Multivariate dimension reduction and kernel methods for biological data integration

Sébastien Déjean, Jérôme Mariette
Kim-Anh Lê Cao, Nathalie Vialaneix

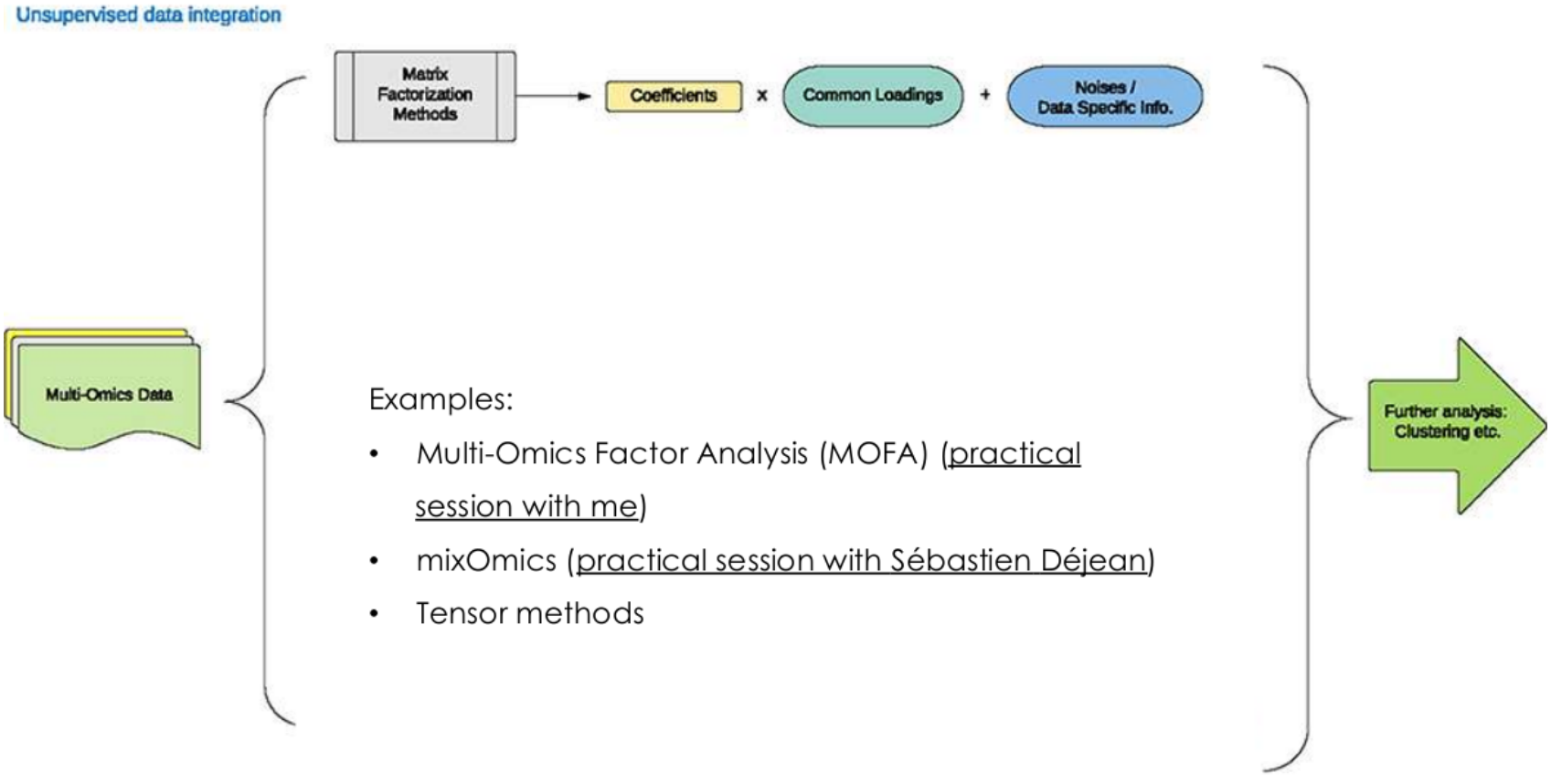
February 26, 2019

Multivariate methods

Kernel methods

Conclusion

Multivariate methods



Principal Component Analysis

PCA: the workhorse for linear multivariate statistical analysis is an (almost) compulsory first step in exploratory data analysis to:

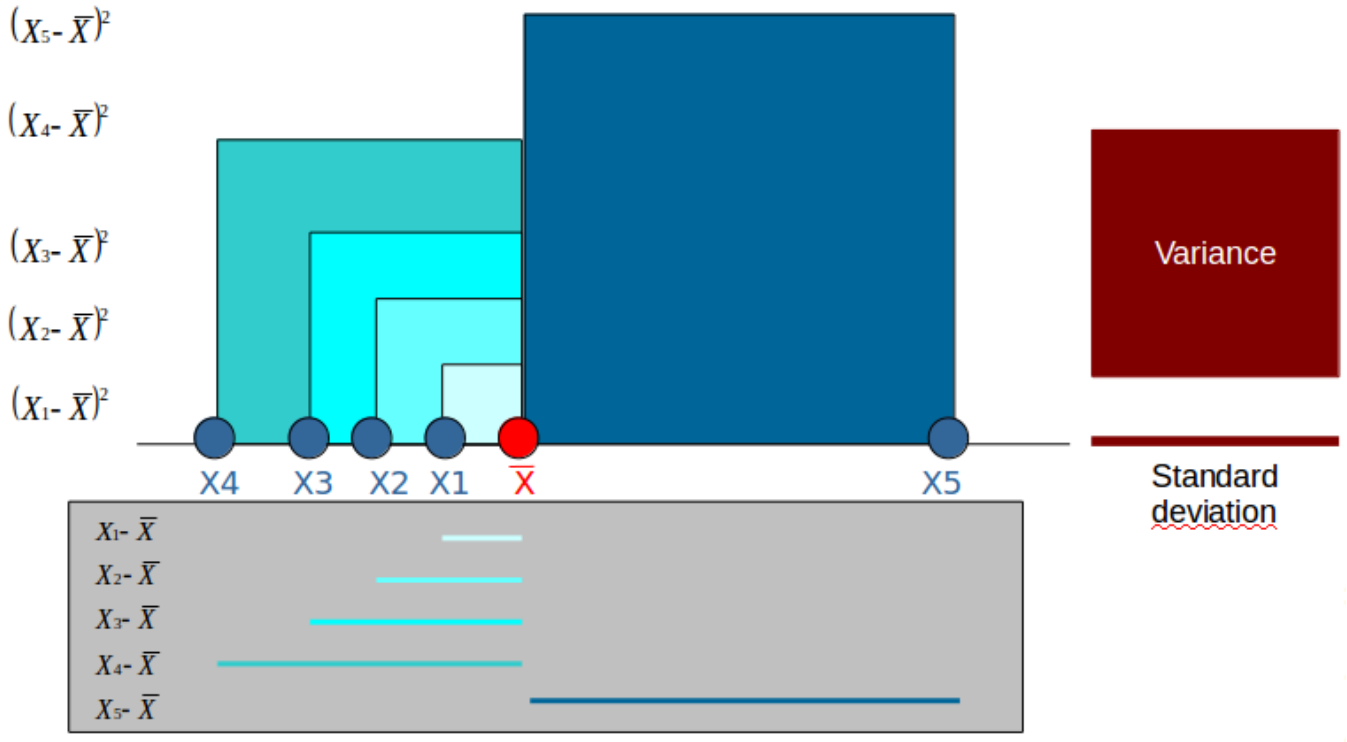
- Understand the underlying data structure
- Identify bias, experimental errors, batch effects.

Original variables are replaced by artificial variables (principal components) which explain as much information as possible from the original data and are orthogonal (covariance=0).

In PCA, the variance == information contained in the data.

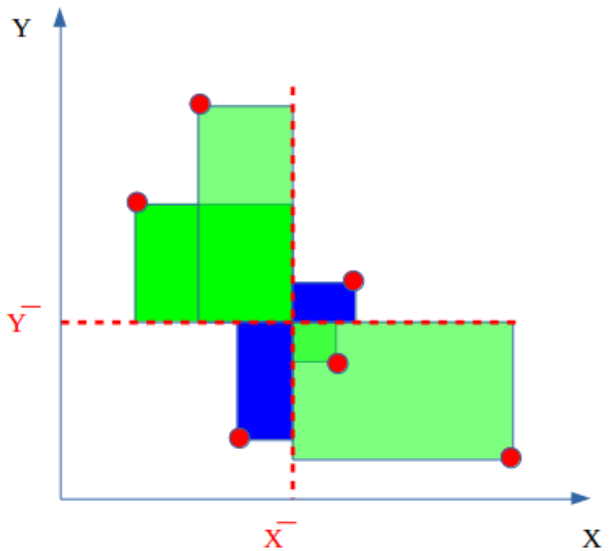
Prerequisites: Variance

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



Prerequisites: Covariance-Correlation

$$Cov(X, Y) = \frac{1}{N} \sum_{i=1}^{N} (X_i - \bar{X})(Y_i - \bar{Y})$$



$$Cor(X, Y) = \frac{Cov(X, Y)}{\sigma_X \sigma_Y}$$

Prerequisites: Linear combinations of variables

		Height			Weight		Linear combination
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

We write the linear combination as a matrix product:

Linear combination $= \mathbf{Xa}$, where X is a $(n \times p)$ matrix and a is a vector of length p)

$\rightarrow$ challenge: optimise the coefficients assigned to each variable

Now a 'larger' data set: the body data set

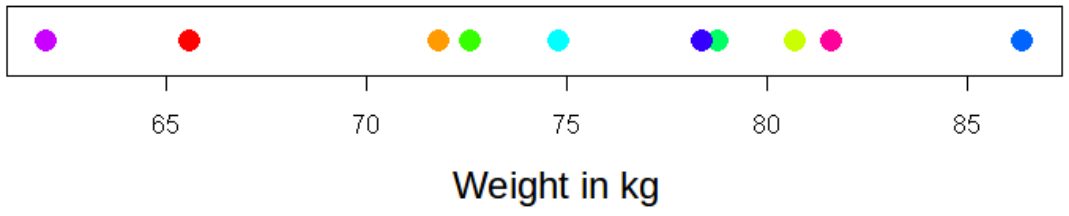
	Shoulder girth	Chest girth	Waist girth	Height	Weight
1	106.2	89.5	71.5	65.6	174.0
2	110.5	97.0	79.0	71.8	175.3
3	115.1	97.5	83.2	80.7	193.5
4	104.5	97.0	77.8	72.6	186.5
5	107.5	97.5	80.0	78.8	187.2
6	119.8	99.9	82.5	74.8	181.5
7	123.5	106.9	82.0	86.4	184.0
8	120.4	102.5	76.8	78.4	184.5
9	111.0	91.0	68.5	62.0	175.0
10	119.5	93.5	77.5	81.6	184.0

$\rightarrow$ Graphical overview of these data?

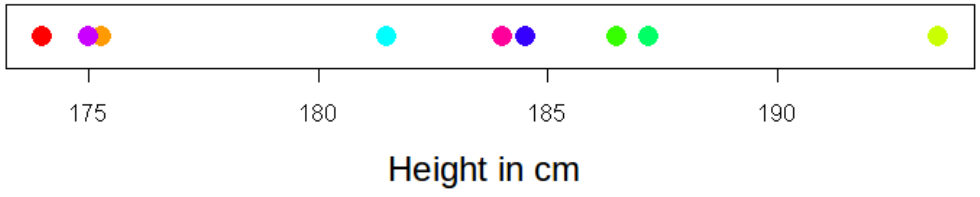
$\rightarrow$ Are all variables needed to summarise the information?

Standard plots in 1D

Weight 65.6 71.8 80.7 72.6 78.8 74.8 86.4 78.4 62.0 81.6

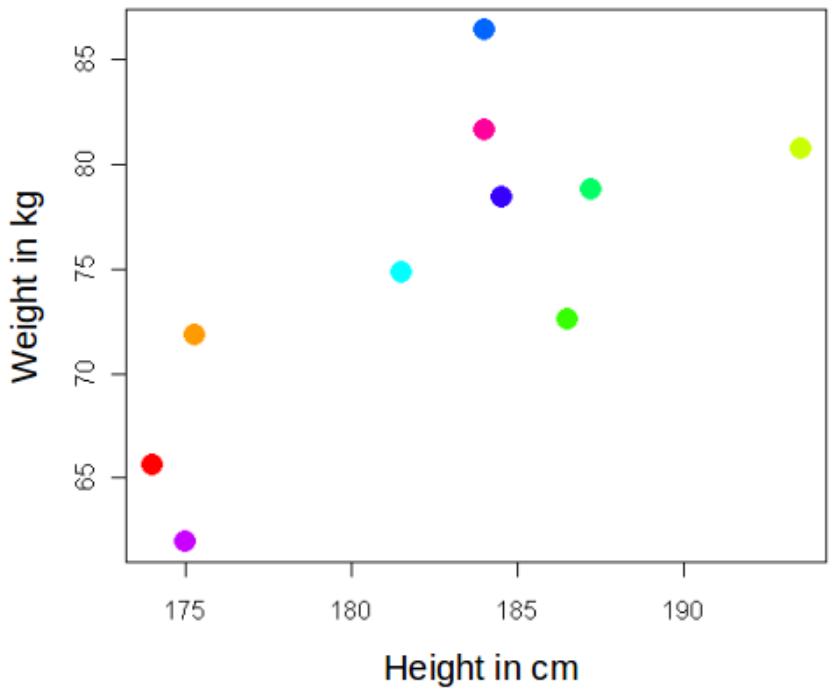


Height 174.0 175.3 193.5 186.5 187.2 181.5 184.0 184.5 175.0 184.0



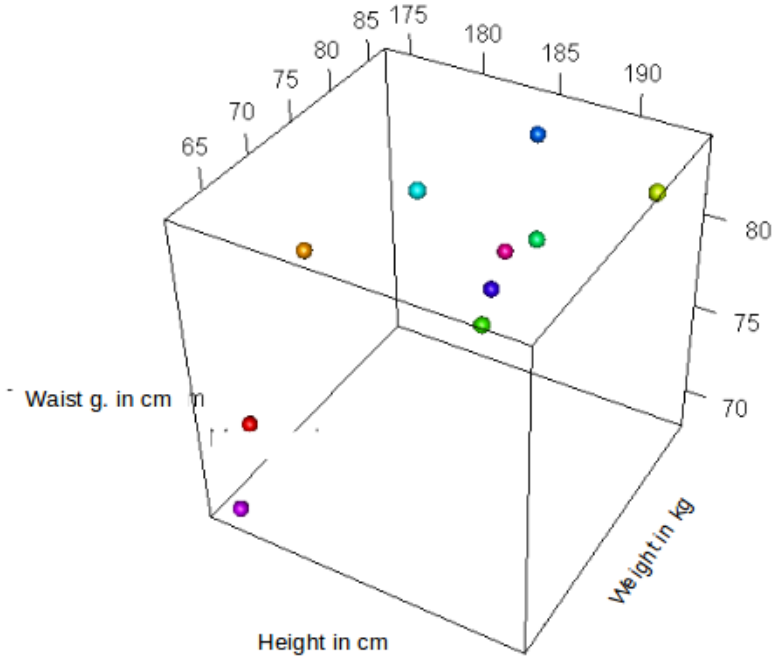
Standard plots in 2D

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

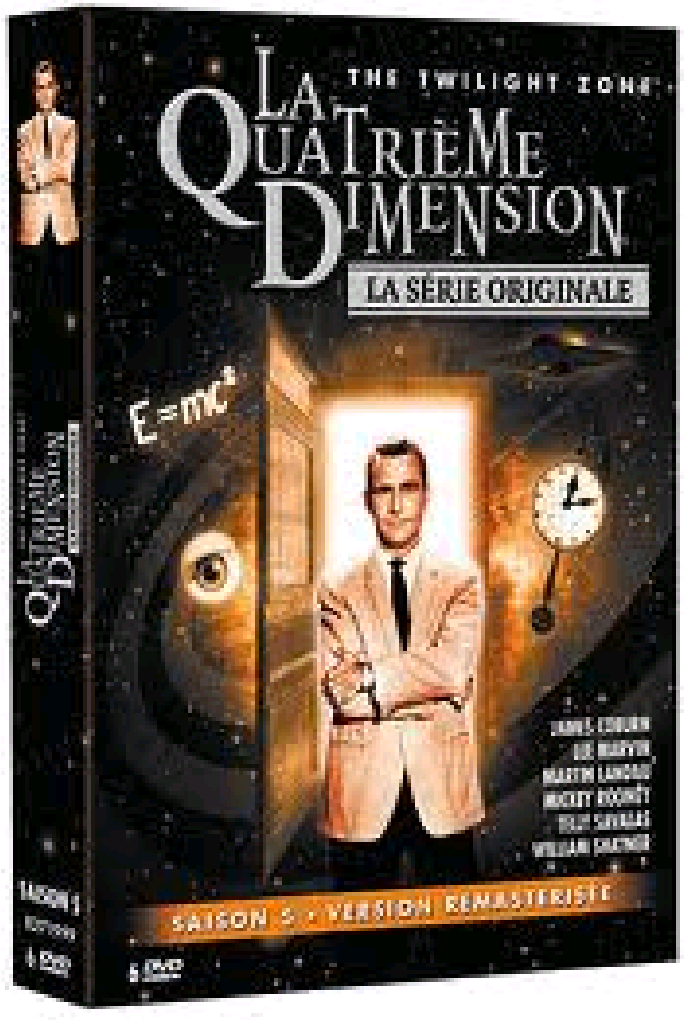


Standard plots in 3D

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

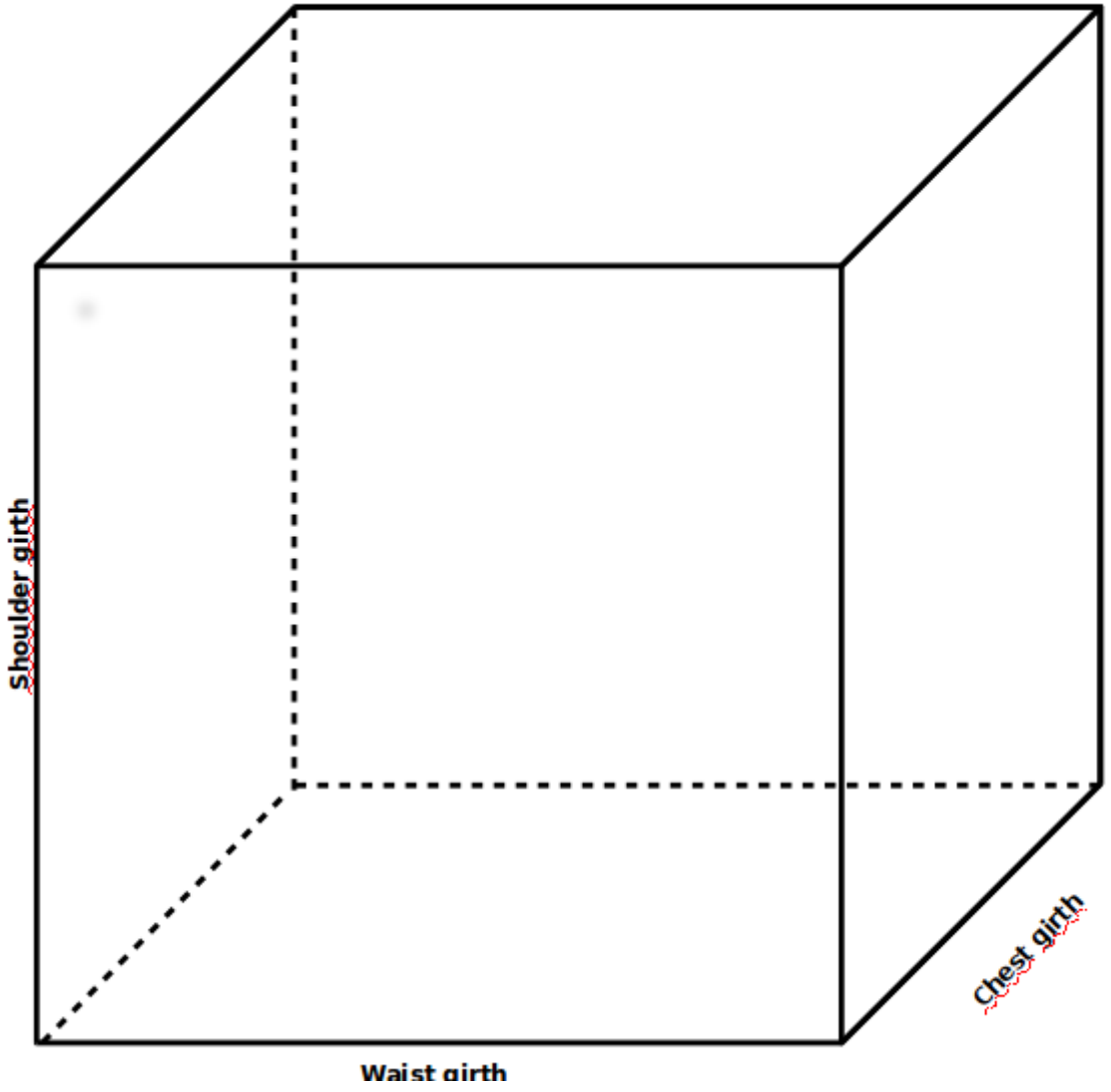


Standard plots in 4D???

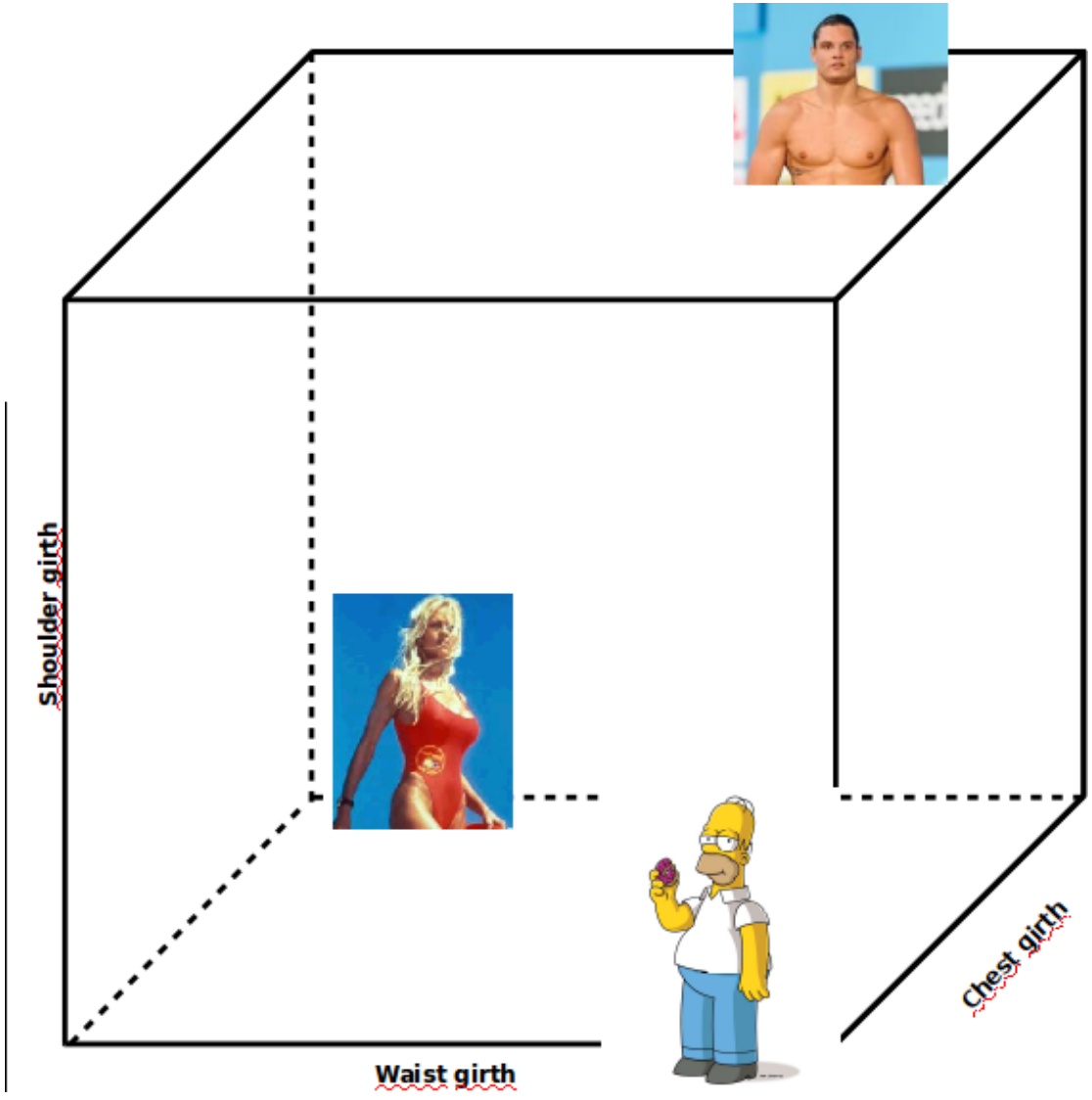


Original title: *The Twilight Zone*

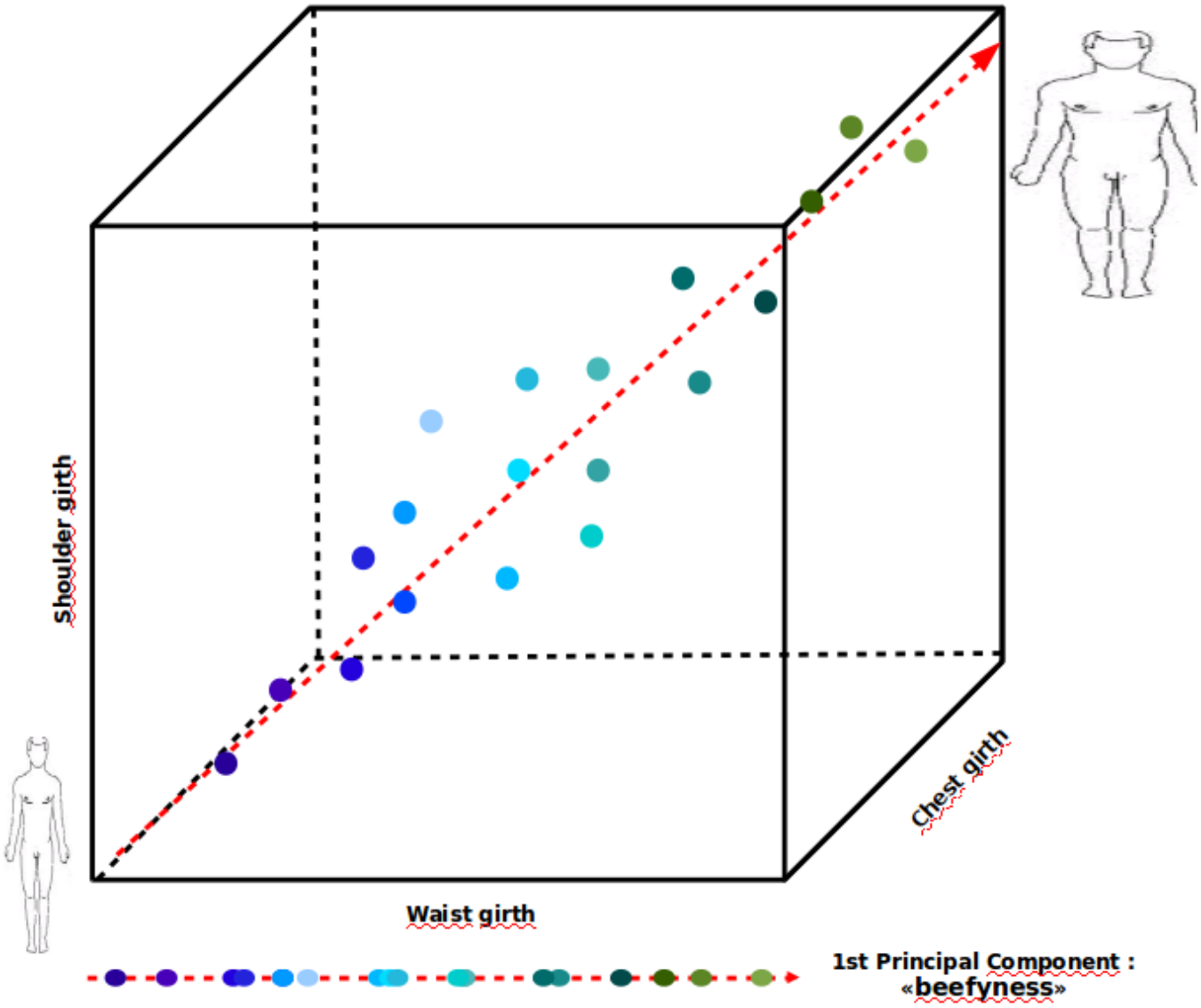
Go back to 3D



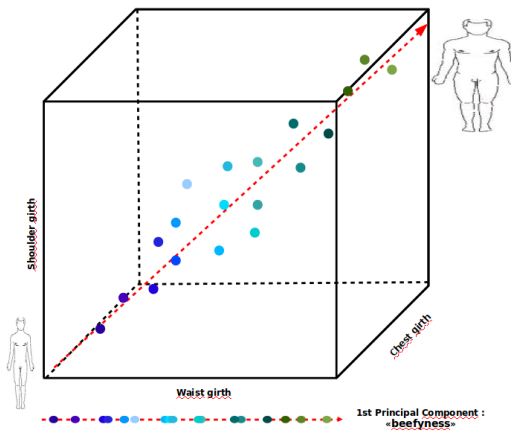
Go back to 3D



PCA: the 'trick'



PCA: the 'trick'



Summary. The measurements are strongly correlated. Indeed, a person with a high shoulder girth should also have high chest girth (with few exceptions!). Thus, information brought by these 5 variables are redundant.

