTRL
GM	CTRL
GM	COND
GM	COND
GM	COND
GM	COND
GM	COND

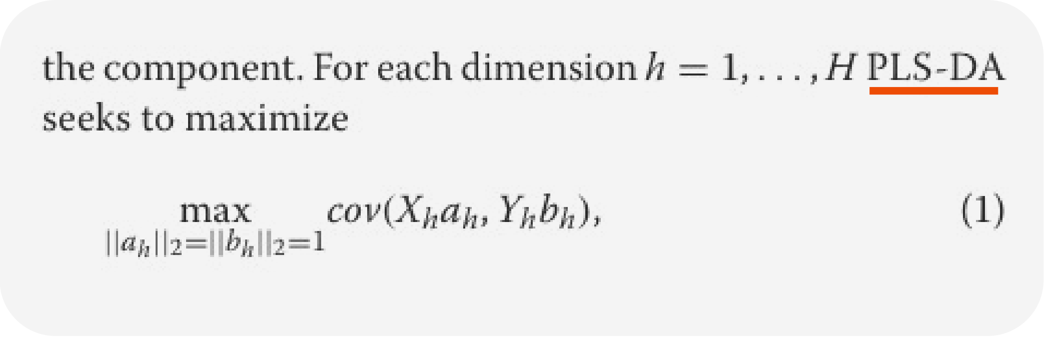
While PLS-DA ignores the data group structure inherent to each independent study, it can give satisfactory results when the between groups variance is smaller than the within group variance.



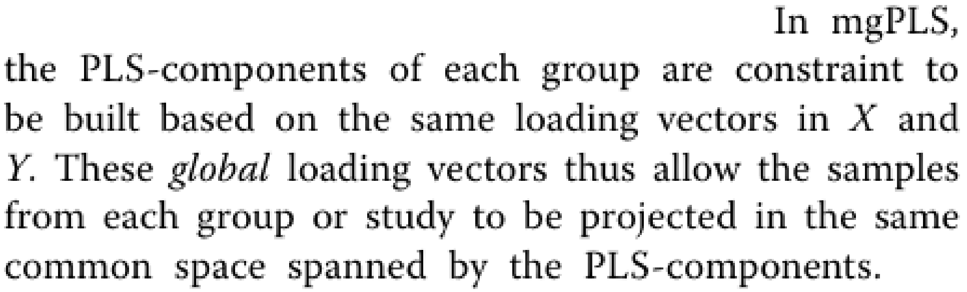
Vertical data integration

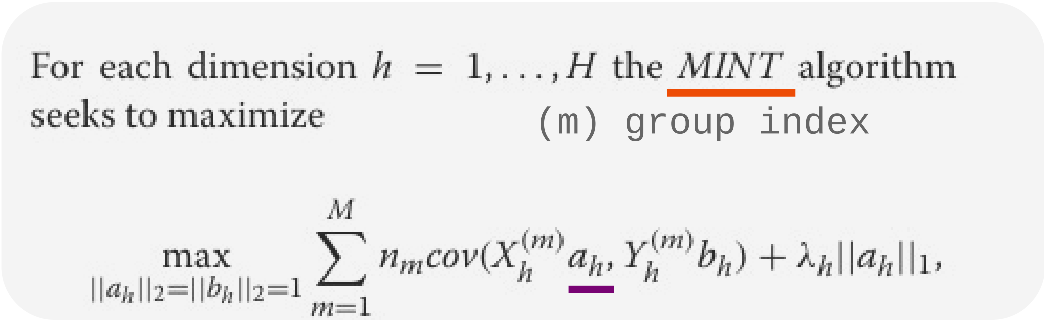

MINT: a multivariate integrative method to identify reproducible molecular signatures across independent experiments and platforms

Florian Rohart[1], Aida Eslami[2], Nicholas Matigian[1], Stéphanie Bougeard[3] and Kim-Anh Lê Cao[1*]


the component. For each dimension $h = 1, \ldots, H$ PLS-DA seeks to maximize

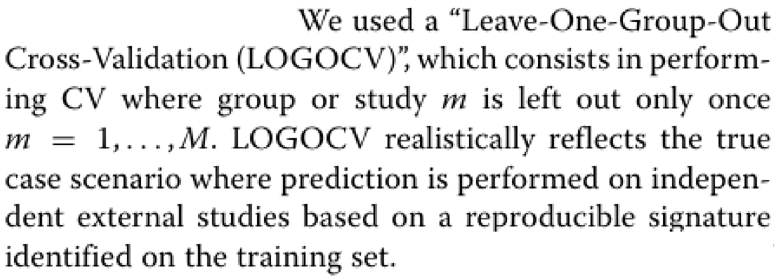
$$\max_{||a_h||_2=||b_h||_2=1} cov(X_h a_h, Y_h b_h), \qquad (1)$$


In mgPLS, the PLS-components of each group are constraint to be built based on the same loading vectors in X and Y. These *global* loading vectors thus allow the samples from each group or study to be projected in the same common space spanned by the PLS-components.


For each dimension $h = 1, \ldots, H$ the *MINT* algorithm seeks to maximize

(m) group index

$$\max_{||a_h||_2=||b_h||_2=1} \sum_{m=1}^{M} n_m cov(X_h^{(m)} \underline{a_h}, Y_h^{(m)} b_h) + \lambda_h ||a_h||_1,$$


We used a "Leave-One-Group-Out Cross-Validation (LOGOCV)", which consists in performing CV where group or study m is left out only once $m = 1, \ldots, M$. LOGOCV realistically reflects the true case scenario where prediction is performed on independent external studies based on a reproducible signature identified on the training set.

a_h: vector of common loadings

Examples

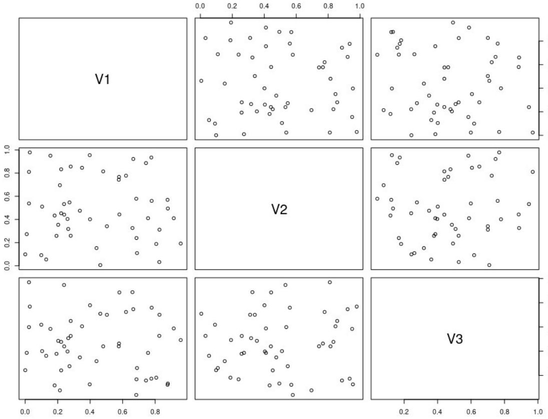
- Simulated toy examples for PCA, PLS-DA, PLS (2 blocks), multi-block PLS-DA

- `liver.toxicity` dataset (included in the package): practical session

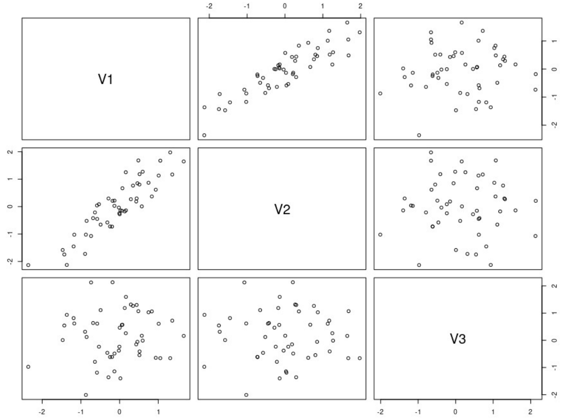
PCA: simulated examples

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

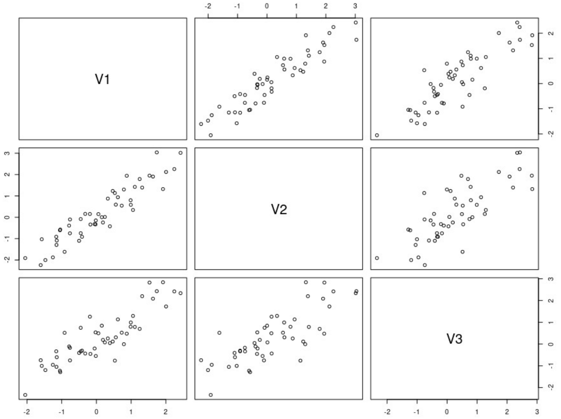
Case 1)
{V1} - {V2} - {V3}



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



Case 3)
{V1 - V2 - V3}



Pearson Correlation matrices

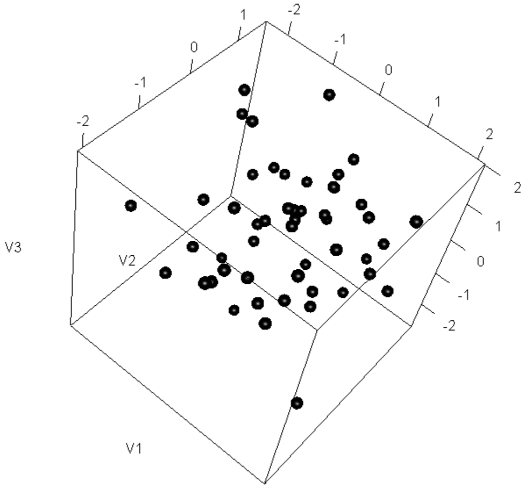
```
1)   V1    V2    V3
V1  1.00 -0.05 -0.12
V2 -0.05  1.00  0.06
V3 -0.12  0.06  1.00
```

```
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
```

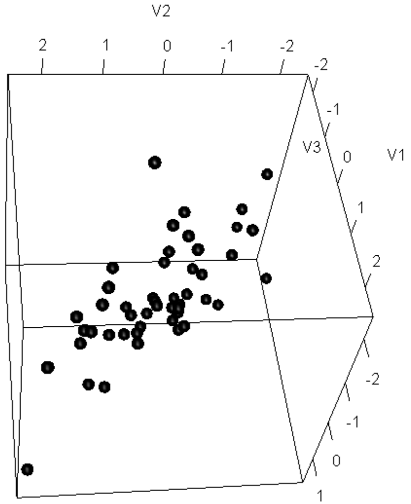
```
3)   V1   V2   V3
V1  1.00 0.93 0.87
V2  0.93 1.00 0.79
V3  0.87 0.79 1.00
```

PCA: simulated examples

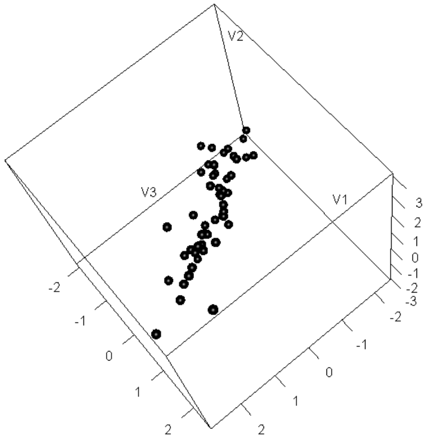
Case 1)



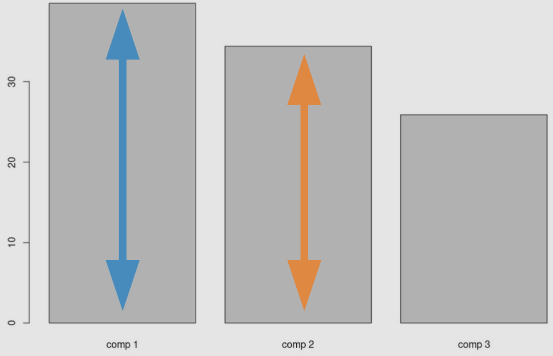
Case 2)



Case 3)



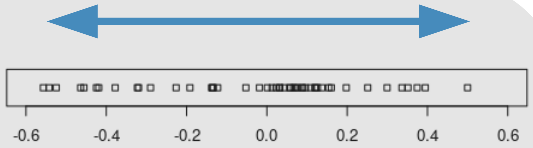
PCA: simulated examples

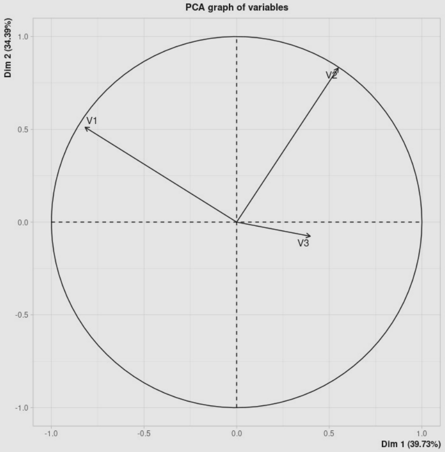


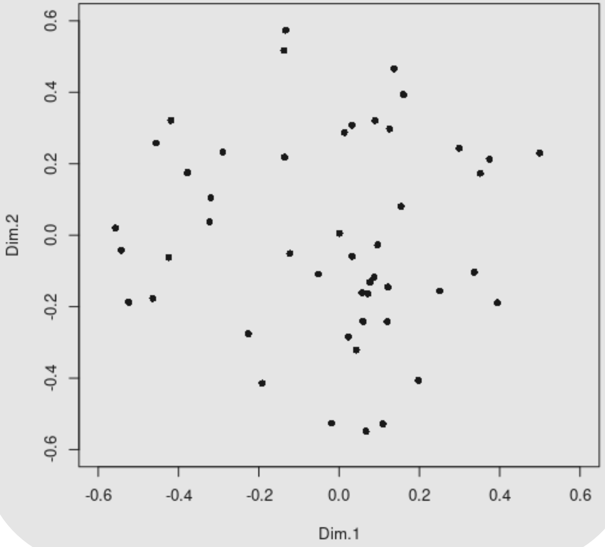
39.7% 34.4% 25.9%

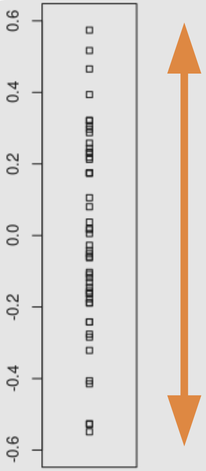
Loadings

```
    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
```





