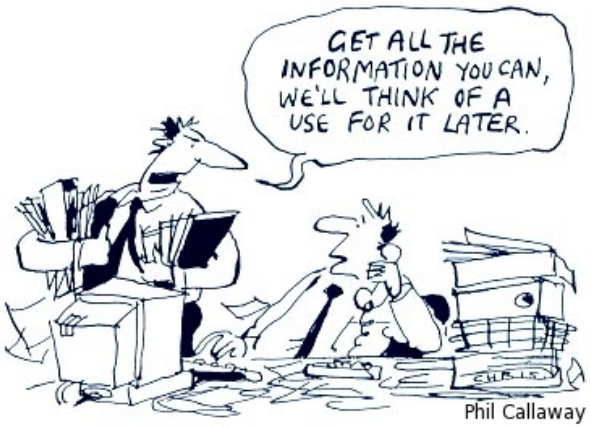
Multivariate analysis and 'omics' integration with

Illustration on some biological studies

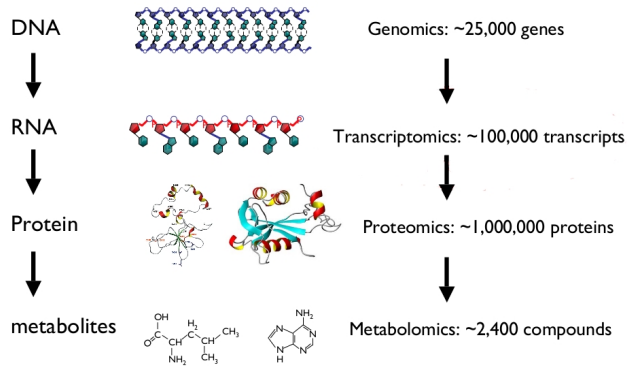
Kim-Anh Lê Cao

Queensland Facility for Advanced Bioinformatics
The University of Queensland

The issue with integrative systems biology



- Unlimited quantity of data ($n << p$ problem)
- Data from multiple sources



→ Efficient and biologically relevant statistical methodologies are needed to combine the information in these heterogeneous data sets.

Single Omics analysis

- Do we observe a 'natural' separation between the different groups of patients?
- Can we identify potential biomarker candidates predicting the status of the patients?

Integrative Omics analysis

- Can we identify a subset of correlated genes and proteins from matching data sets?
- Can we predict the abundance of a protein given the expression of a small subset of genes?
- Do two matching omics data set contain the same information?
- How to take into account a repeated measurement design?

Data setting

n = number of patients
p, q = number of biological features (genes, proteins ..)

Single Omics analysis

- one omic data set X(n x p)
- for a supervised analysis, Y vector indicating the class of the patients

Integrative Omics analysis

- two matching omics data sets (measured on the same patients)
- X (n x p) and Y (n x q) (unsupervised analysis)

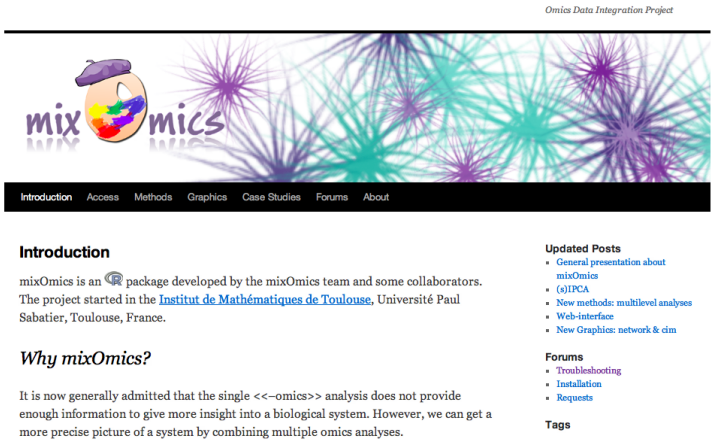
Linear multivariate approaches enable:

- Dimension reduction
  $\rightarrow$ project the data in a smaller subspace
- To handle multicollinear, irrelevant, missing variables
- To capture experimental and biological variation

In particular, in `mixOmics`, focus is on:

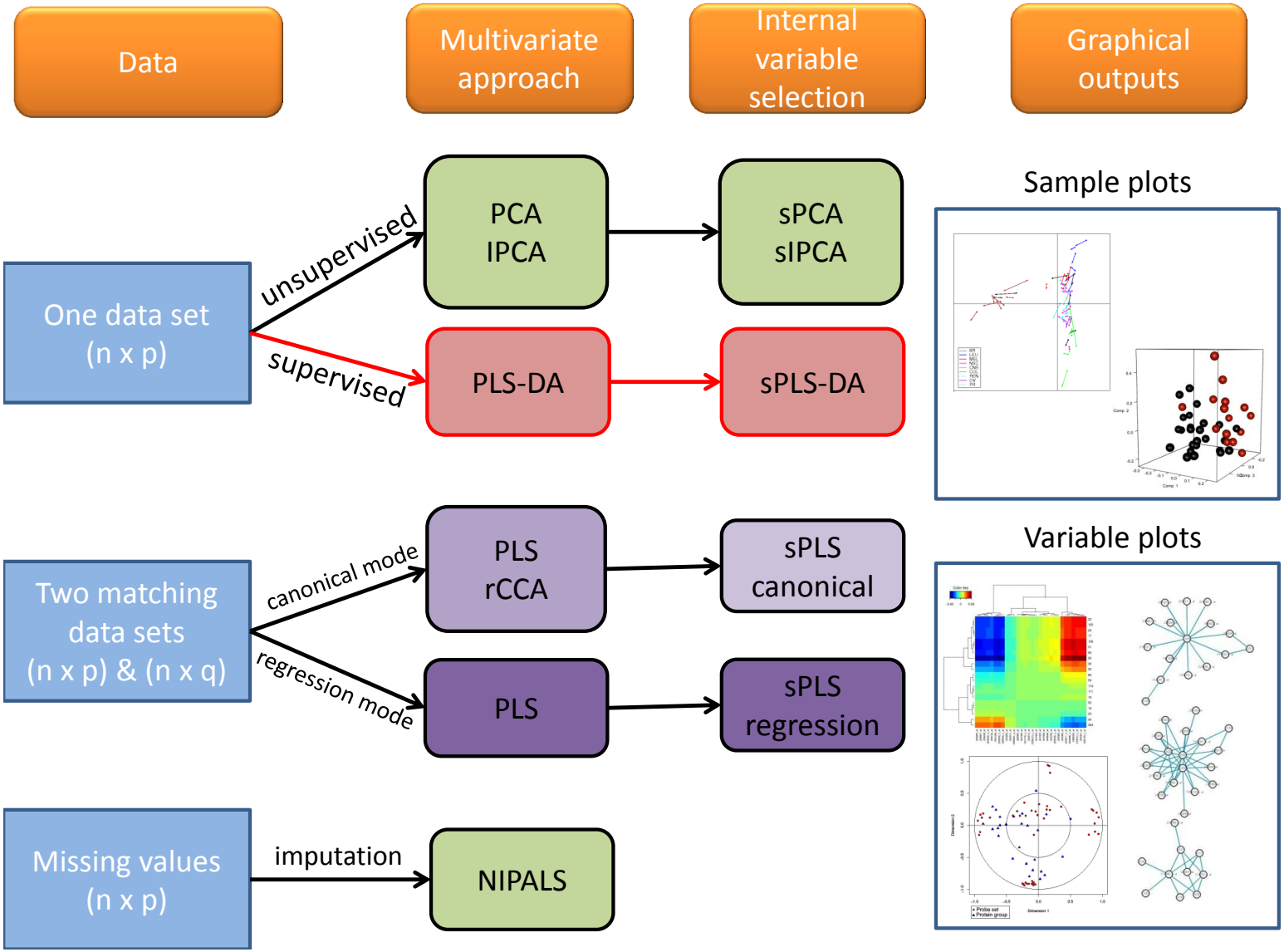
- Data integration
- Variable selection
- Computationally efficient methodologies for large biological data sets
- Interpretable graphical outputs

mixOmics is an R package dedicated to the exploration and the integrative analyses of high dimensional biological data sets.