Graphically in 3D (variables shoulder, chest and waist girths), there are empty areas in the cube: a variable (dotted arrow) calculated as a combination of these 3 variables is sufficient to represent the individuals with a minimal loss in information. All points are located along this direction that is the first principal component.

Algebra: a linear combination of variables

Seek for the best directions in the data that account for most of the variance. Objective function:

$$\max_{||\boldsymbol{a}||=1} var(X\boldsymbol{a})$$

Each principal component $\boldsymbol{t}$ is a linear combination of the original variables ($\boldsymbol{t} = X\boldsymbol{a}$):

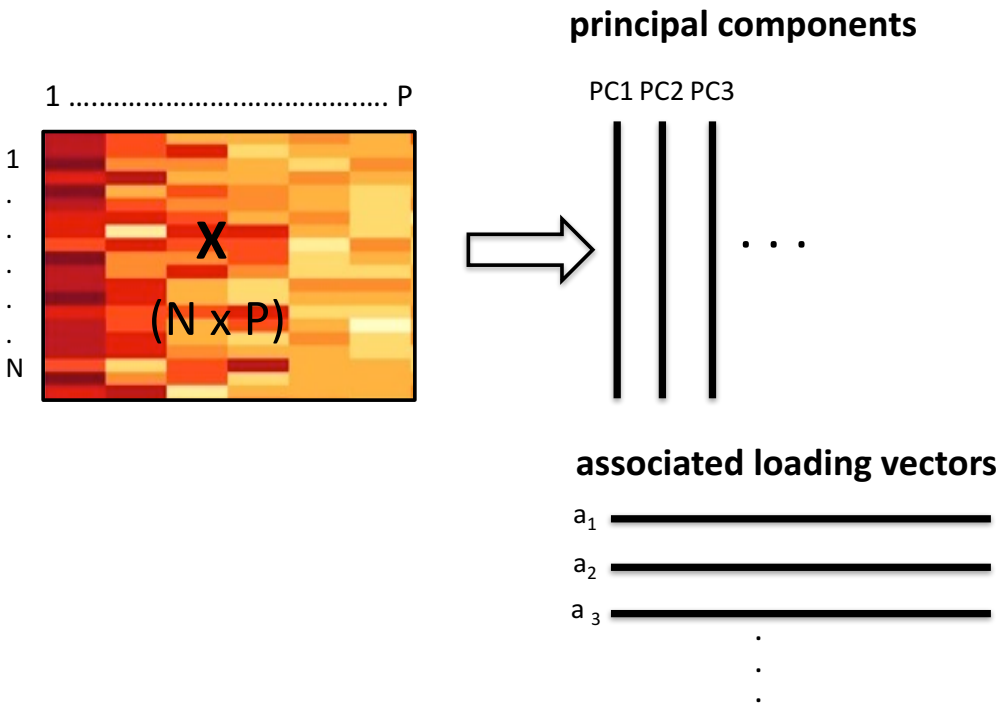
$$\boldsymbol{t} = a_1\boldsymbol{x}^1 + a_2\boldsymbol{x}^2 + \cdots + a_p\boldsymbol{x}^p$$

- $\boldsymbol{X}$ is a $n \times p$ data matrix with $\{\boldsymbol{x}^1, \ldots, \boldsymbol{x}^p\}$ the p variable profiles.
- $\boldsymbol{t}$ is the **first** principal component with max. variance
- $\{a_1, \ldots, a_p\}$ are the weights in the linear combination

The data are projected into a smaller subspace

- Each principal component is orthogonal to each other to ensure that no redundant information is extracted.
- The new PCs form a a smaller subspace of dimension $<< p$.
- Each value in the principal component corresponds to a score for each sample
  $\rightarrow$ we project each sample into a new subspace spanned by the PCs
- Approximate representation of the data points in a low dimensional space
- Summarize the information related to the variance

Data compression with PCA is a matrix decompos



- Components are linear combinations of original variables, and orthogonal to each other.
- Loading vectors indicate the weight (importance) of each variable in the linear combination.

Back to the body data set

Data

	s.g	c.g	w.g	w	h
H 1	106.2	89.5	71.5	65.6	174.0
H 2	110.5	97.0	79.0	71.8	175.3
H 3	115.1	97.5	83.2	80.7	193.5
H 4	104.5	97.0	77.8	72.6	186.5
H 5	107.5	97.5	80.0	78.8	187.2
H 6	119.8	99.9	82.5	74.8	181.5
H 7	123.5	106.9	82.0	86.4	184.0
H 8	120.4	102.5	76.8	78.4	184.5
H 9	111.0	91.0	68.5	62.0	175.0
H 10	119.5	93.5	77.5	81.6	184.0
F 1	105.0	89.0	71.2	67.3	169.5
F 2	100.2	94.1	79.6	75.5	160.0
F 3	99.1	90.8	77.9	68.2	172.7
F 4	107.6	97.0	69.6	61.4	162.6
F 5	104.0	95.4	86.0	76.8	157.5
F 6	108.4	91.8	69.9	71.8	176.5
F 7	99.3	87.3	63.5	55.5	164.4
F 8	91.9	78.1	57.9	48.6	160.7
F 9	107.1	90.9	72.2	66.4	174.0
F 10	100.5	97.1	80.4	67.3	163.8
Mean	108.1	94.2	75.3	70.6	174.4
Var.	68.6	37.5	50.8	85.7	109.3

Covariance matrix

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

$$68.64 + 37.51 + 50.77 + 85.71 + 109.31 = 351.94$$

351.94 represents the quantity of information contained in the data.

Go back to the body data set

Coefficients (optimally calculated) to build principal components

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

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

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

PC3 = ...
```

Coordinates of the individuals on the PCs

	Dim1	Dim2	Dim3	Dim4	Dim5
H1	-6.50	-4.48	-0.37	-1.03	1.27
H2	4.40	2.04	0.81	1.87	1.38
H3	22.66	-5.94	-6.18	0.11	1.97
H4	7.78	-5.24	-8.38	4.10	-1.74
H5	13.73	-2.67	-8.02	0.82	-2.15
H6	15.67	-0.15	4.49	2.33	4.40
H7	26.99	3.19	6.29	0.04	-3.08
H8	18.41	-3.43	5.63	1.09	-1.96
H9	-6.25	-8.48	4.97	0.79	1.86
H10	16.78	-3.67	1.99	-7.08	1.22
F1	-8.83	-0.78	0.28	-3.02	0.07
F2	-7.28	15.41	-2.31	-3.00	-2.35
F3	-6.45	2.25	-7.60	0.95	1.15
F4	-12.51	2.68	8.91	4.27	-1.53
F5	-3.65	20.76	-0.30	-2.45	1.99
F6	-0.63	-4.62	0.34	-3.46	-2.80
F7	-23.61	-5.07	2.20	1.19	-1.15
F8	-37.50	-9.07	-1.33	-1.89	-0.02
F9	-4.98	-3.61	0.33	-0.50	1.02
F10	-8.24	10.89	-1.74	4.86	0.44
Mean	0	0	0	0	0
Var.	255.7	60.2	23.5	8.61	4.0

Covariance matrix between PCs

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