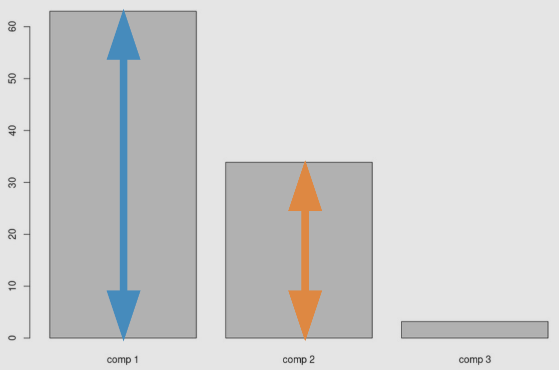




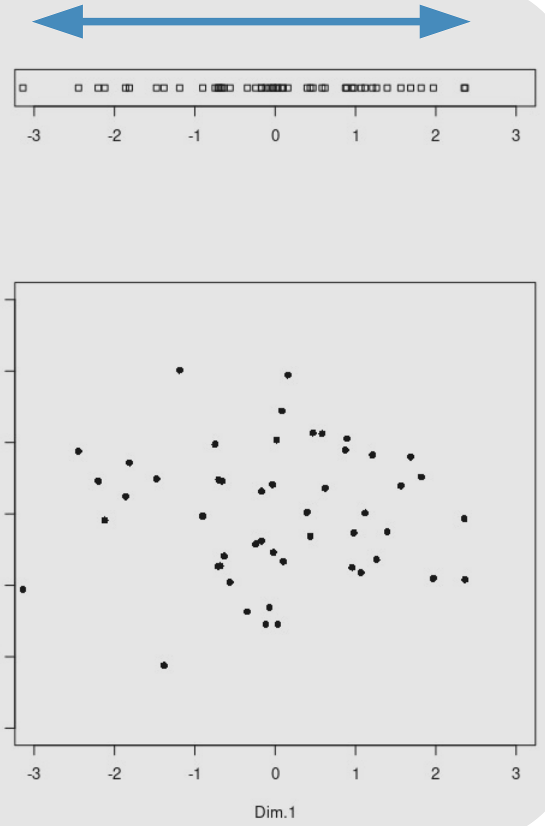
PCA: simulated examples

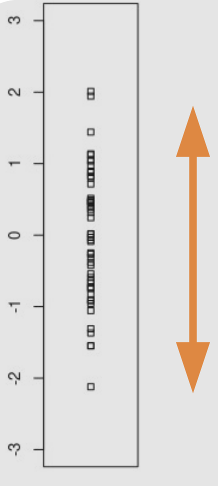
Loadings

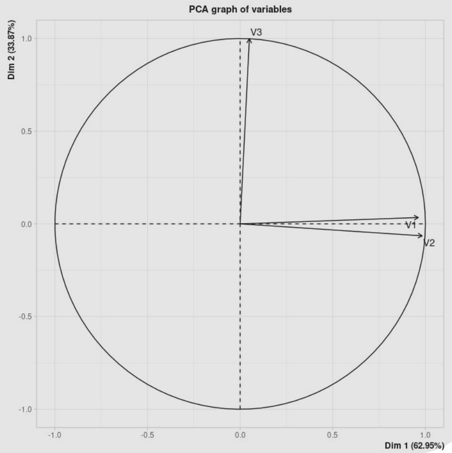
```
   Dim.1 Dim.2 Dim.3
V1  0.77  0.03  0.22
V2  0.97 -0.06 -0.17
V3  0.05  0.91 -0.02
```



62.9% 33.9% 3.2%

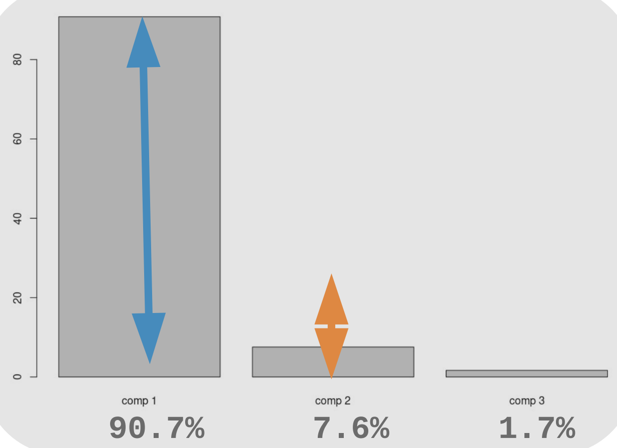


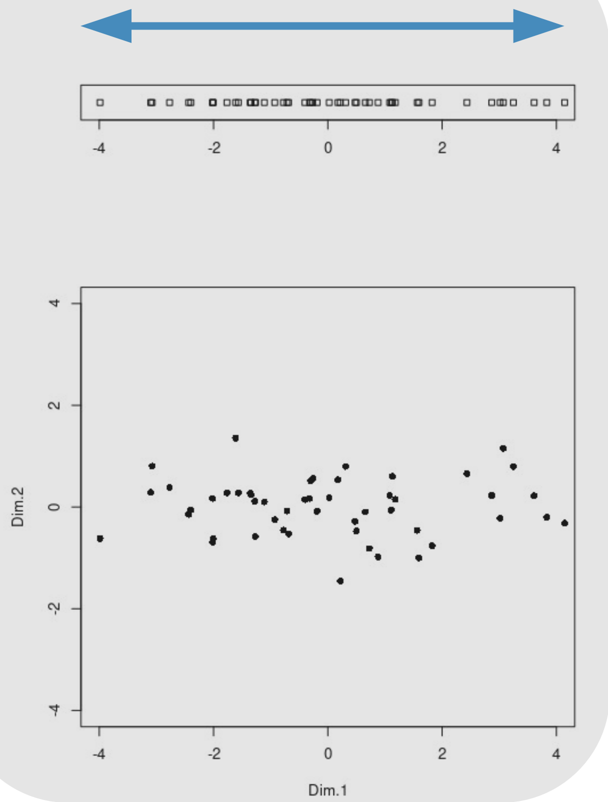


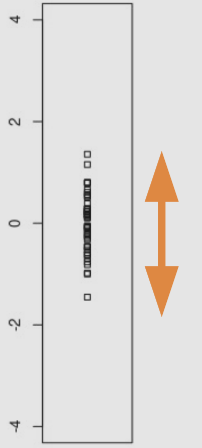


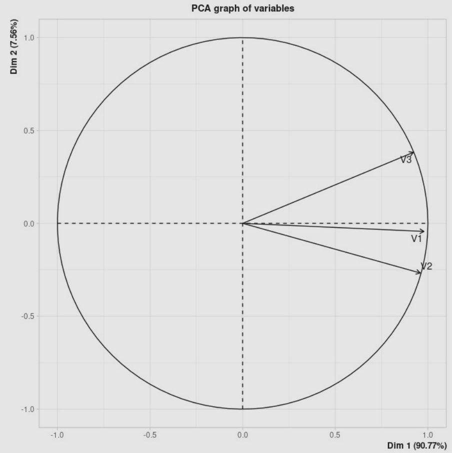
PCA: simulated examples

```
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
```









Discriminant analysis

Explore a data set composed of numerical variables and one categorical variable in order to separate the individuals based on their membership to the levels of the categorical variable

- Linear Discriminant Analysis (LDA): standard method (*underneath, same as PCA: matrix algebra*), needs more individuals than variables
- Projection to Latent Structure - Discriminant Analysis (PLS-DA): (*underneath, PLS algorithm*)

DA: simulated example

- 50 observations
- 4 variables :
  - 3 numerical: V1, V2, V3
  - 1 categorical: Group (A / B)

- Univariate analysis

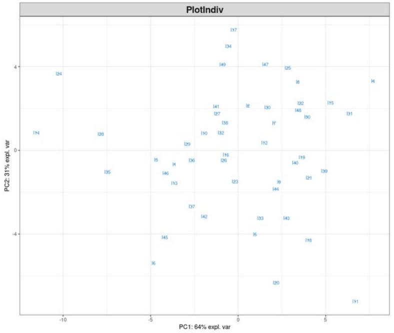
```
            V1      V2      V3

Mean      0.502   0.351 -0.076

Variance 17.28    8.44   1.49
```

- Bivariate analysis (Pearson correlation)

```
       V1     V2    V3
V1  1.000 -0.039 0.079
V2 -0.039  1.000 0.043
V3  0.079  0.043 1.000
```

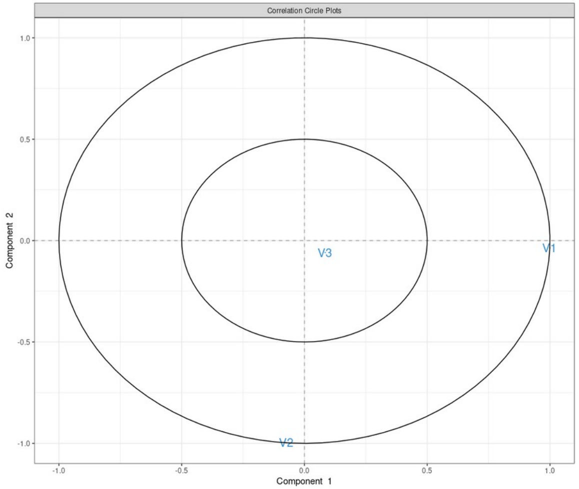
```
     V1    V2    V3 Group
   2.88  1.71  1.19     A
  -0.78 -1.33 -1.04     B
  -0.82 -1.63 -1.96     B
   5.62 -2.23  0.90     A
  -0.30 -0.91 -1.26     B
  -4.10  5.94  1.88     A
   7.98 -3.39  1.08     A
  -6.89  1.78 -0.85     B
   3.48  3.48 -1.83     B
  -3.11  2.07  1.72     A
   2.44 -1.10  1.36     A
   3.10 -3.76  0.95     A
   6.79 -1.71 -0.97     B
  -1.97  3.19 -1.15     B
  -7.35  0.00 -1.02     B
  -2.07  1.00 -1.77     B
   2.97  6.52  0.91     A
  -2.39  0.22 -1.34     B
  -3.54  1.68 -1.38     B
  -1.53 -0.41  1.84     A
  -1.27  3.62 -0.41     B
   3.69 -3.13  1.46     A
  -3.13  1.19  0.69     A
 -11.07  0.08  0.67     A
   1.95 -0.12  0.77     A
  -0.22  0.57  0.80     A
  -0.28  0.87 -0.64     B
  -0.51 -0.41 -1.29     B
   0.92 -1.84  1.02     A
  -0.08 -5.41  0.84     A
   2.09 -1.77 -1.16     B
   2.78  2.10 -1.23     B
   1.65  4.30  0.67     A
  -9.95 -2.80  0.91     A
   1.82 -3.83  0.04     B
  -0.56 -3.67 -1.64     B
  -0.27 -4.57 -1.76     B
   3.91 -2.13  1.67     A
   5.50  1.11 -1.33     B
   1.94  3.53 -0.73     B
   4.72  4.40  0.86     A
   7.56  7.20  0.68     A
   3.84 -1.72 -0.99     B
   3.82  0.80 -1.03     B
  -4.18  1.02  0.54     A
  -3.44  4.76 -1.56     B
   4.41 -1.42 -1.82     B
   4.59  1.43  0.75     A
   4.11  0.46  1.25     A
   0.37  1.83  0.89     A
```

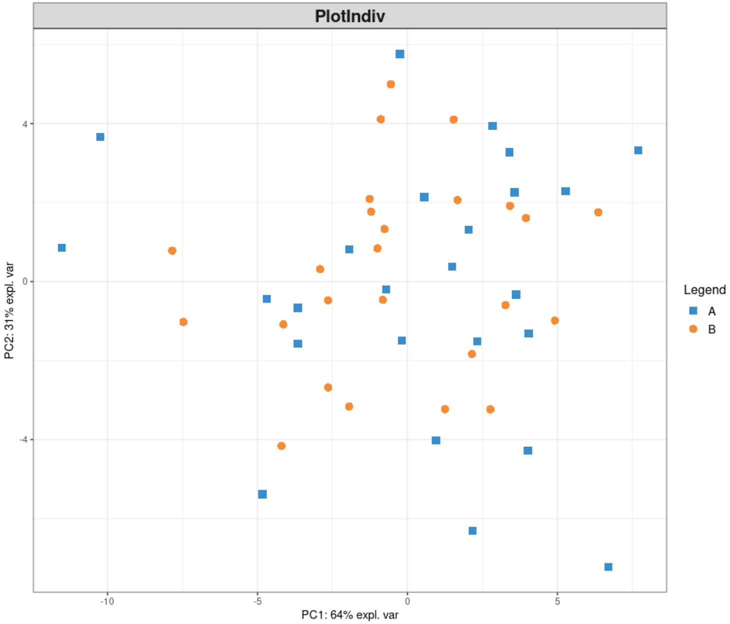
DA: simulated example



- Multivariate **unsupervised** analysis (PCA)

```
Loadings   PC1     PC2     PC3
V1        0.998 -0.052  0.026
V2       -0.053 -0.998  0.023
V3        0.025 -0.025 -0.999
```