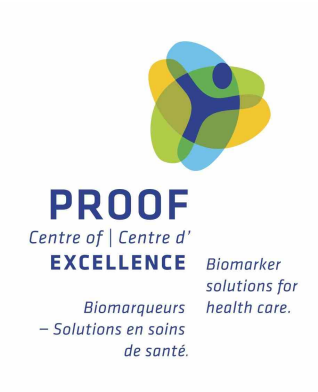


- Website
  - R tutorials
  - Newsletter, User Forum
- Web Interface
  - User friendly interface
  - Comprehensive results page

-**Lê Cao et al. (2009)** integRomics/**mixOmics**: an R package to unravel relationships between two omics data sets, *Bioinformatics*


Data
Multivariate approach
Internal variable selection
Graphical outputs
One data set (n x p)
unsupervised
supervised
PCA IPCA
PLS-DA
sPCA sIPCA
sPLS-DA
Two matching data sets (n x p) & (n x q)
canonical mode
regression mode
PLS rCCA
PLS
sPLS canonical
sPLS regression
Missing values (n x p)
imputation
NIPALS
Sample plots
Variable plots

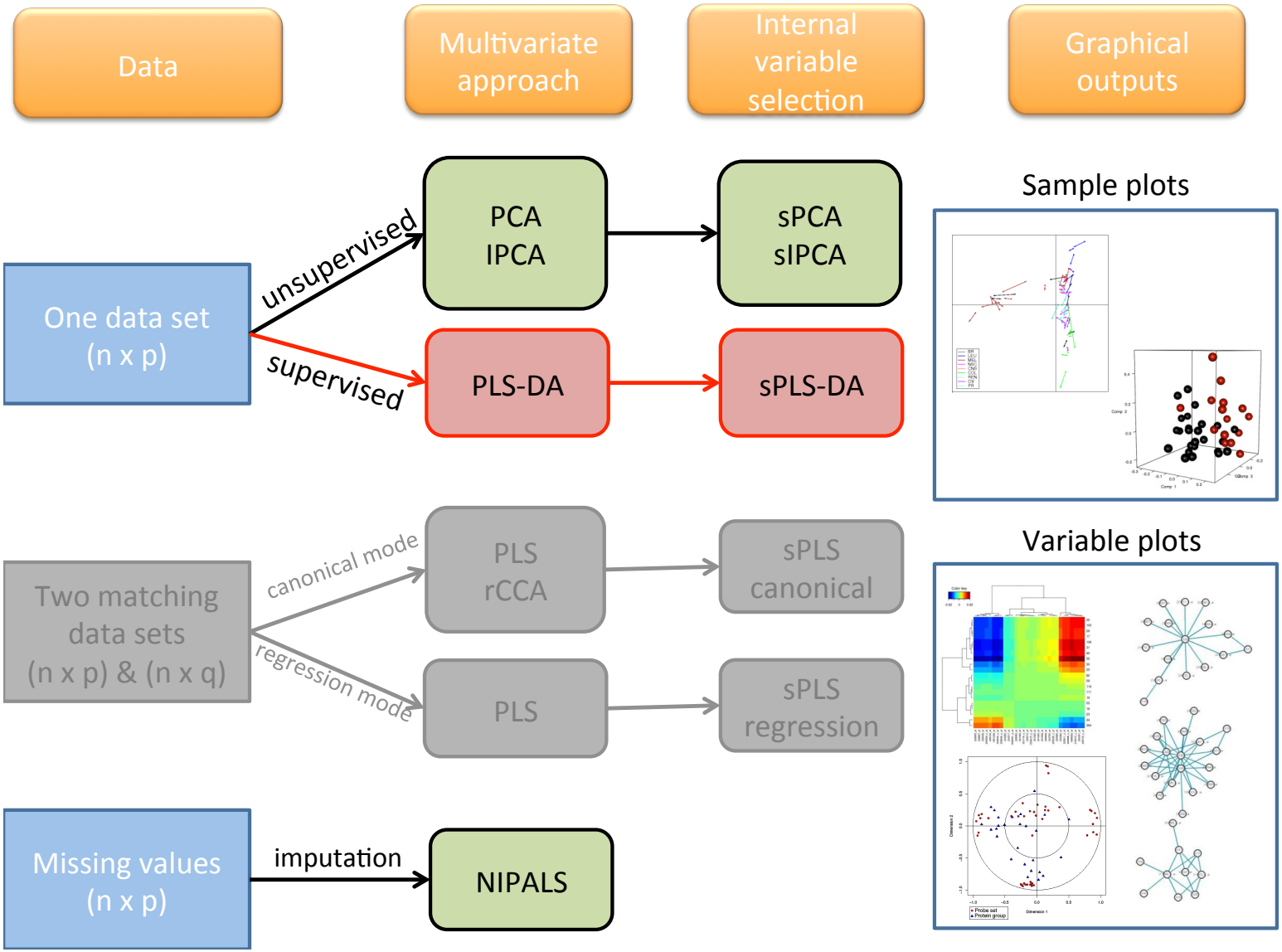
The PROOF Center



Prevention of Organ Failure Centre of Excellence is a not-for-profit organization that develops and implements blood-based biomarker tests to better manage patients with heart, lung and kidney failure and prevent disease progression.

Kidney transplant study: patients with kidney deficiency received a kidney transplant. Each Acute Rejection patient (AR) is matched with a Non Rejection patient (NR).

- The data: n = 40
  1. Genomics (Affymetrix, $p = 27,306$)
  2. Proteomics (iTRAQ, $q = 140$)


Data
Multivariate approach
Internal variable selection
Graphical outputs
One data set (n x p)
unsupervised
supervised
PCA IPCA
PLS-DA
sPCA sIPCA
sPLS-DA
Two matching data sets (n x p) & (n x q)
canonical mode
regression mode
PLS rCCA
PLS
sPLS canonical
sPLS regression
Missing values (n x p)
imputation
NIPALS
Sample plots
Variable plots

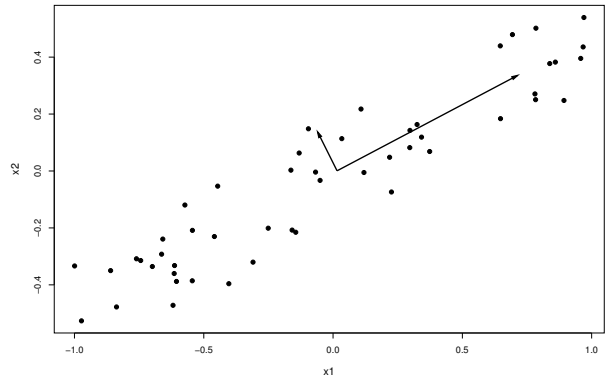
Single Omics analysis

- Do we observe a 'natural' separation between the different groups of patients?
- Can we identify potential biomarker candidates predicting the status of the patients?

Principal Component Analysis: PCA

Seek the best directions in the data that account for most of the variability

→ principal components: artificial variables that are linear combinations of the original variables:



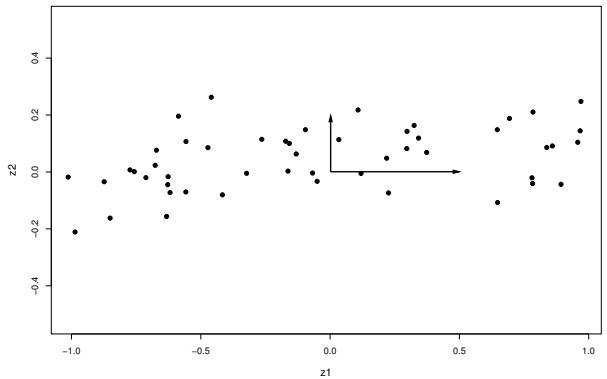
$$\underset{(n)}{\boldsymbol{c}} = \underset{(nxp)}{\boldsymbol{X}} \quad \underset{(p)}{\boldsymbol{v}}$$

- $\boldsymbol{c}$ is a linear function of the elements of $\boldsymbol{X}$ having maximal variance
- $\boldsymbol{v}$ is called the associated loading vector

Principal components cont.

The new PCs form a vectorial subspace of dimension $< p$

Project the data on these new axes.



$\rightarrow$ approximate representation of the data points in a lower dimensional space

Problem:
Interpretation difficult with very large number of (possibly) irrelevant variables

PCA

Objective function:

$$\max_{||\boldsymbol{v_h}||=1} var(X_h \boldsymbol{v_h}), \qquad h = 1 \ldots H$$

Several ways of solving PCA:

- Eigenvalue problem: $S\boldsymbol{v} = \lambda \boldsymbol{v}$; $\boldsymbol{c} = X\boldsymbol{v}$

  $S =$ variance covariance matrix or correlation matrix if X is scaled
- Singular Value Decomposition (SVD): X = UDV'; C = UD

  $D =$diagonal matrix with $\sqrt{\lambda_j}$;

  $\boldsymbol{u_j}(\boldsymbol{v_j})$ are eigenvectors of $\frac{1}{n}XX'$ $(\frac{1}{n}X'X)$
- NIPALS algorithm

Independent Component Analysis (ICA):

