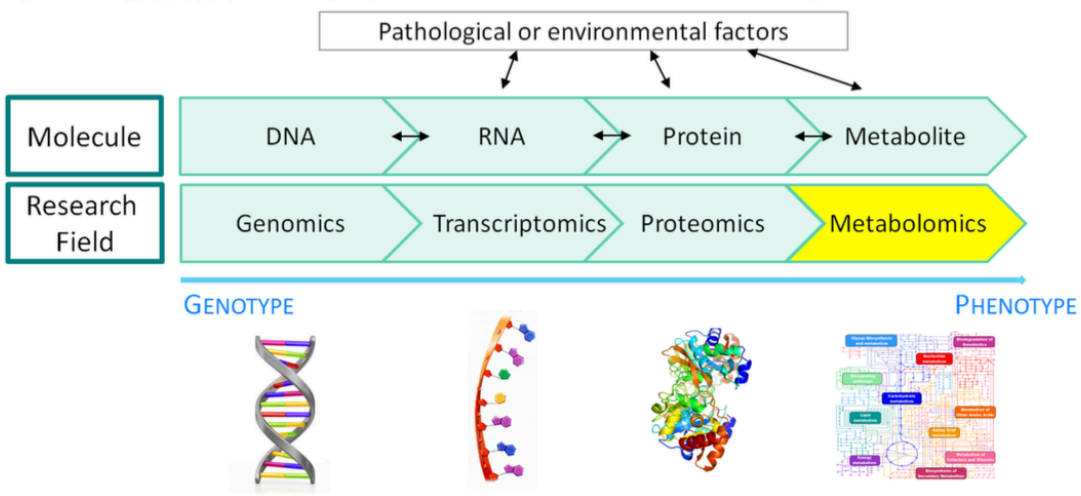


How can statisticians contribute in the analysis of (high-dimensional) OMICS data ?

The practical case of metabolomic spectral data.

Bernadette Govaerts

Institute of statistics biostatistics and actuarial sciences

ISBA – IMMAQ – Université catholique de Louvain

Looking for inspiration….

- ISBA work ?
- Data Sciences ?
- Big data ?
- Adapted to the public ?
- Coherent with my profile

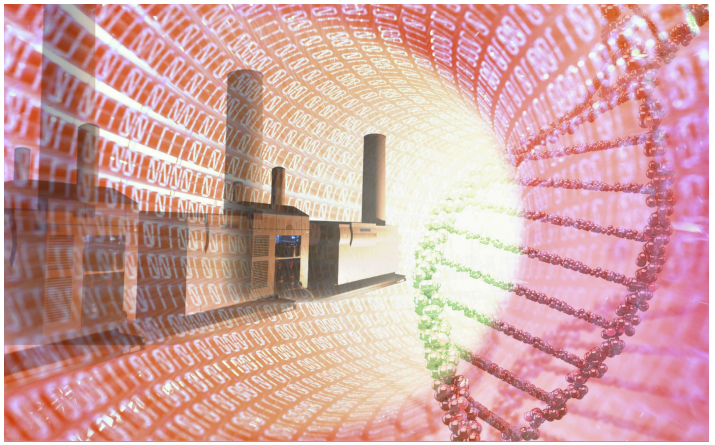
Omics data and metabolomics data and possible statistical contributions

Omics data world

Current projects

- Data preprocessing
- Analysis of complex designs
- Biomarkers discovery

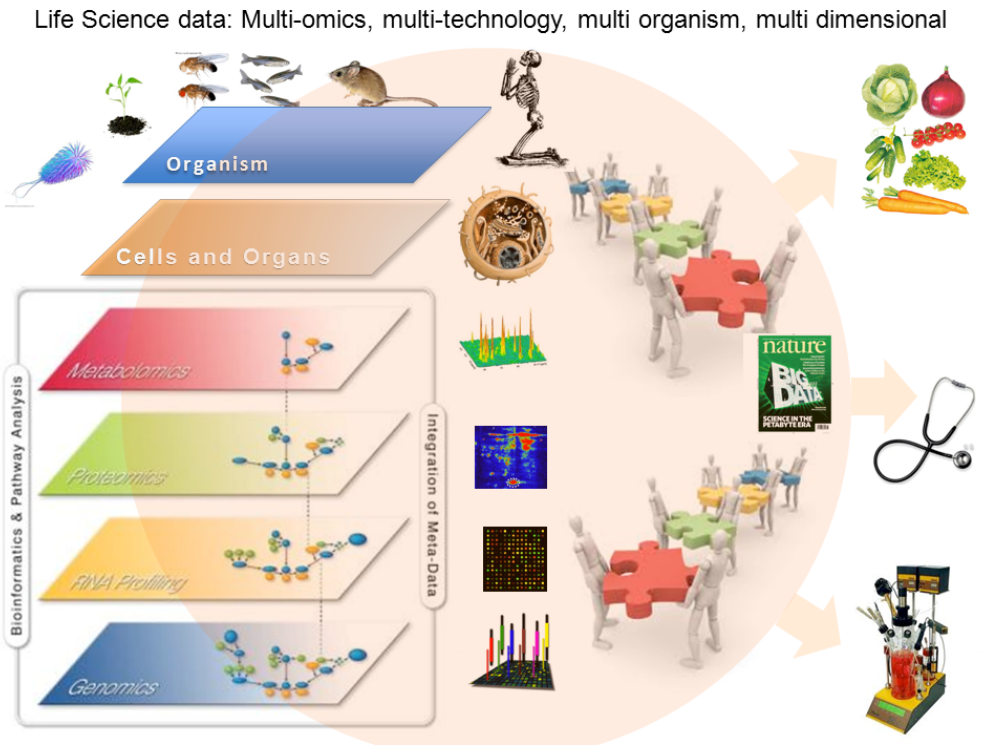
Conclusion



Omics data world

The -omics sciences levels

Omics aims at the collective characterization and quantification of pools of biological molecules that translate into the structure, function, and dynamics of an organism or organisms.