```
255.66 + 60.18 + 23.48 + 8.61 + 4.01
= 351.94
```

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

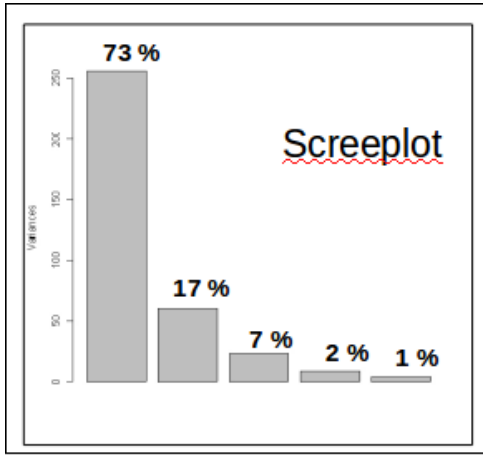
The same quantity of information (**351.94**) is kept but it is ``optimally'' allocated.

Choosing the parameters in PCA

How many principal components to choose to summarize most of the information?

We can obtain as many components as the rank of the matrix X

- Proportion of explained variance / cumulative prop.
- Screeplot of eigenvalues: any elbow?
- Sample plot: makes sense?

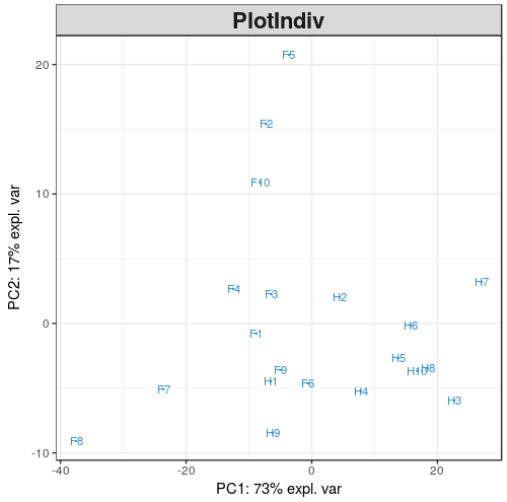


Cumulative proportion of explained variance for the 5 principal components:

PC1	PC1 to 2	PC1 to 3	PC1 to 4	PC1 to 5
0.73	0.90	0.97	0.99	1

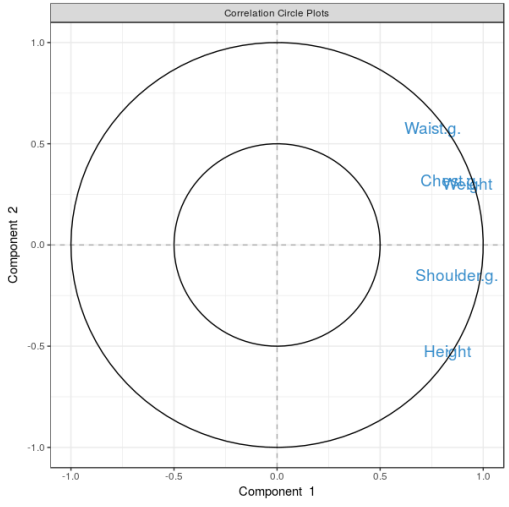
PCA is a visualisation tool

Sample plot



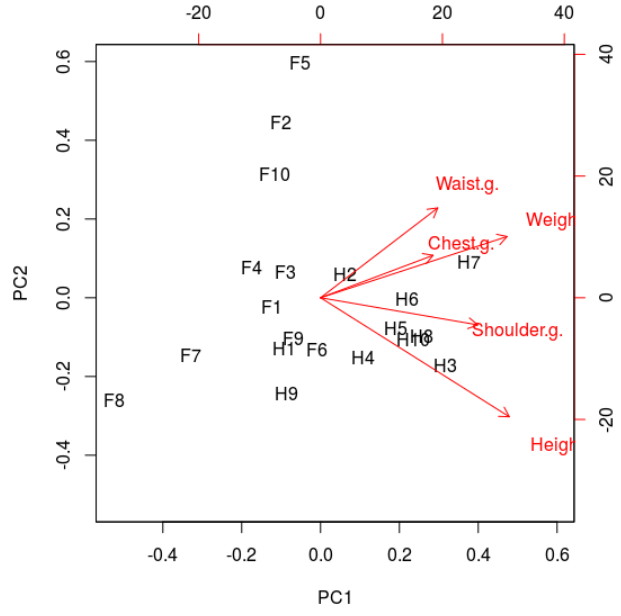
+

Variable plot



=

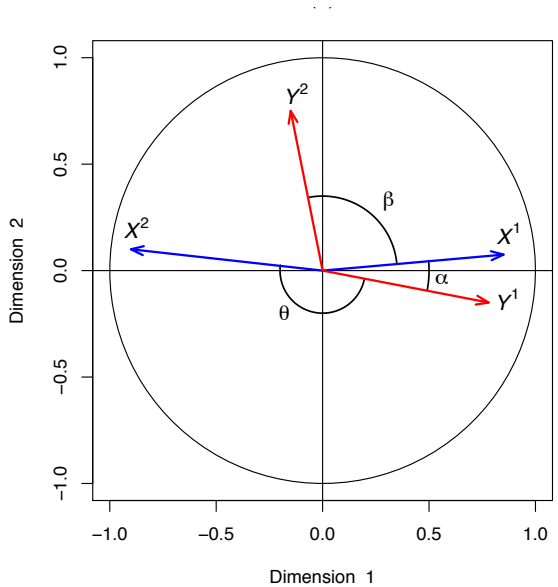
Biplot



Variable representation

To obtain the coordinate of each variable: calculate the correlation between the original data and each PC

- correlation between the variable and the PC = cos(angle) between the variable vector and the PC
- correlation between two variables = cos(angle) between 2 vectors



- data centered and scaled in PCA
- cos(α) close to 1 $\rightarrow$ cor > 0
- cos(β) close to 0 $\rightarrow$ cor $\simeq 0$
- cos(β) close to -1 $\rightarrow$ cor < 0

PCA summary

- PCA is a matrix decomposition technique that allows dimension reduction.
- Perform a PCA first to understand the sources of variation in your data.
- Always report the % explained variance per component.
- PCA can highlight 'batch effect' in the data and can be used to check that batch-effect removal techniques are efficient.

- *Should I scale my data before performing PCA?* (`scale = TRUE`)
  - Without scaling: a variable with high variance will solely drive the first principal component
  - With scaling: one noisy variable with low variability will be assigned the same variance as other meaningful variables

- *Can I perform PCA with missing values?*
  - NIPALS (Non-linear Iterative PArtial Least Squares - implemented in mixOmics) can impute missing values but must be built on many components. The proportion of NAs should not exceed 20% of total data.
    The best thing to do about missing data is not to have any.
    Gertrude Cox, 1900-1978, American statistician

Can PCA fail?

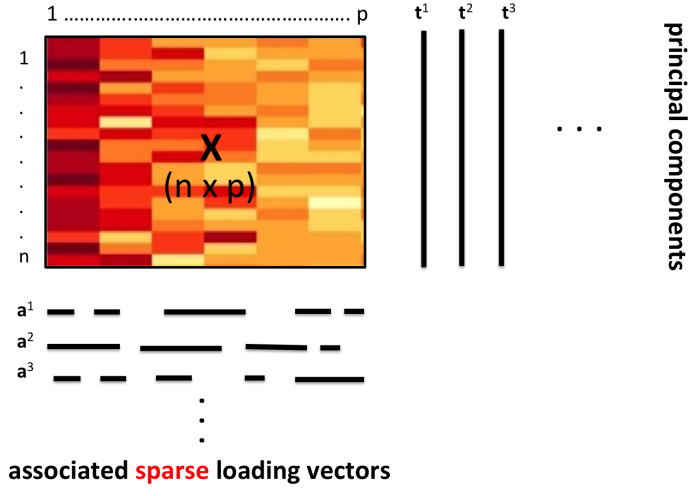
The problem lies in your data not in PCA!

- When the biological question may not be related to the highest variance in the data
  → Independent Component Analysis (ICA) or variants (see IPCA in mixOmics).
- When there are too many noisy variables which contribute to the variance
  → sparse PCA (see next)
- When samples are **not** independent (e.g. time course data, repeated measures) subject variation > the time variation
  → PCA multilevel approach (in mixOmics)

Going sparse: principle

- Large number of variables: noisy / irrelevant contribute to the variance $\rightsquigarrow$ PCA difficult to visualise and understand
- Clearer signal if some of the variable weights $\{a_1, \ldots, a_p\}$ were set to 0 for the 'irrelevant' variables ($\sim$ smallest weights)

$$\boldsymbol{t} = 0 * \boldsymbol{x}^1 + a_2 \boldsymbol{x}^2 + \cdots + 0 * \boldsymbol{x}^p$$



$\rightsquigarrow$ Sparse PCA, sparse PLSDA, sparse PLS...

Supervised analysis

Aim: To seek for a linear combination of variables to characterise or separate two or more classes of samples.

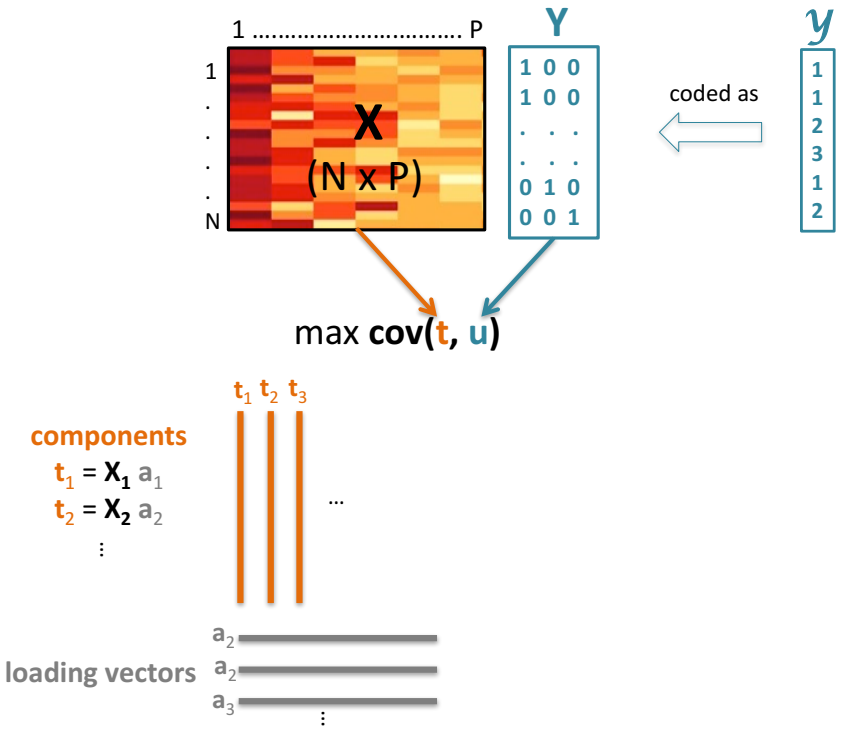
Result of a linear multivariate classifier:

- Dimensionality reduction prior to classification.
- A classifier able to predict the class of a new sample based on a linear combination of features.

Multivariate classification approaches:

- Fisher's Linear Discriminant Analysis (LDA)
- Partial Least Squares Discriminant Analysis (PLS-DA)

PLS-DA includes sample group information



- decomposition of the data matrix X in relation with the outcome y with a set of components and loading vectors for dimension reduction
- Outcome y transformed internally into a dummy matrix (see Table 4.1)

The problem to solve is:

$$\max_{||\boldsymbol{a}||=1,||\boldsymbol{b}||=1} cov(X\boldsymbol{a}, Y\boldsymbol{b})$$

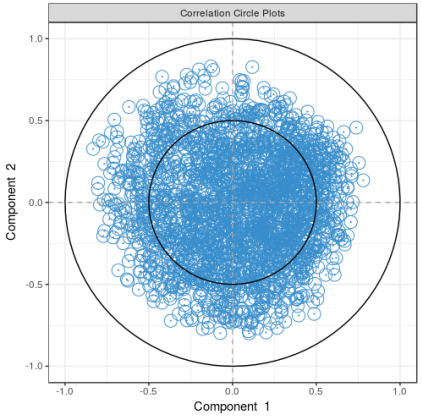
$\boldsymbol{t} = X\boldsymbol{a}$ and $\boldsymbol{u} = Y\boldsymbol{b}$ are the PLS-DA components.

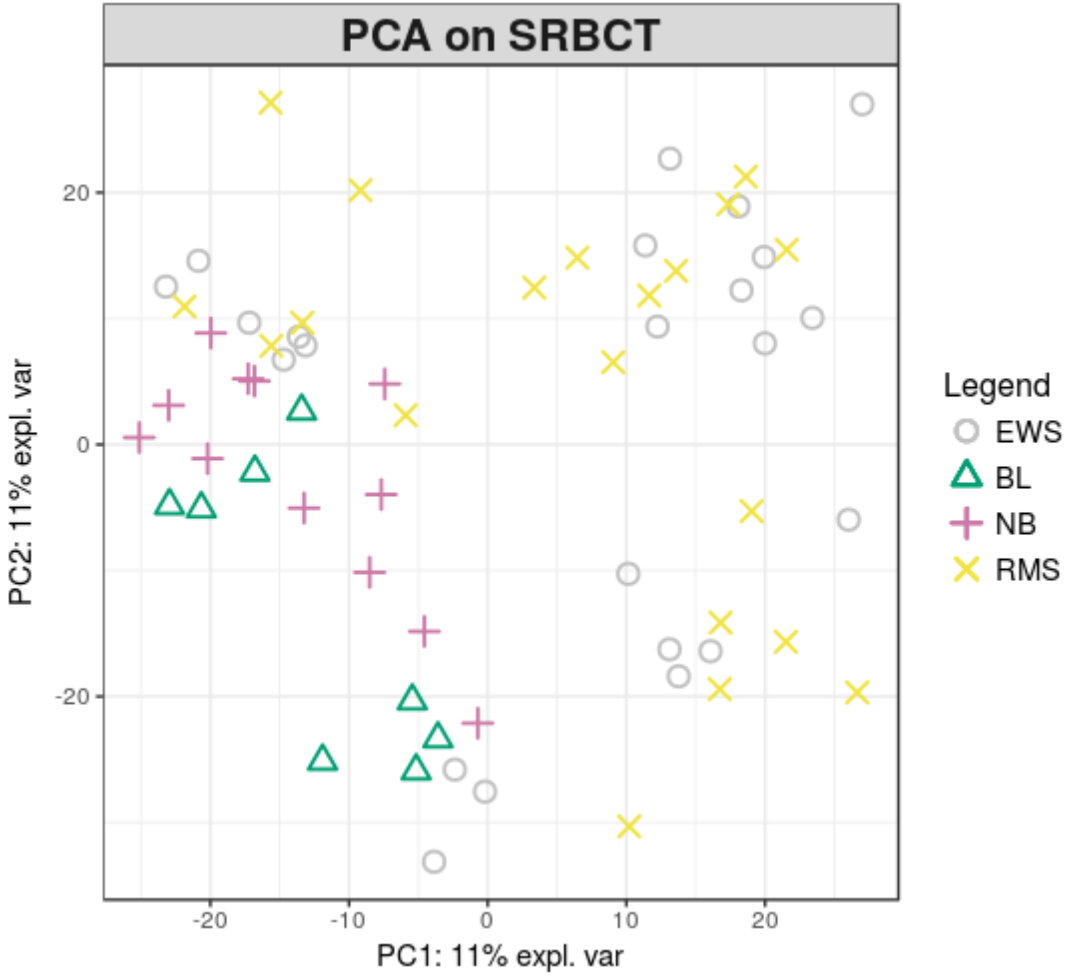
Example: SRBCT data set

- **63** samples
- expression of **2308** genes
- class tumour of each sample, **4** classes: 23 Ewing Sarcoma (EWS), 8 Burkitt Lymphoma (BL), 12 neuroblastoma (NB), 20 rhabdomyosarcoma (RMS)