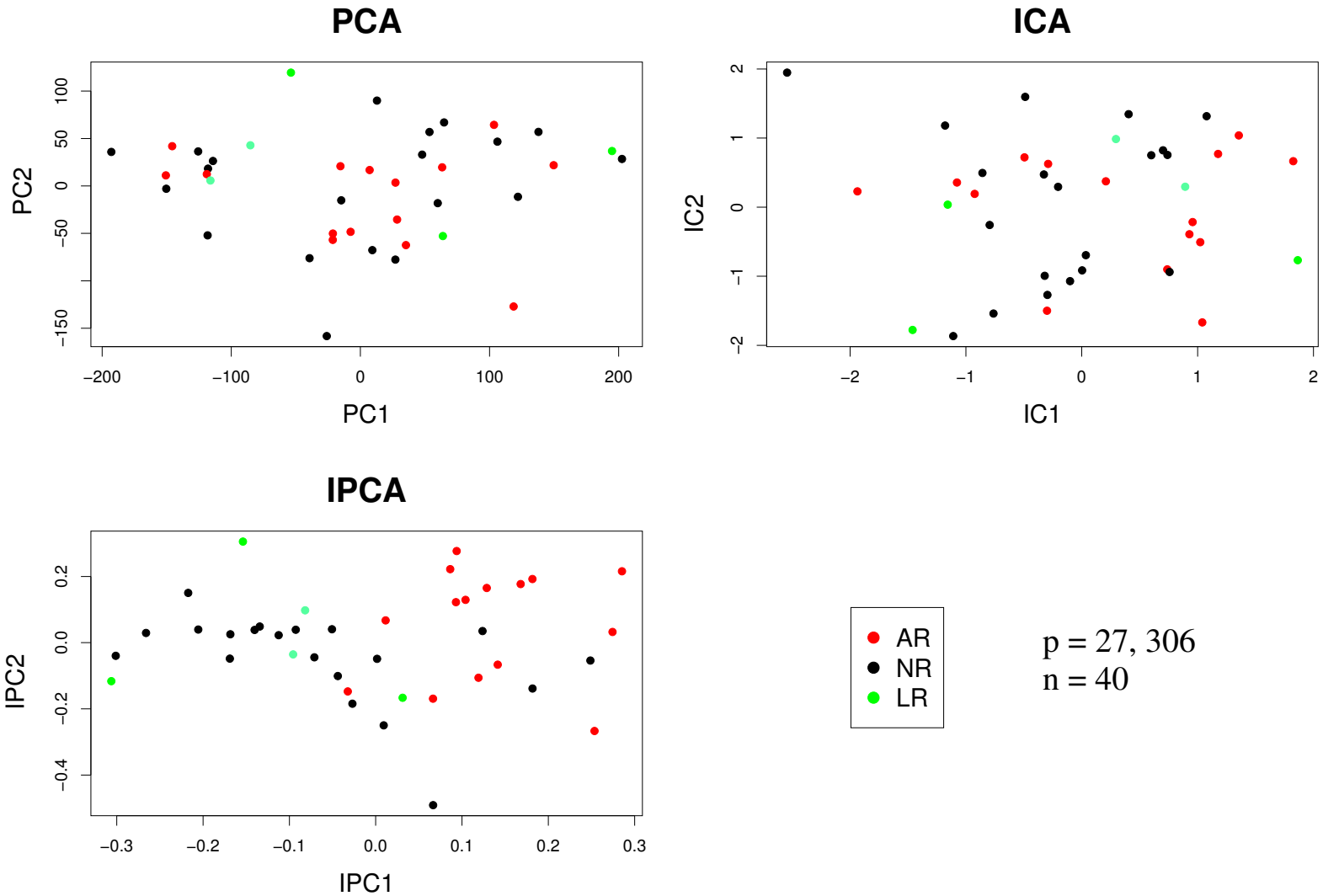
- assumes non Gaussian data distribution ($\neq$ PCA).
- 'blind source' signal separation.
- seeks for a set of independent components ($\neq$ PCA).

IPCA is based on Independent Component Analysis (ICA):

- Combines the advantages of both PCA and ICA.
- The PCA loadings are transformed via ICA to obtain independent loading vectors and independent principal components.

Yao, F., Coquery, J. and Lê Cao, K-A. 2012 Independent Principal Component Analysis for biologically meaningful dimension reduction of large biological data sets, *BMC Bioinformatics*.

Illustration on the genomics data

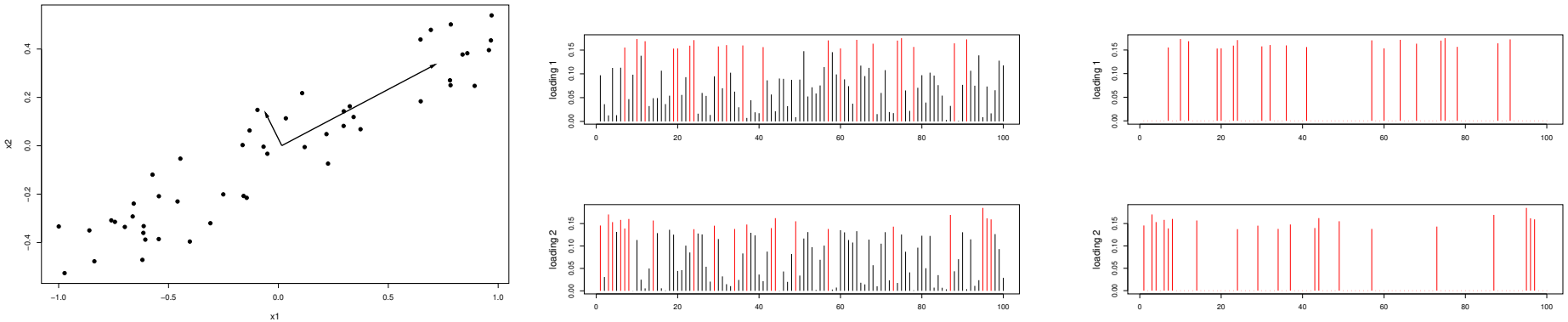


sparse Principal Component Analysis: sPCA

Principal components

loading vectors (PCA)

sparse loading vectors (sPCA)



The principal components are linear combinations of the original variables, variables weights are defined in the associated loading vectors.

sparse PCA computes the sparse loading vectors to remove irrelevant variables using lasso penalizations (**Shen & Huang 2008**, *J. Multivariate Analysis*).

sparse Principal Component Analysis: sPCA

sparse PCA: sparse loading vectors to remove noisy or irrelevant variables which determine the principal components

$\rightarrow$ Solving PCA through least squares problem (SVD) allows to include regularization parameters

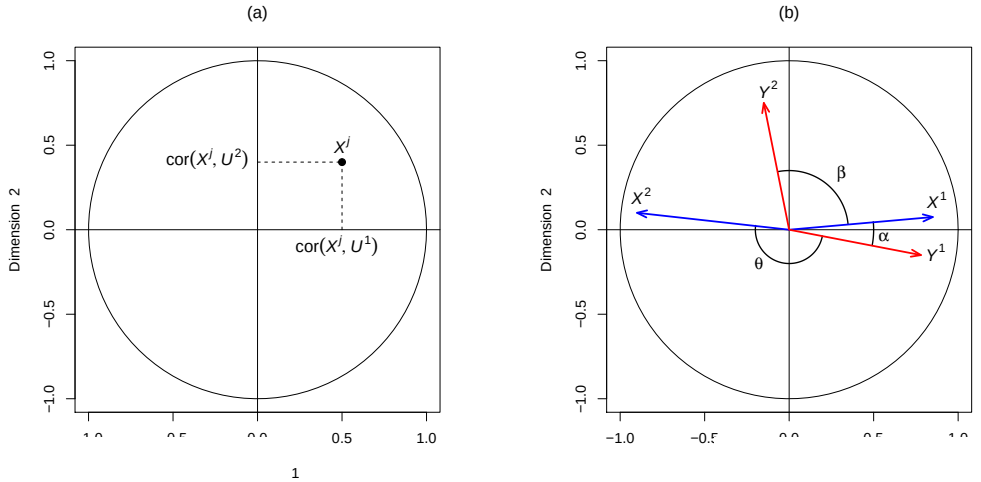
$$\min_{\mathbf{v_h}} ||X_h - \mathbf{u}_h \mathbf{v}_h^T||_F^2 + P_\lambda(\mathbf{u}_h)$$

P_λ is a penalty function with tuning regularization parameter λ

$\rightarrow$ use Lasso penalization
$\rightarrow$ obtain sparse loading vectors, with very few non-zero elements

Shen, H., Huang, J.Z. 2008. Sparse principal component analysis via regularized low rank matrix approximation, *J. Multivariate Analysis*.