Omics sciences generate enormous amount of data,

many data analysis challenges are transversal

BUT data acquisition technologies are very specific

The special case of metabolomic data

Techniques used to measure metabolomic profiles : spectroscopy (MS – H-NMR)

Final data : 1D or 2D spectra. Basic 1D H^1-NMR spectra = 32768 points

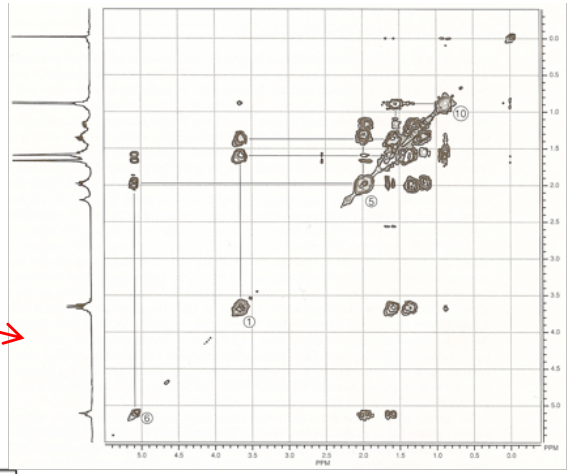
One molecule : multiple peaks in the spectra

Sampling

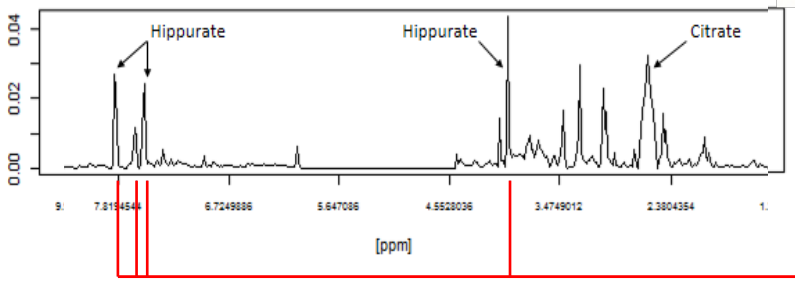
Data Acquisition

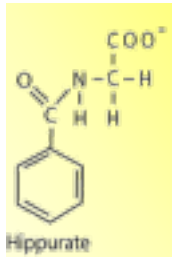
Data Preprocessing

2D Spectra



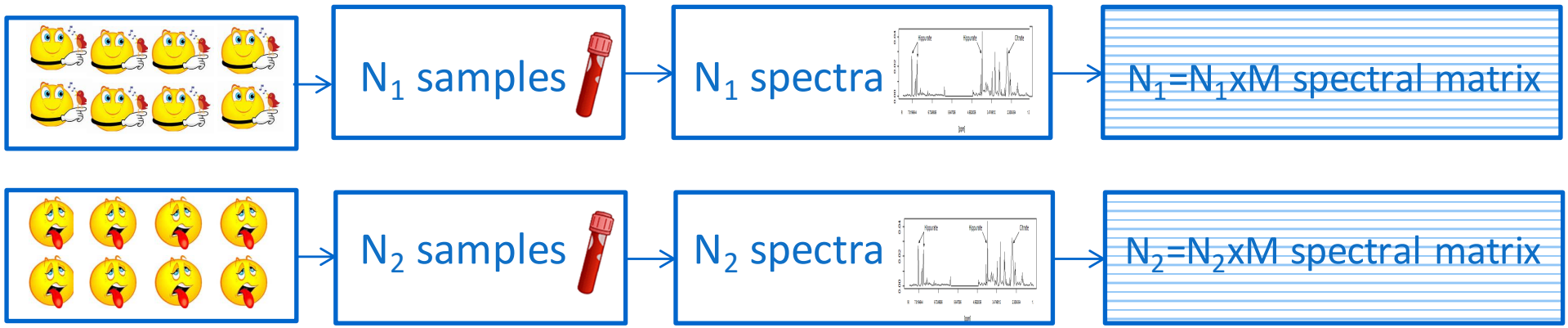
1D Spectra





Typical metabolomic study and questions (1D spectra)

Two groups of subjects : healthy and ill.

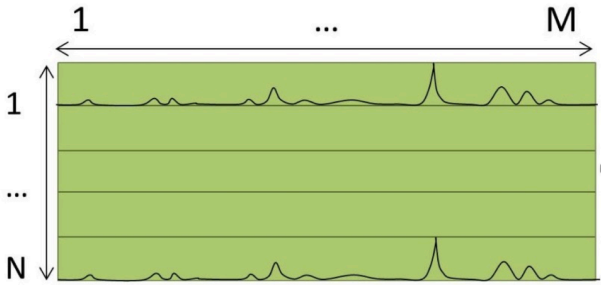


Two main goals

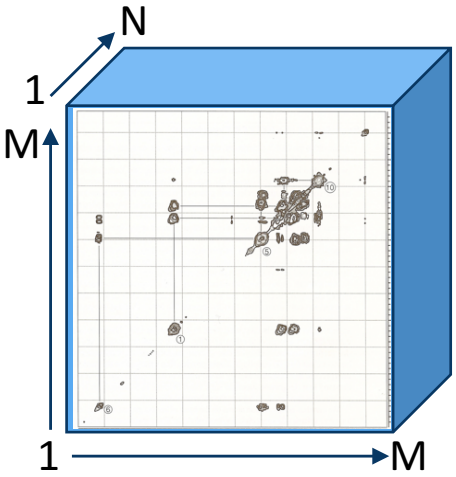
- Biomarkers discovery : Find combinations of molecules (biomarkers) that differentiate the 2 groups and understand corresponding chemical pathway.
- Diagnostic kit development : Develop predictive models to classify a subject in a group on the basis of the found biomarkers.