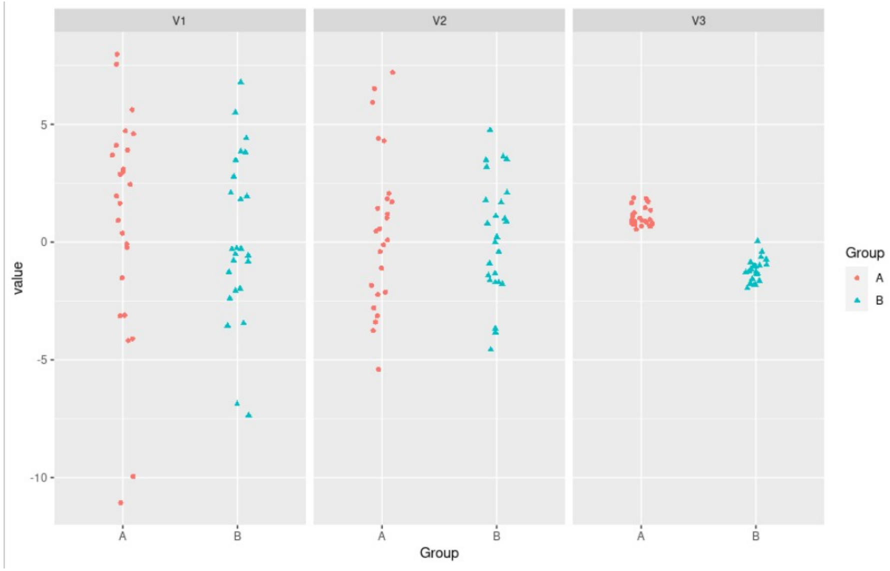


DA: simulated example

- Linear Discriminant Analysis

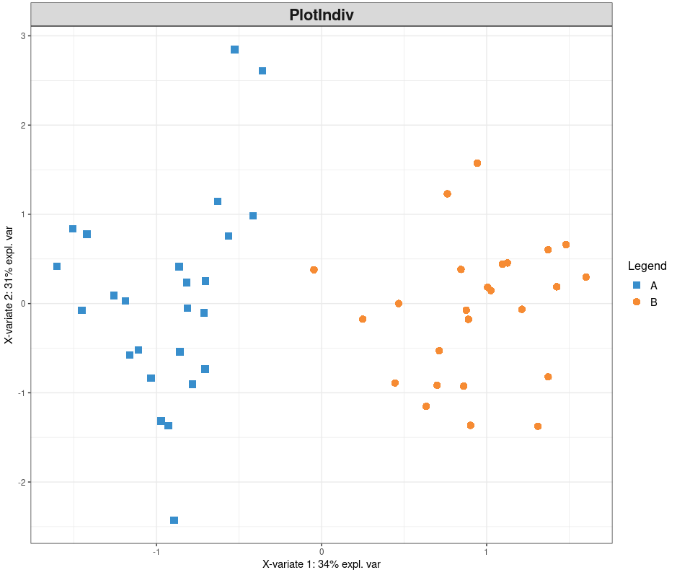
Loadings	LD1
V1	-0.007
V2	-0.011
V3	**-2.295**

V3 is highly involved in the discrimination of the two groups (even with a small variance)



- PLS Discriminant Analysis

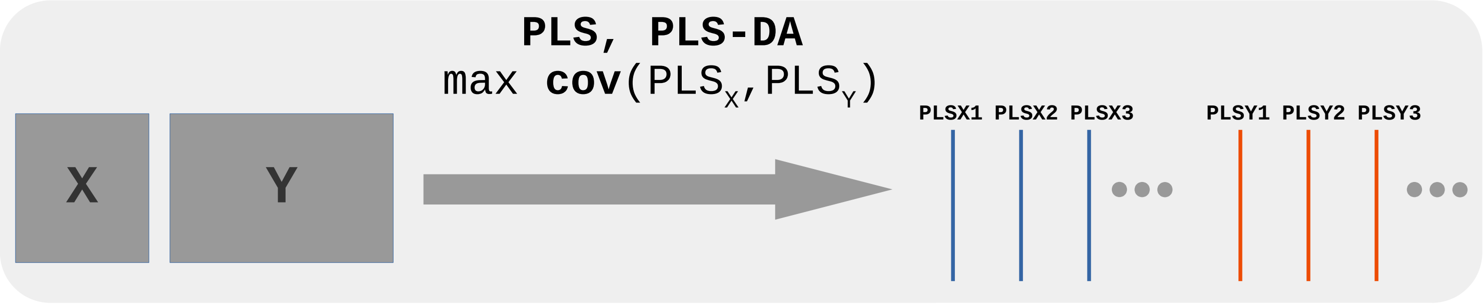
Loadings	PLSDA1	PLSDA2	PLSDA3
V1	-0.088	-0.888	-0.452
V2	-0.053	-0.449	0.892
V3	**-0.995**	0.103	-0.008



Two-block integration

Unravel the relationships between two types of variables measured on the same matching samples

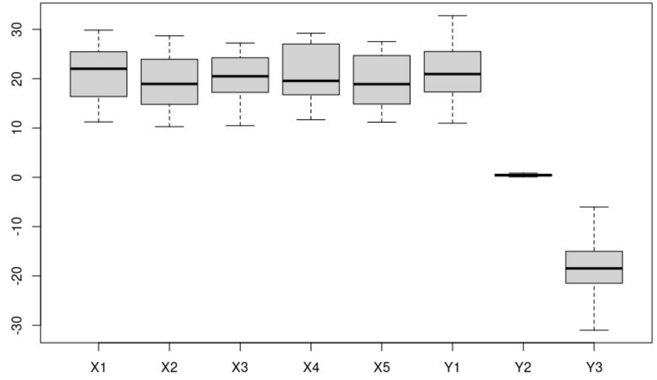
- Understand the correlation/covariance structure between two data sets
- Select co-regulated biological entities across samples
- Methods: Projection to Latent Structures (**PLS, maximize covariance**)



Two-block: simulated example

- 20 observations
- 2 sets of numerical variables:
  - X: 5 variables
  - Y: 3 variables

- Univariate analysis

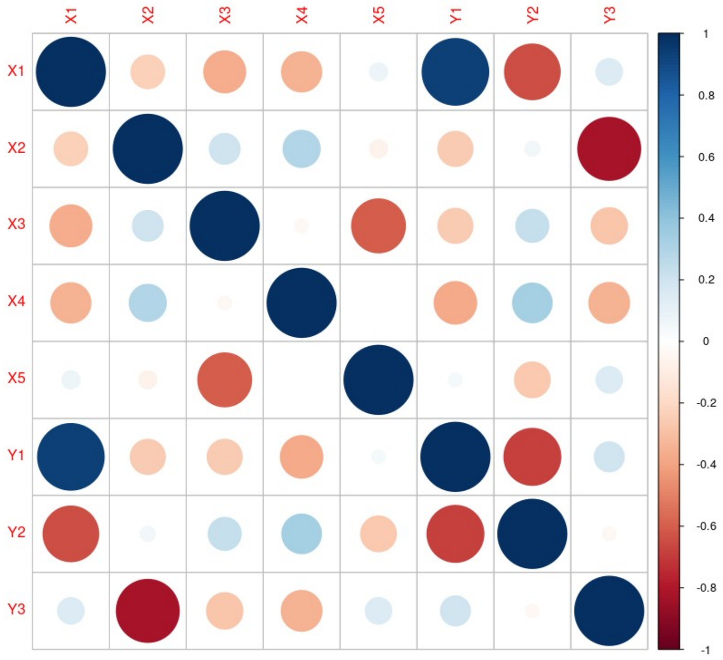


```
 X1   X2   X3   X4   X5           Y1   Y2    Y3
15.3 28.7 26.4 28.3 18.7         16.1  0.7 -31.0
17.4 14.2 22.9 15.9 24.3         16.2  0.6 -14.8
21.5 23.0 25.7 19.2 18.0         22.1  0.2 -18.3
28.2 12.5 21.1 16.6 16.5         25.9  0.3 -18.6
14.0 15.3 20.6 23.0 25.1         16.9  0.7 -13.0
28.0 17.7 25.8 15.2 14.1         31.9  0.5 -16.4
28.9 10.3 10.5 19.6 24.2         28.2  0.2  -6.0
23.2 17.6 19.5 25.3 12.4         21.1  0.7 -18.9
22.6 27.4 24.6 11.7 14.9         23.7  0.1 -25.9
11.2 16.8 23.9 27.5 12.9         11.0  0.9 -15.7
14.1 19.6 19.6 16.8 14.8         18.9  0.6 -21.8
13.5 22.0 27.2 26.8 11.2         13.5  0.6 -17.2
23.7 19.9 18.8 16.9 22.8         25.1  0.3 -15.2
17.7 13.7 14.9 16.7 27.5         17.7  0.5 -10.9
25.4 26.5 11.4 19.5 25.6         23.9  0.5 -20.2
20.0 23.4 12.0 27.8 25.9         20.3  0.2 -21.1
24.4 25.9 16.3 27.3 19.1         20.7  0.5 -31.0
29.8 12.2 20.4 17.8 18.2         32.8  0.1 -14.5
17.6 24.5 23.2 25.5 26.2         17.9  0.3 -29.4
25.5 18.2 18.1 29.2 22.1         29.9  0.2 -20.1
```

Two-block: simulated example

- Bivariate analysis

```
      X1    X2    X3    X4    X5    Y1    Y2    Y3
X1  1.00 -0.24 -0.37 -0.34  0.07  0.93 -0.65  0.15
X2 -0.24  1.00  0.20  0.29 -0.07 -0.26  0.05 -0.83
X3 -0.37  0.20  1.00 -0.04 -0.61 -0.26  0.23 -0.28
X4 -0.34  0.29 -0.04  1.00  0.00 -0.38  0.33 -0.35
X5  0.07 -0.07 -0.61  0.00  1.00  0.04 -0.27  0.15
Y1  0.93 -0.26 -0.26 -0.38  0.04  1.00 -0.68  0.19
Y2 -0.65  0.05  0.23  0.33 -0.27 -0.68  1.00 -0.04
Y3  0.15 -0.83 -0.28 -0.35  0.15  0.19 -0.04  1.00
```



Main effects: (X1, Y1) positively correlated, (X1, Y2) and (X2, Y3) negatively correlated

Two-block: simulated example

Projection to Latent Structure

Loadings

(2 sets, one for each dataset)

```
    CCA.X1 CCA.X2  CCA.X3
X1 -0.73   -0.58   -0.27
X2  0.42   -0.80    0.35
X3  0.30   -0.13   -0.49
X4  0.42   -0.08   -0.17
X5 -0.17    0.06    0.73


    CCA.Y1 CCA.Y2  CCA.Y3
Y1 -0.71   -0.48   -0.88
Y2  0.51    0.45   -0.24
Y3 -0.49    0.75   -0.41
```

What is underneath?
An iterative algorithm

Two-block: simulated example

```
PLS.X1 PLS.X2 PLS.X3     PLS.Y1 PLS.Y2 PLS.Y3
  2.38  -0.78   0.21       2.11  -0.85  -0.07
 -0.33   1.15   0.33       0.70   1.31   0.55
  0.44  -0.64  -0.44      -0.52  -0.40  -0.18
 -1.64   0.13  -1.39      -0.90  -0.35   0.49
  0.57   1.39   0.81       0.76   1.57  -0.32
 -0.97  -0.66  -1.71      -1.35  -0.31  -1.12
 -2.55   0.61   0.32      -2.29   0.92  -0.02
  0.09  -0.06  -1.28       0.75   0.66  -0.30
  0.10  -1.32  -0.36      -0.42  -1.61   0.42
  1.96   1.33  -0.93       1.97   1.75  -0.23
  0.67   0.78  -0.05       0.94   0.14   0.00
  2.26   0.28  -1.21       1.11   0.74  -0.37
 -0.82  -0.27   0.49      -0.91   0.11  -0.15
 -0.92   1.37   1.41      -0.09   1.45   0.58
 -0.85  -1.23   1.83      -0.02  -0.03   0.19
  0.26  -0.23   1.78      -0.22  -0.66   0.43
  0.33  -1.22   0.39       1.24  -1.17   0.61
 -1.88   0.01  -1.26      -2.42  -0.73  -0.45
  1.13  -0.33   1.10       0.90  -1.38   0.70
 -0.25  -0.33  -0.05      -1.35  -1.16  -0.77
```