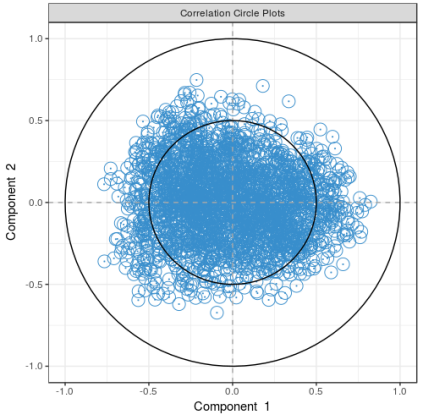
Khan et al. (2001). Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nature Medicine 7(6)

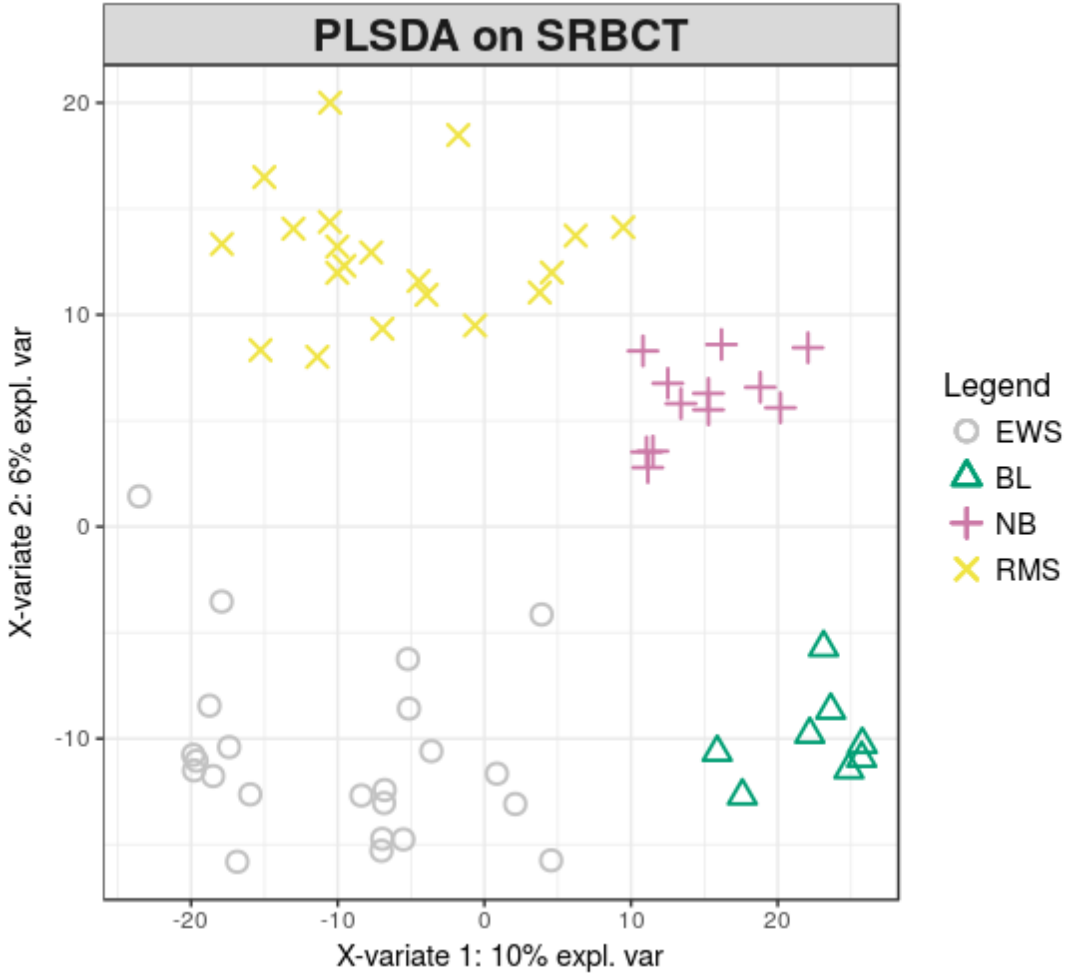
Example: PCA first!





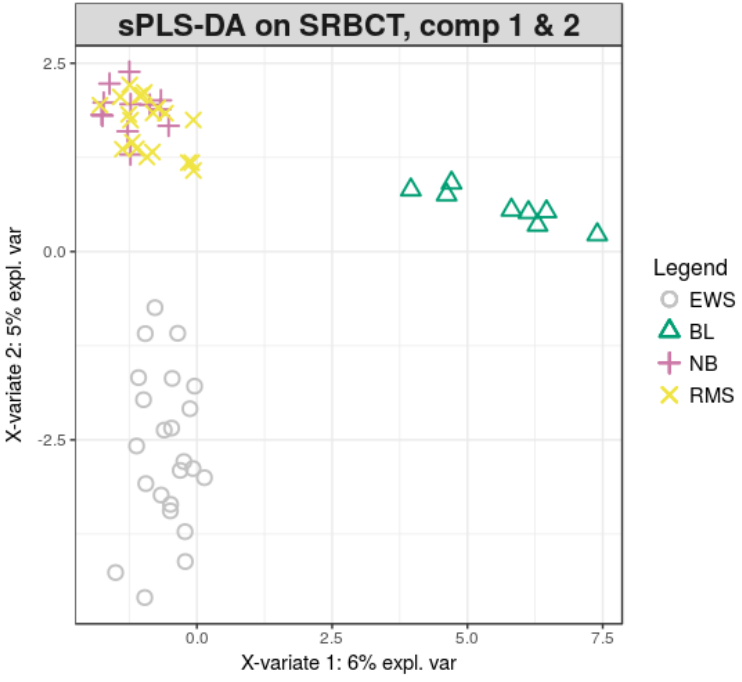
Example: PLSDA

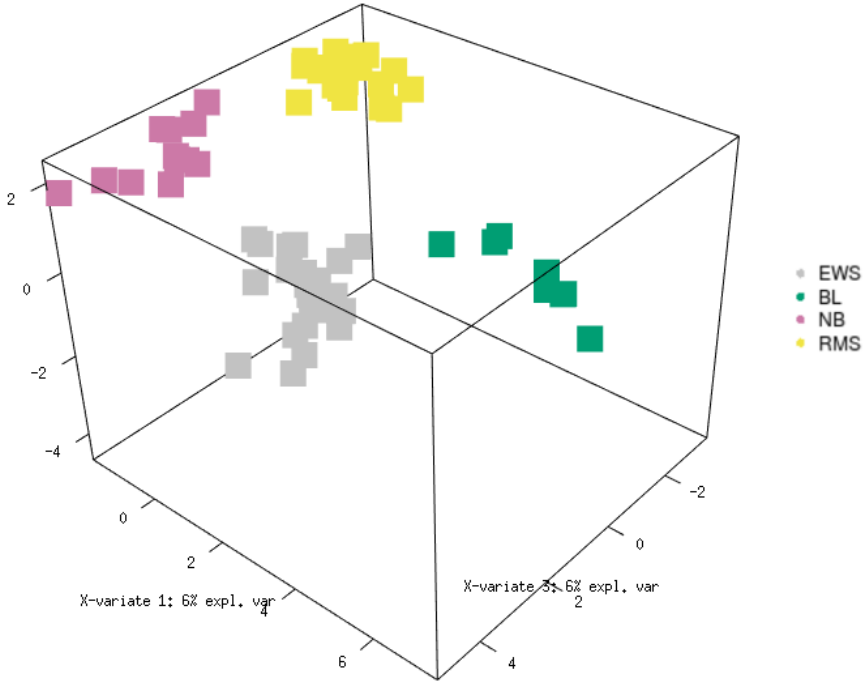




Example: Sparse PLSDA

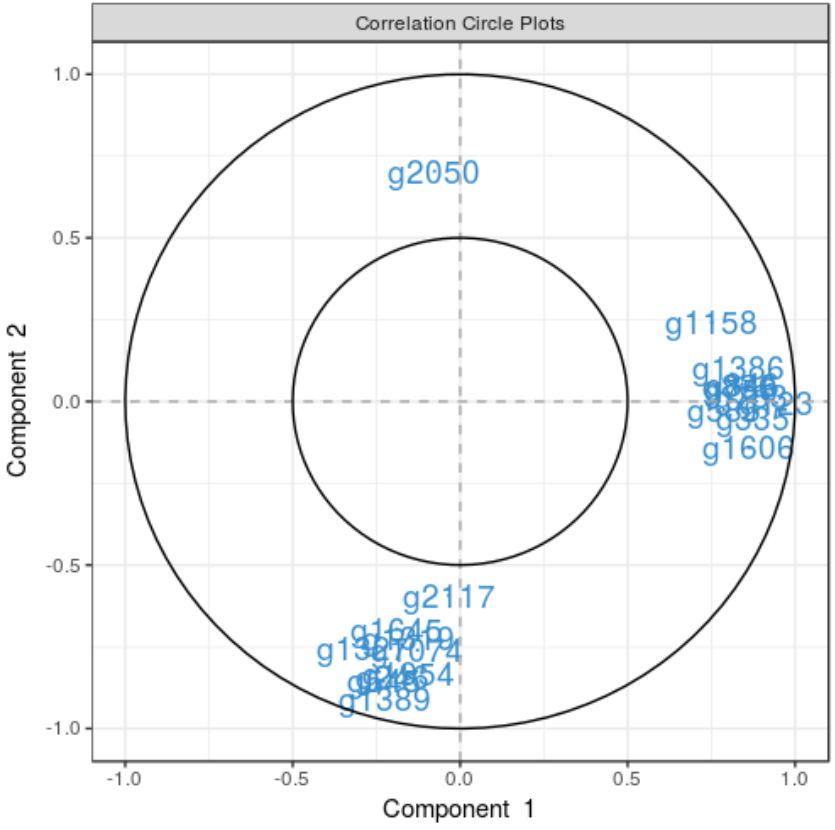
Sample plots

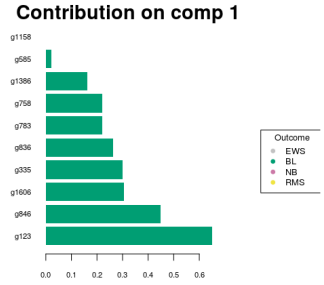


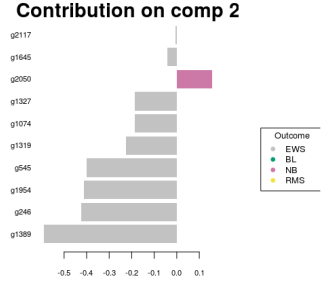


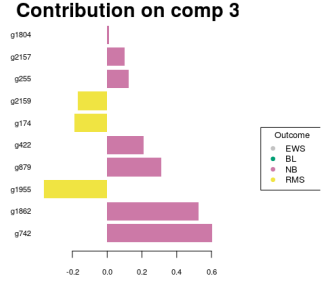
Example: Sparse PLSDA

Variable plots



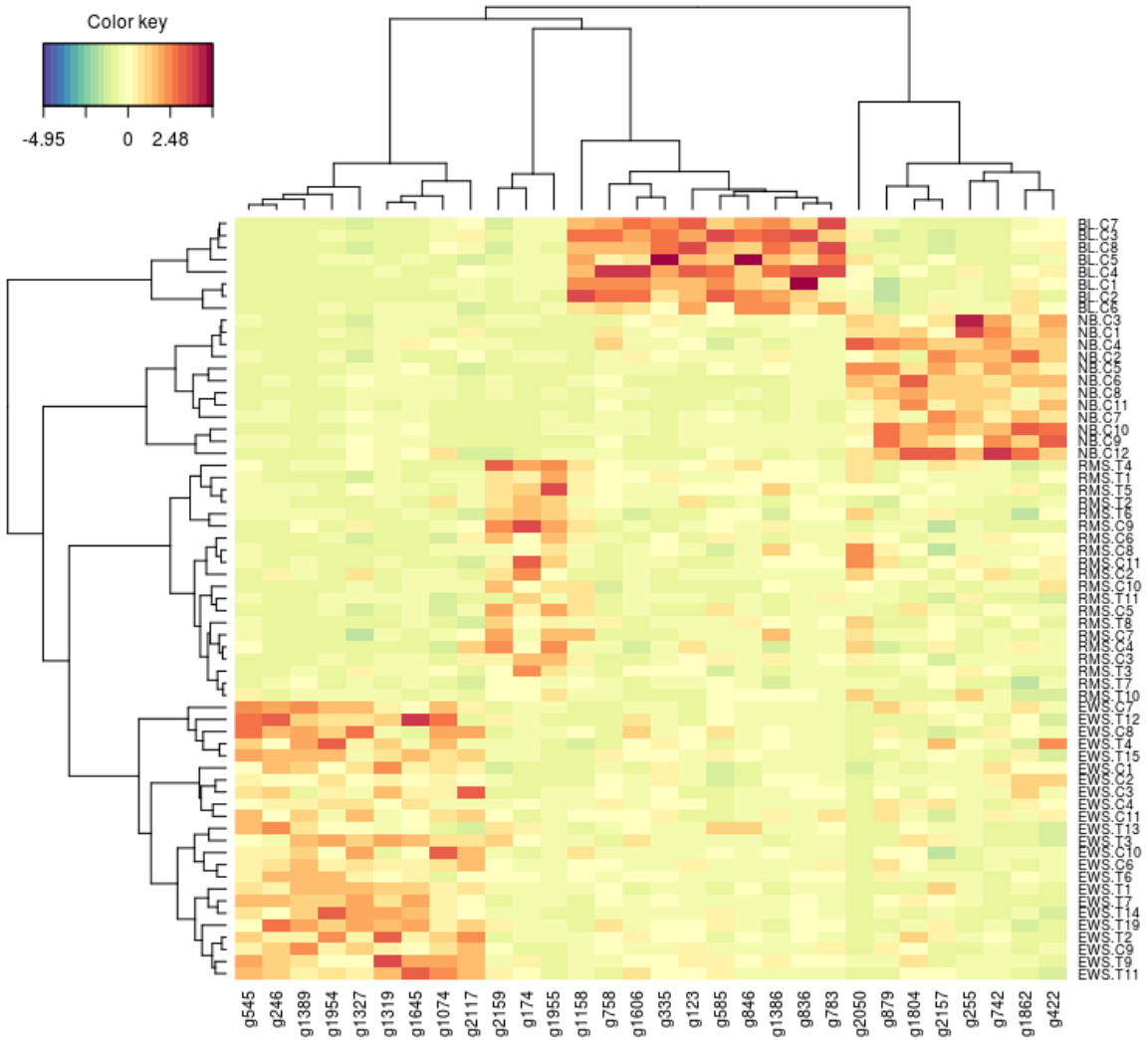






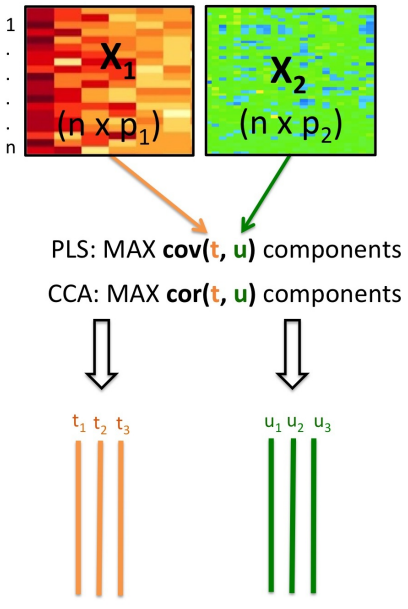
Example: Sparse PLSDA

Another variable plot



Two-blocks integration

Aim: Unravel the relationships between two omics data sets



Multivariate two-blocks integration approaches:

- Canonical Correlation Analysis (CCA), maximise the correlation between linear combination of variables in each data set
- Projection to Latent Structure / Partial Least Squares (PLS), maximise the covariance between linear combination of variables in each data set

Sparse PLS: select co-regulated biological entities across samples

Example: `nutrimouse` data set

- **40** mice: **2** genotypes (WT / PPARα) x **5** diets(*) x **4** replicates

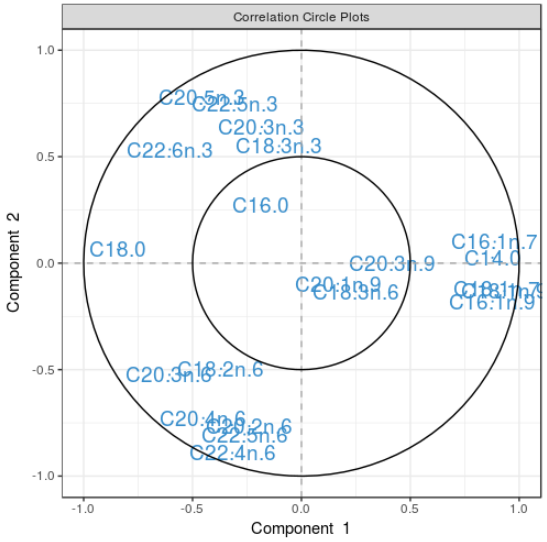
  (*) Oils used for experimental diets preparation were corn and colza oils (50/50) for a reference diet (`REF`), hydrogenated coconut oil for a saturated fatty acid diet (`COC`), sunflower oil for an Omega6 fatty acid-rich diet (`SUN`), linseed oil for an Omega3-rich diet (`LIN`) and corn/colza/enriched fish oils for the `FISH` diet (43/43/14)

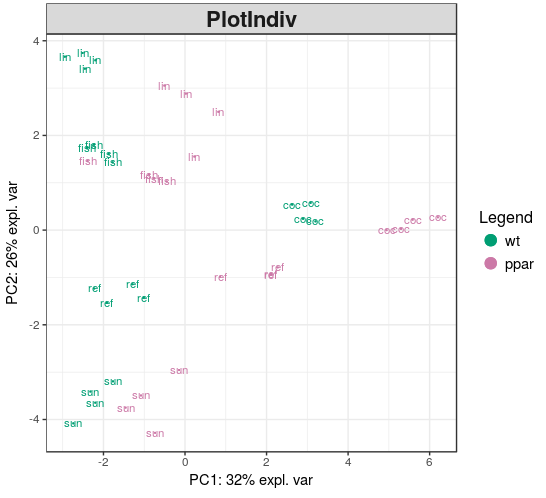
- **2** data sets acquired in liver:
  - expression of **120** genes
  - concentration of **21** fatty acids

Martin, P. G. P. et al. (2007). Novel aspects of PPAR-mediated regulation of lipid and xenobiotic metabolism revealed through a multrigenomic study. Hepatology, 54

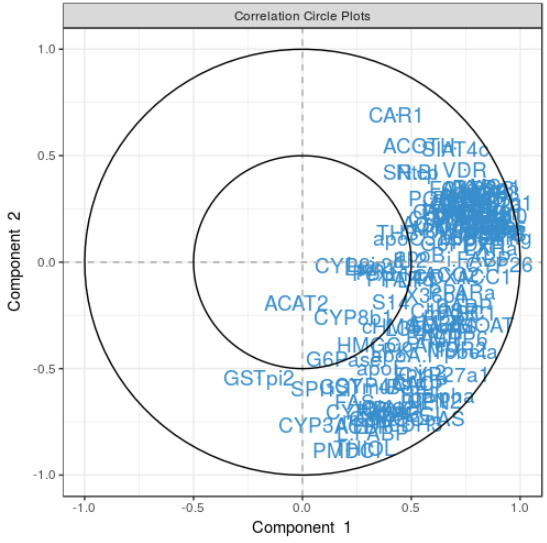
PCA first!

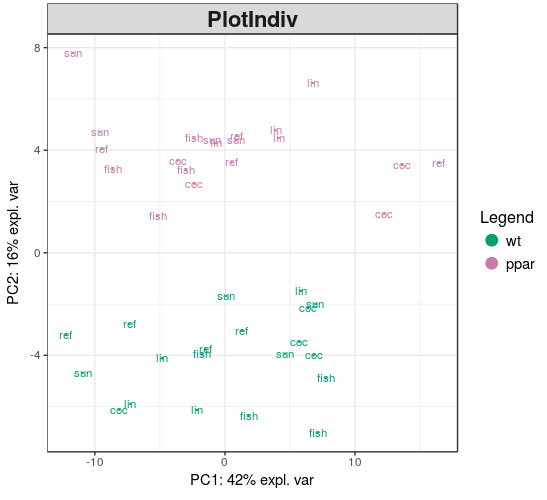
Lipids





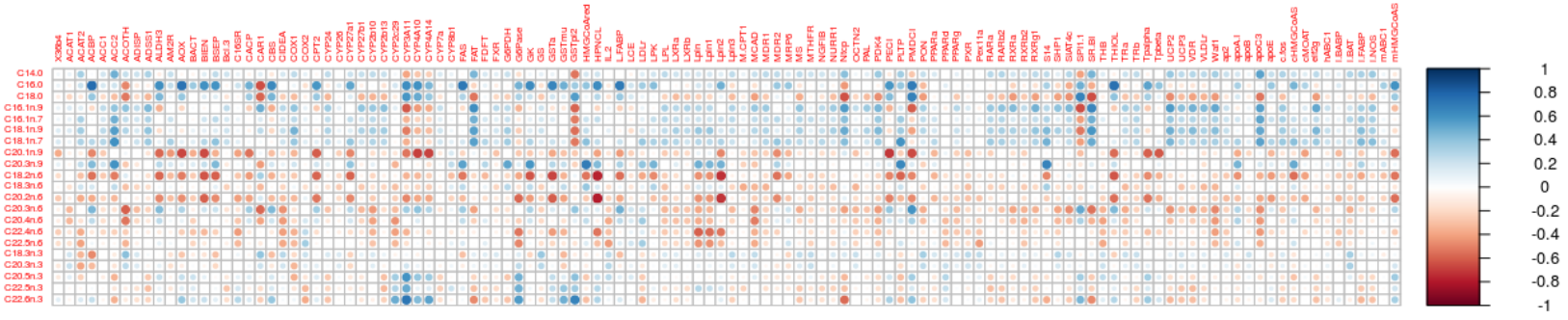
Genes





Relationships between lipids and genes?

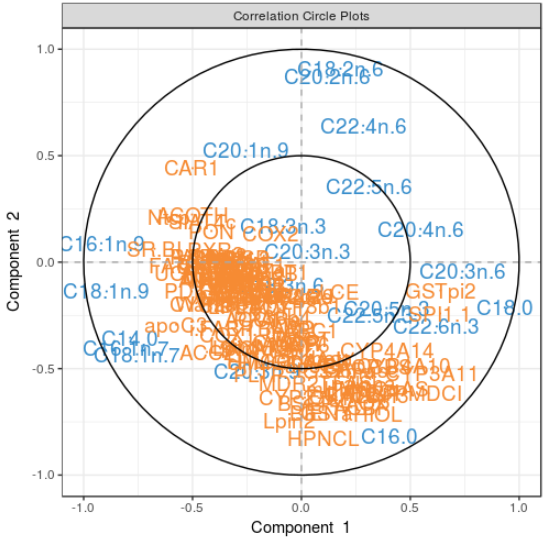
Pairwise correlations

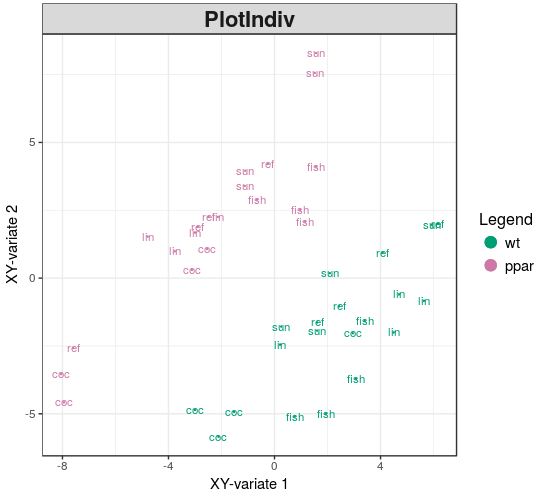


Package `corrplot`

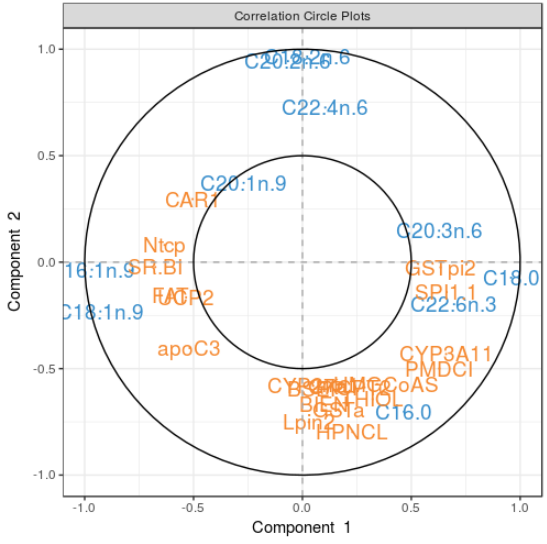
PLS - SPLS

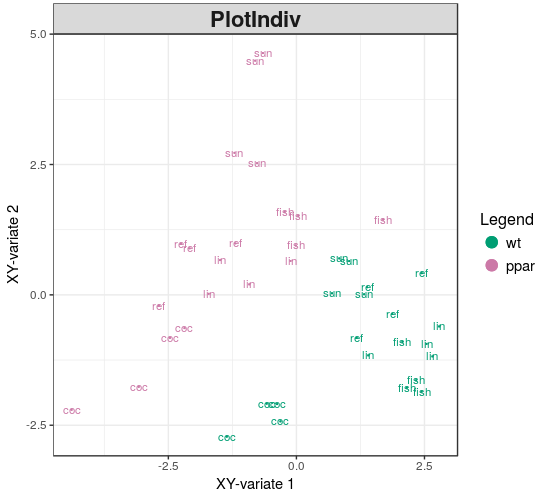
PLS



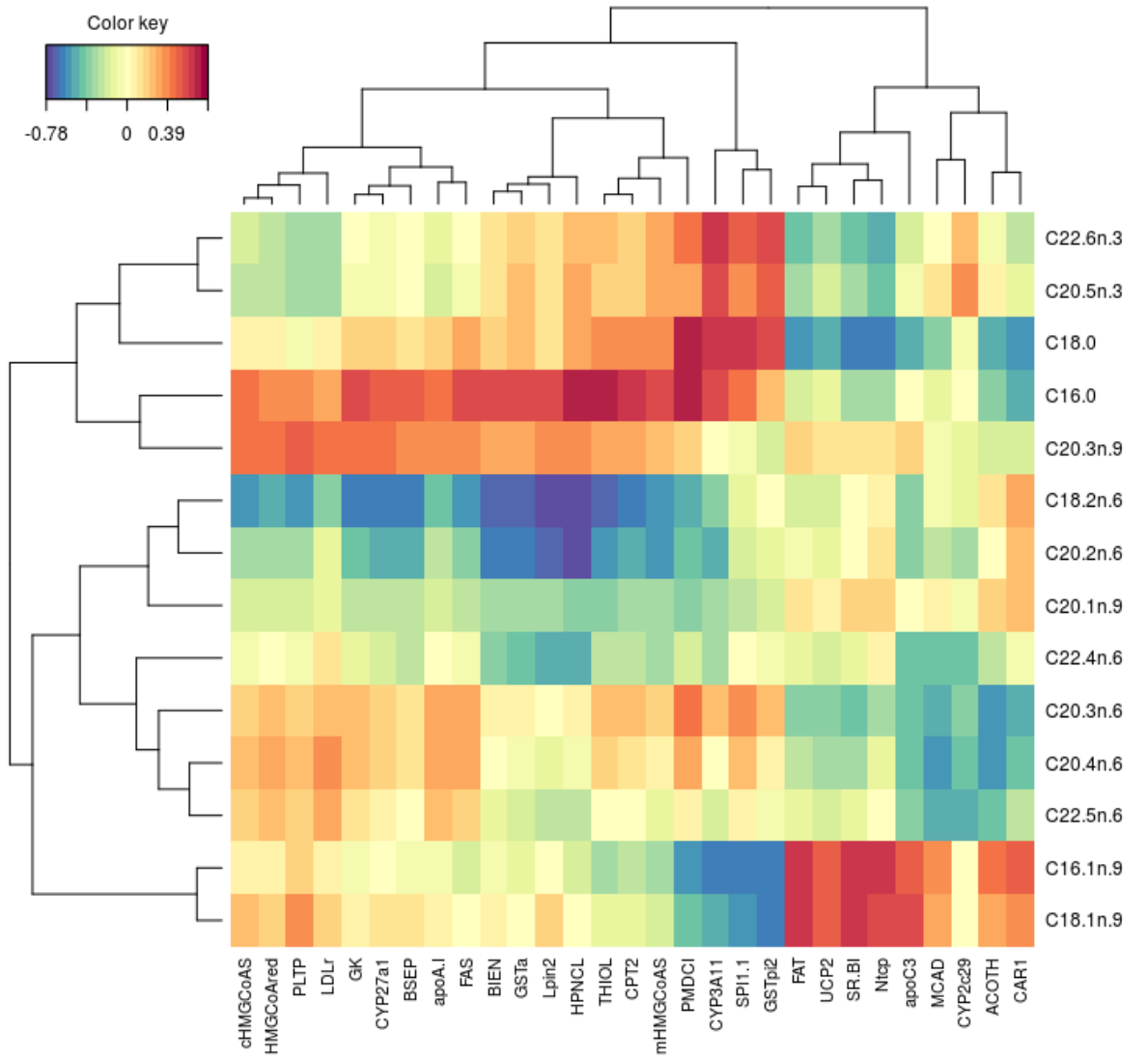


SPLS



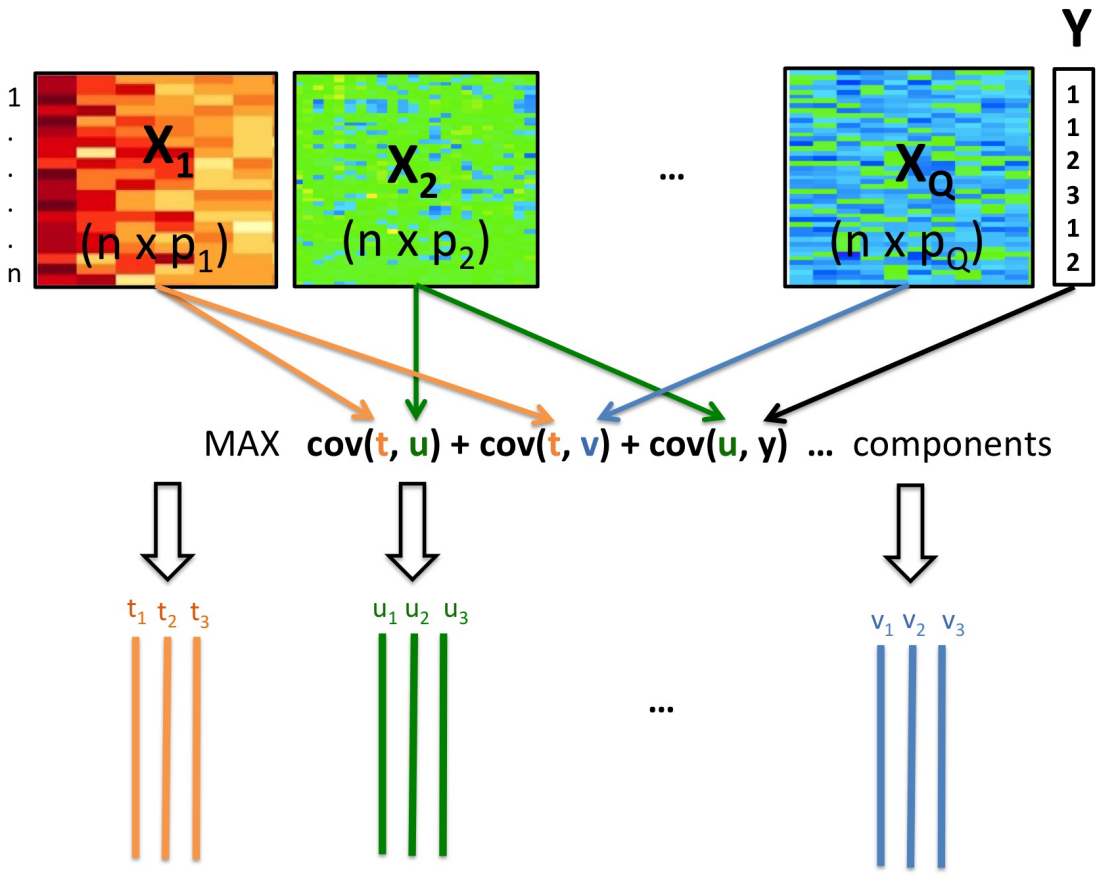


Variable representation



$N-$integration: a set of component **per** data set

Block-PLSDA maximises the (weighted) **sum of covariances** between each pair of data sets and an outcome



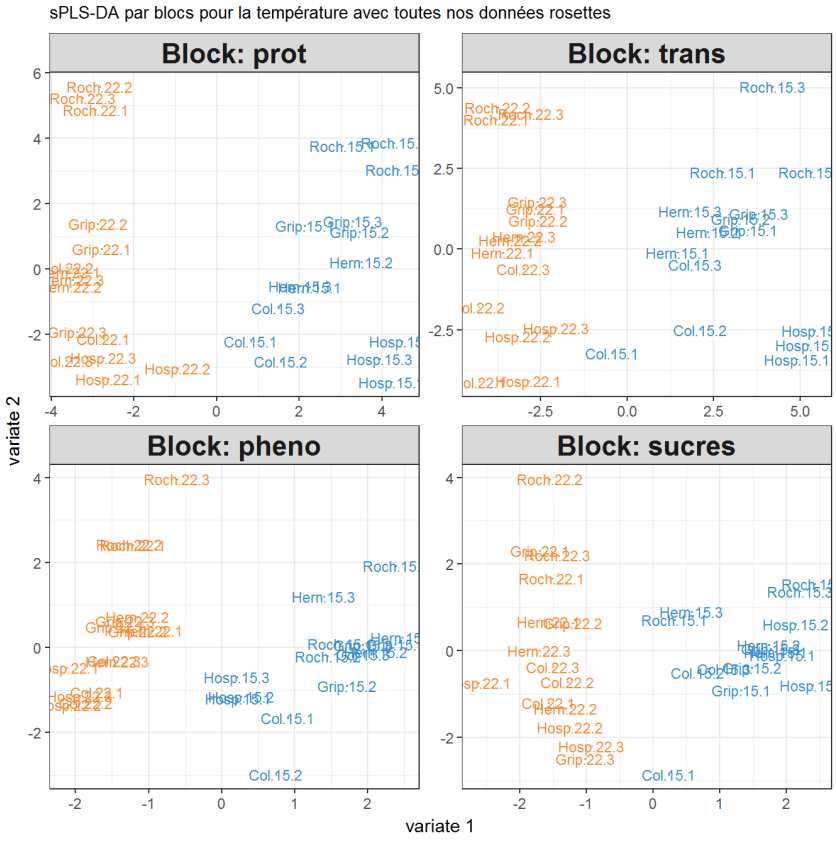
Example: `Wallomics` data set

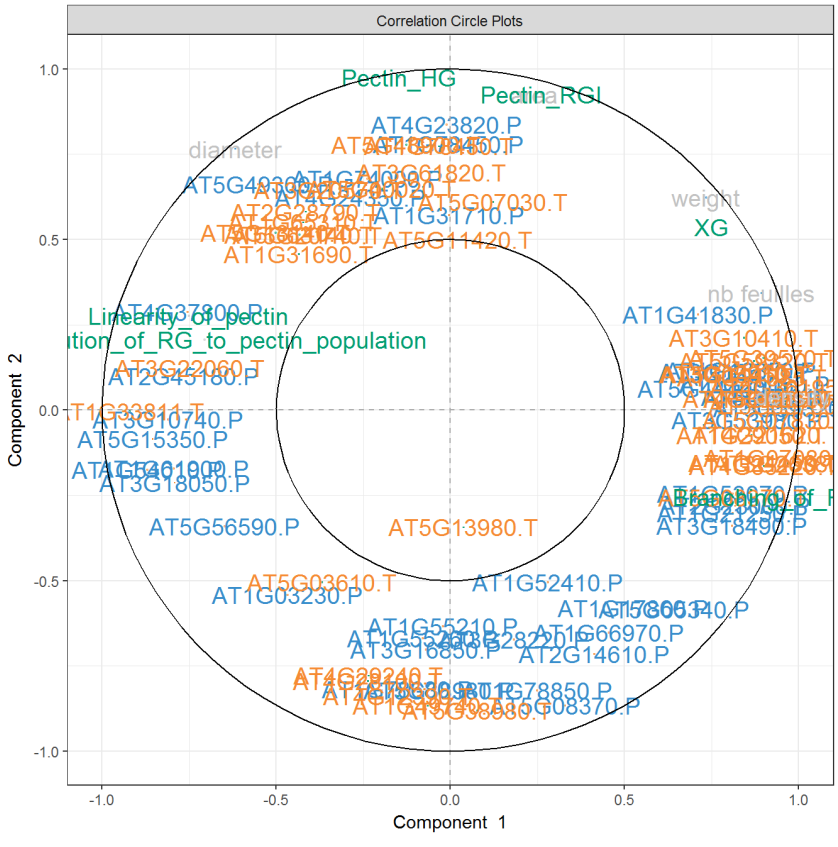
- **30** samples: **5** ecotypes (`Roch, Grip, Hern, Hosp`) x **2** temperatures x **3** replicates
- **4** data sets: phenomics (9), metabolomics (7), proteomics ($\sim$ 400), transcriptomics ($\sim$ 20000)

H. Duruflé, M. Selmani, P. Ranocha, E. Jamet, C. Dunand, S. Déjean (2018). A powerful framework for an integrative study with heterogeneous omics data: from univariate statistics to multi-block analysis, doi: https://doi.org/10.1101/35792, bioRxiv

Example: a supervised sparse multi-block analysis

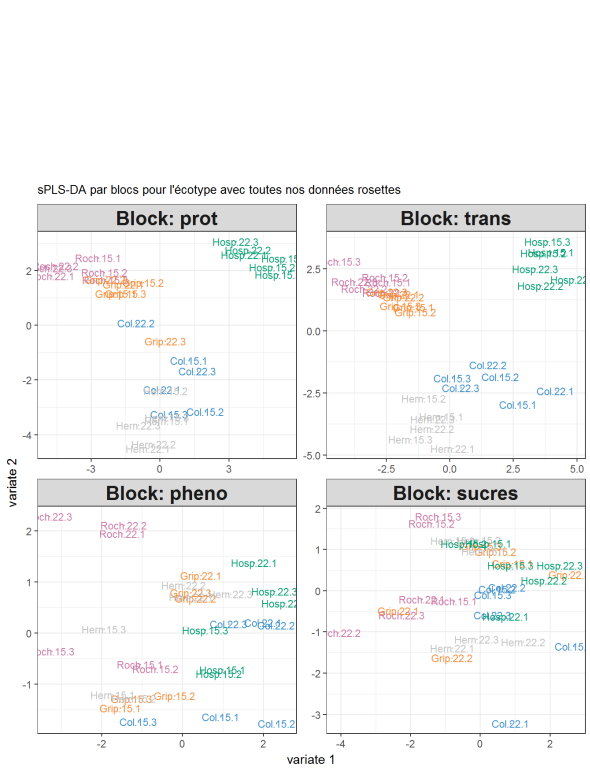
Temperature

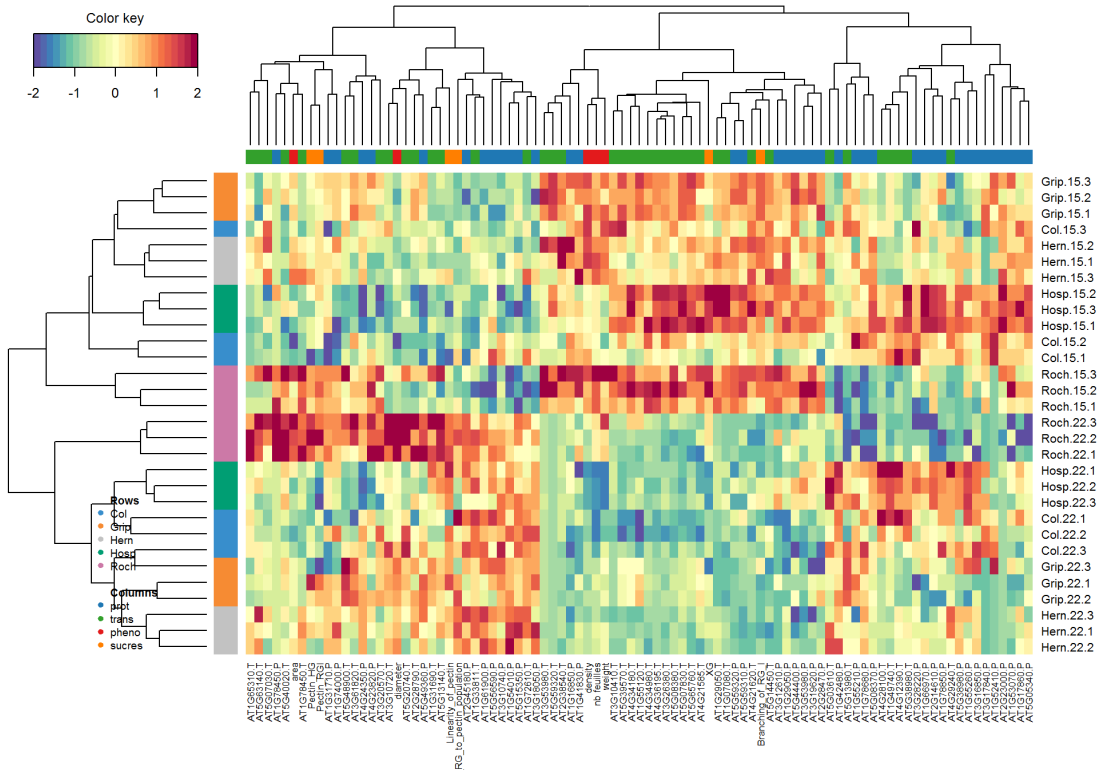




Example: a supervised sparse multi-block analysis

Ecotype





To put it in a nutshell

- Multivariate linear methods enables to answer a wide range of biological questions: data exploration, classification, integration of multiple data sets
- Variable selection (sparse)

Principles

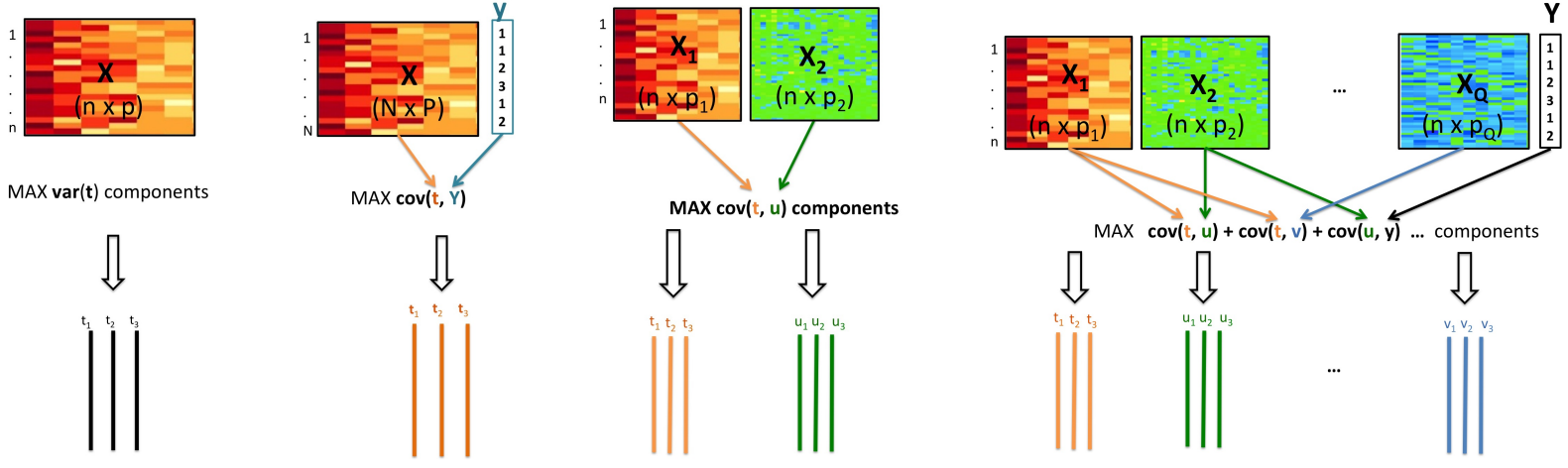
PCA $\max var(aX) \rightarrow a?$

PLS $\max cov(aX, bY) \rightarrow a, b?$

CCA $\max cor(aX, bY) \rightarrow a, b?$

PLSDA $\rightarrow$ PLS

Multi-blocks $\max \sum cov(a_i X_i, b_j X_j) \rightarrow a_i, b_i?$



Practical session

1. Run the method: `MyResult <- pca(X)`
2. Represent individuals: `plotIndiv(MyResult)`
3. Represent variables: `plotVar(MyResult)`

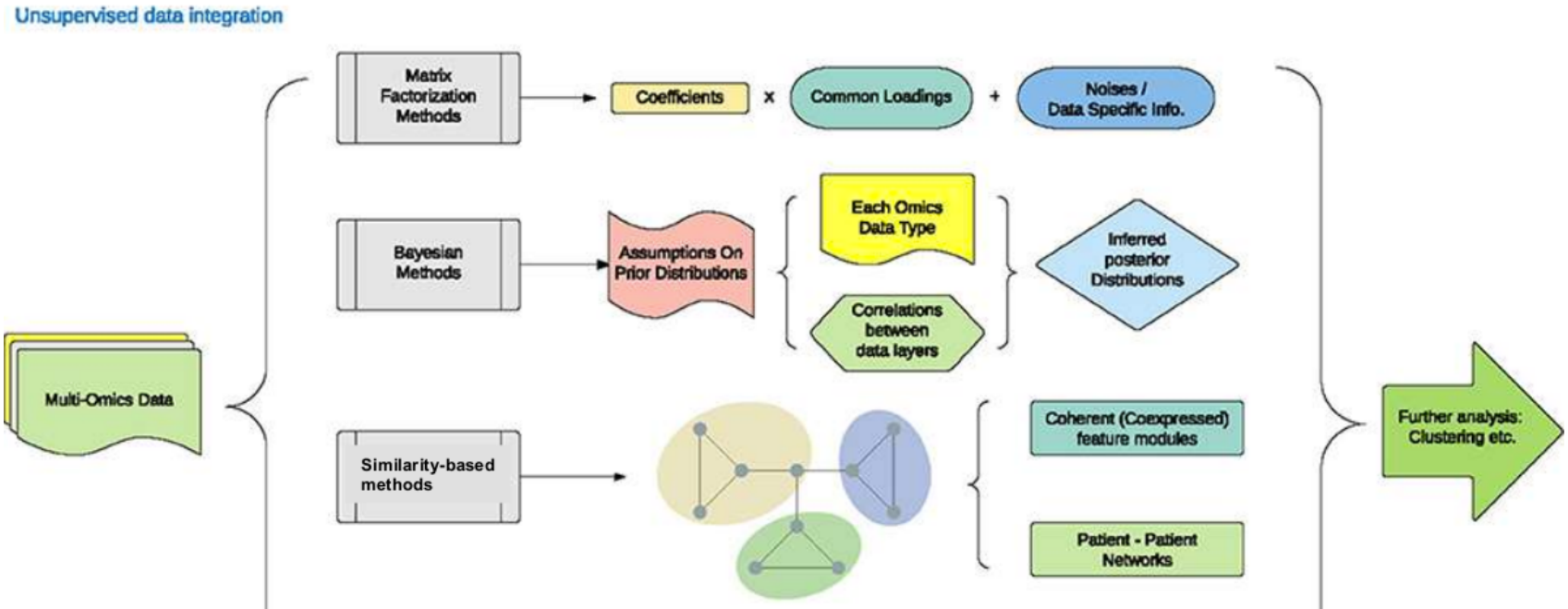
X. Read the help files: `?pca, ?plotIndiv, ?plotVar...`

Multivariate methods

Kernel methods

Conclusion

Kernel methods



Examples:

Network-based methods:

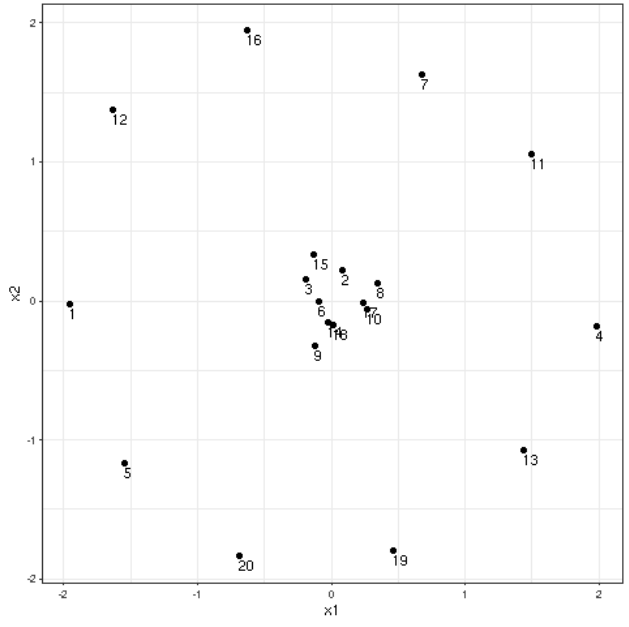
- Similarity between macromolecules: Anais Baudot tomorrow
- Similarity between samples: Similarity Network Fusion SNF

Kernel methods: practical session Jérôme Mariette

Prerequisites: dot product

	x_1	x_2
1	-1.96	-0.02
2	0.08	0.22
3	-0.19	0.16
4	1.98	-0.19