Metabolomic data properties

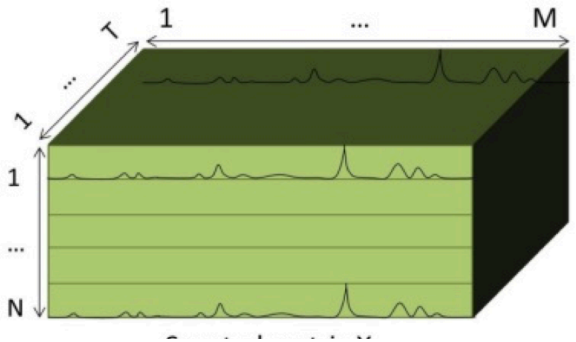
Typical data



1D Spectra matrices



2D Spectra array



3 or + spectral array

Advantages ☺

- Well structured
- Low missing values
- Big but not very big data …
- Experimental (controled design)

Difficulties ☹

- Expensive data. N often < 100
- Very wide data tables N << M
- High collinearity between variables
- Biological material with high variability (and instability of features)
- High instrumental noise

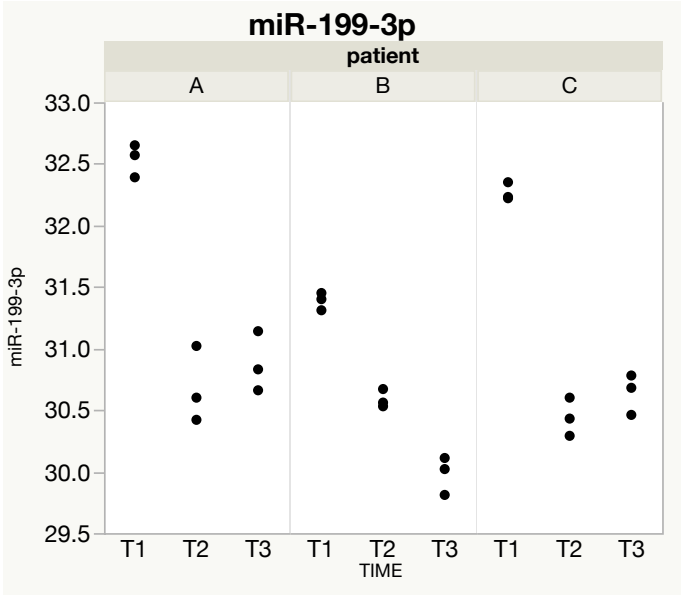
Illustration of biological and instrumental variation

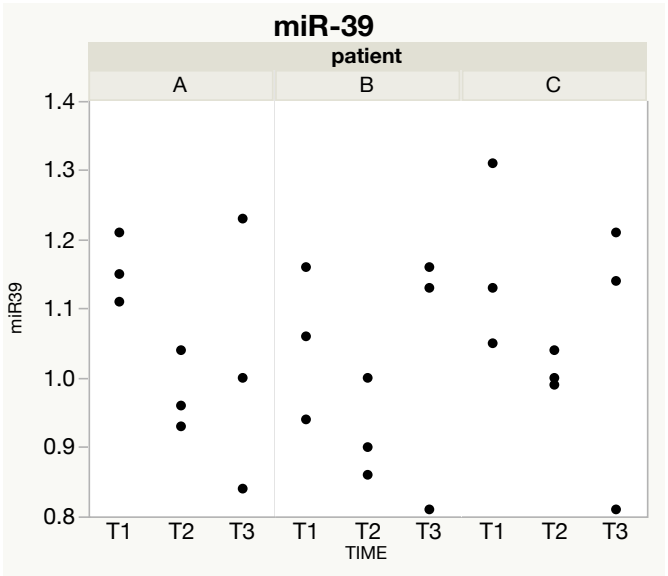
In Omics data, day to day variation in a patient and instrumental noise may be much higher than patient to patient variation !

Example of q-PCR – transcriptomic micro RNA (miR) data

Design : 3 Patients – 3 blood samples per patient – 3 instrumental replicates

Gene expression (Normalized)





And the goal finally is to find a difference between patients !….

Classical metabolomic data analysis methods

Origin : Chemometrics

Principle

Linear Dimension reduction methods

Methods : PCA, PLS, O-PLS, O-PLS-DA, ICA…

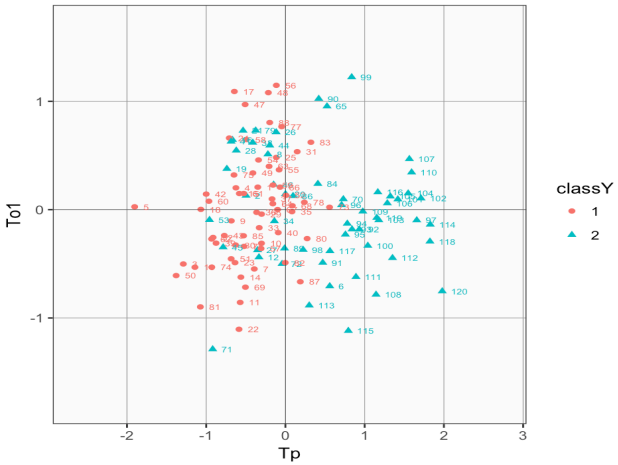
Avantages

- Visual
- Simple to interpret by practitioners

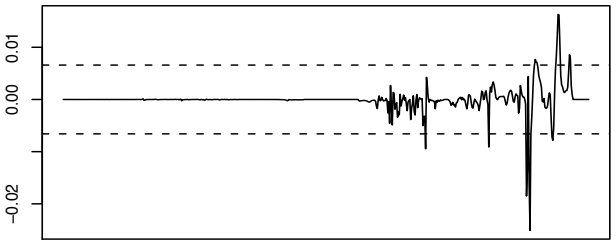
Limitations

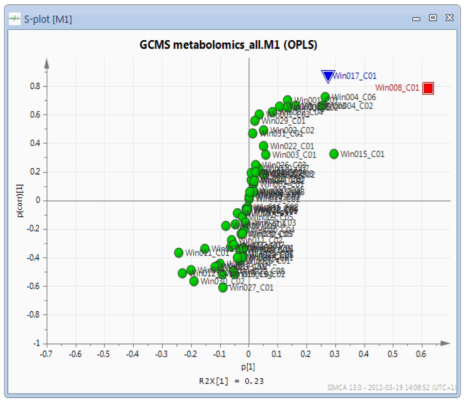
- Only adapted to very simple designs (e.g. 2 classes)
- No direct feature selection
- Few statistical testing methods
- Linearity not always adapted to biological data
- Low emphasize on data pre-treatment

O-PLS-DA Score plot

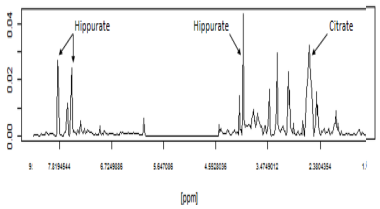


Loading plot and S-plot





Where can the statisticians help ? And ISBA Work



DOE : Design (and analyse) complex multifactor metabolomic experiments – sample size estimation

Evaluate and compare the quality of spectral data sets.