Illustration: PCA and sPCA



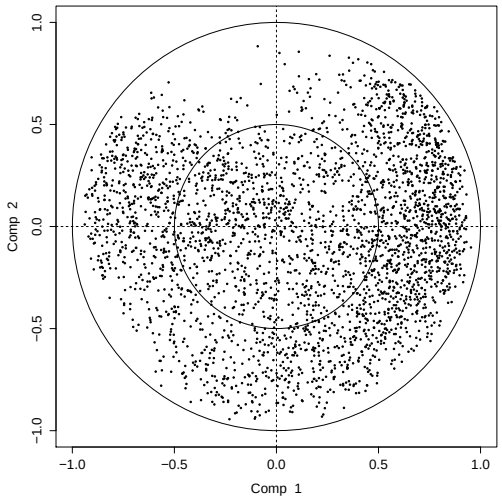


Figure: PCA

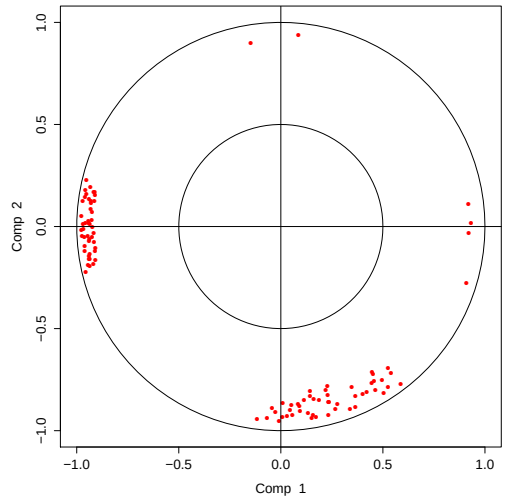
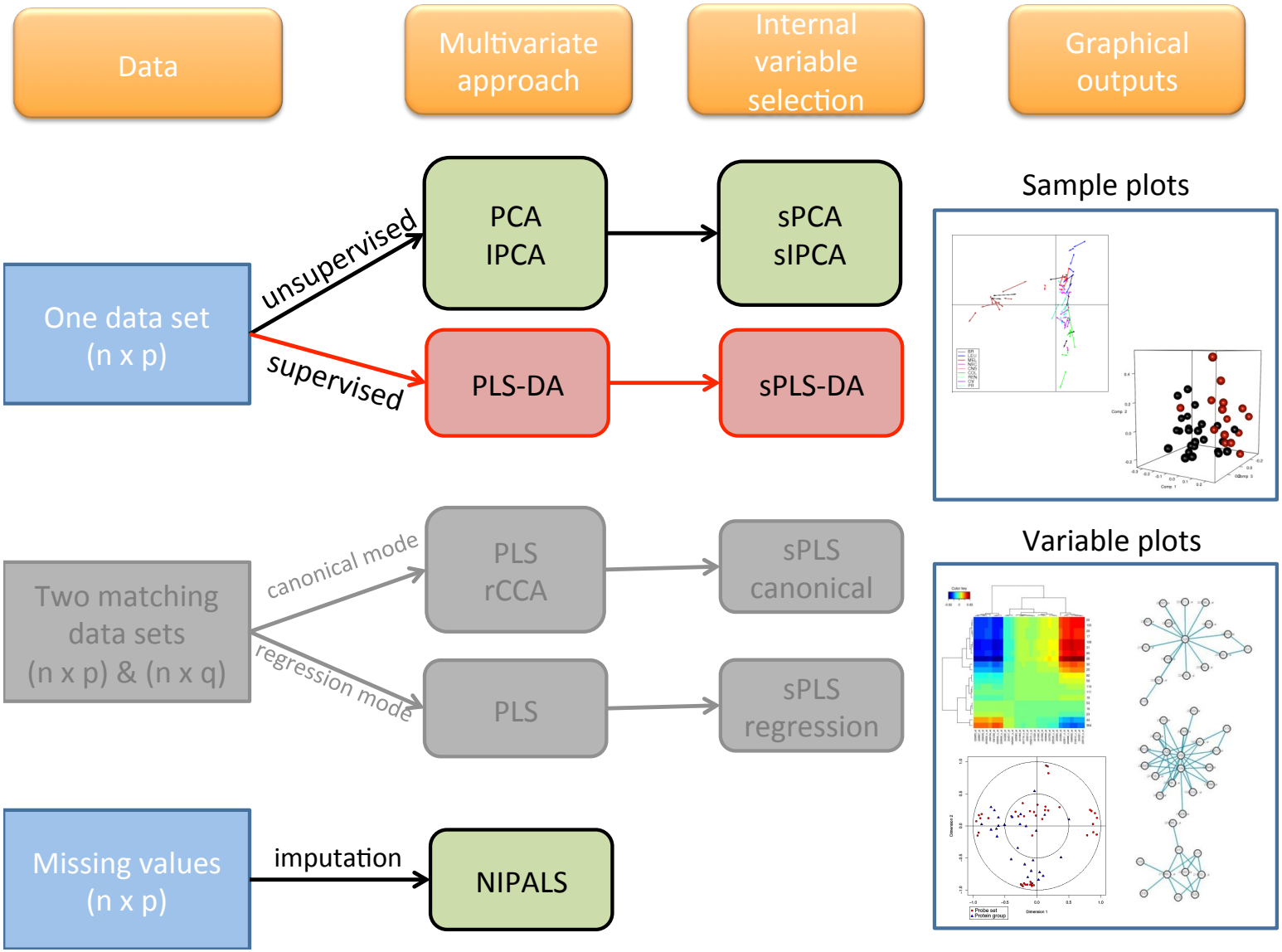


Figure: sPCA

Discriminant Analysis



PLS - Discriminant Analysis

- Similarly to Linear Discriminant Analysis, classical PLS-DA looks for the best components to separate the sample groups.
- As opposed to PCA/ICA methods, it is a supervised approach.

Objective function:

$$\max_{||\boldsymbol{u}_h||=1,||\boldsymbol{v}_h||=1} cov(X_h \boldsymbol{u}_h, Y \boldsymbol{v}_h) \qquad h = 1 \ldots H$$

Y is the qualitative response matrix (dummy block matrix)

Lê Cao K-A., Boitard S. and Besse P. (2011) Sparse PLS Discriminant Analysis: biologically relevant feature selection and graphical displays for multiclass problems, *BMC Bioinformatics*, 12:253.

sparse PLS - Discriminant Analysis

Include variable selection in PLS-DA via L_1 penalization on the loading vectors.

Let $M_h = X_h^T Y_h$,

$$\min_{\mathbf{u_h}} ||M_h - \mathbf{u}_h \mathbf{v}_h^T||_F^2 + P_\lambda(\mathbf{u}_h), \qquad h = 1 \ldots H$$

- use Lasso penalization (P_λ is a penalty function with regularization parameter λ),
- sparse loading vector u_h enables variable selection,
- sPLS-DA searches for discriminative variables that can help separating the sample groups,
- evaluate the discriminative power of the variable selection using cross-validation.

Illustration of sPLS-DA

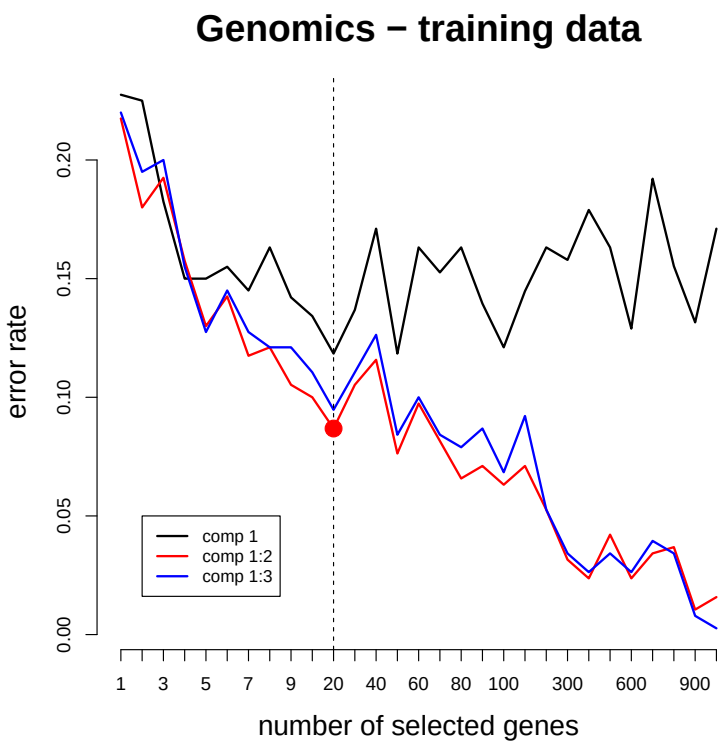


Figure: Tuning the number of var. to select with cross-validation

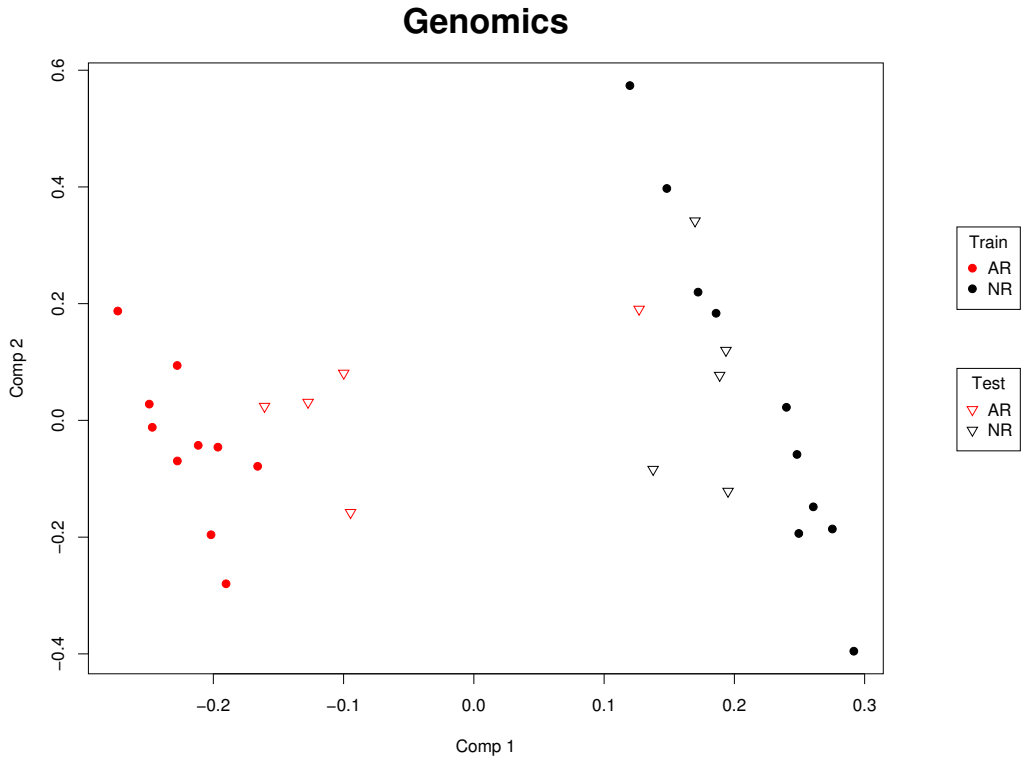
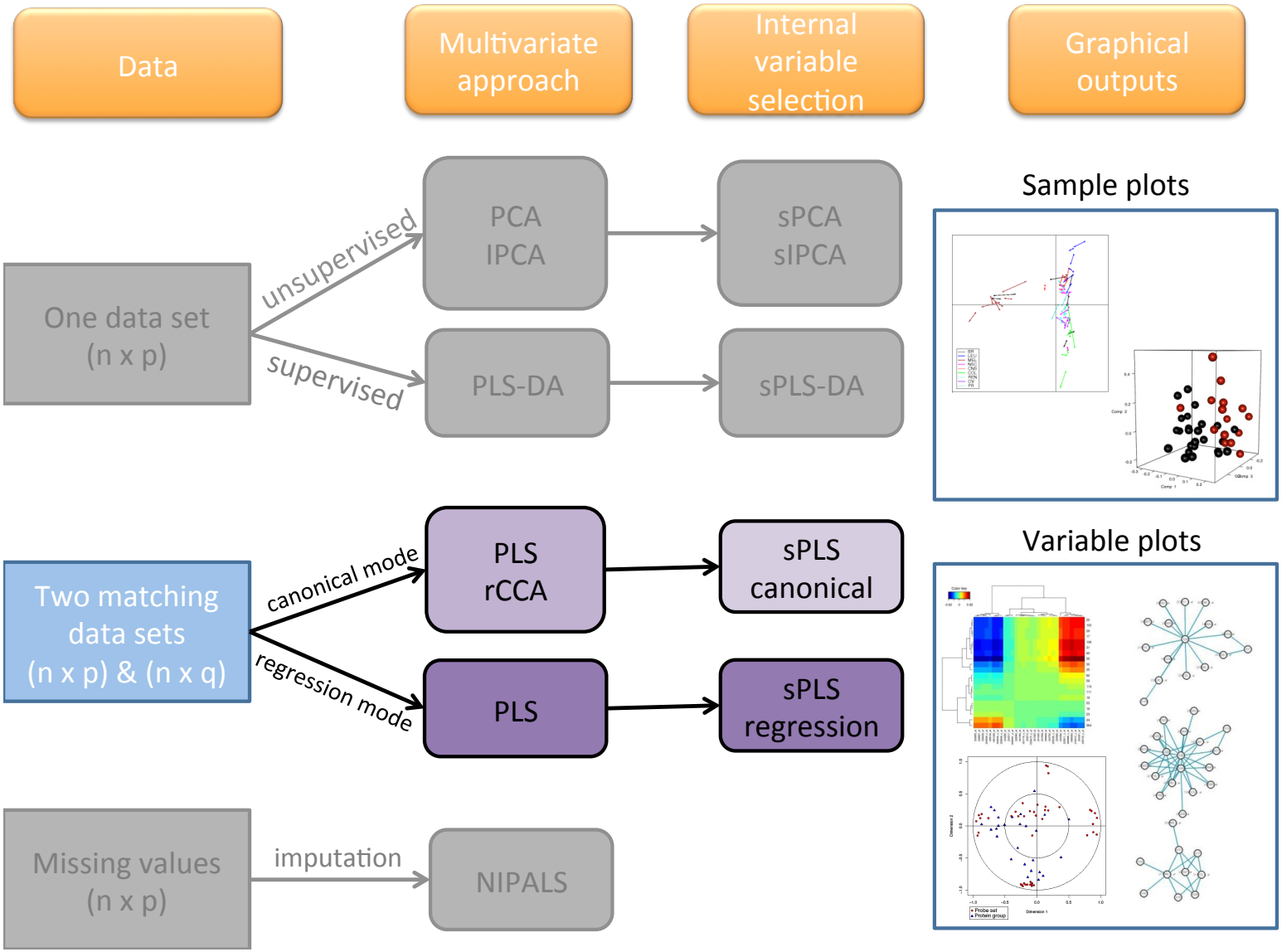