5	-1.55	-1.17
6	-0.09	-0.00
7	0.68	1.62
8	0.35	0.13
9	-0.12	-0.32
10	0.26	-0.06
11	1.50	1.05
12	-1.63	1.38
13	1.44	-1.08
14	-0.02	-0.15
15	-0.13	0.33
16	-0.63	1.95
17	0.24	-0.02
18	0.02	-0.18
19	0.46	-1.80
20	-0.68	-1.84

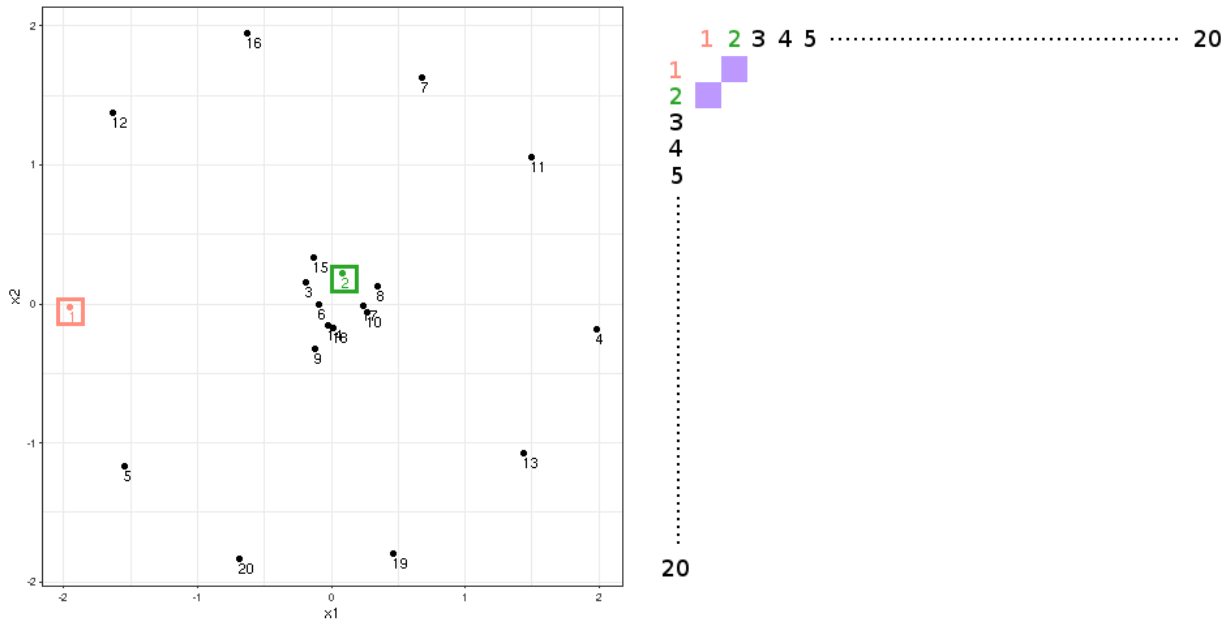
$$K_{ij} = x_1^i x_1^j + x_2^i x_2^j$$



Prerequisites: dot product

	x_1	x_2
1	-1.96	-0.02
2	0.08	0.22
3	-0.19	0.16
4	1.98	-0.19
5	-1.55	-1.17
6	-0.09	-0.00
7	0.68	1.62
8	0.35	0.13
9	-0.12	-0.32
10	0.26	-0.06
11	1.50	1.05
12	-1.63	1.38
13	1.44	-1.08
14	-0.02	-0.15
15	-0.13	0.33
16	-0.63	1.95
17	0.24	-0.02
18	0.02	-0.18
19	0.46	-1.80
20	-0.68	-1.84

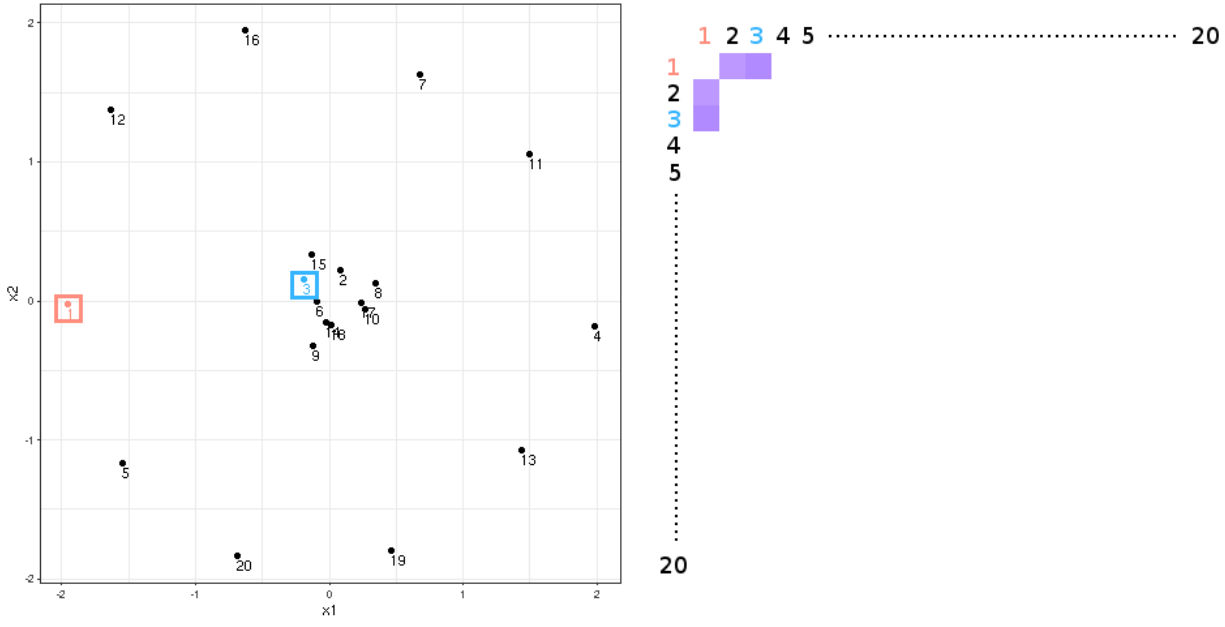
$$K_{12} = -1.96 \times 0.08 + (-0.02) \times 0.22 = \text{-0.16}$$



Prerequisites: dot product

	x_1	x_2
1	-1.96	-0.02
2	0.08	0.22
3	-0.19	0.16
4	1.98	-0.19
5	-1.55	-1.17
6	-0.09	-0.00
7	0.68	1.62
8	0.35	0.13
9	-0.12	-0.32
10	0.26	-0.06
11	1.50	1.05
12	-1.63	1.38
13	1.44	-1.08
14	-0.02	-0.15
15	-0.13	0.33
16	-0.63	1.95
17	0.24	-0.02
18	0.02	-0.18
19	0.46	-1.80
20	-0.68	-1.84

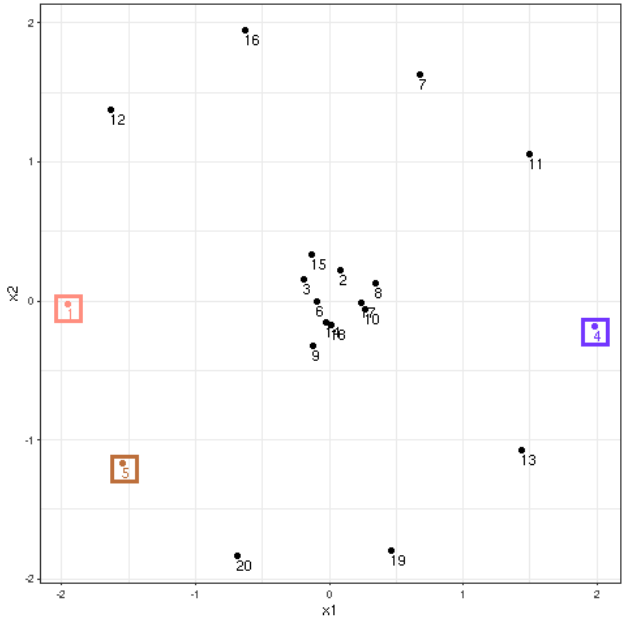
$$K_{13} = -1.96 \times (-0.19) + (-0.02) \times 0.16 = 0.37$$

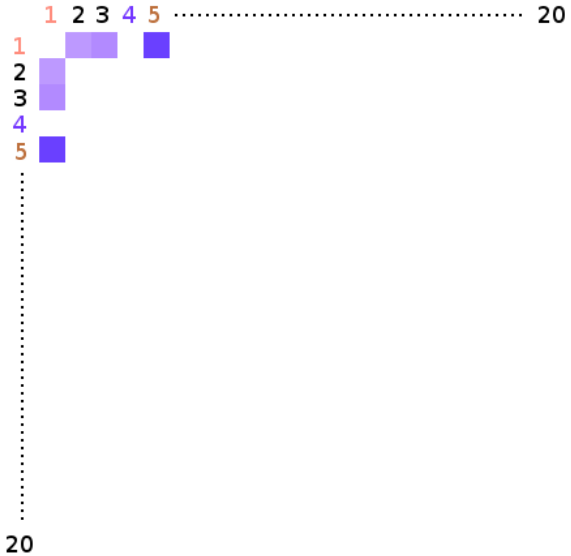


Prerequisites: dot product

	x_1	x_2
1	-1.96	-0.02
2	0.08	0.22
3	-0.19	0.16
4	1.98	-0.19
5	-1.55	-1.17
6	-0.09	-0.00
7	0.68	1.62
8	0.35	0.13
9	-0.12	-0.32
10	0.26	-0.06
11	1.50	1.05
12	-1.63	1.38
13	1.44	-1.08
14	-0.02	-0.15
15	-0.13	0.33
16	-0.63	1.95
17	0.24	-0.02
18	0.02	-0.18
19	0.46	-1.80
20	-0.68	-1.84

$$K_{14} = -1.96 \times 1.98 + (-0.02) \times (-0.19) = -3.88$$
$$K_{15} = -1.96 \times (-1.55) + (-0.02) \times (-1.17) = 3.06$$

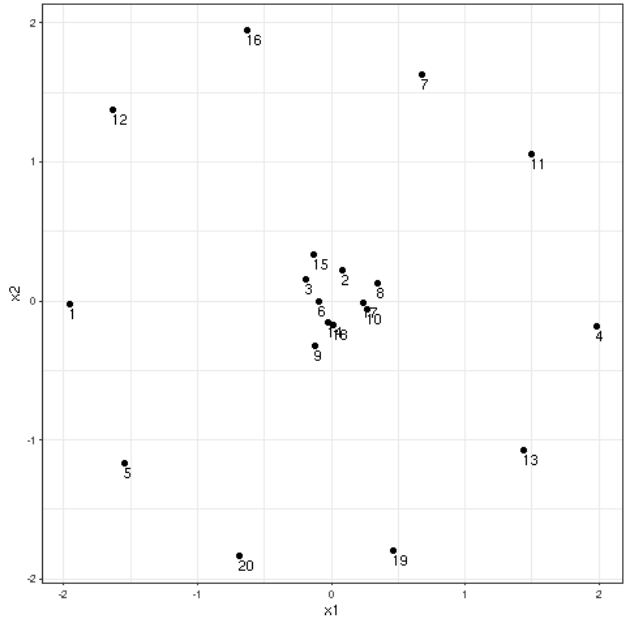


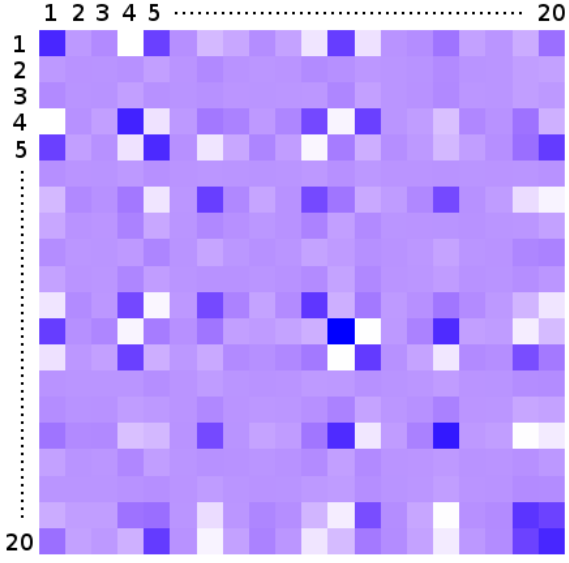


Prerequisites: dot product

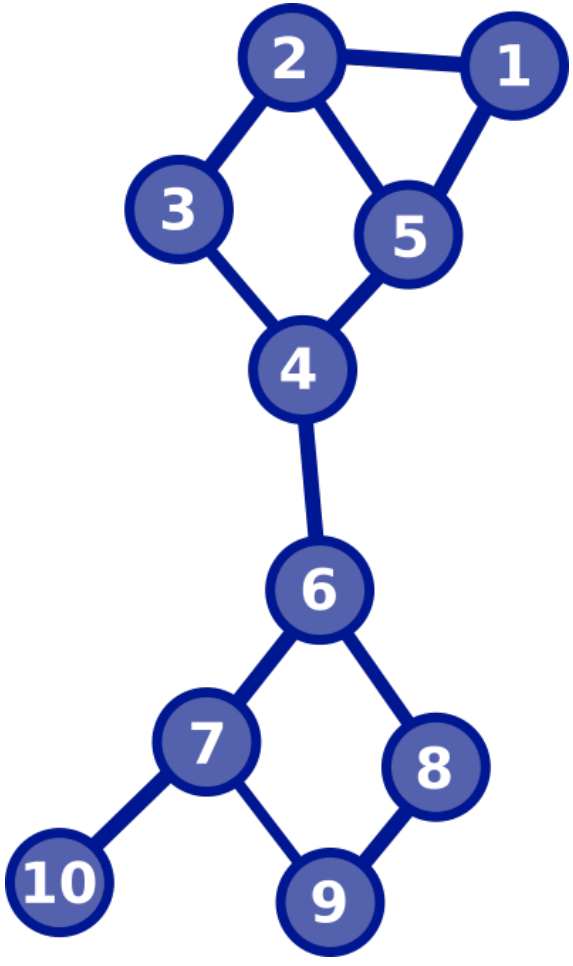
	x_1	x_2
1	-1.96	-0.02
2	0.08	0.22
3	-0.19	0.16
4	1.98	-0.19
5	-1.55	-1.17
6	-0.09	-0.00
7	0.68	1.62
8	0.35	0.13
9	-0.12	-0.32
10	0.26	-0.06
11	1.50	1.05
12	-1.63	1.38
13	1.44	-1.08
14	-0.02	-0.15
15	-0.13	0.33
16	-0.63	1.95
17	0.24	-0.02
18	0.02	-0.18
19	0.46	-1.80
20	-0.68	-1.84

$K = xx^T$ is a kernel : linear kernel





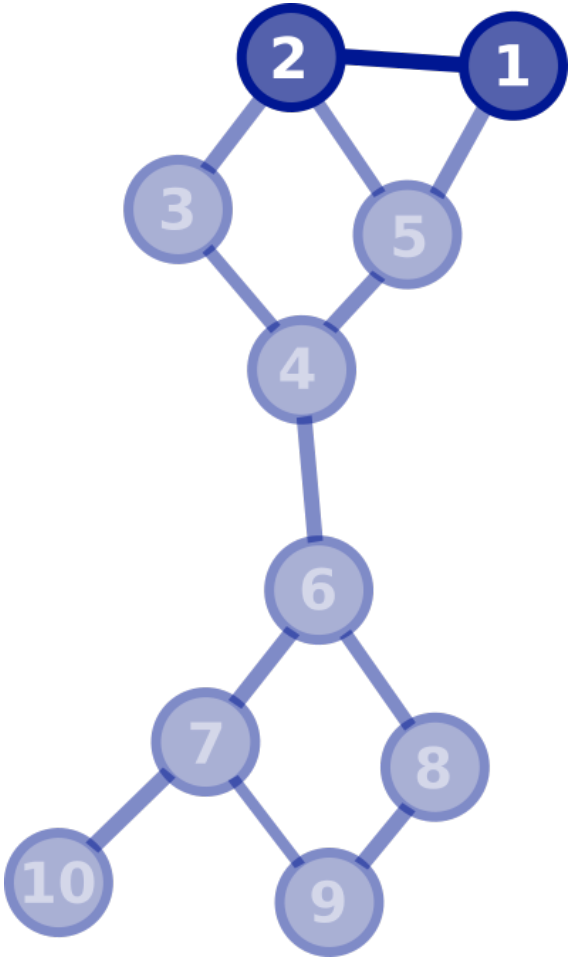
Prerequisites: dissimilarity measure



Shortest-Path dissimilarity

	1	2	3	4	5	6	7	8	9	10
1										
2										
3										
4										
5										
6										
7										
8										
9										
10										

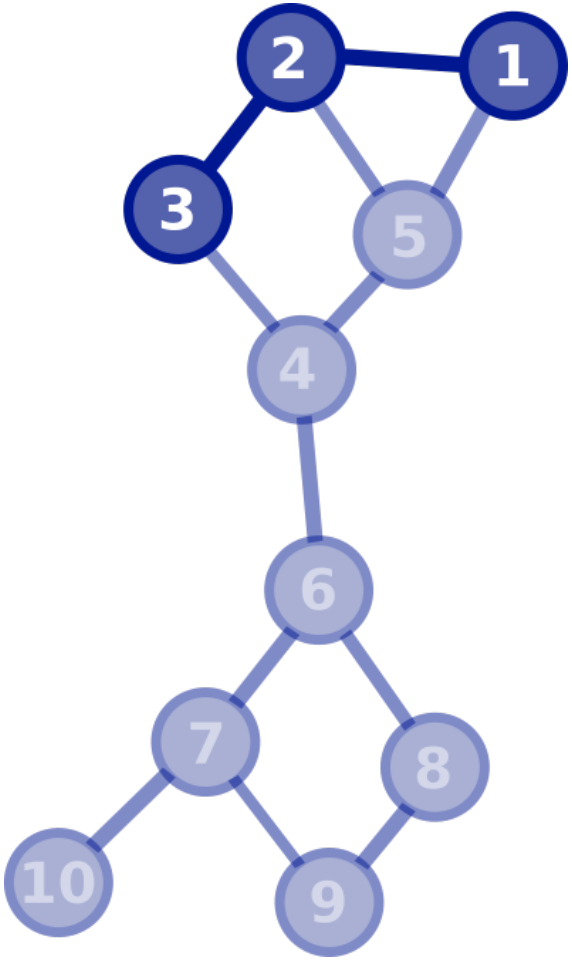
Prerequisites: dissimilarity measure



Shortest-Path dissimilarity

	1	2	3	4	5	6	7	8	9	10
1		1								
2	1									
3										
4										
5										
6										
7										
8										
9										
10										

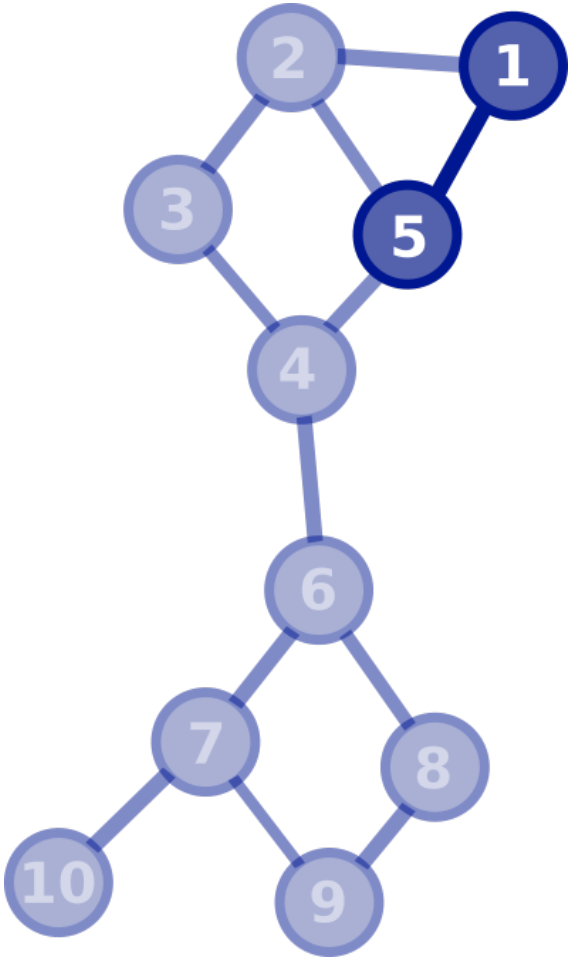
Prerequisites: dissimilarity measure



Shortest-Path dissimilarity

	1	2	3	4	5	6	7	8	9	10
1		1	2							
2	1									
3	2									
4										
5										
6										
7										
8										
9										
10										

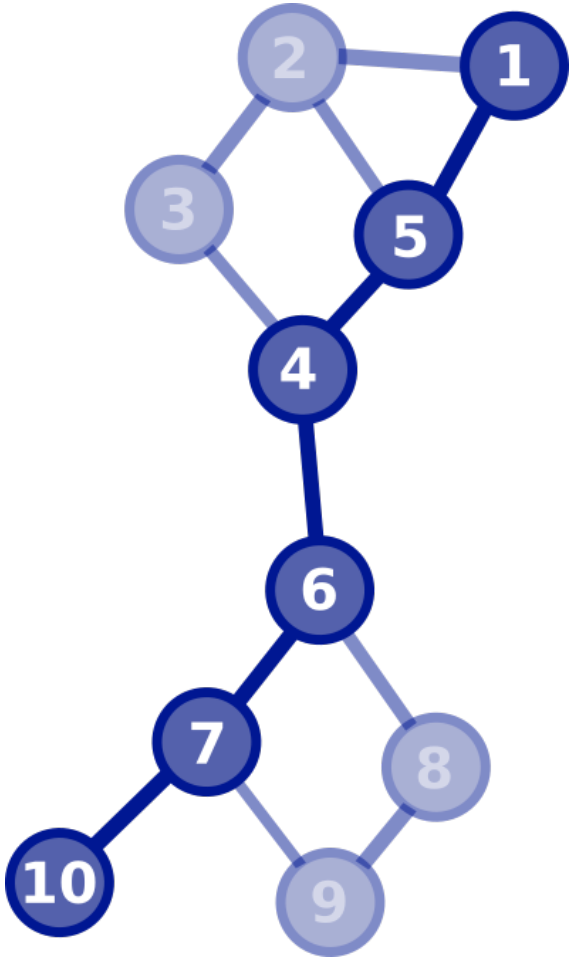
Prerequisites: dissimilarity measure



Shortest-Path dissimilarity

	1	2	3	4	5	6	7	8	9	10
1		1	2	2	1					
2	1									
3	2									
4	2									
5	1									
6										
7										
8										
9										
10										

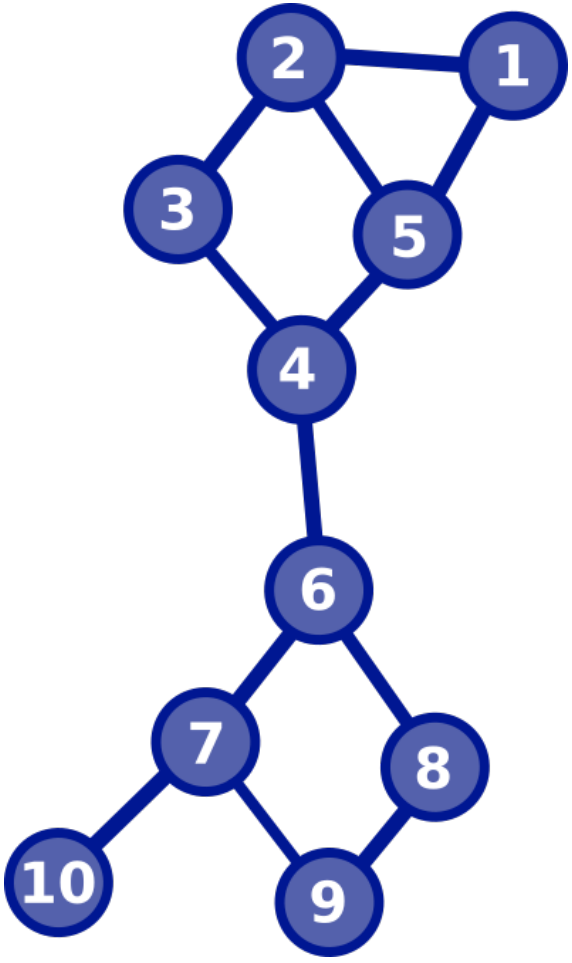
Prerequisites: dissimilarity measure



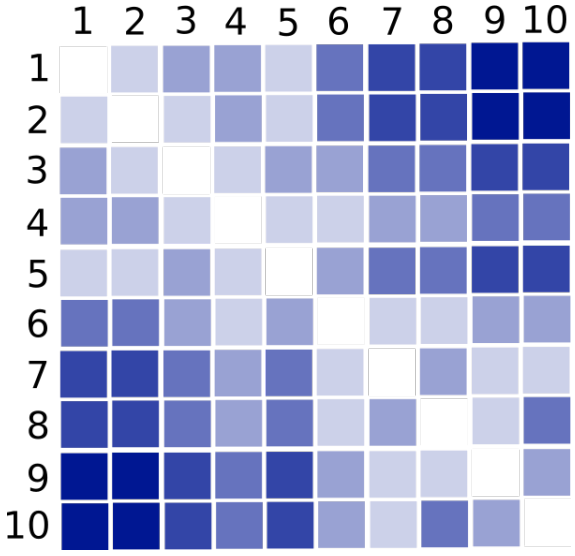
Shortest-Path dissimilarity

	1	2	3	4	5	6	7	8	9	10
1		1	2	2	1	3	4	4	5	5
2	1									
3	2									
4	2									
5	1									
6	3									
7	4									
8	4									
9	5									
10	5									

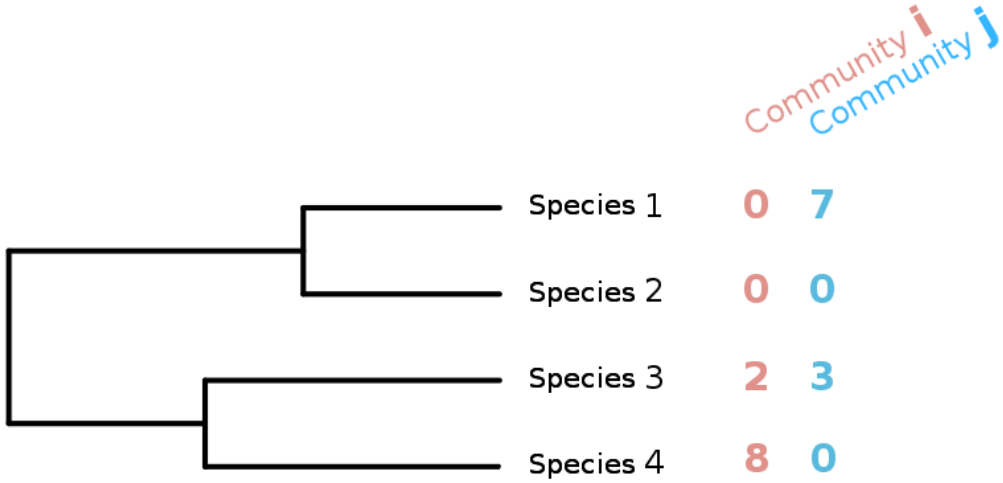
Prerequisites: dissimilarity measure



Shortest-Path dissimilarity



Prerequisites: dissimilarity measure



Phylogenetic kernel

- Based on the UniFrac distance [Lozupone and Knight, 2005] ;
- Diversity fraction specific to community i and j weighted by the evolution distance between species:

$$d_{UF}(x_i, x_j) = \frac{\sum_{b=1}^{B} l_b \left(\mathbb{I}_{\{r_{ib}>0,\, r_{jb}=0\}} + \mathbb{I}_{\{r_{jb}>0,\, r_{ib}=0\}} \right)}{\sum_{b=1}^{B} l_b \mathbb{I}_{\{r_{ib}+r_{jb}>0\}}}$$

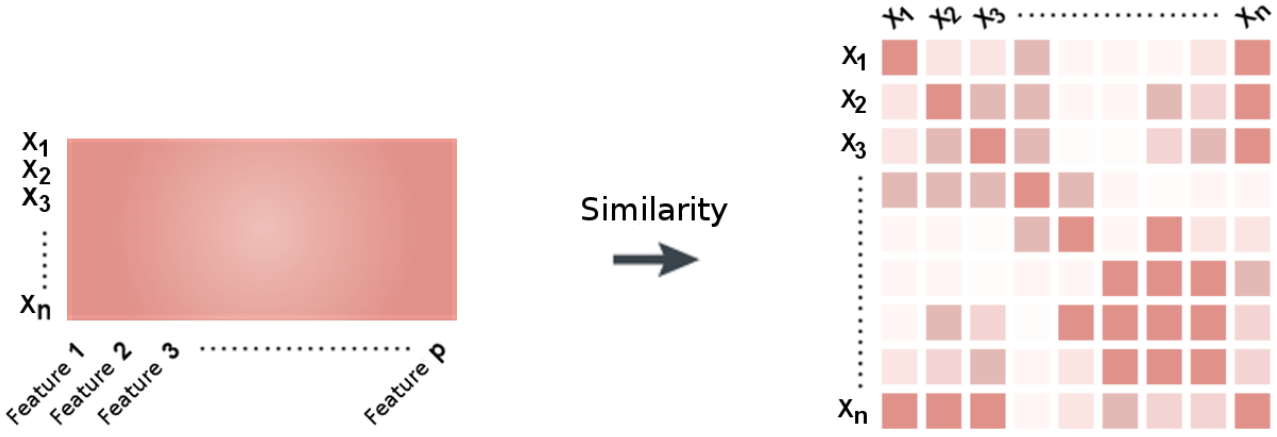
Prerequisites: dissimilarity measure

Phylogenetic kernel

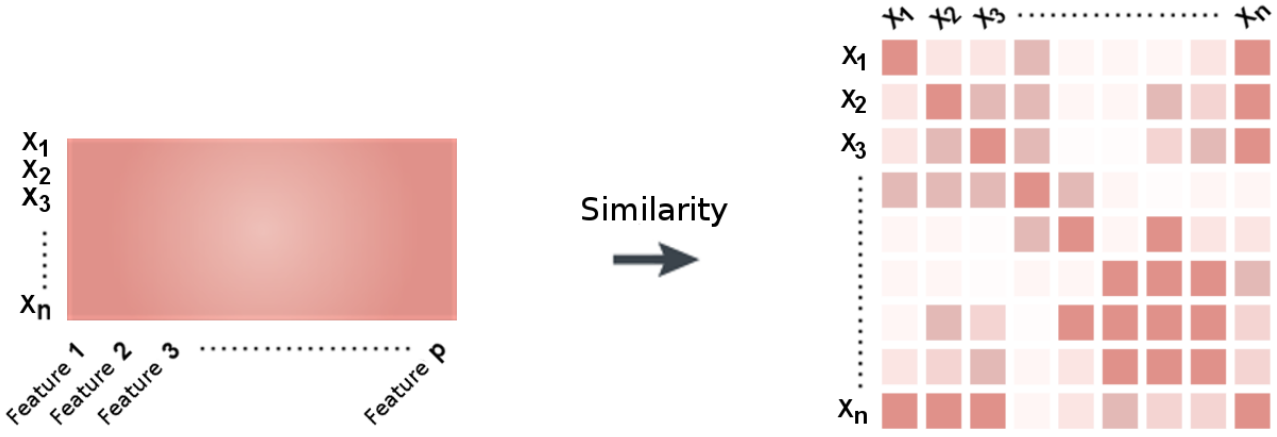
- Based on the UniFrac distance [Lozupone and Knight, 2005] ;
- Diversity fraction specific to community i and j weighted by the evolution distance between species:

$$d_{UF}(x_i, x_j) = \frac{\sum_{b=1}^{B} l_b (\mathbb{I}_{\{r_{ib}>0,\, r_{jb}=0\}} + \mathbb{I}_{\{r_{jb}>0,\, r_{ib}=0\}})}{\sum_{b=1}^{B} l_b \mathbb{I}_{\{r_{ib}+r_{jb}>0\}}}$$

Prerequisites: kernels



Prerequisites: kernels

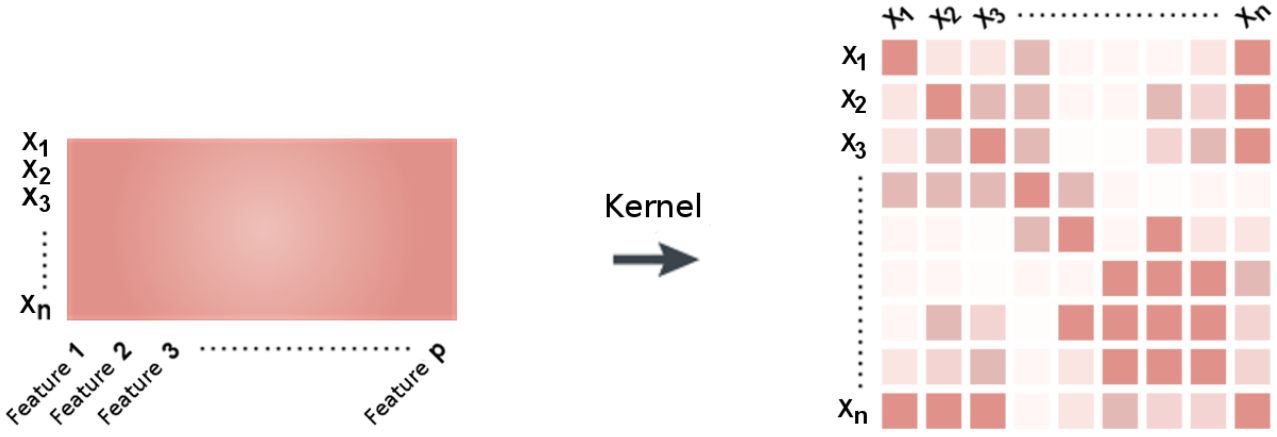


Desired mathematical properties for the similarity