Spectral data pretreatment based on complex mathematical and statistical algorithms

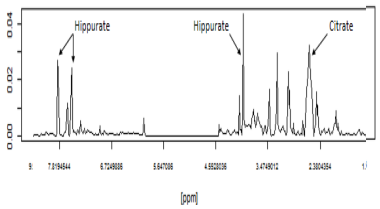
Data analysis

- Find biomarkers ("feature selection")
- Develop Prediction models
- Integrate multi-omics data (metabo, geno, clinical…)

Current medical omics projects

- Metabiose → Endometriosis (SPWallonia with ULG)
- DMLA (FNRS with ULG and University of Bordeau)

Where can the statisticians help ? And ISBA Work



DOE : Design (and analyse) complex multifactor metabolomic experiments – sample size estimation

2

Evaluate and compare the quality of spectral data sets.

1

Spectral data pretreatment based on complex mathematical and statistical algorithms

Data analysis

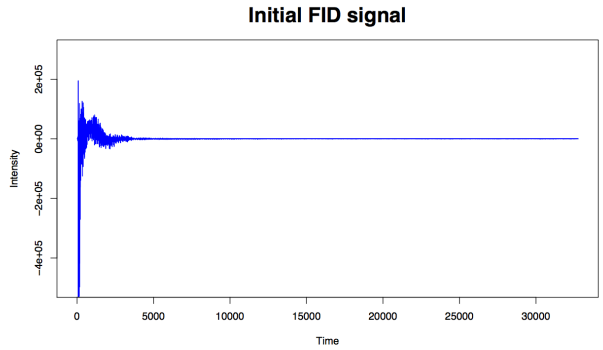
- Find biomarkers (“feature selection”)

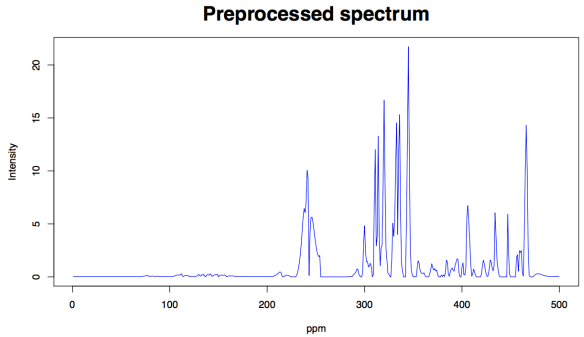
3

Data preatreatment – the needs

The minimal needs

Transform the signal from time to frequency domain





Problems

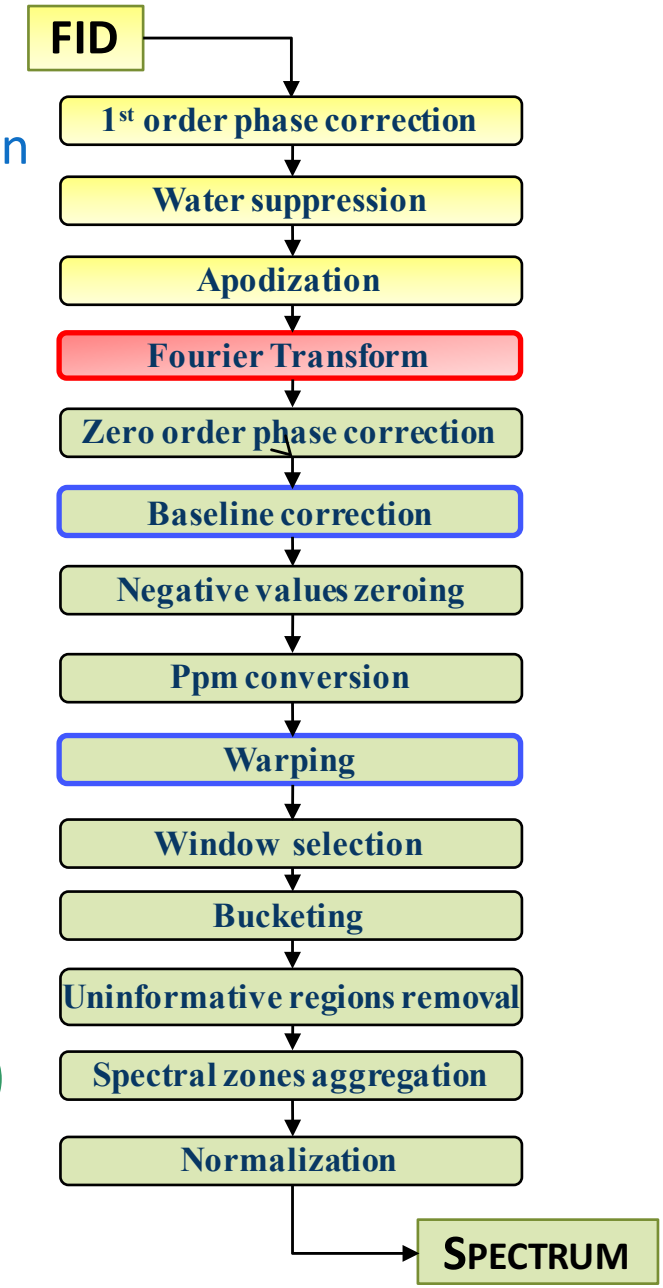
Many sources affect the quality of the final spectra

R library developped at ISBA (M. Martin)

- Full pretreatment workflow
- Include several innovant algorithms
- Many inputs by Paul Eilers

Default solutions provided by Brucker (and others)

Black box manual solutions with limited features



Data pretreatment – Examples of PEPS algorithms

Use of non parametric smoothing methods in spectral pretreatment.