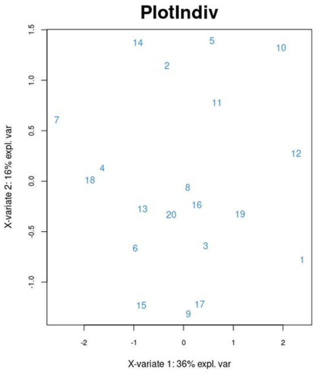
`cor(PLS.X1, PLS.Y1) =  0.86`

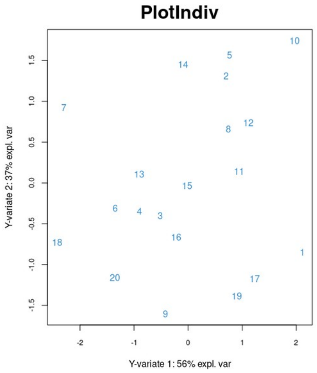
`cov``(PLS.X1, PLS.Y1)` **`=  1.46`**

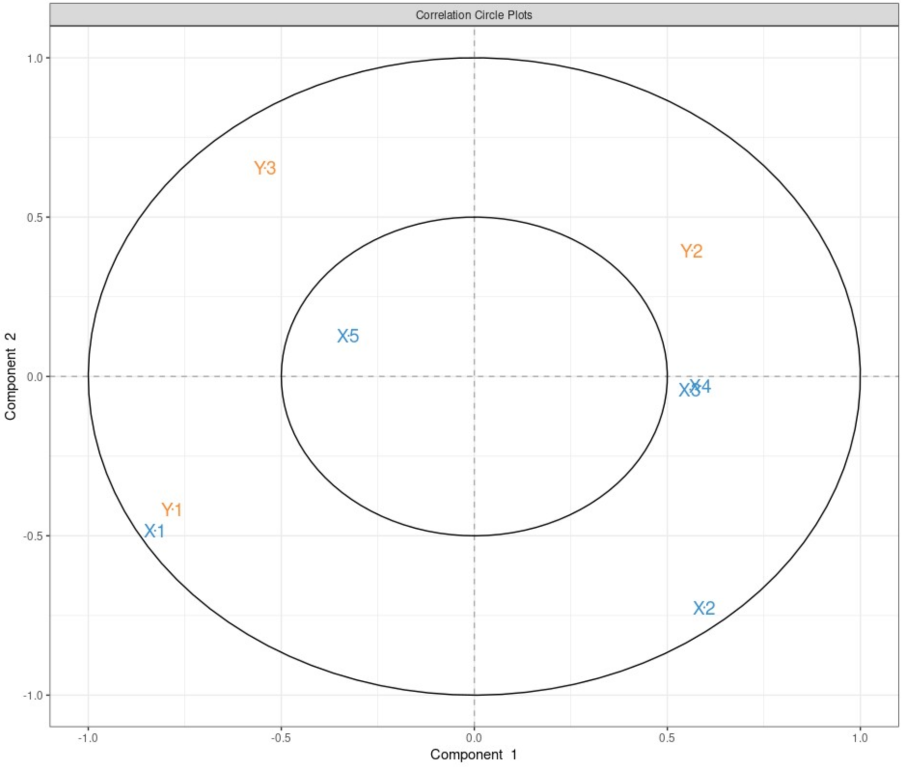
Two-block: simulated example

```
PLS.X1 PLS.X2 PLS.X3
 2.38  -0.78   0.21
-0.33   1.15   0.33
 0.44  -0.64  -0.44
-1.64   0.13  -1.39
 0.57   1.39   0.81
-0.97  -0.66  -1.71
-2.55   0.61   0.32
 0.09  -0.06  -1.28
 0.10  -1.32  -0.36
 1.96   1.33  -0.93
 0.67   0.78  -0.05
 2.26   0.28  -1.21
-0.82  -0.27   0.49
-0.92   1.37   1.41
-0.85  -1.23   1.83
 0.26  -0.23   1.78
 0.33  -1.22   0.39
-1.88   0.01  -1.26
 1.13  -0.33   1.10
-0.25  -0.33  -0.05
```

```
PLS.Y1 PLS.Y2 PLS.Y3
 2.11  -0.85  -0.07
 0.70   1.31   0.55
-0.52  -0.40  -0.18
-0.90  -0.35   0.49
 0.76   1.57  -0.32
-1.35  -0.31  -1.12
-2.29   0.92  -0.02
 0.75   0.66  -0.30
-0.42  -1.61   0.42
 1.97   1.75  -0.23
 0.94   0.14   0.00
 1.11   0.74  -0.37
-0.91   0.11  -0.15
-0.09   1.45   0.58
-0.02  -0.03   0.19
-0.22  -0.66   0.43
 1.24  -1.17   0.61
-2.42  -0.73  -0.45
 0.90  -1.38   0.70
-1.35  -1.16  -0.77
```





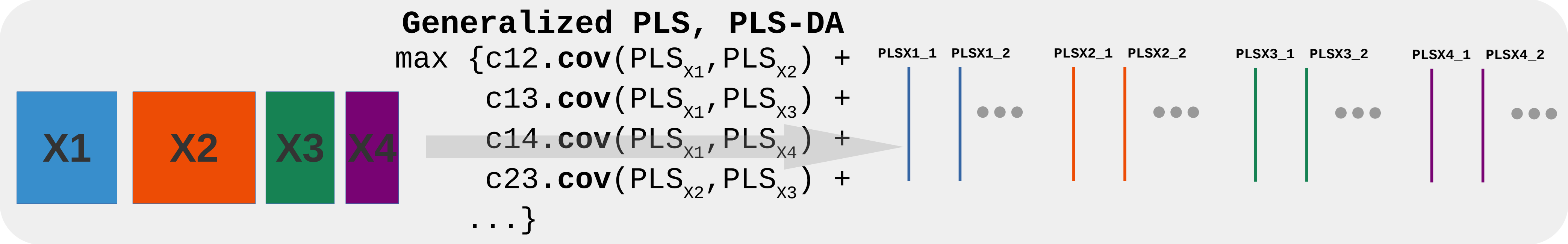


Variable plot clearly highlights the correlation structure between the two datasets: (X1, Y1) positively correlated, (X1, Y2) and (X2, Y3) negatively correlated

Multi-block integration

Unravel the relationships between **more than two** types of variables measured on the same matching samples

- Understand the relationships structure between several data sets
- Select co-regulated biological entities across samples
- Method: Multi-block PLS



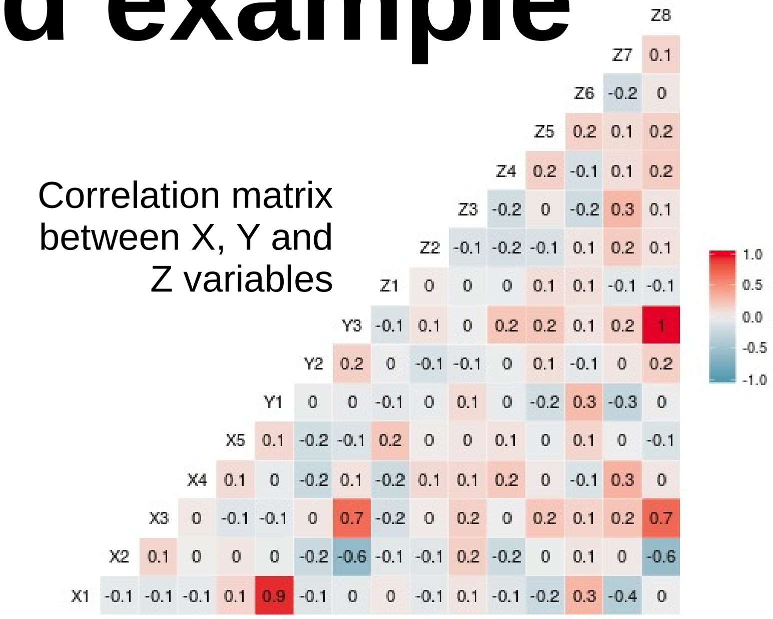
Generalized PLS, PLS-DA

$$\begin{aligned}\max \{ & c12.\mathbf{cov}(PLS_{X1},PLS_{X2}) + \\ & c13.\mathbf{cov}(PLS_{X1},PLS_{X3}) + \\ & c14.\mathbf{cov}(PLS_{X1},PLS_{X4}) + \\ & c23.\mathbf{cov}(PLS_{X2},PLS_{X3}) + \\ & ...\}\end{aligned}$$

Multi-block: simulated example

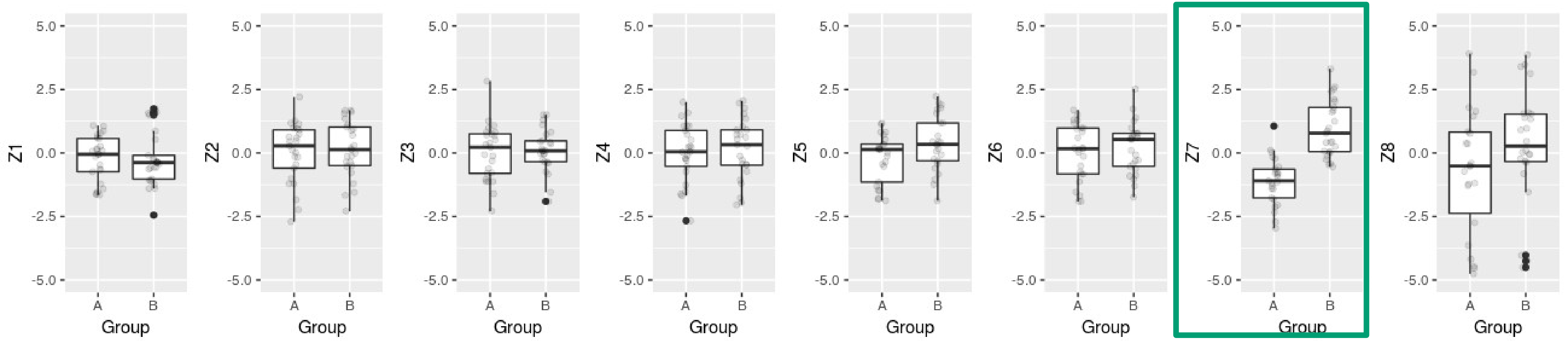
- 50 observations
- 3 sets of numerical variables:
  - X: 5 variables
  - Y: 3 variables
  - Z: 8 variables
- 1 categorical variable Group A/B

Correlation matrix between X, Y and Z variables



	X2	X3	X4	X5	Y1	Y2	Y3	Z1	Z2	Z3	Z4	Z5	Z6	Z7	Z8
X1	-0.1	-0.1	-0.1	0.1	0.9	-0.1	0	0	-0.1	0.1	-0.1	-0.2	0.3	-0.4	0
X2		0.1	0	0	0	-0.2	-0.6	-0.1	-0.1	0.2	-0.2	0	0.1	0	-0.6
X3			0	-0.1	-0.1	0	0.7	-0.2	0	0.2	0	0.2	0.1	0.2	0.7
X4				0.1	0	-0.2	0.1	-0.2	0.1	0.1	0.2	0	-0.1	0.3	0
X5					0.1	-0.2	-0.1	0.2	0	0	0.1	0	0.1	0	-0.1
Y1						0	0	-0.1	0	0.1	0	-0.2	0.3	-0.3	0
Y2							0.2	0	-0.1	-0.1	0	0.1	-0.1	0	0.2
Y3								-0.1	0.1	0	0.2	0.2	0.1	0.2	1
Z1									0	0	0	0.1	0.1	-0.1	-0.1
Z2										-0.1	-0.2	-0.1	0.1	0.2	0.1
Z3											-0.2	0	-0.2	0.3	0.1
Z4												0.2	-0.1	0.1	0.2
Z5													0.2	0.1	0.2
Z6														-0.2	0
Z7															0.1

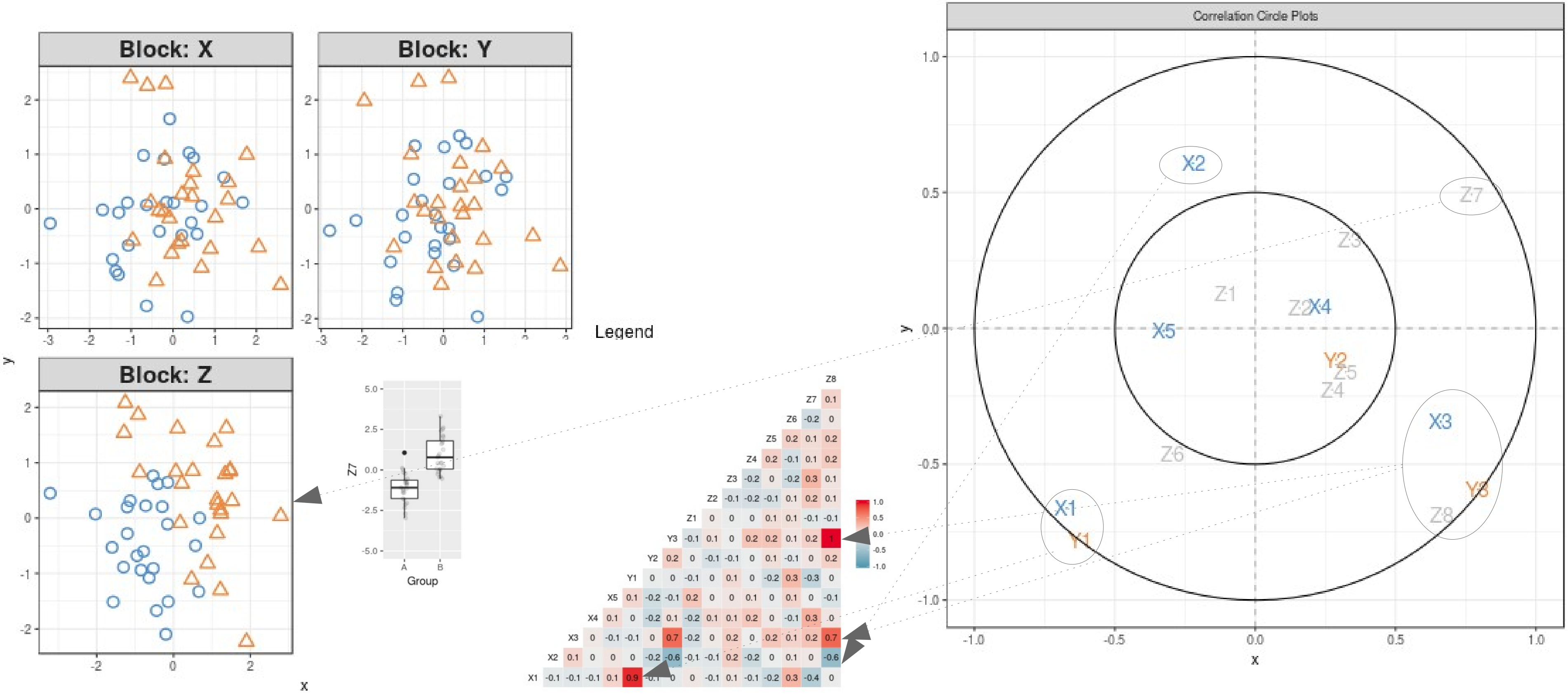
Boxplot of the variables Z according to the group