Function $K : \mathcal{G} \times \mathcal{G} \to \mathbb{R}$ st:

- symmetry: $K(x_i, x_j) = K(x_j, x_i)$;
- and positivity: $\forall m \in \mathbb{N}$, $\forall x_1, ..., x_m \in \mathcal{G}$, $\forall \alpha_1, ..., \alpha_m \in \mathbb{R}$, $\sum_{i,j=1}^{m} \alpha_i \alpha_j K(x_i, x_j) \geq 0$;

Prerequisites: kernels

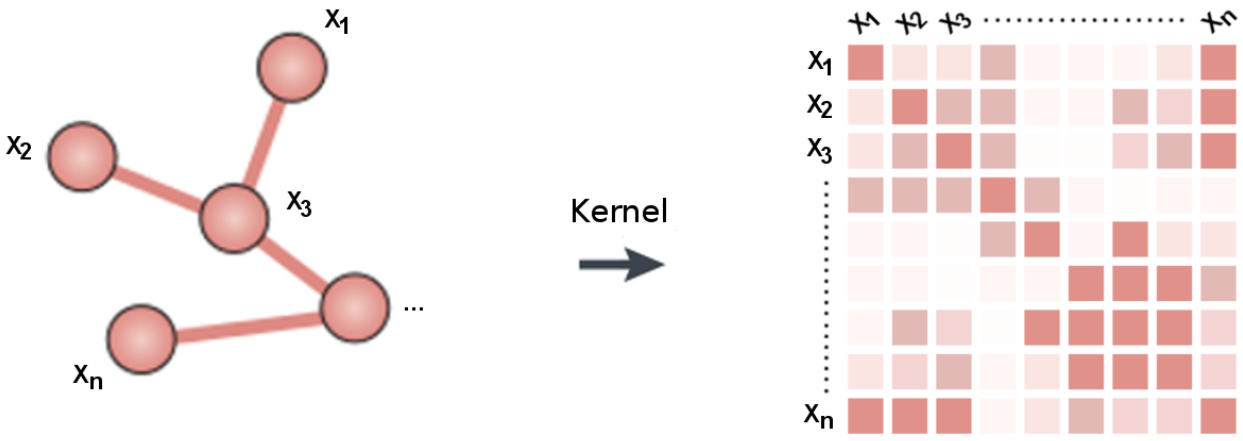


Desired mathematical properties for the similarity

Function $K : \mathcal{G} \times \mathcal{G} \to \mathbb{R}$ st:

- symmetry: $K(x_i, x_j) = K(x_j, x_i)$;
- and positivity: $\forall m \in \mathbb{N}$, $\forall x_1, ..., x_m \in \mathcal{G}$, $\forall \alpha_1, ..., \alpha_m \in \mathbb{R}$, $\sum_{i,j=1}^{m} \alpha_i \alpha_j K(x_i, x_j) \geq 0$;

Prerequisites: kernels

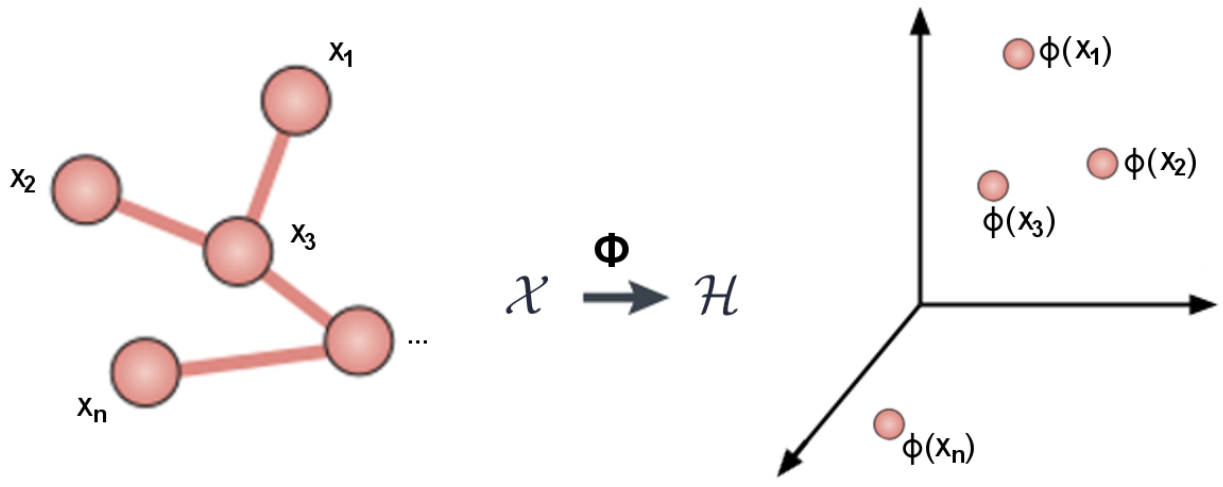


Desired mathematical properties for the similarity

Function $K : \mathcal{G} \times \mathcal{G} \to \mathbb{R}$ st:

- symmetry: $K(x_i, x_j) = K(x_j, x_i)$;
- and positivity: $\forall m \in \mathbb{N}$, $\forall x_1, ..., x_m \in \mathcal{G}$, $\forall \alpha_1, ..., \alpha_m \in \mathbb{R}$, $\sum_{i,j=1}^{m} \alpha_i \alpha_j K(x_i, x_j) \geq 0$;

Prerequisites: kernels

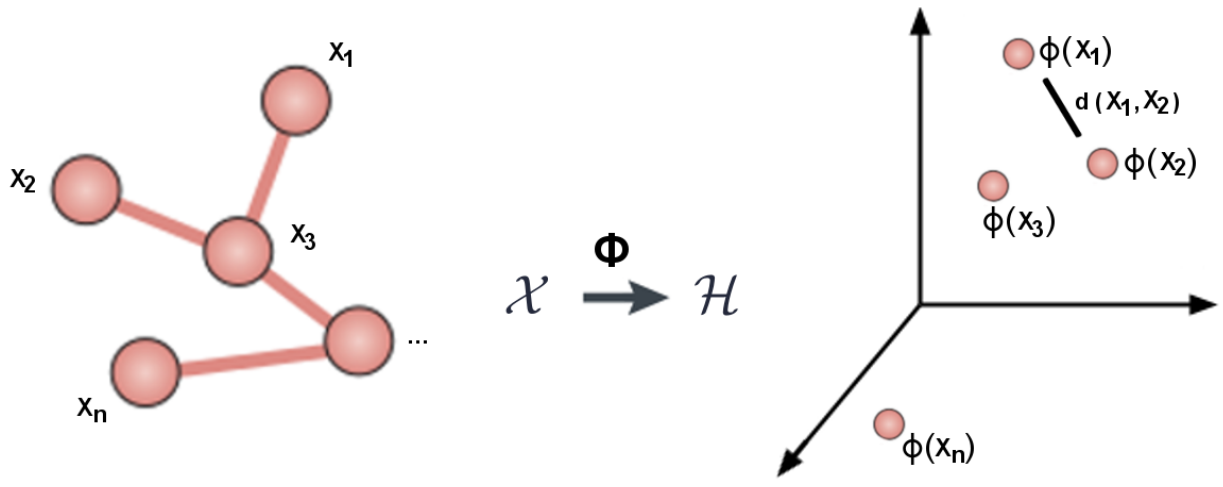


Desired mathematical properties for the similarity

Function $K : \mathcal{G} \times \mathcal{G} \to \mathbb{R}$ st:

- symmetry: $K(x_i, x_j) = K(x_j, x_i)$;
- and positivity: $\forall m \in \mathbb{N}$, $\forall x_1, ..., x_m \in \mathcal{G}$, $\forall \alpha_1, ..., \alpha_m \in \mathbb{R}$, $\sum_{i,j=1}^{m} \alpha_i \alpha_j K(x_i, x_j) \geq 0$;

Prerequisites: kernels



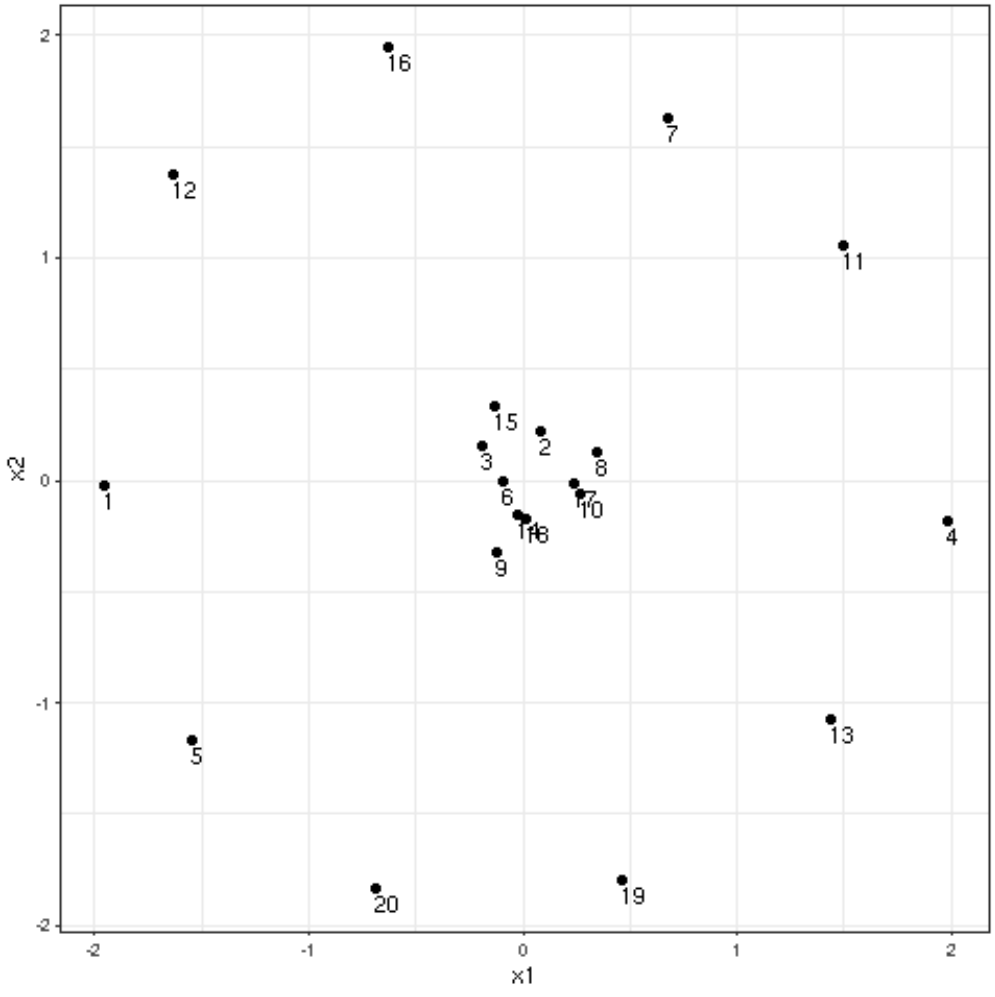
Desired mathematical properties for the similarity

Function $K : \mathcal{G} \times \mathcal{G} \to \mathbb{R}$ st:

- symmetry: $K(x_i, x_j) = K(x_j, x_i)$;
- and positivity: $\forall m \in \mathbb{N}$, $\forall x_1, ..., x_m \in \mathcal{G}$, $\forall \alpha_1, ..., \alpha_m \in \mathbb{R}$, $\sum_{i,j=1}^{m} \alpha_i \alpha_j K(x_i, x_j) \geq 0$;

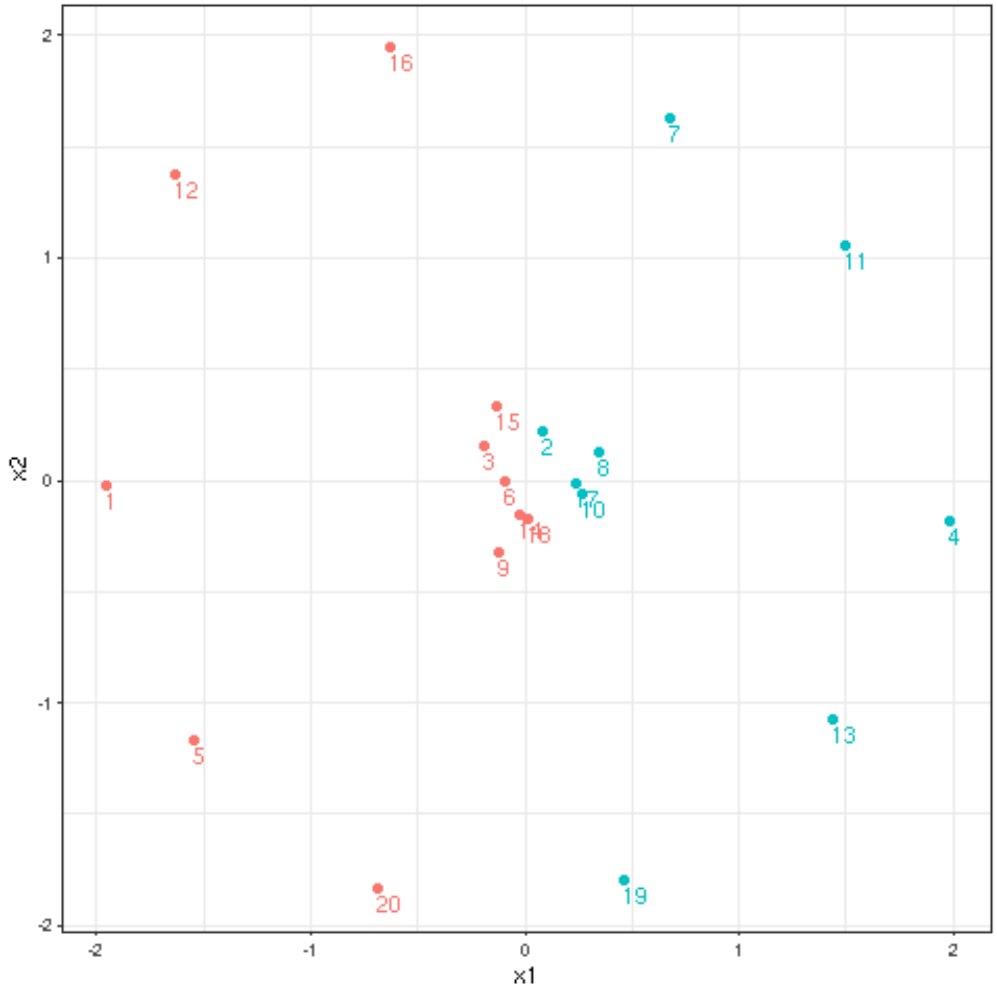
Prerequisites: kernels

	x_1	x_2
1	-1.96	-0.02
2	0.08	0.22
3	-0.19	0.16
4	1.98	-0.19
5	-1.55	-1.17
6	-0.09	-0.00
7	0.68	1.62
8	0.35	0.13
9	-0.12	-0.32
10	0.26	-0.06
11	1.50	1.05
12	-1.63	1.38
13	1.44	-1.08
14	-0.02	-0.15
15	-0.13	0.33
16	-0.63	1.95
17	0.24	-0.02
18	0.02	-0.18
19	0.46	-1.80
20	-0.68	-1.84



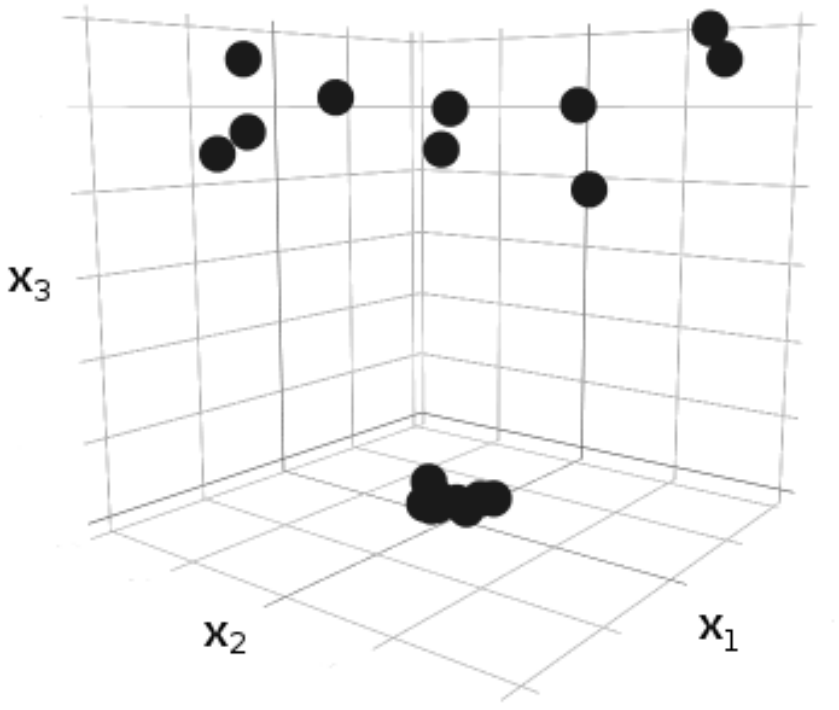
Prerequisites: kernels

	x_1	x_2
1	-1.96	-0.02
2	0.08	0.22
3	-0.19	0.16
4	1.98	-0.19
5	-1.55	-1.17
6	-0.09	-0.00
7	0.68	1.62
8	0.35	0.13
9	-0.12	-0.32
10	0.26	-0.06
11	1.50	1.05
12	-1.63	1.38
13	1.44	-1.08
14	-0.02	-0.15
15	-0.13	0.33
16	-0.63	1.95
17	0.24	-0.02
18	0.02	-0.18
19	0.46	-1.80
20	-0.68	-1.84



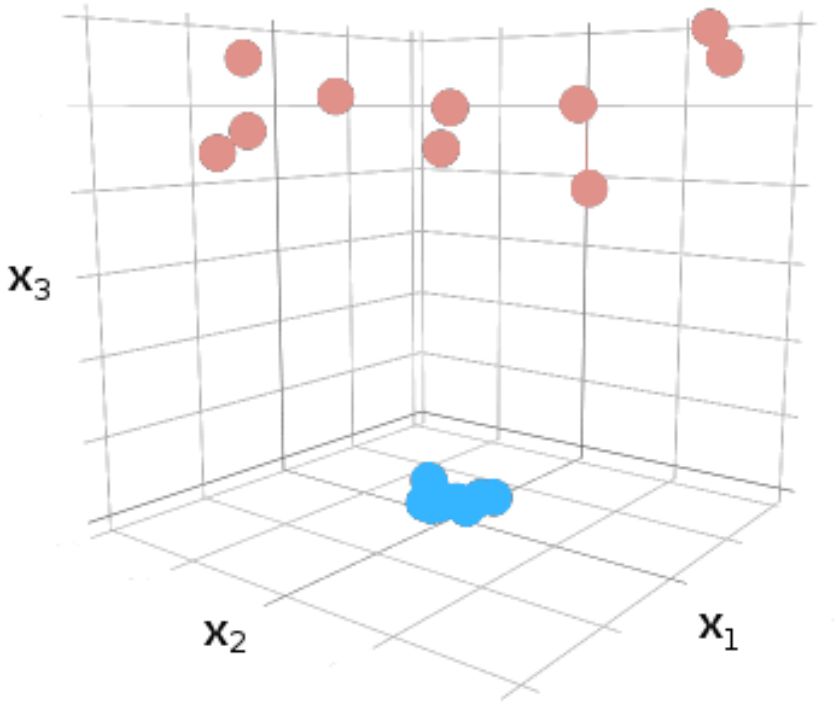
Prerequisites: kernels

	x_1	x_2	$x_3 = x_1^2 + x_2^2$
1	-1.96	-0.02	3.83
2	0.08	0.22	0.05
3	-0.19	0.16	0.06
4	1.98	-0.19	3.96
5	-1.55	-1.17	3.77
6	-0.09	-0.00	0.01
7	0.68	1.62	3.11
8	0.35	0.13	0.14
9	-0.12	-0.32	0.12
10	0.26	-0.06	0.08
11	1.50	1.05	3.36
12	-1.63	1.38	4.55
13	1.44	-1.08	3.23
14	-0.02	-0.15	0.02
15	-0.13	0.33	0.13
16	-0.63	1.95	4.19
17	0.24	-0.02	0.06
18	0.02	-0.18	0.03
19	0.46	-1.80	3.45
20	-0.68	-1.84	3.85



Prerequisites: kernels

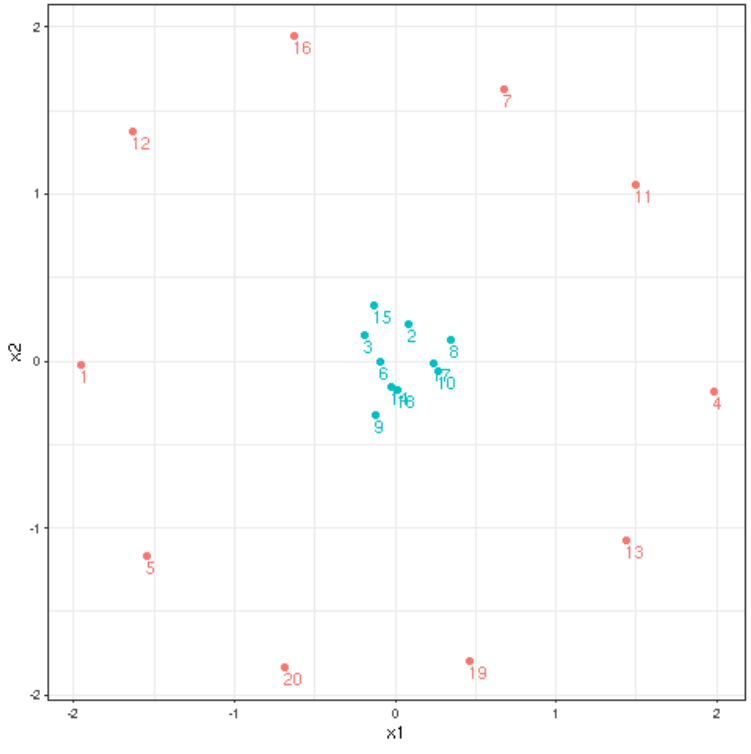
	x_1	x_2	$x_3 = x_1^2 + x_2^2$
1	-1.96	-0.02	3.83
2	0.08	0.22	0.05
3	-0.19	0.16	0.06
4	1.98	-0.19	3.96
5	-1.55	-1.17	3.77
6	-0.09	-0.00	0.01
7	0.68	1.62	3.11
8	0.35	0.13	0.14
9	-0.12	-0.32	0.12
10	0.26	-0.06	0.08
11	1.50	1.05	3.36
12	-1.63	1.38	4.55
13	1.44	-1.08	3.23
14	-0.02	-0.15	0.02
15	-0.13	0.33	0.13
16	-0.63	1.95	4.19
17	0.24	-0.02	0.06
18	0.02	-0.18	0.03
19	0.46	-1.80	3.45
20	-0.68	-1.84	3.85



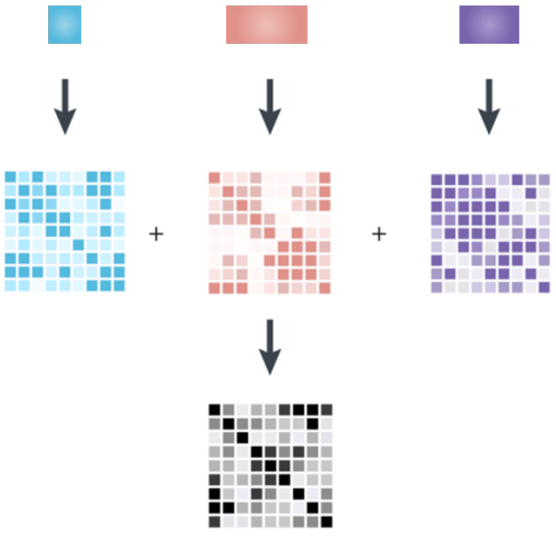
Prerequisites: kernels

	x_1	x_2
1	-1.96	-0.02
2	0.08	0.22
3	-0.19	0.16
4	1.98	-0.19
5	-1.55	-1.17
6	-0.09	-0.00
7	0.68	1.62
8	0.35	0.13
9	-0.12	-0.32
10	0.26	-0.06
11	1.50	1.05
12	-1.63	1.38
13	1.44	-1.08
14	-0.02	-0.15
15	-0.13	0.33
16	-0.63	1.95
17	0.24	-0.02
18	0.02	-0.18
19	0.46	-1.80
20	-0.68	-1.84

Gaussian kernel : $K_{ij} = \exp(-\gamma \|x_i - x_j\|^2_{\mathbb{R}^p})$



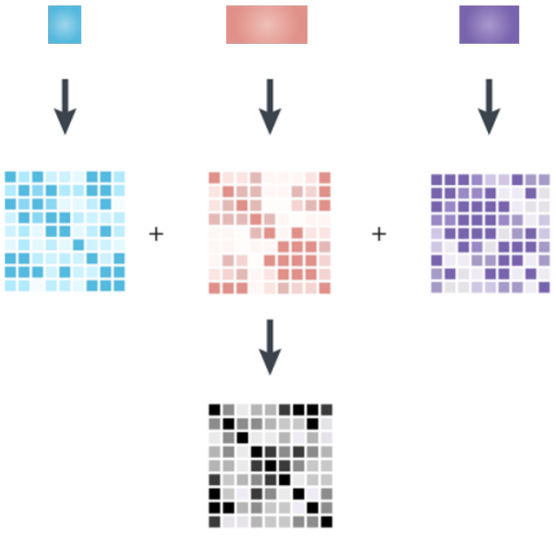
Prerequisites: kernels



Practical interests

- Represent a natural framework to **integrate** multiple datasets ;

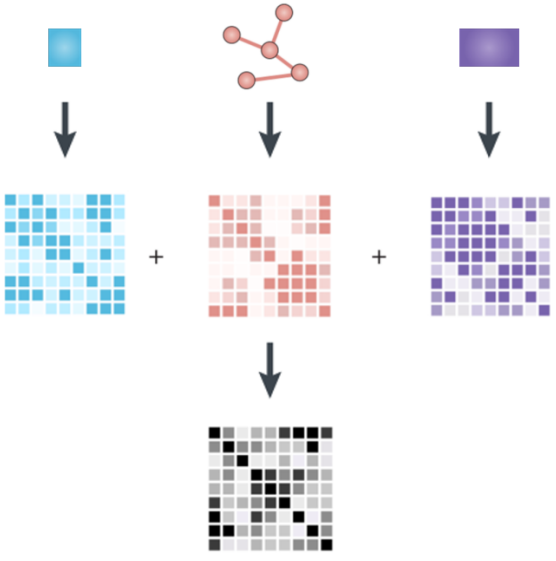
Prerequisites: kernels



Practical interests

- Represent a natural framework to **integrate** multiple datasets ;
- Allow to analyse **heterogenous** datasets ;

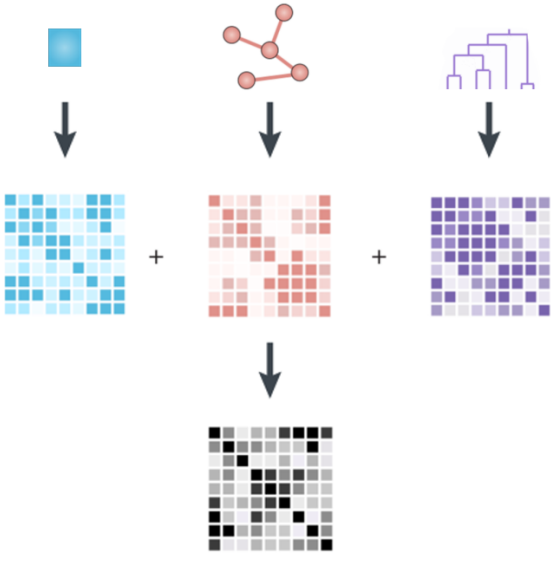
Prerequisites: kernels



Practical interests

- Represent a natural framework to **integrate** multiple datasets ;
- Allow to analyse **heterogenous** datasets ;

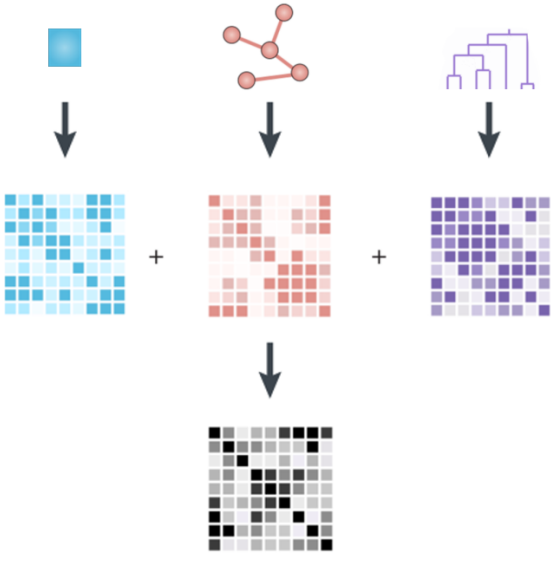