Baseline correction

Asymetric smoothed LS

$$Q = V + \lambda R = \sum_{j=1}^{m} \psi_j (F_j - Z_j)^2 + \lambda \sum_{j=3}^{m} (\Delta^2 Z_j)^2$$

$$\text{if } F_j > Z_j,\ \psi_j = p,\ \text{else } \psi_j = (1-p)$$

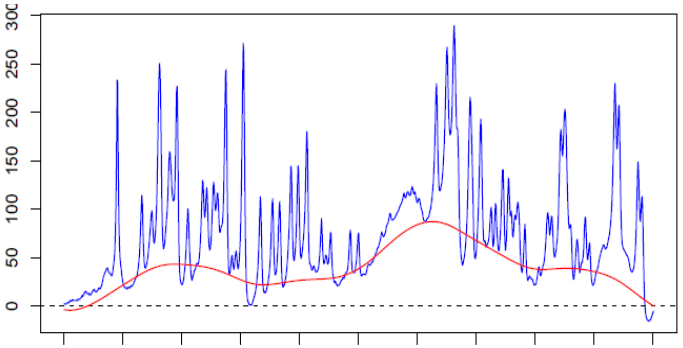
Warping

Smoothed distortion of frequencis axis

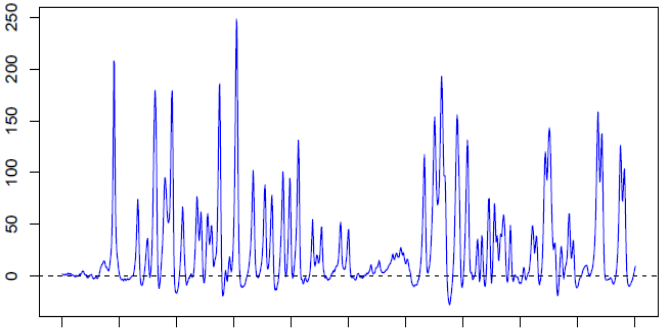
$$Q = \sum_{j=1}^{m} \left[G_j - F_j^w\right]^2 + \lambda \sum_{l=3}^{L} (\Delta^2 \alpha_l)^2 + \kappa \sum_{l=1}^{L} \alpha_l^2$$

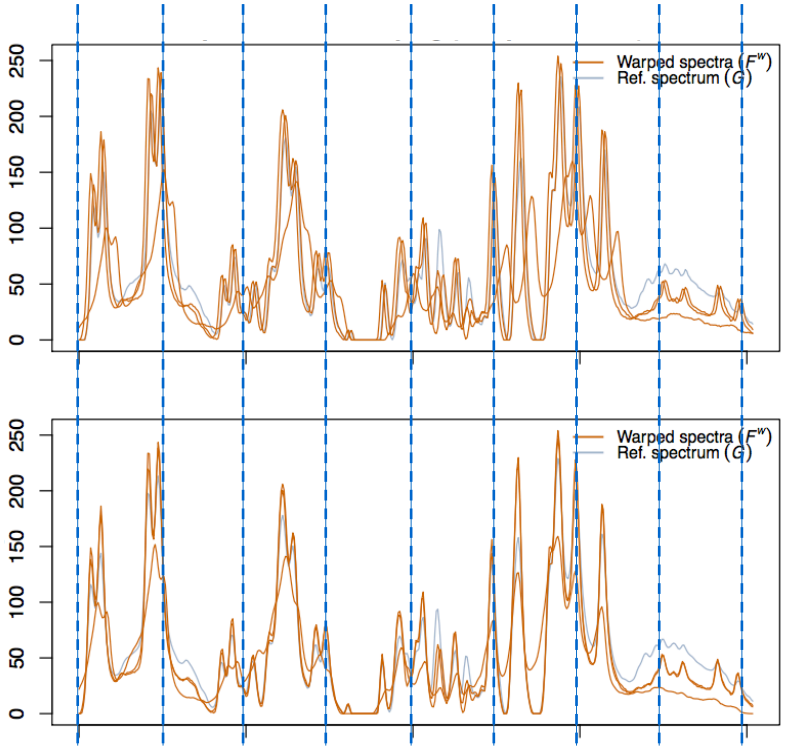
$$w(\nu) = \sum_{k=0}^{K} \beta_k \nu^k + \sum_{l=1}^{L} \alpha_l B_l$$

Before



After

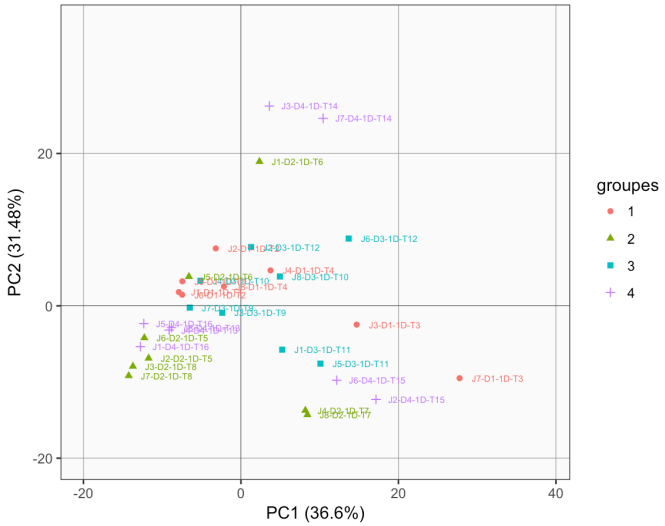




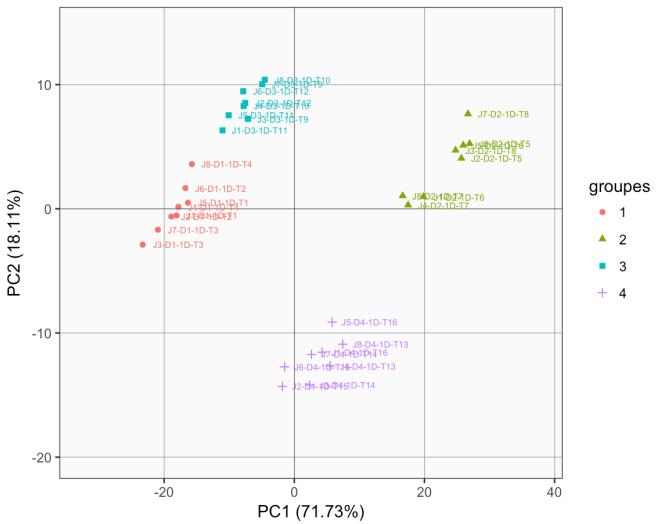
Data pretreatment – What can we gain ? Where are we now ?

What can we gain with PEPS ?