Figure: Predicting the class of the test set samples

Parameters to tune

- Number of components:
  -PCA, IPCA: explained variance
  -IPCA: kurtosis value
  -PLS-DA: $K - 1$
- Lasso penalization λ^h, $(h = 1, \ldots, H)$:
  -sPCA, sIPCA: sparsity degree, stability analysis, permutations, cluster analysis
  -sPLS-DA: classification error rate with cross-validation

$\rightarrow$ the biologist will also help choosing these parameters!


Data
Multivariate approach
Internal variable selection
Graphical outputs
One data set (n x p)
unsupervised
supervised
PCA
IPCA
PLS-DA
sPCA
sIPCA
sPLS-DA
Two matching data sets (n x p) & (n x q)
canonical mode
regression mode
PLS rCCA
PLS
sPLS canonical
sPLS regression
Missing values (n x p)
imputation
NIPALS
Sample plots
Variable plots

Integrative Omics analysis

- Can we identify a subset of correlated genes and proteins from matching data sets?
- Can we predict the abundance of a protein given the expression of a small subset of genes?
- Do two matching omics data set contain the same information?

- Partial Least Squares regression maximises the covariance between each linear combination (components) associated to each data set
- sparse PLS has been developed to include variable selection from both data sets
- Two modes are proposed to model the relationship between the two data sets (`mode='regression'` or `mode = 'canonical'`)

Lê Cao K-A., Rossouw D., Robert-Granié C. and Besse P. 2008. A Sparse PLS for Variable Selection when Integrating Omics data. *SAGMB* **7**(1).

Lê Cao K.-A., Martin P.G.P, Robert-Granié C. and Besse, P. 2009. Sparse Canonical Methods for Biological Data Integration: application to a cross-platform study. *BMC Bioinformatics*, 10:34.

Integration of two data sets

Aims:

- unravel the correlation stucture between two data sets
- select co-regulated biological entities across samples

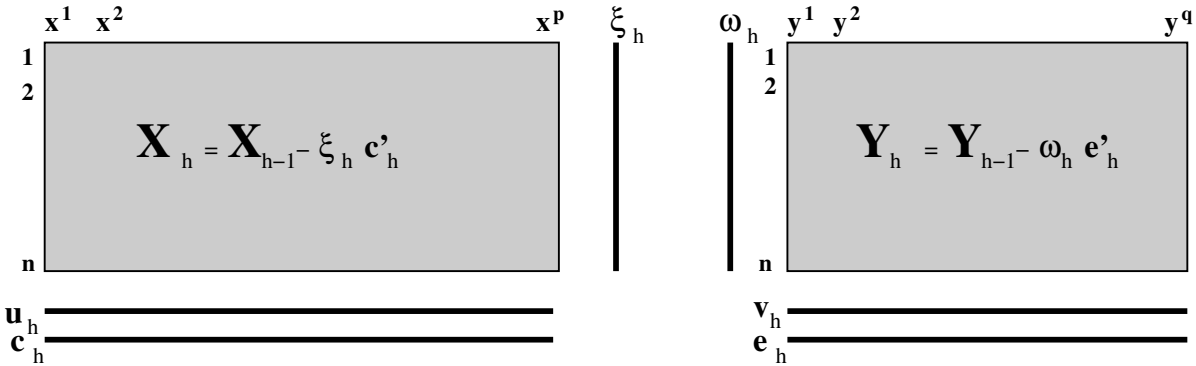
→ select and integrate in a one step procedure the different types of data

PLS objective function:

$$\max_{||u_h||=1,||v_h||=1} cov(X_h \mathbf{u}_h, Y_h \mathbf{v}_h), \qquad h = 1 \ldots H$$

where X (n x p) is the transcriptomics data set and Y (n x q) is the proteomics data set

Partial Least Squares



For each iteration h, $h = 1..H$, decompose X and Y into:

1. Loadings vectors $\mathbf{u}_h$ and $\mathbf{v}_h$, p- and q- dimensional vectors
2. Latent variables ξ_h and ω_h, n-dimensional vectors
3. Regression of X_{h-1} and Y_{h-1} on ξ_h and ω_h, reg. coeff. $\mathbf{c}_h$ and $\mathbf{e}_h$
4. Residual matrices: deflation step of X_{h-1} and Y_{h-1}

sparse PLS-SVD

Use the PLS-SVD variant that directly gives the latent variables and loading vectors and low rank rank approximation.

Let $M_h = X_h^T Y_h$, sparse PLS solves the optimization problem:

$$\min_{\mathbf{u_h}, \mathbf{v_h}} ||M_h - \mathbf{u}_h \mathbf{v}_h'||_F^2 + P_{\lambda_1}(\mathbf{u}_h) + P_{\lambda_2}(\mathbf{v}_h)$$

where P_λ is a penalty function

$\rightarrow$ obtain simultaneously sparse loadings $\mathbf{u}_h$ and $\mathbf{v}_h$
$\rightarrow$ simultaneous variable selection in both data sets

Illustration of sparse PLS: sample plot

sPLS aims at selecting correlated variables (genes, proteins) across the same samples by performing a multivariate regression.
Regression: explain the protein abundance w.r.t the gene expression "$\Rightarrow$ relationship".

- The latent variables (PLS components) are determined based on the selected genes and proteins $\rightarrow$ give more insight into the samples similarities.
- Unsupervised approach