Prerequisites: kernels



Practical interests

- Represent a natural framework to **integrate** multiple datasets ;
- Allow to analyse **heterogenous** datasets ;

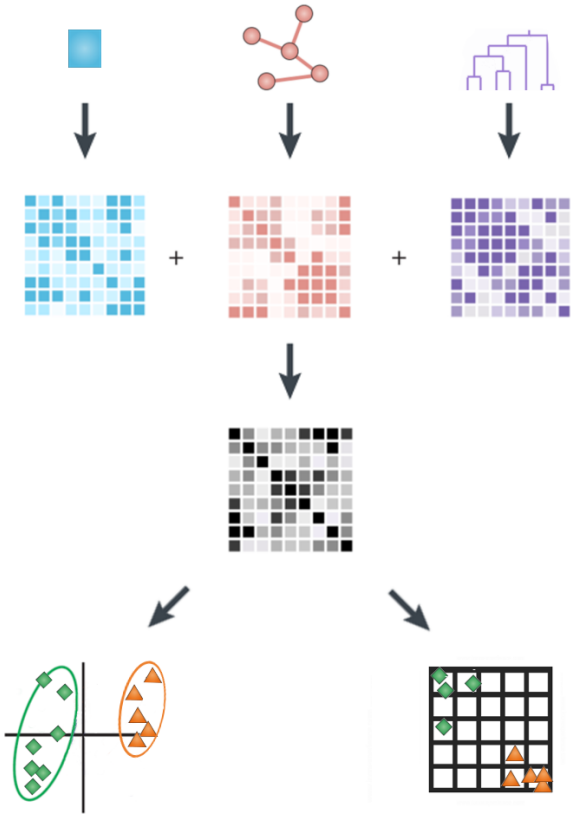
Prerequisites: kernels



Practical interests

- Represent a natural framework to **integrate** multiple datasets ;
- Allow to analyse **heterogenous** datasets ;
- Give acces to a large number of similarity / dissimilarity measures ;

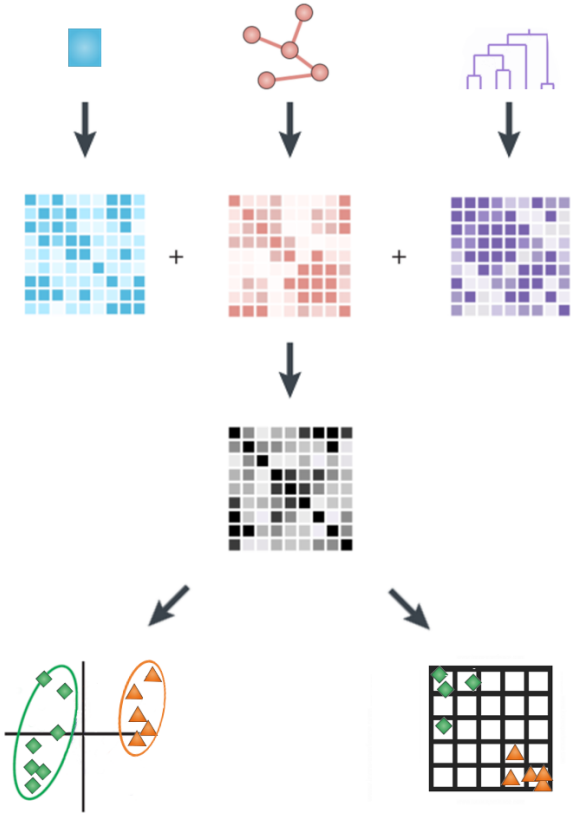
Prerequisites: kernels



Practical interests

- Represent a natural framework to **integrate** multiple datasets ;
- Allow to analyse **heterogenous** datasets ;
- Give acces to a large number of similarity / dissimilarity measures ;
- Allow to apply a **large panel of methods** (kernel trick) : PCA, SOM, linear model, supervised classification, clustering, . . .

Prerequisites: kernels



Practical interests

- Represent a natural framework to **integrate** multiple datasets ;
- Allow to analyse **heterogenous** datasets ;
- Give acces to a large number of similarity / dissimilarity measures ;
- Allow to apply a **large panel of methods** (kernel trick) : PCA, SOM, linear model, supervised classification, clustering, . . .

Drawbacks

- **Algorithm complexity** ;
- Loss of **model interpretability** ;

Exploratory analysis: kernel PCA

Standard Principal Component Analysis (PCA)

- Projection of high dimensional dataset in a small dimensional space
- Designed so as to keep most of the data variability
- Axes interpretable from a variable and from an observation point of view (axes are linear combinations of the original variables)

Exploratory analysis: kernel PCA

Standard Principal Component Analysis (PCA)

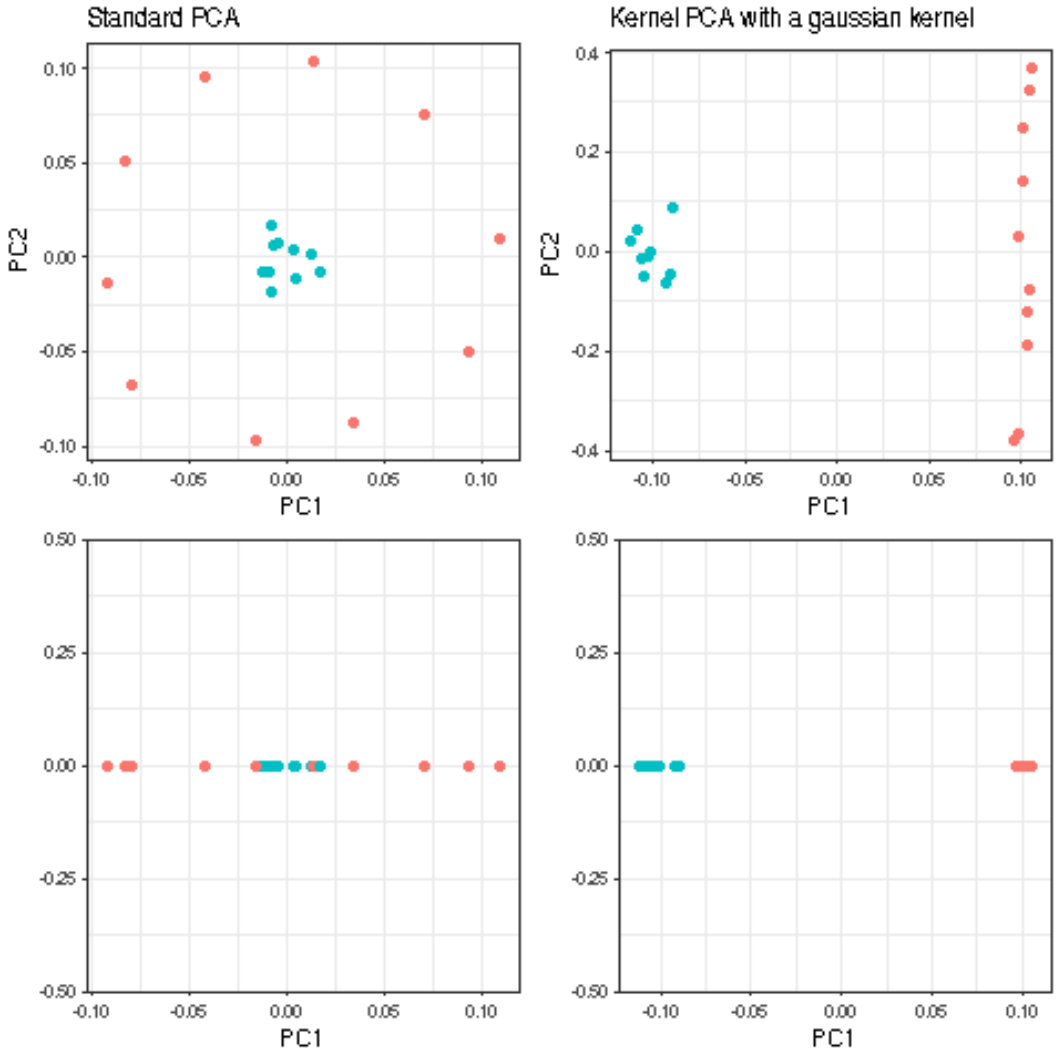
- Projection of high dimensional dataset in a small dimensional space
- Designed so as to keep most of the data variability
- Axes interpretable from a variable and from an observation point of view (axes are linear combinations of the original variables)

K-PCA [Schölkopf et al., 1998]

- PCA in the feature space (corresponds to a non linear projection of the original data in the original space)

Exploratory analysis: kernel PCA

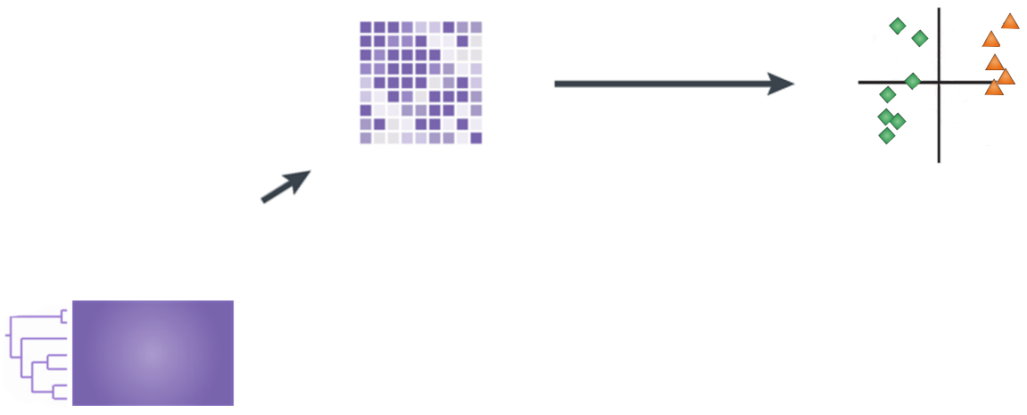
	x_1	x_2
1	-1.96	-0.02
2	0.08	0.22
3	-0.19	0.16
4	1.98	-0.19
5	-1.55	-1.17
6	-0.09	-0.00
7	0.68	1.62
8	0.35	0.13
9	-0.12	-0.32
10	0.26	-0.06
11	1.50	1.05
12	-1.63	1.38
13	1.44	-1.08
14	-0.02	-0.15
15	-0.13	0.33
16	-0.63	1.95
17	0.24	-0.02
18	0.02	-0.18
19	0.46	-1.80
20	-0.68	-1.84



Exploratory analysis: kernel PCA

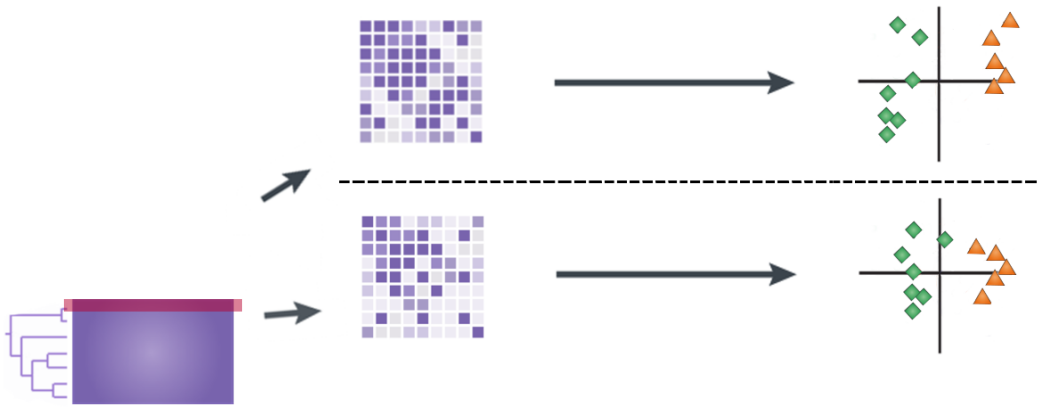
- Generic approach based on random permutations to assess variables influence.

Exploratory analysis: kernel PCA



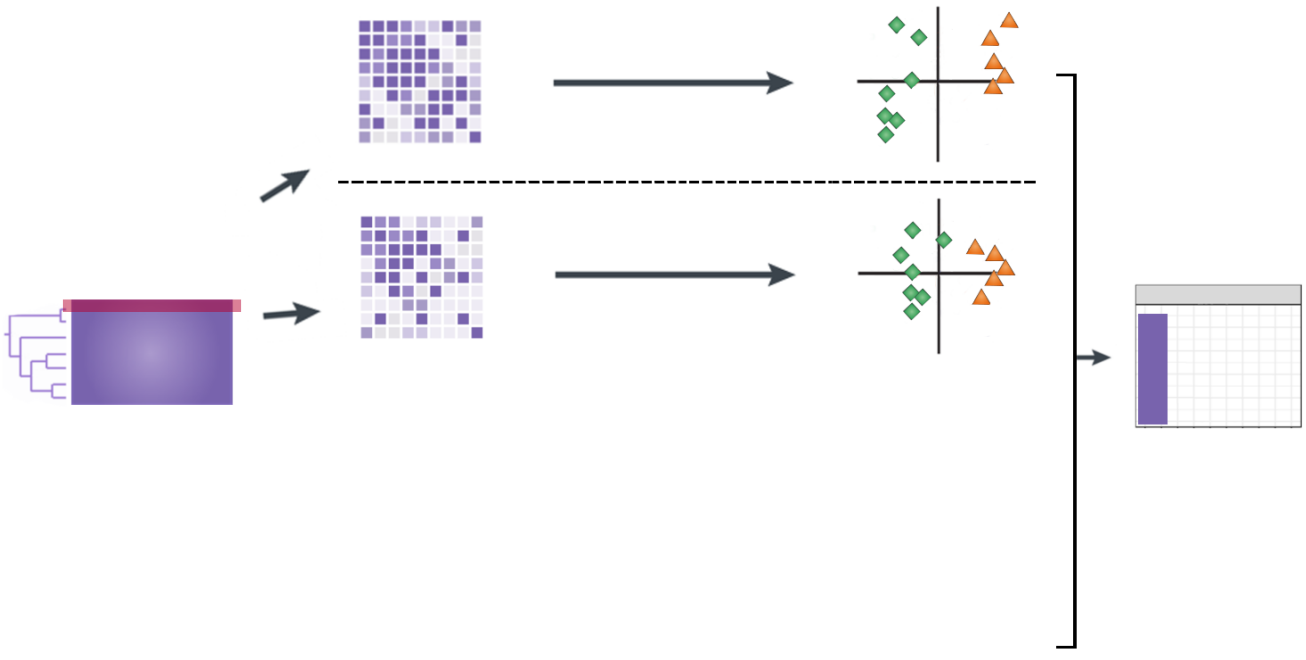
- Compute kernel K ;
- Kernel PCA.

Exploratory analysis: kernel PCA



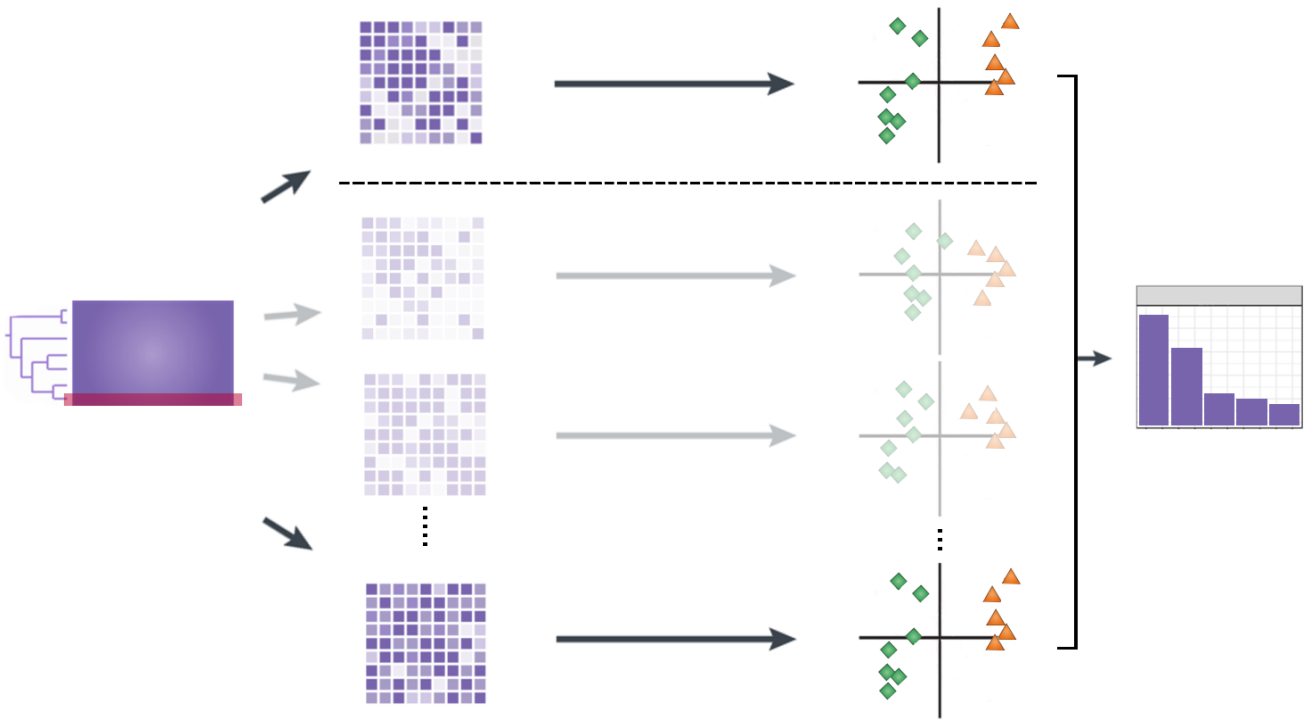
- Variable 1 permutation ;
- Compute kernel $\tilde{K}^1$ and the kernel PCA.

Exploratory analysis: kernel PCA



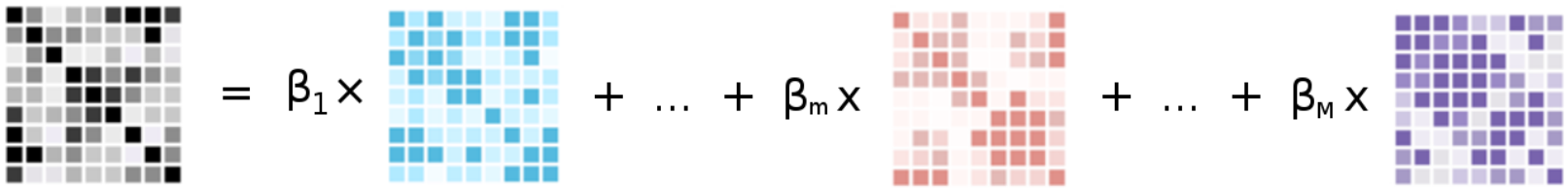
- Compute the Crone and Crosby distance [Crone and Crosby, 1995] between K and $\tilde{K}^1$ PCA sub-spaces.

Exploratory analysis: kernel PCA



- Permute all variables and compute the Crone and Crosby distance between K and $(\tilde{K}^j)_j$ PCA sub-spaces.

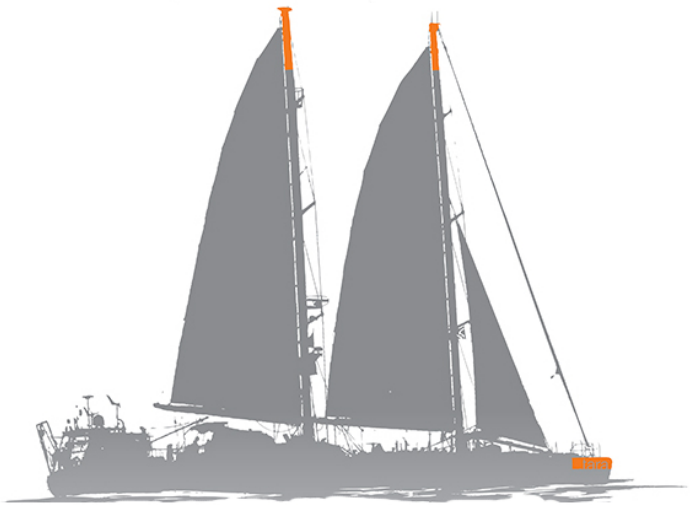
Integration: multiple kernel learning



$$K^{(*)} = \sum_{m=1}^{M} \beta_m K^{(m)} \text{ avec } \beta_m \geq 0 \text{ et } \sum_{m=1}^{M} \beta_m = 1$$

- **Naive approach**: $\beta_m = \frac{1}{M}$
- **Supervised** framework: β_m chosen to **minimise the prediction error** [Gönen and Alpaydin, 2011]
- **Unsupervised** framework: combine M kernels dedicated to datasets taking values in an arbitrary space.

Example: *TARA* oceans datasets



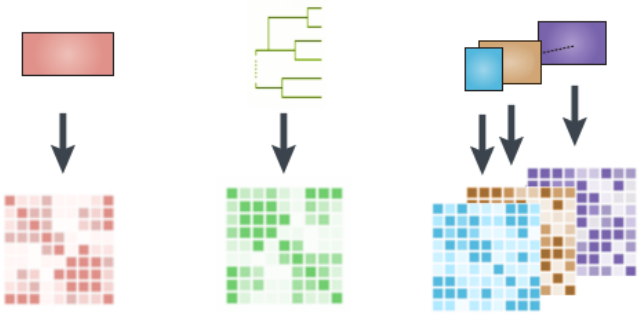
TARA OCEANS

EXPEDITIONS
www.taraexpeditions.org

The 2009-2013 expedition

- 48 samples
- 2 depth: surface (SRF) and deep chlorophyll maximum (DCM)
- 31 geographic localisation

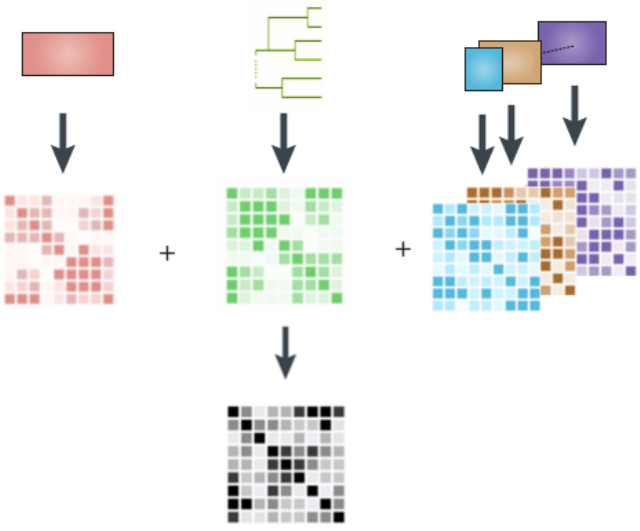
Example: *TARA* oceans datasets



8 *TARA* Oceans datasets :

- phychem physico-chemical data $\Rightarrow$ **linear kernel**.
- pro.phylo prokaryote phylogenetic tree $\Rightarrow$ kernel based on the **weighted Unifrac** distance.
- pro.NOGs prokaryotic functional composition $\Rightarrow$ kernel based on the **Bray-Curtis** dissimilarity.
- euk.pina, euk.nano, euk.micro and euk.meso : eukaryotic composition splited in 4 groups $\Rightarrow$ kernel based on the **Bray-Curtis** dissimilarity.
- vir.VCs : viral composition $\Rightarrow$ kernel based on the **Bray-Curtis** dissimilarity.