Basic pretreatment



PEPS pretreatment



	BI	WI	DunnW	RandW	AdjRandW	Mean.Q2
HumSerum_PEPS_Full	88.01	11.99	0.6406	1	1	0.9479
HumSerum_PEPS_Basic	20.54	79.46	0.3055	0.6714	0.09719	0.5036

Current state of PEPS R Library

- Available on Gitup
- Part of W4m a Galaxy web suite for metabolomic data analysis

Workflow4metabolomics

Complex (metabol)-omics designs analysis

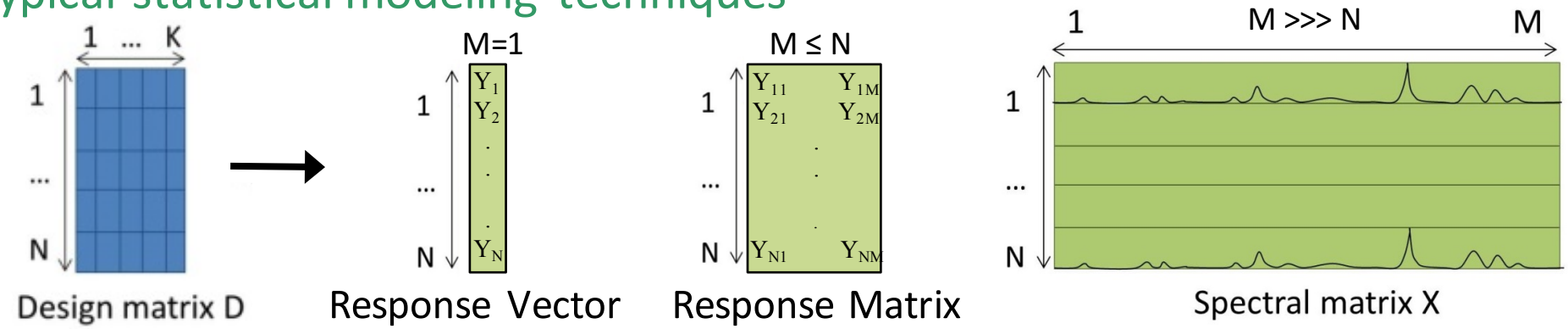
Standard design : Classify between 2 groups

When are more complex analysis needed ?

- Complex medical studies : longitudinal or cross-over studies…

  Possible factors : treatment, time, sex, age, etc…
- Designs to study the variance components in omics technologies

Typical statistical modeling techniques



- M=1 : Regression, ANOVA, Linear mixed models.
- M≤N : MANOVA and M-General Linear Model – strict hypothesis
- M>>N : Need for a combination of dimension reduction and statistical modeling

Complex (metabol)-omics designs analysis

Solution used in chemometrics : ASCA, APCA…

Combination of dimension reduction techniques (PCA, PLS, PARAFAC, ComDim) and *balanced simple ANOVA* models. → *too limited for omics sciences*

ISBA interests (M. Thiel and M. Martin)

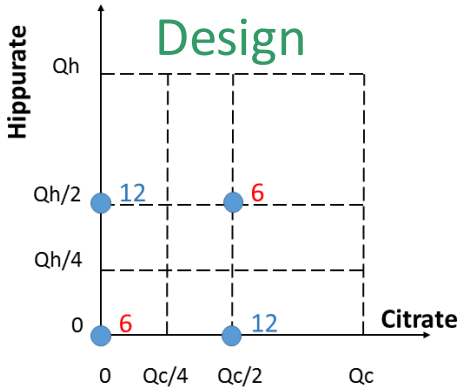
Treat : unbalanced designs and mixed linear models

Deliver : graphical solutions + statistical numerical results.

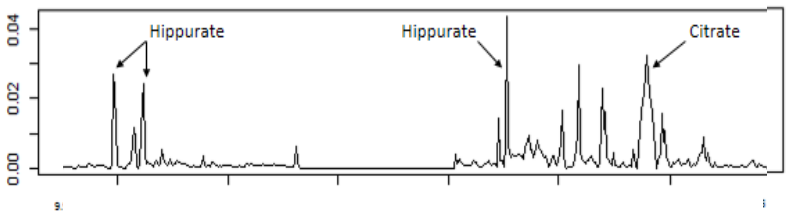
APCA+ approach (M. Thiel)

→ Solves the problem of unbalancing in General Linear Models.

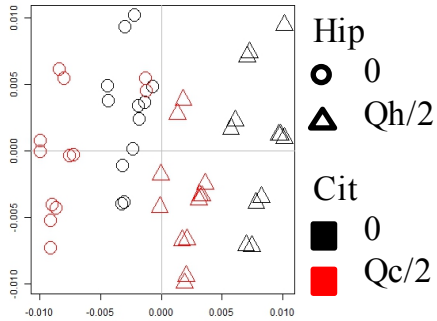
Design



Typical spectra



Classical PCA



ANOVA2 – APCA+ example

Step 1 : Parallel GLM and effect matrix estimation

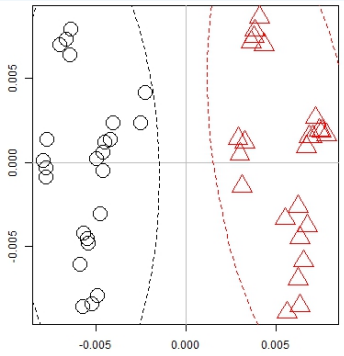
Ex: (1) ANOVA 2 crossed model : $y_{ijk} = \mu_{..} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$

(2) Effect matrix decomposition : $\mathbf{Y} = \mathbf{M}_0 + \mathbf{M}_A + \mathbf{M}_B + \mathbf{M}_{AB} + \mathbf{E}$ (**M** is nxm)

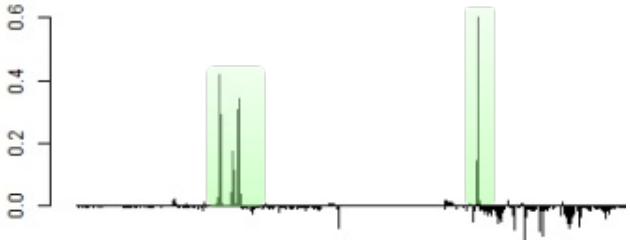
Step 2 : Decompose augmented effect matrices by SVD

$\mathbf{M}_A + \mathbf{E} = T_A P_A'$ $\mathbf{M}_B + \mathbf{E} = T_B P_B'$ $\mathbf{M}_{AB} + \mathbf{E} = T_{AB} P_{AB}'$