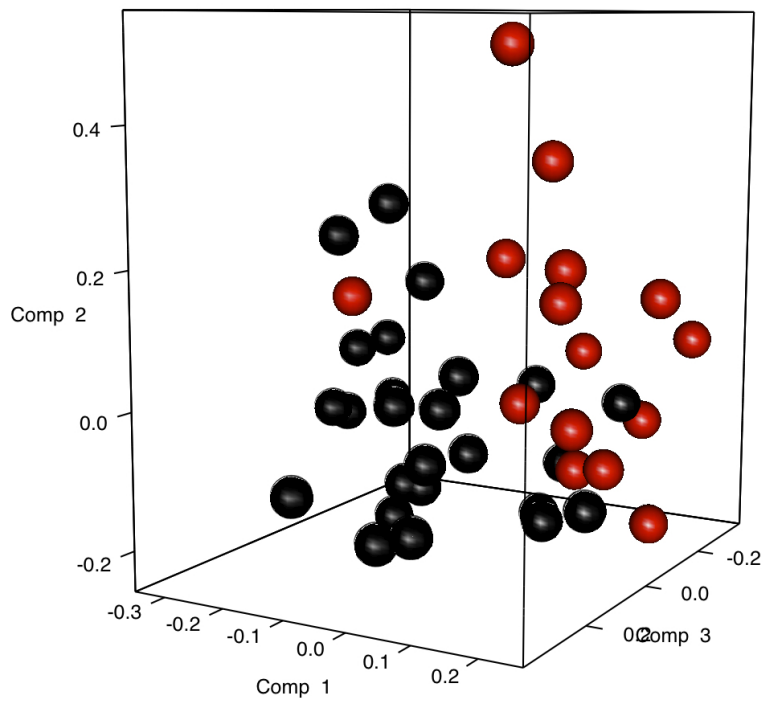
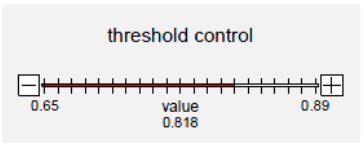
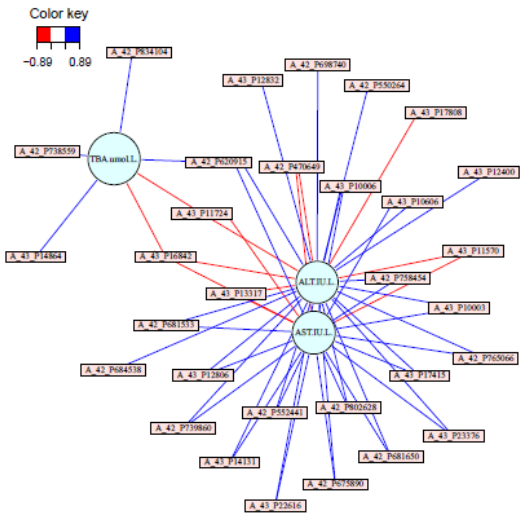
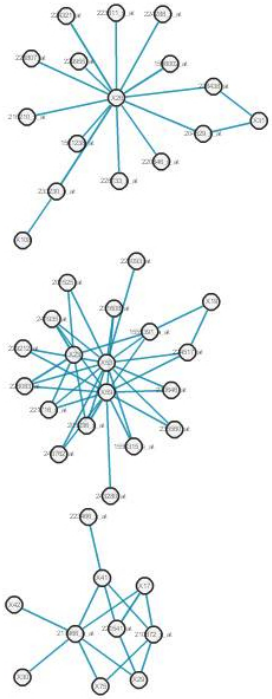


Illustration of sparsePLS: variable plot

Relevance networks are bipartite graphs directly inferred from the (s)PLS components.







González I., Lê Cao K.-A., Davis, M.D. and Déjean S. Visualising association between paired 'omics' data sets. *In revision.*

Illustration of sparsePLS: variable plot

Some other insightful graphical outputs to highlight relationships between 2 data sets:

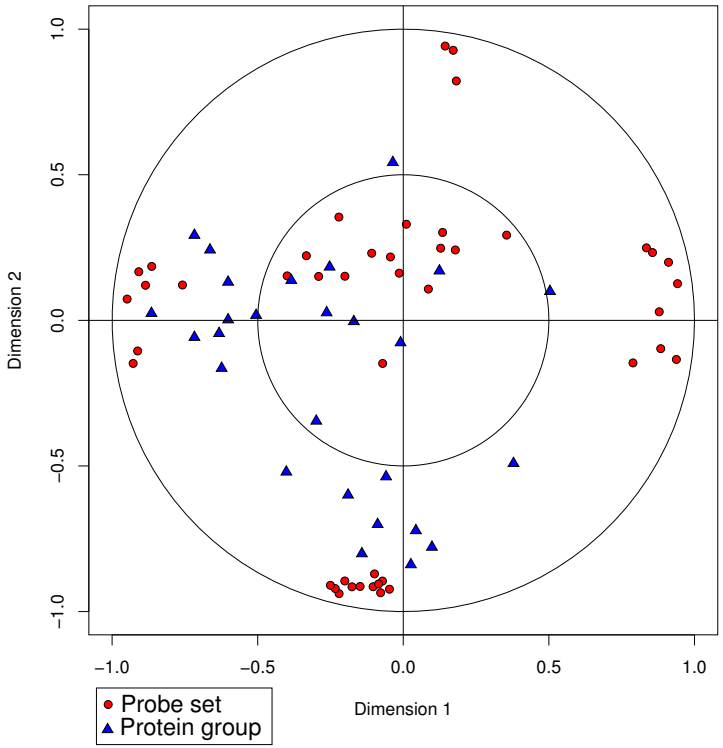


Figure: Correlation circle plots

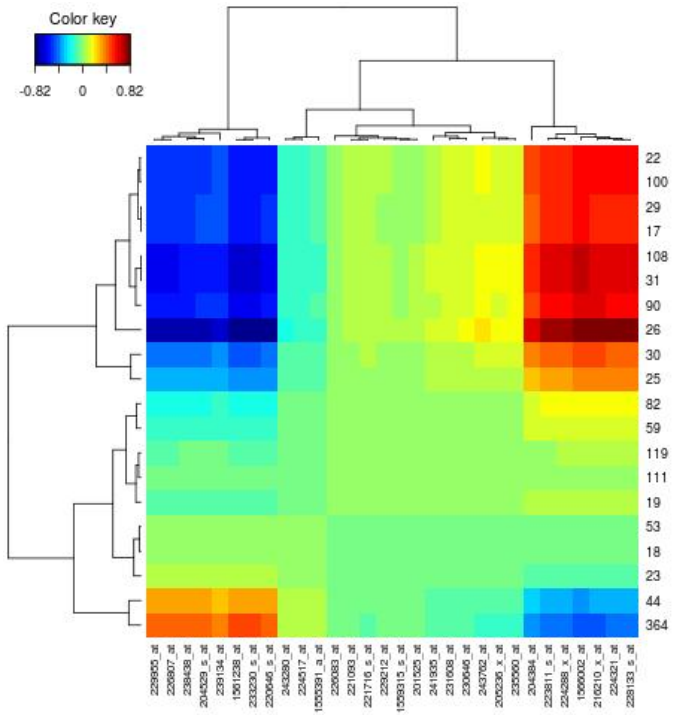


Figure: Clustered Image Maps

sPLS canonical mode: sample plots

Selects correlated variables across the same samples and highlights the correlation structure between the two data sets.
Canonical mode: "$\Leftrightarrow$ relationship"

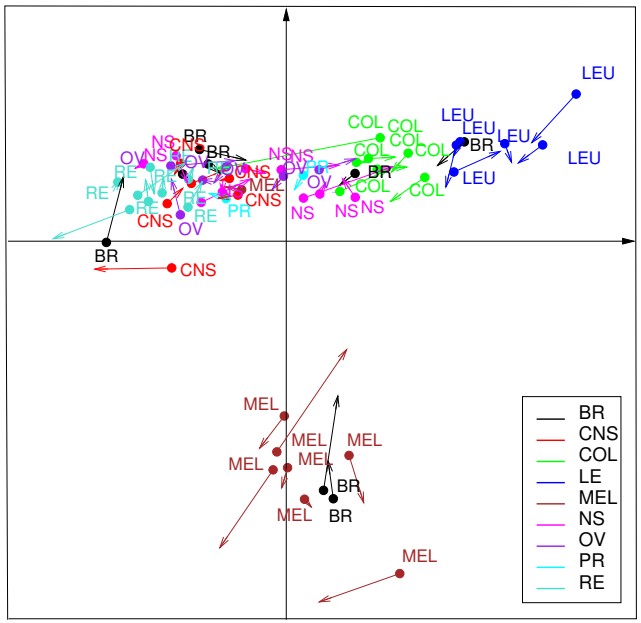


Figure: Arrow plot to highlight the similarities between 2 data sets

- NCI60 cross-platform study
- X = affymetrix, Y = cDNA microarray platforms
- Which are the variable from both sets that contain similar information on the samples?

Parameters to tune

- Number of PLS components:
  -Q_h^2 index
- Lasso penalizations λ_1^h, λ_2^h $(h = 1, \ldots, H)$:
  -regression mode: error prediction with cross-validation
  -canonical mode: maximisation of the covariance, stability analysis, permutations

$\rightarrow$ the biologist will also help choosing these parameters!

Cross-over design: VAC 18 study

- Peripheral blood mononuclear cells obtained before and after vaccination and simulated with 4 different conditions: HIV-LIPO5, GAG+, GAG- and NS.
- Transcripts ($p = 44,000$), Cytokines ($q = 10$), $n = 12$ unique patients

Subject	Stimulation			
After vaccination	LIPO5	GAG+	GAG-	NS
1	×	×	×	×
2	×	×	×	×
⋮	⋮	⋮	⋮	⋮
12	×	×	×	×

→ Take into account the correlation across conditions

from B. Liquet

Split up variation with a mixed model framework

As suggested by Westerhuis et al., 2010:

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

For the expression of variable k, subject s and treatment j,

$$x_{sj}^k = \underbrace{x_{..}^k}_{\text{offset}} + \underbrace{(x_{s.}^k - x_{..}^k)}_{\text{between-sample variation}} + \underbrace{(x_{sj}^k - x_{s.}^k)}_{\text{within-sample variation}}$$

where $\underbrace{(x_{sj}^k - x_{s.}^k)}_{\text{within-sample variation}} = \underbrace{(x_{.j}^k - x_{..}^k)}_{\text{Treatment effect}} + \underbrace{(x_{sj}^k - x_{s.}^k - x_{.j}^k + x_{..}^k)}_{\text{Error}}$

from B. Liquet

Multilevel approach

- Multilevel approach: explains the different parts of variation. The within-sample variation is explained by the Treatment factor.
- Objective: select the genes which can discriminate the 4 stimulations (stimulation effect)

$\rightarrow$ apply sPLS-DA on the within matrix $\boldsymbol{X}_w$ rather than the original data set $\boldsymbol{X}$ to take into account the repeated measures design.

from B. Liquet

Liquet, B. Lê Cao, K-A., Hocini, H., Thiébaut, R. A novel approach for biomarker selection and the integration of repeated measures experiments from two platforms, *submitted*.