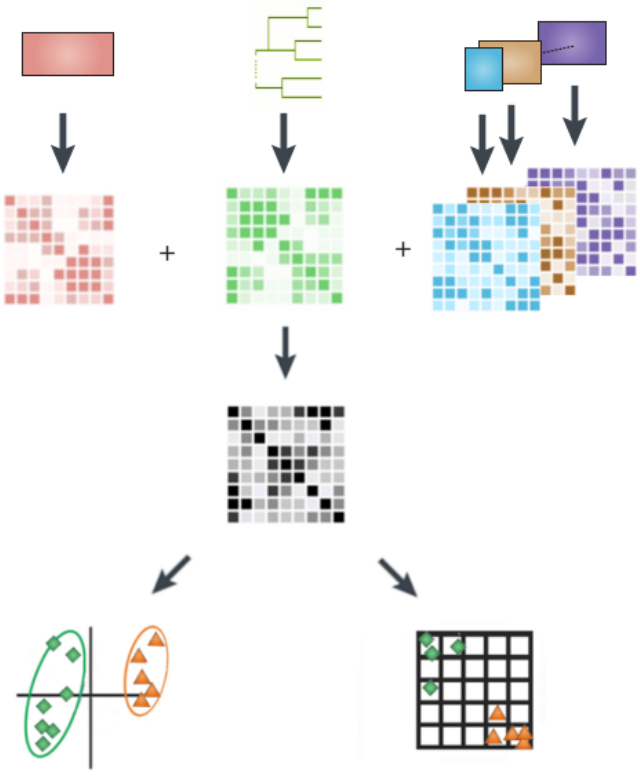
Example: *TARA* oceans datasets



Unsupervised multiple kernel learning de learn the β_m coeffecients :

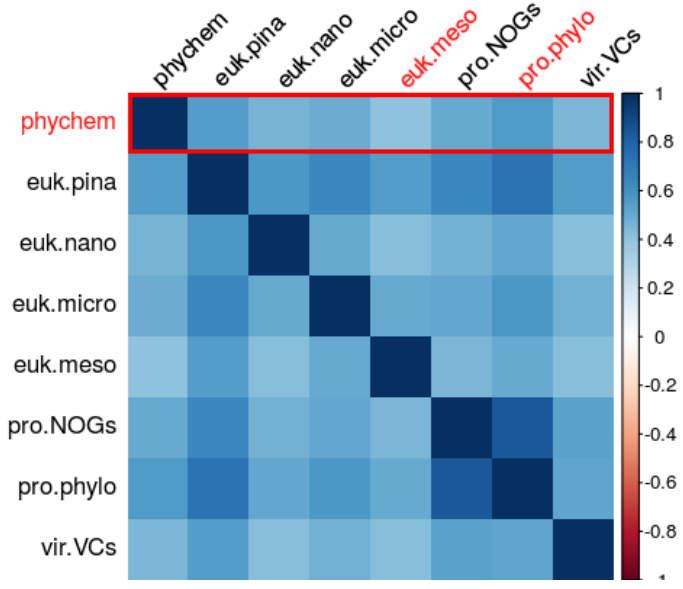
$$K^{(*)} = \sum_{m=1}^{M} \beta_m K^{(m)}.$$

Example: *TARA* oceans datasets



Apply standard data mining methods (clustering, linear model, PCA, . . .) in the feature space.

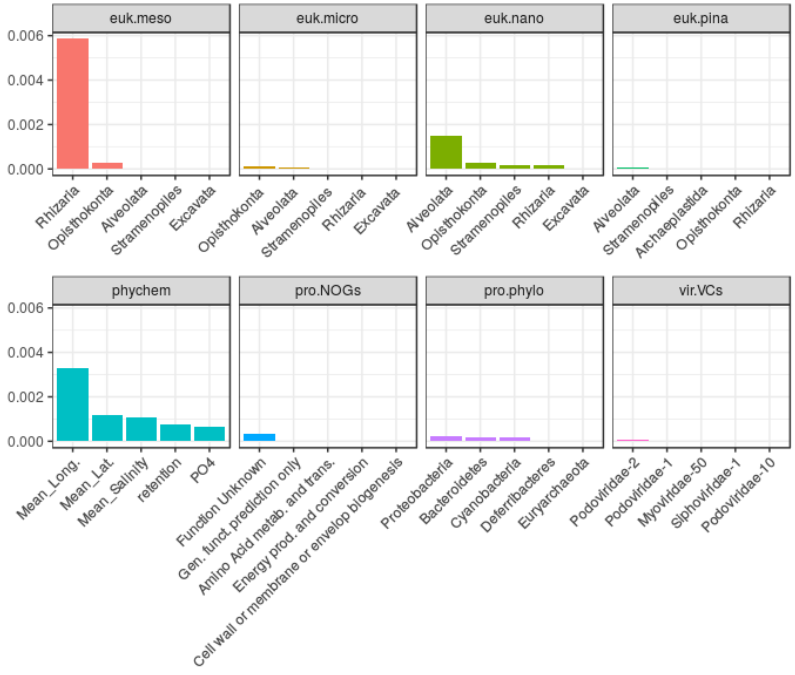
Example: *TARA* oceans datasets

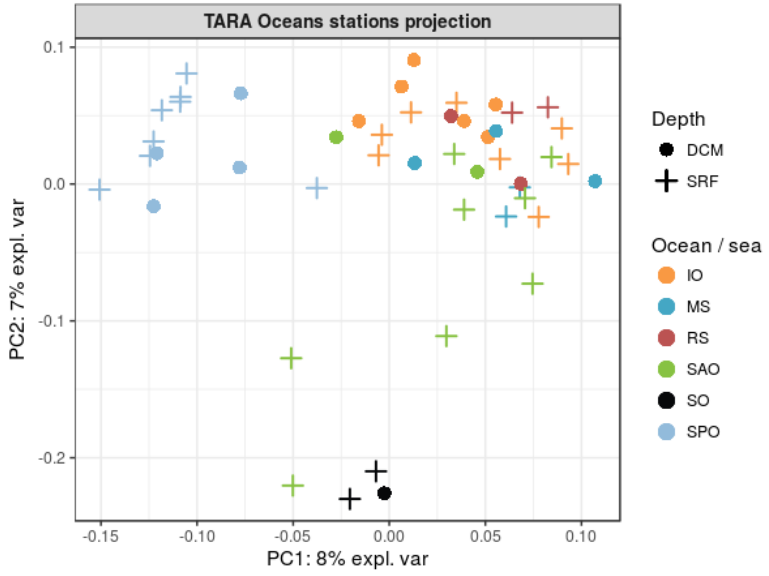


Correlations between kernels (STATIS)

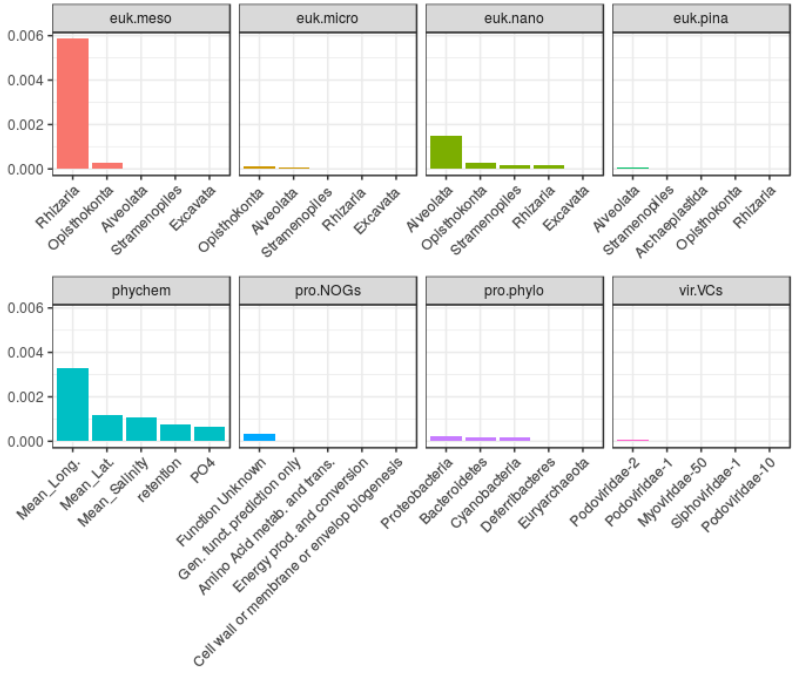
- Stronger correlations between **phychem** and small sizes organisms than large sizes organisms ([de Vargas et al., 2015] and [Sunagawa et al., 2015]).

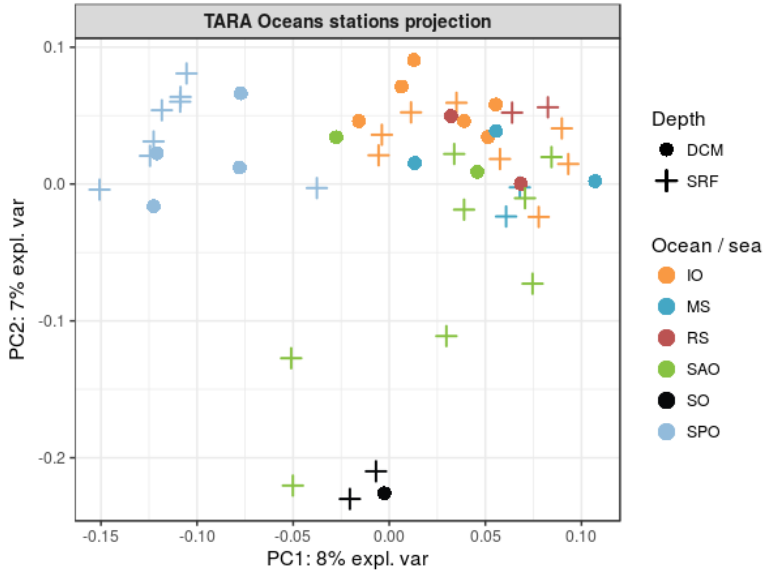
Example: *TARA* oceans datasets





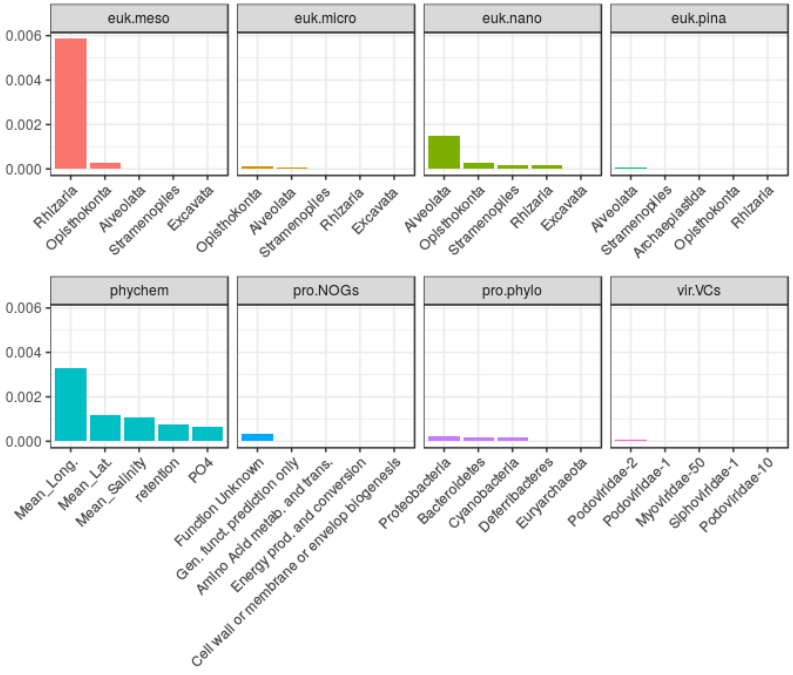
Example: *TARA* oceans datasets

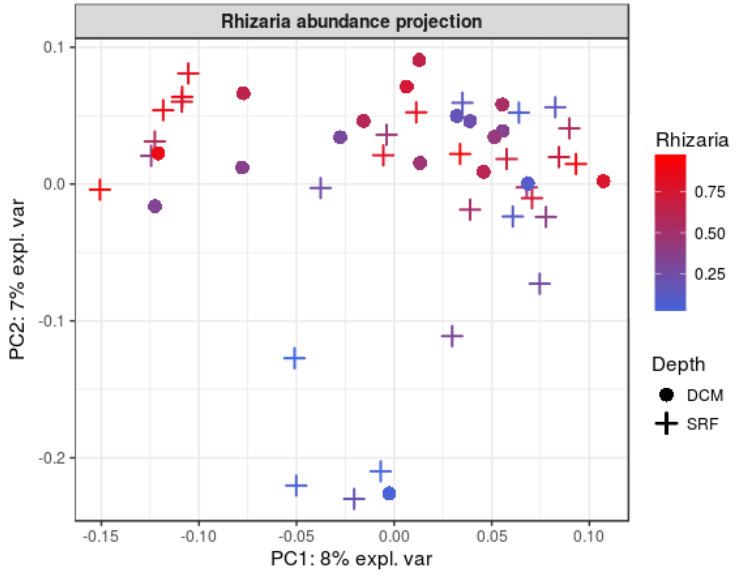




- Large size organisms are the most important: *Rhizaria* and *Alveolata* phyla.

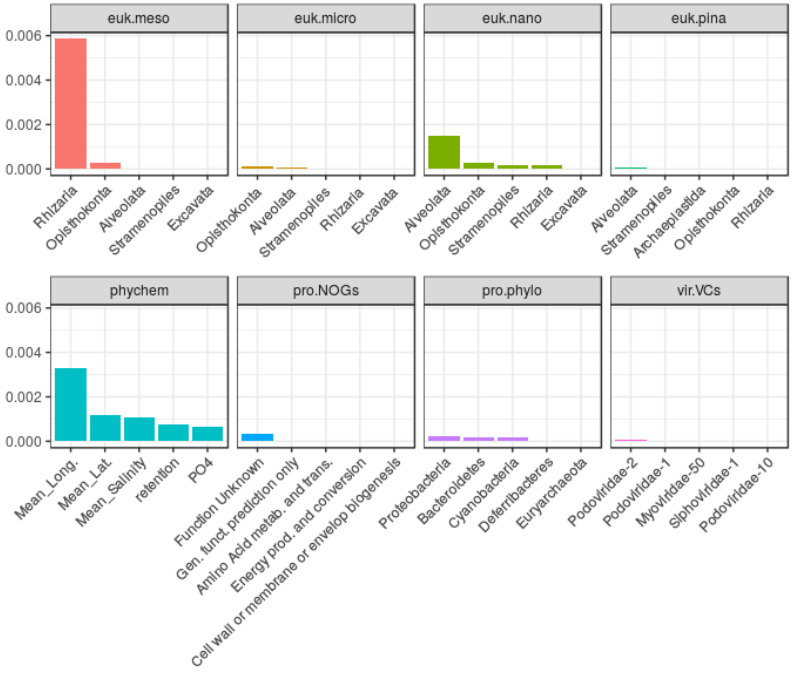
Example: *TARA* oceans datasets

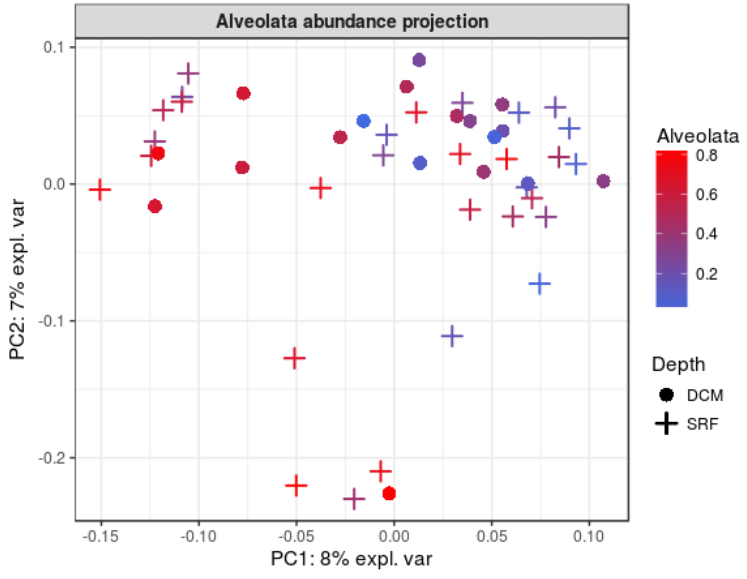




- Large size organisms are the most important: *Rhizaria* and *Alveolata* phyla.
- SO and SPO epipelagic waters mainly differ in terms of *Rhizarians* abundances.

Example: *TARA* oceans datasets





- Large size organisms are the most important: *Rhizaria* and *Alveolata* phyla.
- SO and SPO epipelagic waters mainly differ in terms of *Rhizarians* abundances.
- Both of them differ from the other studied waters in terms of *Alveolata* abundances.

Practical session

1. Compute kernels: `MyKernel <- compute.kernel(X)`
2. Combine kernels: `MyMetaKernel <- combine.kernels(K1=MyKernel, ...)`
3. Run the method: `MyResult <- kernel.pca(MyMetaKernel)`
4. Represent individuals: `plotIndiv(MyResult)`
5. Represent variables: `plotVar.kernel.pca(MyResult)`

X. Read the help files: `?compute.kernel, ?kernel.pca, ?plotIndiv, ...`

Multivariate methods

Kernel methods

Conclusion

Hints

- 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 one data set)
- Address problems explicitly formulated: "I want to integrate my data" is not a problem explicitly formulated
- Clearly identify supervised and unsupervised question and methods to use. "PCA is not a good method, I can't see my clusters..."

[Crone and Crosby, 1995] Crone, L. J. and Crosby, D. S. (1995).
Statistical applications of a metric on subspaces to satellite meteorology.
Technometrics, 37(3):324–328.

[de Vargas et al., 2015] de Vargas, C., Audic, S., Henry, N., Decelle, J., Mahé, P., Logares, R., Lara, E., Berney, C., Le Bescot, N., Probert, I., Carmichael, M., Poulain, J., Romac, S., Colin, S., Aury, J., Bittner, L., Chaffron, S., Dunthorn, M., Engelen, S., Flegontova, O., Guidi, L., Horák, A., Jaillon, O., Lima-Mendez, G., Lukeš, J., Malviya, S., Morard, R., Mulot, M., Scalco, E., Siano, R., Vincent, F., Zingone, A., Dimier, C., Picheral, M., Searson, S., Kandels-Lewis, S., *Tara* Oceans coordinators, Acinas, S., Bork, P., Bowler, C., Gorsky, G., Grimsley, N., Hingamp, P., Iudicone, D., Not, F., Ogata, H., Pesant, S., Raes, J., Sieracki, M. E., Speich, S., Stemmann, L., Sunagawa, S., Weissenbach, J., Wincker, P., and Karsenti, E. (2015).
Eukaryotic plankton diversity in the sunlit ocean.
Science, 348(6237).

[Gönen and Alpaydin, 2011] Gönen, M. and Alpaydin, E. (2011).
Multiple kernel learning algorithms.
Journal of Machine Learning Research, 12:2211–2268.

[Lozupone and Knight, 2005] Lozupone, C. and Knight, R. (2005).
UniFrac: a new phylogenetic method for comparing microbial communities.
Applied and Environmental Microbiology, 71(12):8228–8235.

[Schölkopf et al., 1998] Schölkopf, B., Smola, A., and Müller, K. (1998).
Nonlinear component analysis as a kernel eigenvalue problem.
Neural Computation, 10:1299–1319.

[Sunagawa et al., 2015] Sunagawa, S., Coelho, L., Chaffron, S., Kultima, J., Labadie, K., Salazar, F., Djahanschiri, B., Zeller, G., Mende, D., Alberti, A., Cornejo-Castillo, F., Costea, P., Cruaud, C., d'Oviedo, F., Engelen, S., Ferrera, I., Gasol, J., Guidi, L., Hildebrand, F., Kokoszka, F., Lepoivre, C., Lima-Mendez, G., Poulain, J., Poulos, B., Royo-Llonch, M., Sarmento, H., Vieira-Silva, S., Dimier, C., Picheral, M., Searson, S., Kandels-Lewis, S., *Tara* Oceans coordinators, Bowler, C., de Vargas, C., Gorsky, G., Grimsley, N., Hingamp, P.,

Iudicone, D., Jaillon, O., Not, F., Ogata, H., Pesant, S., Speich, S., Stemmann, L., Sullivan, M., Weissenbach, J., Wincker, P., Karsenti, E., Raes, J., Acinas, S., and Bork, P. (2015).
Structure and function of the global ocean microbiome.
Science, 348(6237).