Hippurate score plot



Hippurate loading plot

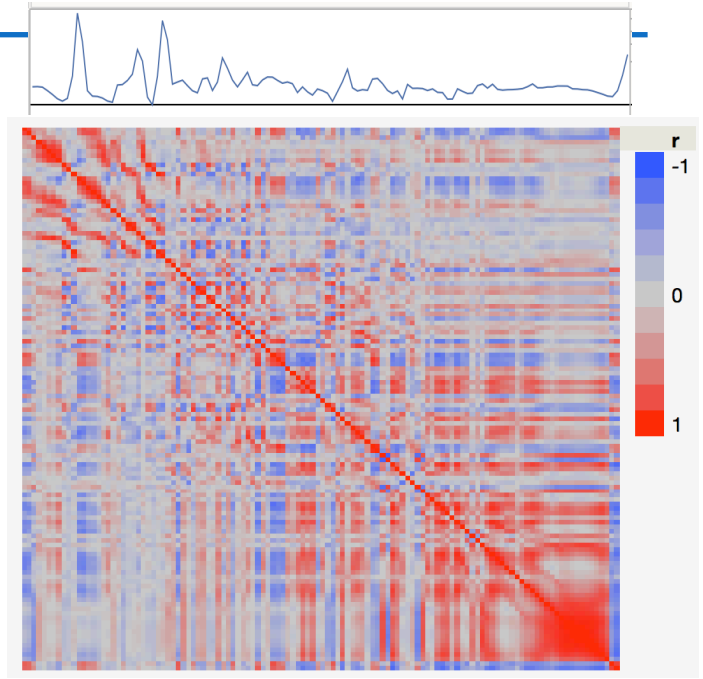


	% Var	P-value
Hippurate	18.1	0.000
Citrate	12.1	0.004
Interaction	1.0	0.950
Error	53.9	
Total	85.1	

Technical difficulties in unbalanced cases

Effect matrices and contribution computation, significance tests setup…

Biomarker discovery – Current interests



State of the art : O-PLS-DA

Needs

- Include feature selection IN the classification method → Sparse methods
- In spectra, correlations make sens
- Graphics results are needed for practitioners

Project 1 : Sparse PLS and O-PLS (B. Féraud)

- Principe : supervised dimension reduction with sparse latent variables
- Show good results in variable selection AND predictive power

Project 2 : Sparse fuzzy group lasso (R. Marion)

- In development : goal : get interpretable selected features.

$$\hat{\beta}^* = \underset{\mathbf{W},\beta^*}{\arg\min} \frac{1}{2n} \|\mathbf{y} - \mathbf{X}^*\beta^*\|_2^2 + \lambda_1 \sum_{k=1}^{K} \|\beta_k^*\|_2 + \lambda_2 \sum_{k=1}^{K} \sum_{j=1}^{m} (w_{jk})^d \left\| \mathbf{x}_j \beta_j - \frac{\sum_{l=1}^{m} (w_{lk})^d \mathbf{x}_l \beta_l}{\sum_{l=1}^{m} (w_{lk})^d} \right\|_2^2$$

Prediction Sparse Fuzzy groups of features

Conclusion

Working in omics data and data sciences is

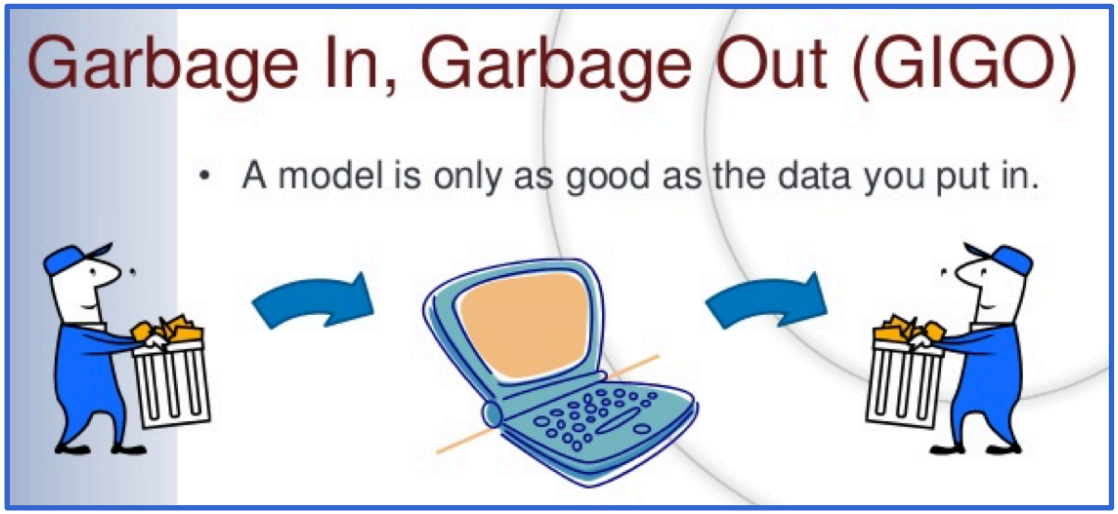
- Fun, exciting, challenging !
- It offers many opportunities to data scientists, statisticians, computer scientists, bioinformaticians, machine learners to develop
  - New methods and algorithms
  - Powerfull computer softwares
  - New theories about high dimensional data

BUT

- Omics data and can be dirty and non informative
- Omics are technical and team work in essential.

Data scientists must be active actors in this context

The crucial step : Data acquisition and preparation



Data scientists must accept to spend time at this early stage of the process.

How can data scientist contribute to the quality of (omics) data ?