Cross-over design

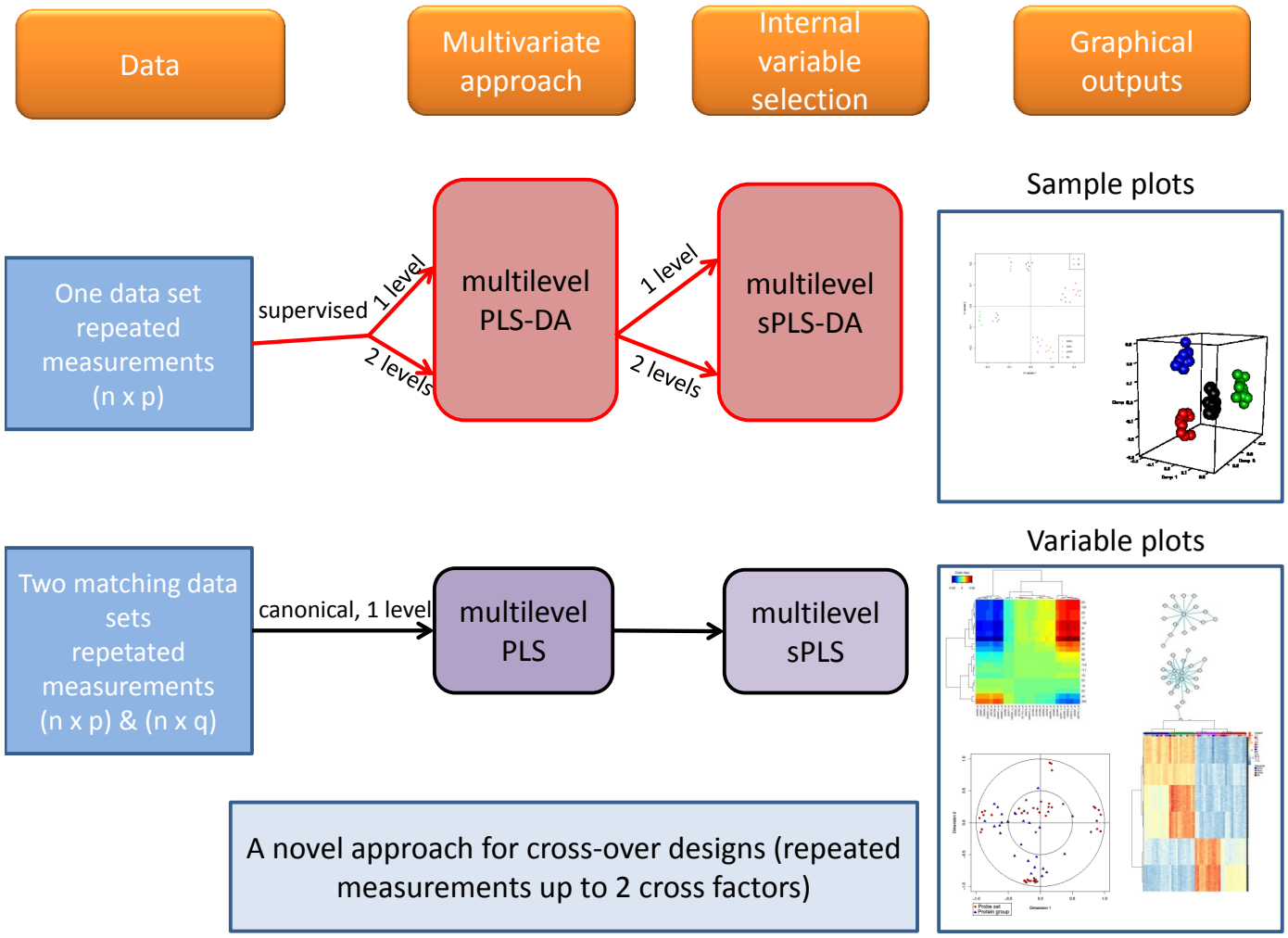


Illustration: multilevel analysis

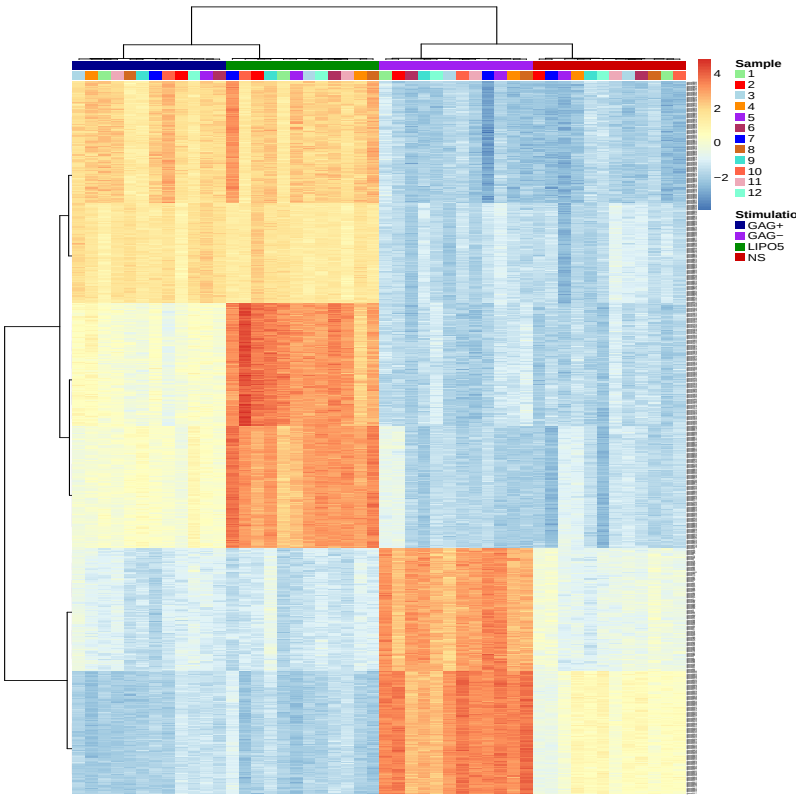


Figure: one level analysis

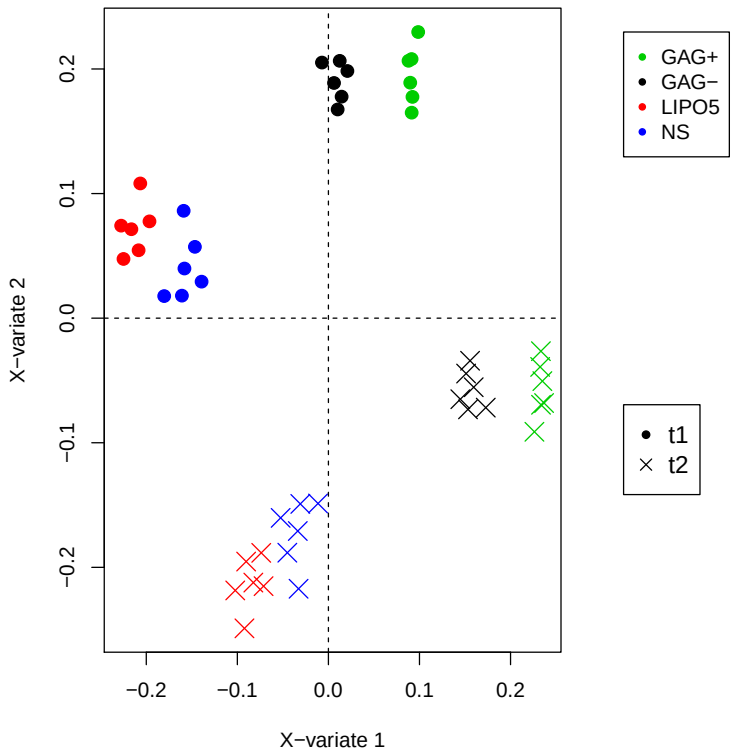


Figure: two level analysis

Regularized CCA

Classical Canonical Correlation Analysis solves the problem

$$\max cor_{a_h,b_h}(Xa_h, Yb_h) \quad \text{s.t.} \quad var(Xa_h) = var(Yb_h) = 1$$

For $n << p + q$, the empirical covariance matrices are ill-conditionned $\rightarrow$ canonical correlations close to 1.

In regularized CCA the covariance matrices are replaced by:

$$Cov(X) + \lambda_1 Id \text{ and } Cov(Y) + \lambda_2 Id$$

González I., Déjean S., Martin P.G.P., Goncalves O., Besse P. and Baccini A. 2009 Highlighting Relationships Between Heteregeneous Biological Data Through Graphical Displays Based On Regularized Canonical Correlation Analysis, *Journal of Biological Systems*, 17 (2).

Multi-block analysis: Regularized Generalised CCA

- RGCCA generalizes rCCA to more than 2 data sets
- Constitutes a general framework for many multi-block data analysis methods
- Objective: seeks linear combinations of block variables:
  (i) block components explain their own block well and/or
  (ii) block components that are assumed to be connected are highly correlated.

Tenenhaus, A., Tenenhaus, M (2011) Regularized Generalised Canonical Correlation Analysis, *Psychometrika*, 76 (2).