Multi-block: simulated example

Three individual plots

Variable plot



															Z8
														Z7	0.1
													Z6	-0.2	0
												Z5	0.2	0.1	0.2
											Z4	0.2	-0.1	0.1	0.2
										Z3	-0.2	0	-0.2	0.3	0.1
									Z2	-0.1	-0.2	-0.1	0.1	0.2	0.1
								Z1	0	0	0	0.1	0.1	-0.1	-0.1
							Y3	-0.1	0.1	0	0.2	0.2	0.1	0.2	1
						Y2	0.2	0	-0.1	-0.1	0	0.1	-0.1	0	0.2
					Y1	0	0	-0.1	0	0.1	0	-0.2	0.3	-0.3	0
				X5	0.1	-0.2	-0.1	0.2	0	0	0.1	0	0.1	0	-0.1
			X4	0.1	0	-0.2	0.1	-0.2	0.1	0.1	0.2	0	-0.1	0.3	0
		X3	0	-0.1	-0.1	0	0.7	-0.2	0	0.2	0	0.2	0.1	0.2	0.7
	X2	0.1	0	0	0	-0.2	-0.6	-0.1	-0.1	0.2	-0.2	0	0.1	0	-0.6
X1	-0.1	-0.1	-0.1	0.1	0.9	-0.1	0	0	-0.1	0.1	-0.1	-0.2	0.3	-0.4	0

Take home message

- Practice on your own data! The best way to understand what a method has to tell you.
- Do not bypass the elementary analyses (univariate, bivariate, multivariate single data set).
- Address problems explicitly formulated: ‘‘I want to integrate my data’’ is not a problem explicitly formulated.
- Clearly identify supervised and unsupervised questions and the methods to use. ‘‘PCA is not a good method, I can’t see my clusters...’’ reveals a misunderstanding of PCA.

What about microbiome data?

Lê Cao KA, Costello ME, Lakis VA, Bartolo F, Chua XY, et al. (2016)
MixMC: A Multivariate Statistical Framework to Gain Insight into Microbial Communities.
PLOS ONE 11(8): e0160169. https://doi.org/10.1371/journal.pone.0160169

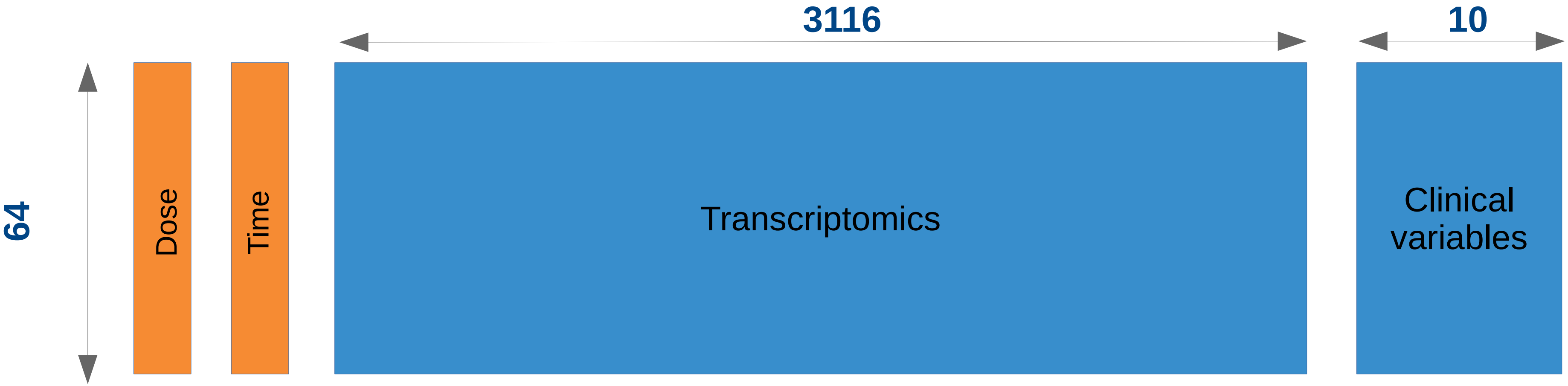
- mixMC: pipeline set up for microbial communities, using some of the standards methods in mixOmics but with a bit of tweaking
- due to the sparse and compositional nature (represent proportions or relative information) of microbiome data, there are specific pre-processing steps which need to be undergone in order to avoid spurious results
- Option 1: some of the functions (`pca()`, `plsda()`) include the argument `logratio = 'CLR'`
- Option 2: use the `logratio.transfo()` function

From the mixOmics newsletter (2023/05/15)
A new recording about time-course multi-omics integration is now available at this page:
`mixomics.org/time-course-integration/`
We are still working actively in this area, especially for **microbiome data**.

Examples

Bushel, P.R., Wolfinger, R.D. & Gibson, G. Simultaneous clustering of gene expression data with clinical chemistry and pathological evaluations reveals phenotypic prototypes. BMC Syst Biol 1, 15 (2007). https://doi.org/10.1186/1752-0509-1-15

```
R> library(mixOmics)
R> data(liver.toxicity)
R> help(liver.toxicity)
```

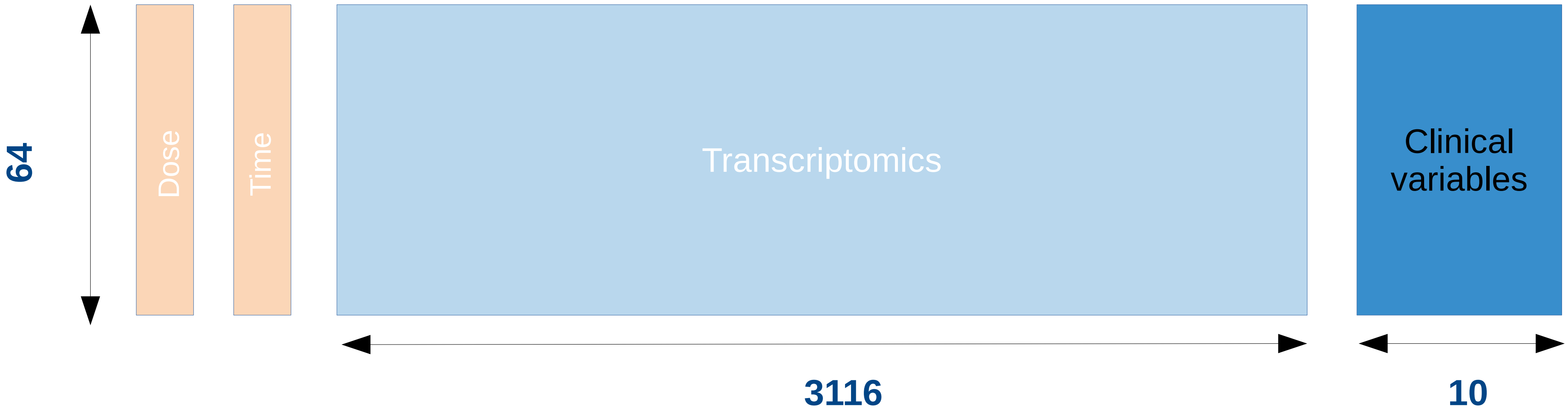


Doses of acetaminophen (**low/high**) and times of necropsies (**6/18/24/48h**)

Expression measure of 3116 genes for the 64 subjects (rats)

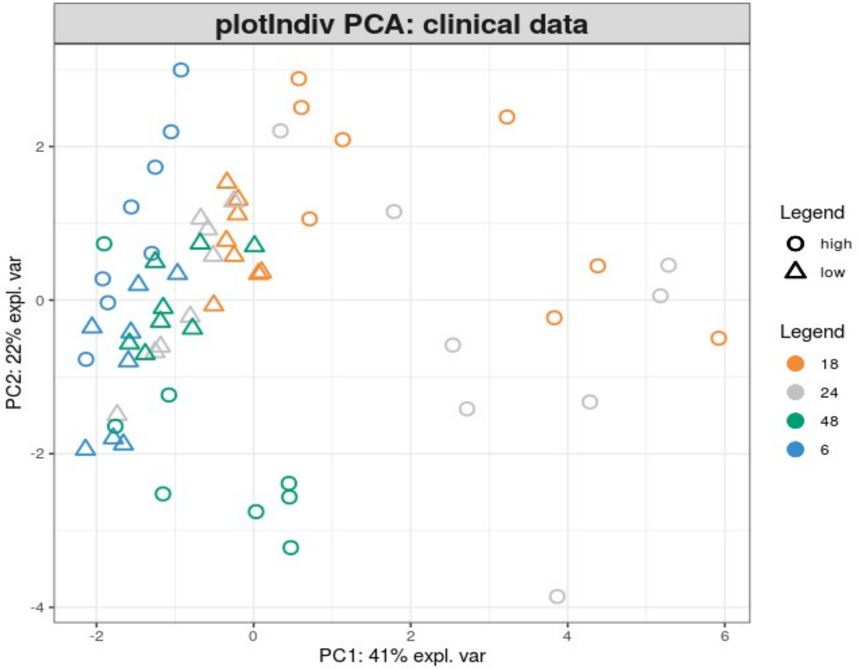
10 clinical variables for the same 64 subjects

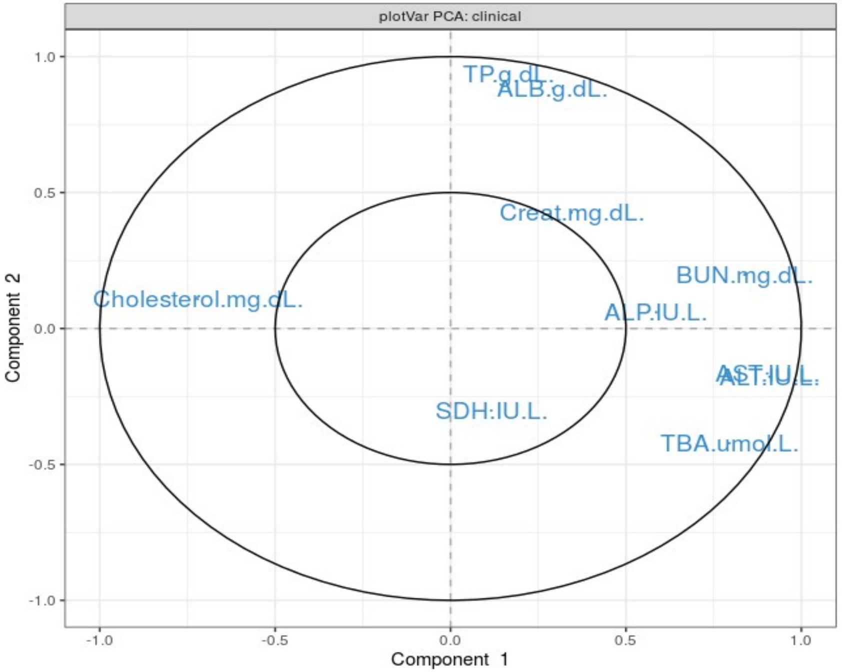
Explore one data set



Question: based on clinical data, do we naturally observe clusters of samples which correspond to the different dose or exposure treatments?