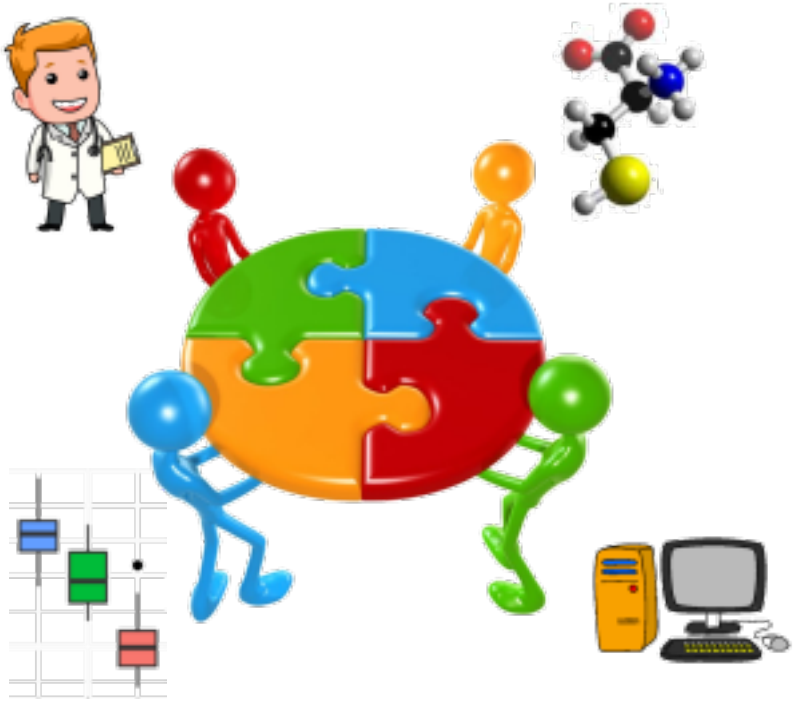
- Be present in the design of the study
- Calculate sample sizes,
- Quantify objectively the sources of variability data before the study
- Propose data pre-proccessing tools to improve the data before analysis

Data sciences is a team work

To succeed in data sciences projects, many skills are needed !

Working in interdisciplinary teams is compulsary.

- Will and capacity of collaboration in the team is needed.
- The difficulty and time involved must be recognized in the academic world.
- Young researchers must be integrated in well organized teams to become quickly efficient.



Thank you for your attention !
Do you have questions ?

Thanks to our partners

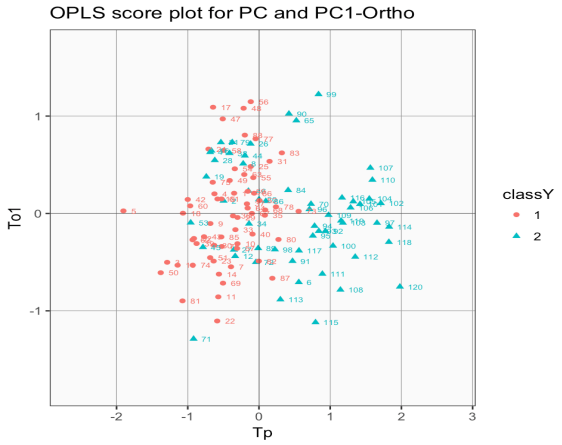
Infrastructure nationale
en métabolomique et fluxomique

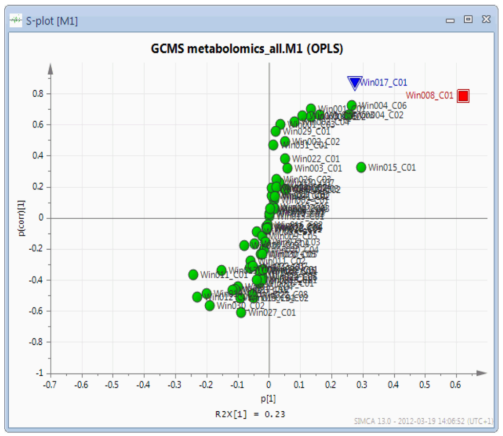
Matériel supplémentaire

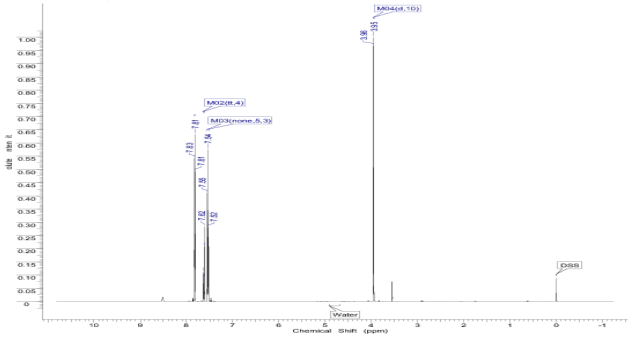
Biomarker discovery in metabolomics – state of the art

3 steps method used by 98% practionners

O-PLS-DA → S-Plot → Metabolites







Supervised classification method

- Response orthogonal information removal
- Dimension reduction
- Score plot

Features selection tools

- Popular S-plot
- Permutation tests
- Importance criteria

Metabolites research

Search of the link between selected features and metabolites

Classical metabolomic data analysis methods

Origin : Chemometrics

Principle

Linear Dimension reduction methods

Methods : PCA, PLS, O-PLS, PLS-DA, ICA…

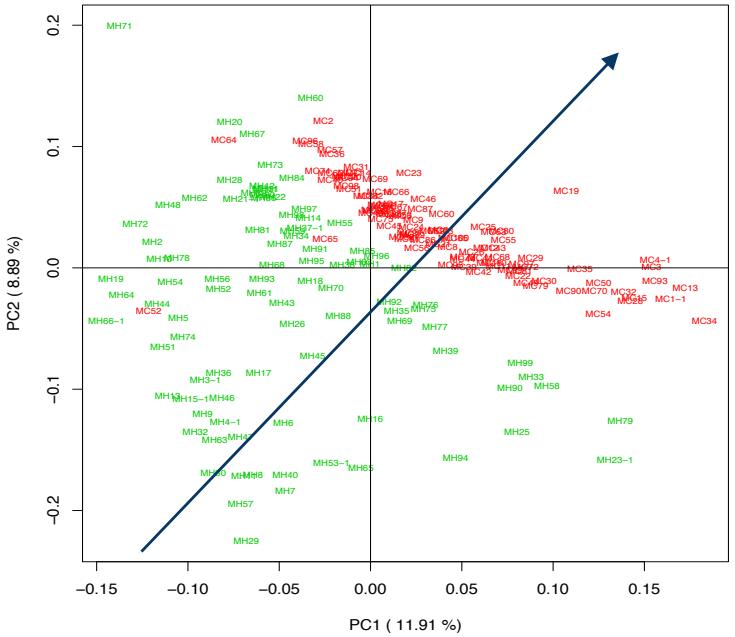
Avantages

- Visual
- Simple to interpret by practitioners

Limitations

- Only adapted to very simple designs (e.g. 2 classes)
- No direct feature selection
- Few statistical testing methods
- Linearity not always adapted to biological data
- Low emphasize on data pre-treatment

Score plot



Loading plot