RGCCA

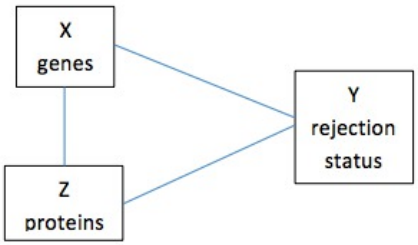
For J blocks of variables $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_j$, the design matrix $\boldsymbol{C} = \{c_{j,k}\}$, the function g and the shrinkage constants $\tau_1, \ldots, \tau_J$,
RGCCA optimizes the problem:

$$\max_{\boldsymbol{a}_1,\ldots,\boldsymbol{a}_J} \sum_{j,k=1,j\neq k}^{J} c_{kj} g(Cov(\boldsymbol{X}_j \boldsymbol{a}_j, \boldsymbol{X}_k \boldsymbol{a}_k))$$

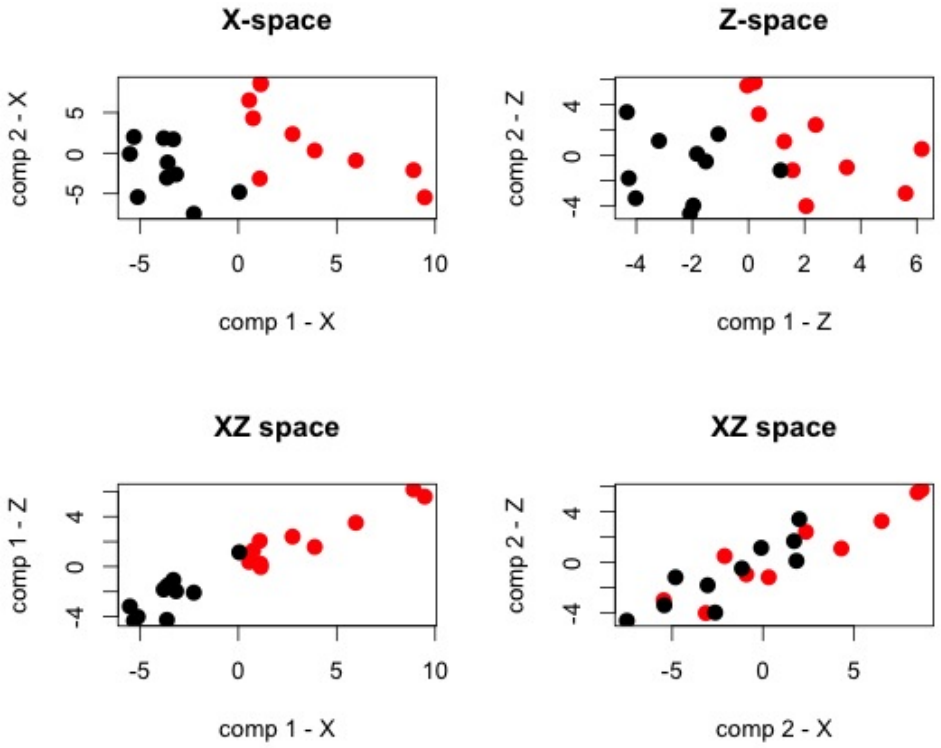
subject to the constraints $\tau_j ||a_j||^2 + (1-\tau_j) Var(\boldsymbol{X}_j a_j) \quad j = 1, \ldots, J$,
where the $\boldsymbol{a}_j$ are the loading vectors associated to each block j.

Similar to the sPLS, L_1 penalizations can be applied to the loading vectors to obtain a sparse version of RGCCA (sRGCCA, *in preparation*).

Illustration: design and sample plot



design	X	Z	Y
X	0	1	1
Z	1	0	1
Y	1	1	0



- Based on PLS path modelling: decide connexions btw blocks
- Also choose the g function depending on the type of relationship between loading vectors

Figure: PROOF data

Illustration: variable profile plot

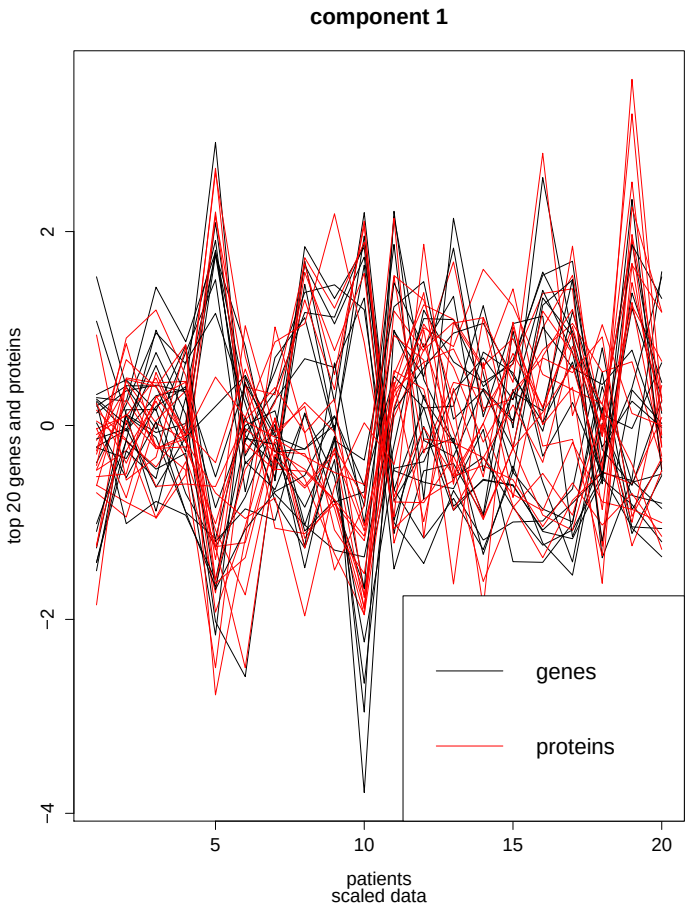


Figure: Loading vectors: comp. 1

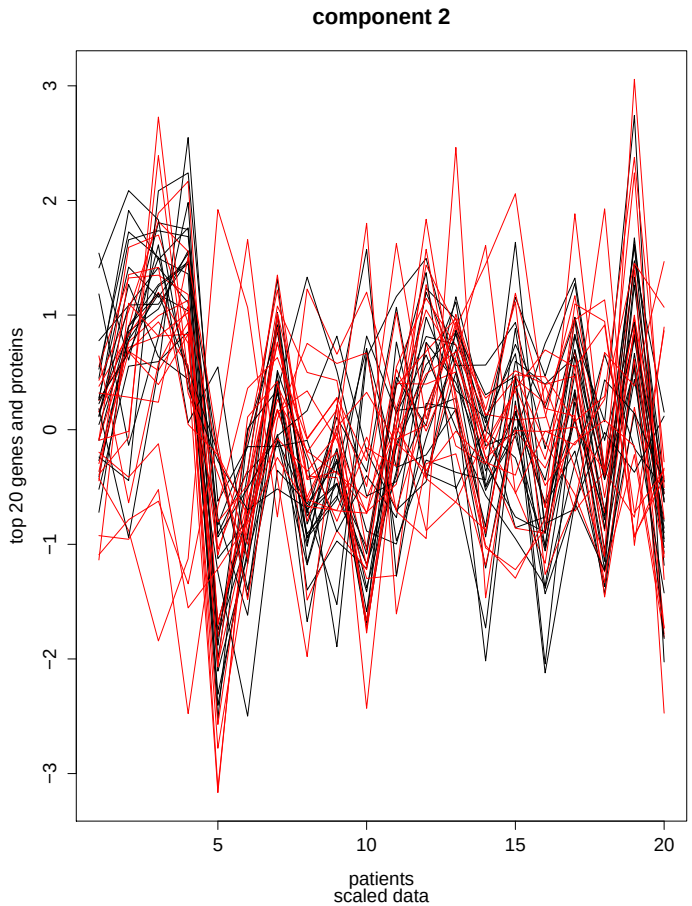


Figure: Loading vectors: comp. 2

These multivariate integrative approaches are:

- flexible and can answer various types of questions.
- can highlight the potential of the data.
- enable to generate new biological hypotheses to be further investigated.

:

- currently implements 6 different methodologies plus their sparse variants
- proposes comprehensive graphical outputs
- includes a web-associated interface

Future work includes:

- Cross-platform comparison and time-course experiments integration

Acknowledgements

mixOmics team

Sébastien Déjean	Univ. Tlse
Ignacio González	Univ. Tlse
Xin Yi Chua	QFAB

PROOF Project

Oliver Günther	PROOF
Scott Tebutt	PROOF

VAC18 Project

Benoît Liquet	Univ. Bordeaux 2
RodolpheThiébaut	Univ. Bordeaux 2

RGCCA

Arthur Tenenhaus	Supelec Paris
Vincent Frouin	CEA

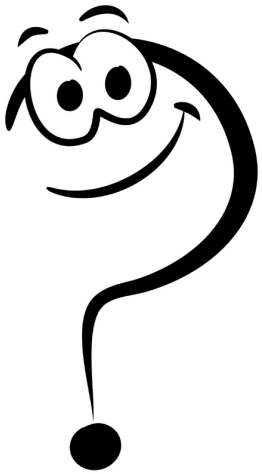
Other contributors to mixOmics

Jeff Coquery	QFAB, Sup'Biotech
Eric Fangzhou Yao	QFAB, Shanghai Univ
Mourad Larbi	QFAB, Univ. Nice
Pierre Monget	QFAB, CESI

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

Questions?

Watch out the new look of our web interface!
(in progress)



http://mixomics.qfab.org

mixomics@math.univ-toulouse.fr