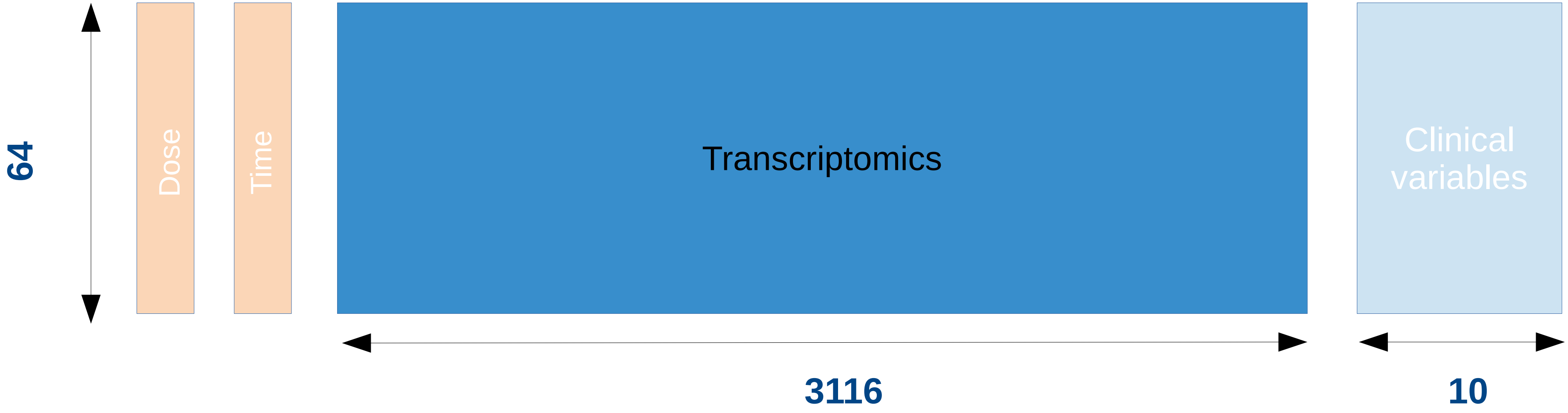
PCA on clinical variables





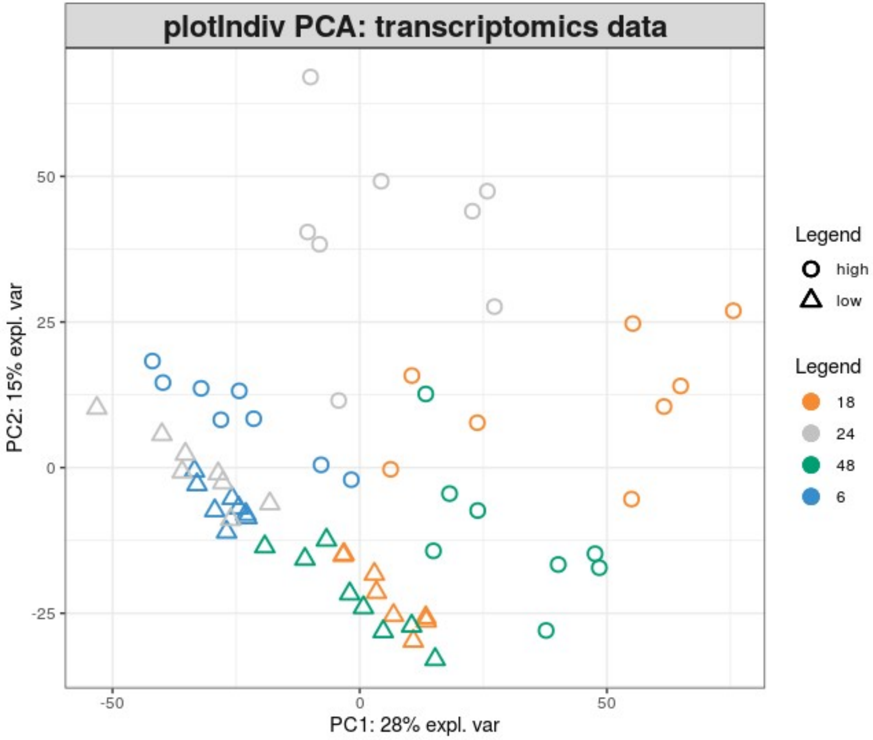
Answer: so, so...

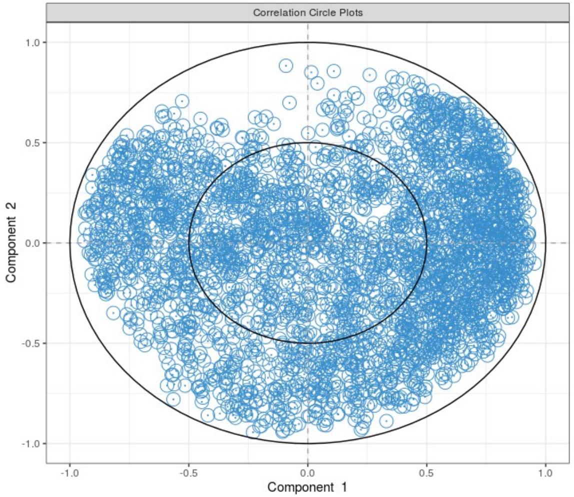
Explore another data set



Question: based on transcriptomics data, do we naturally observe clusters of samples which correspond to the different dose or exposure treatments?

PCA on transcriptomics data

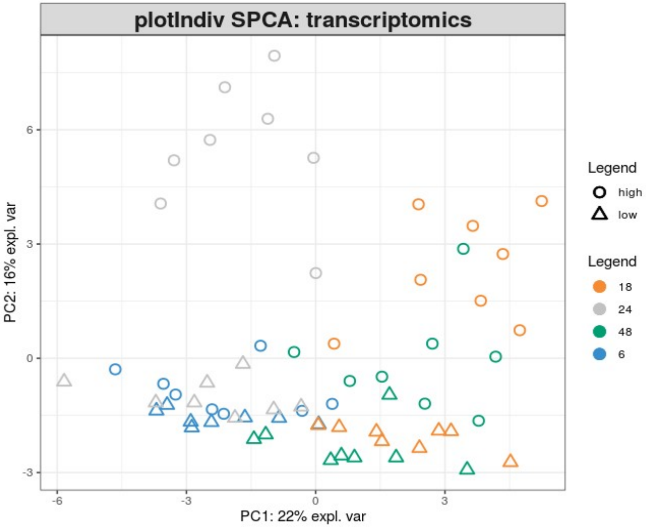


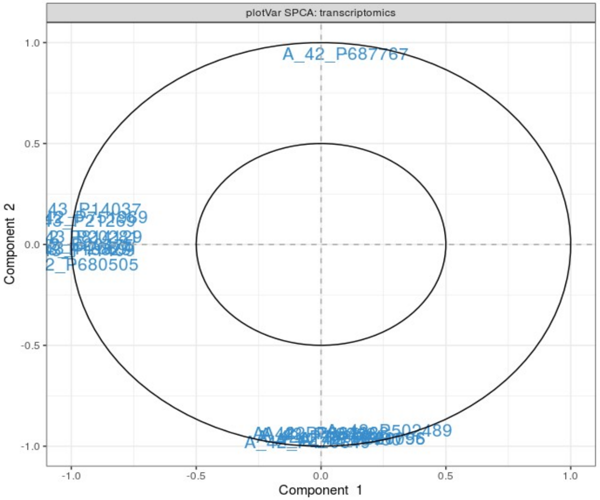


Answer: dose effect appears clearly as well as trends in time effect...

Too many genes? Sparse PCA

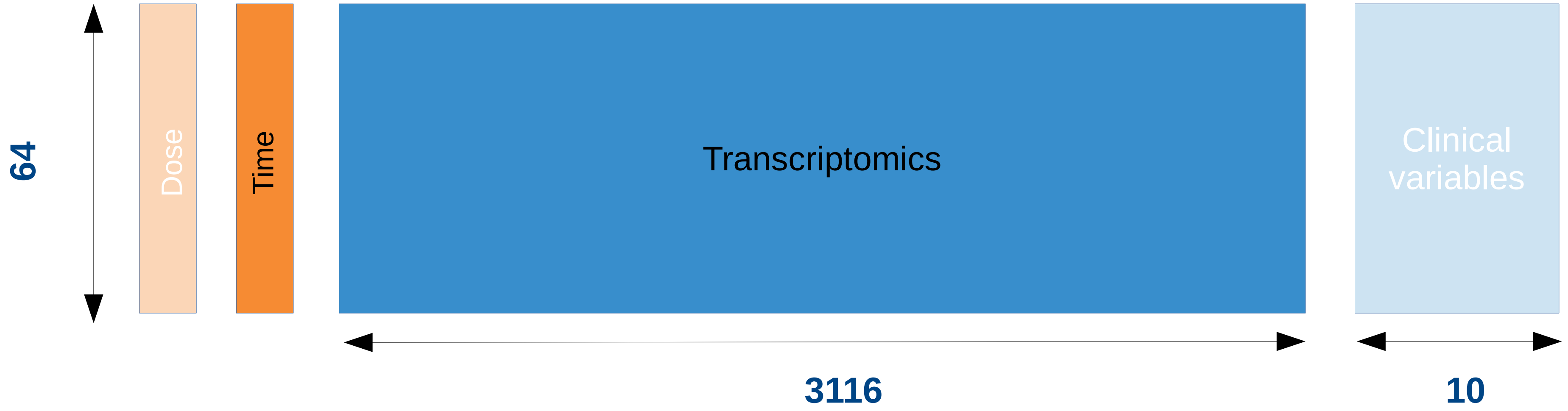
Question: based on transcriptomics data, do we naturally observe clusters of samples which correspond to the different doses or exposure treatments **when we select some genes highly involved in the variability of the data**?





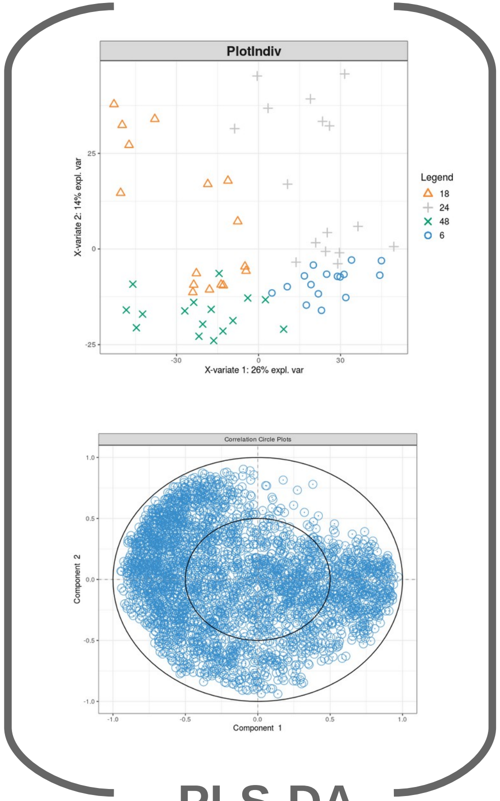
Answer: behaviour roughly similar when considering every gene or not.

Supervised analysis: transcriptomics / time

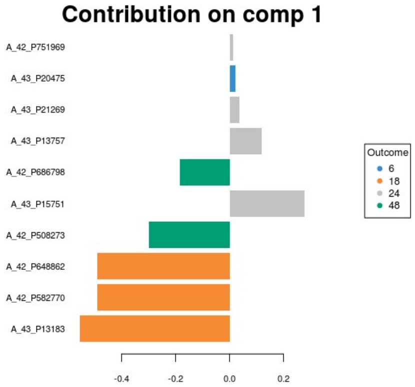


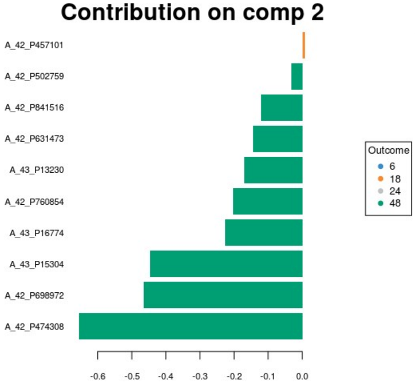
Question: Based on transcriptomics data, can we identify a molecular signature that characterizes the different treatment times?

(S)PLS-DA transcript. / time

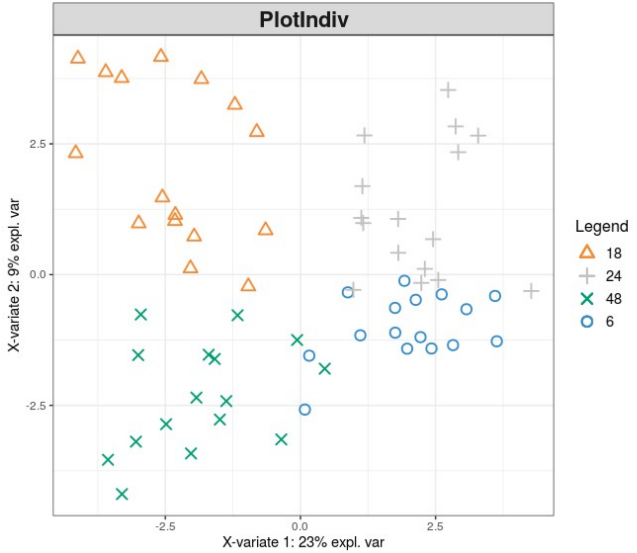


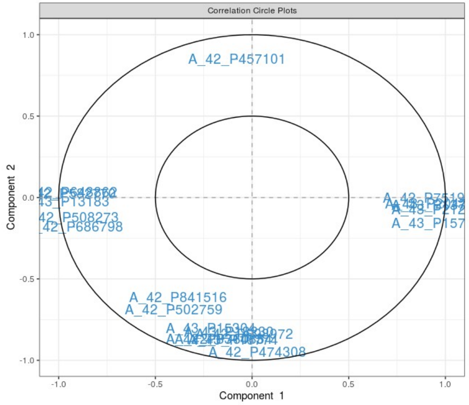
PLS-DA





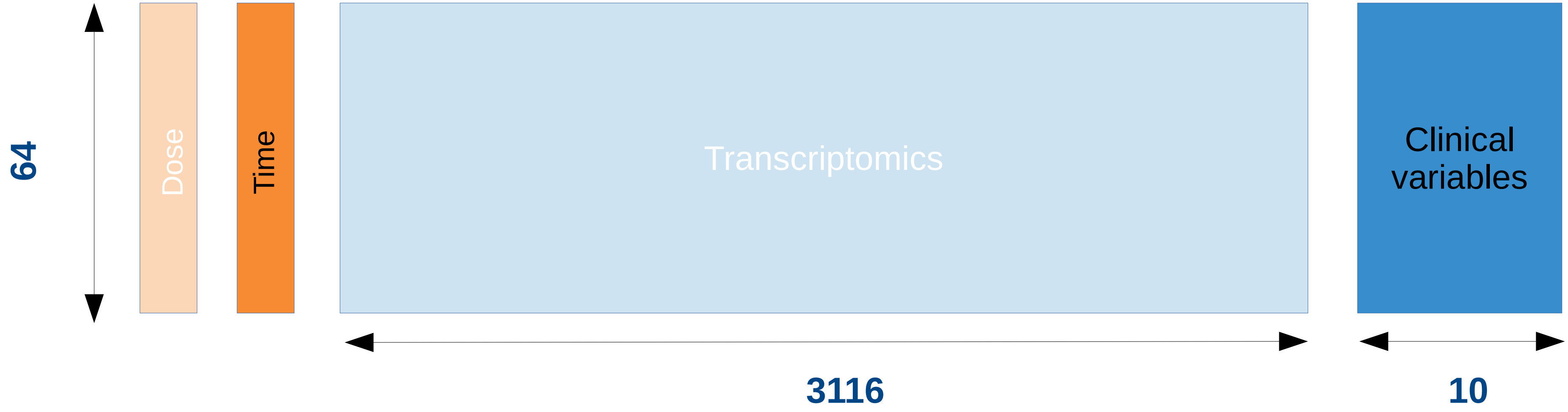
S-PLS-DA





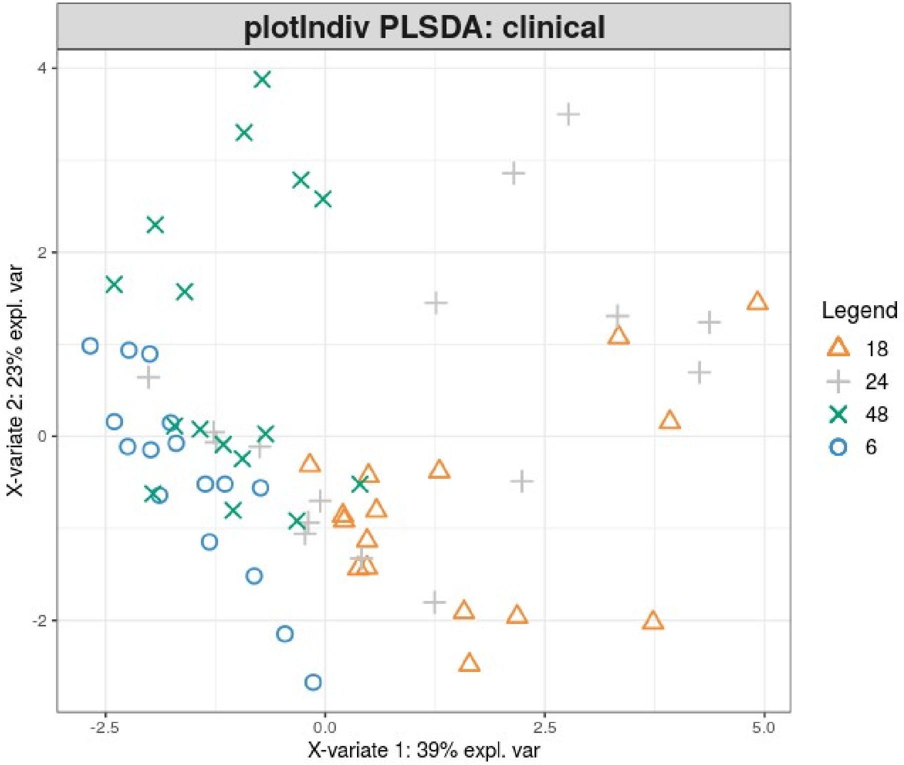
Answer: Probably, something to investigate...

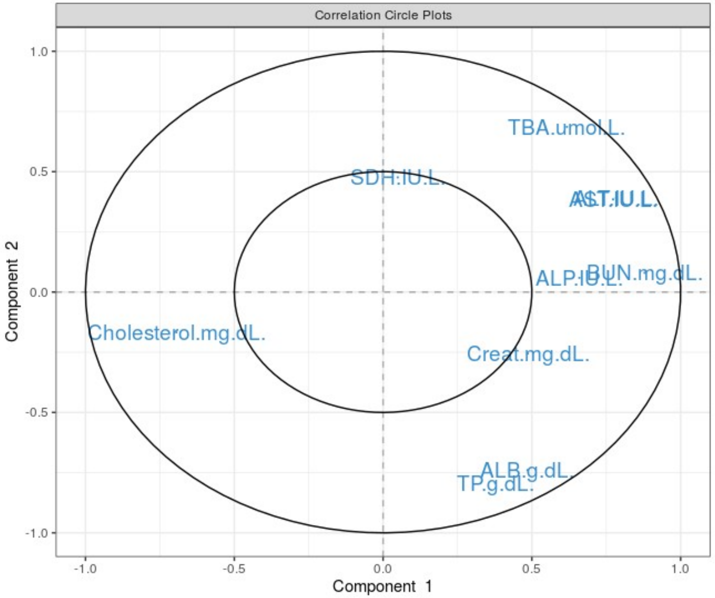
Supervised analysis: clinical / time



Question: Do we observe a better discrimination with the clinical data?

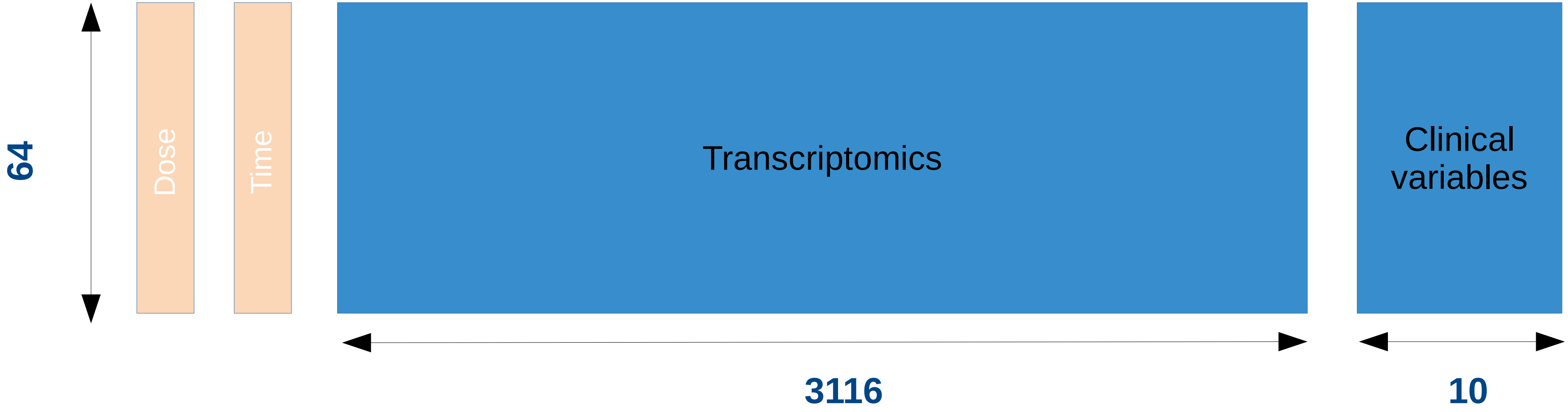
PLS-DA clinical / time





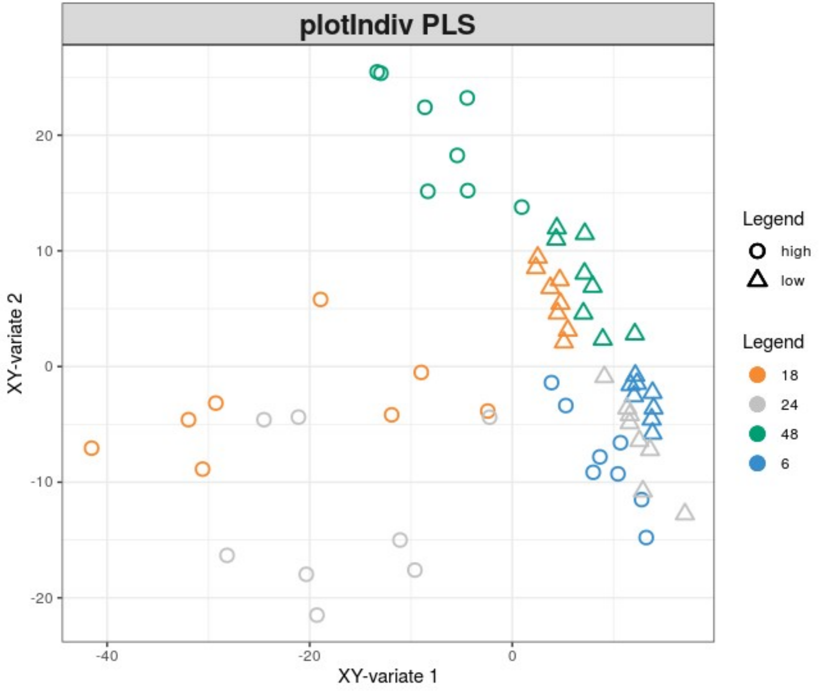
Answer: not as good as transcriptomics.

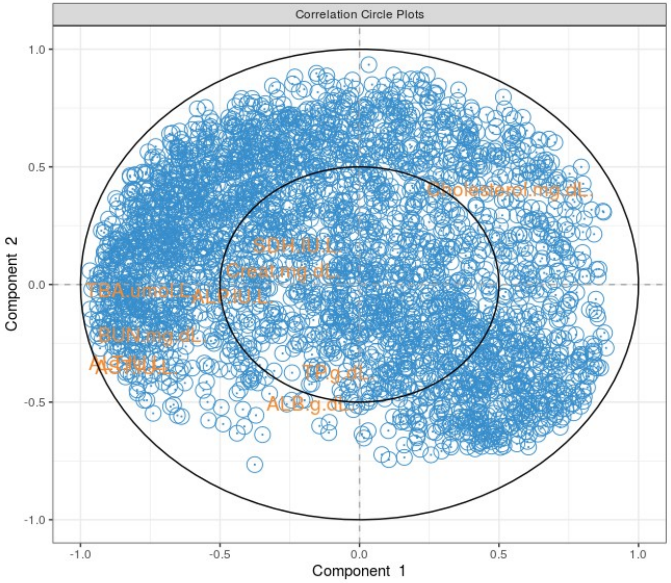
Unravel relationships between 2 datasets



Question: Can we unravel relationships between transcriptomics data and clinical data ?

PLS: transcriptomics / clinic

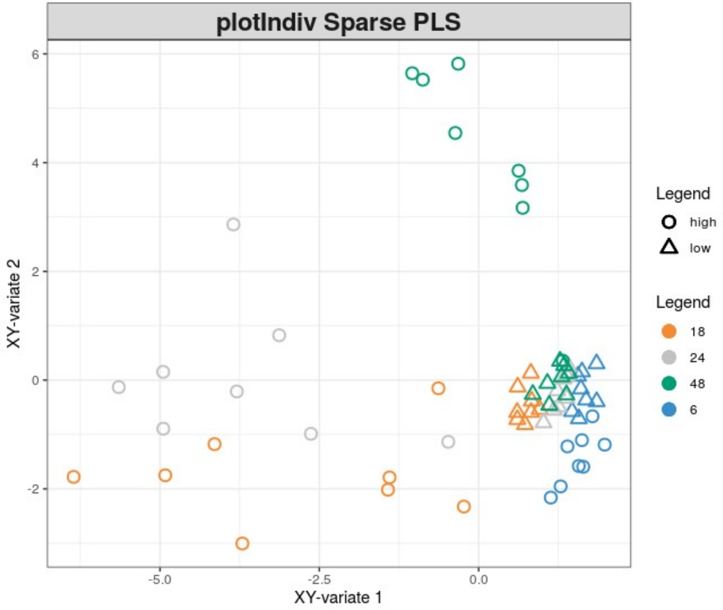


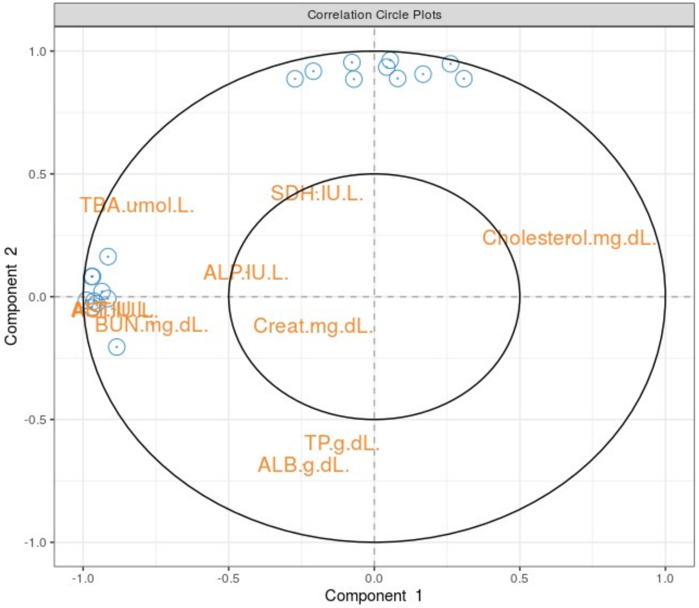


Answer: well, well... to be investigated

Sparse PLS: transcriptomics / clinic

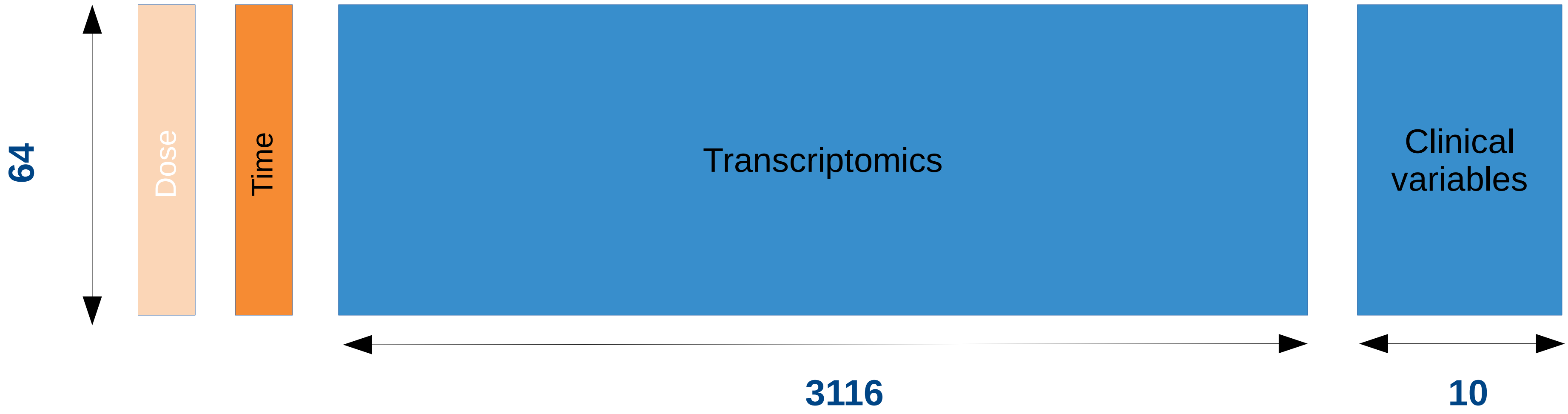
Question: Can we unravel relationships between transcriptomics data and clinical data? **What are the genes that characterize these relationships**?





Answer: interesting trends on the individual plot and few genes involved.

Multi-blocks supervised analysis



Question: Does the integration of the clinical and transcriptomics datasets bring better insight into the biological similarities between samples within the same treatment dose?

Investigation carried out with two design matrices

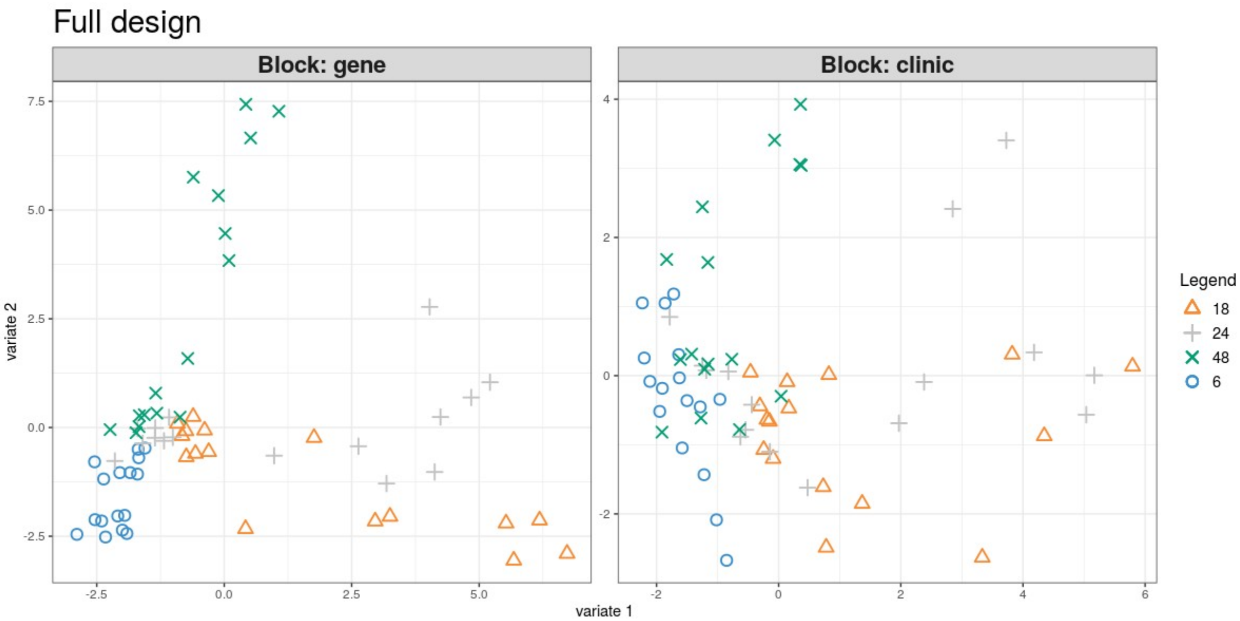
Full design

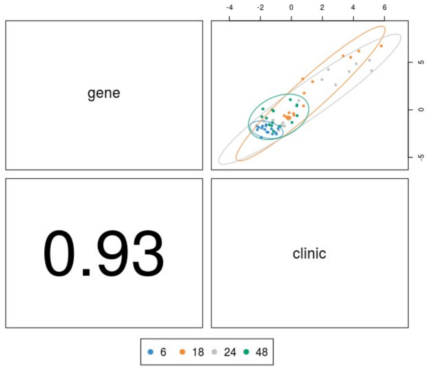
	Tr.	Cl.	Time
Trans.	0	**1**	1
Clinic.	**1**	0	1
Time	1	1	0

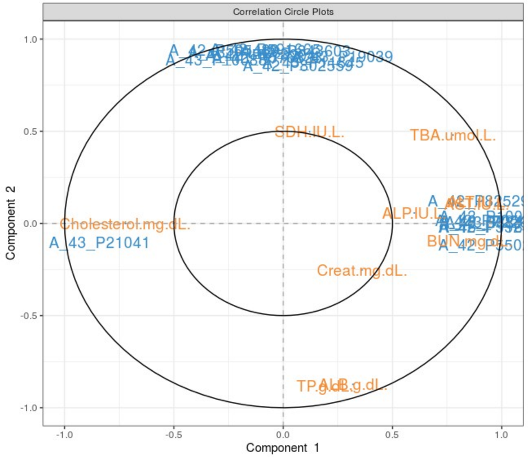
DA-oriented design

	Tr.	Cl.	Time
Trans.	0	**0.1**	1
Clinic.	**0.1**	0	1
Time	1	1	0

Multi-blocks sparse PLS-DA: transcriptomics / clinic / time

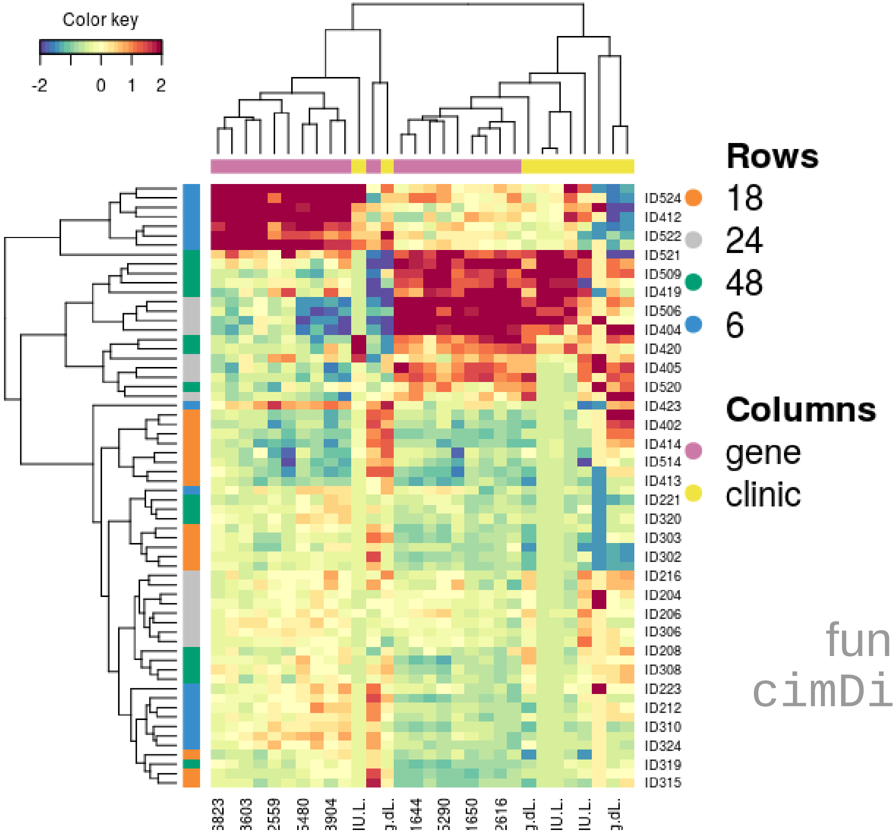






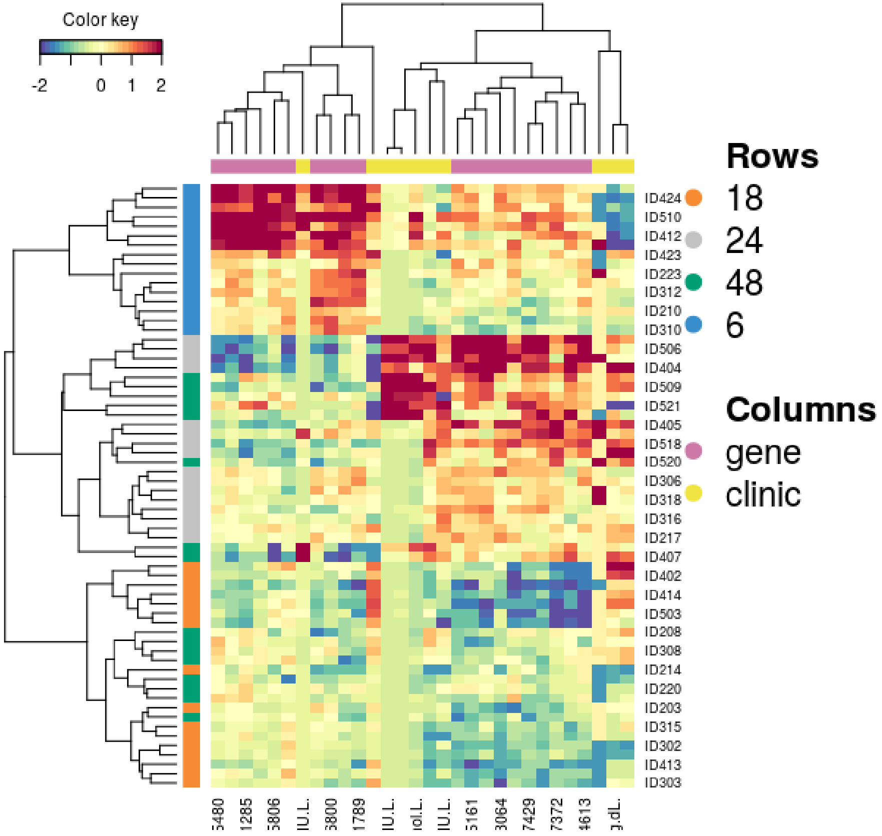
Answer: results to be investigated...

Multi-blocks sparse PLS-DA: transcriptomics / clinic / time



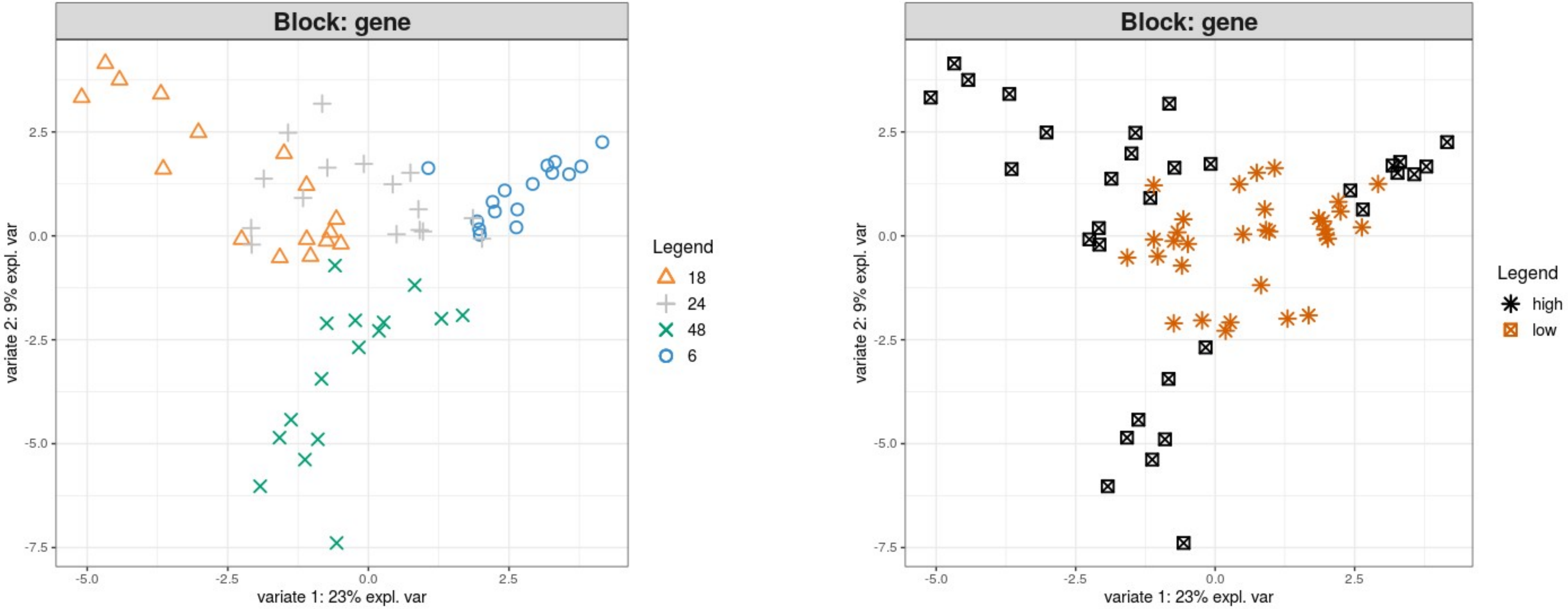
Full design

function
`cimDiablo()`



DA-oriented design

Multi-blocks sparse PLS-DA: transcriptomics / clinic / time



DA-oriented design