Combination of Independent Component Analysis and statistical modeling for the identification of metabonomic biomarkers

Réjane Rousseau (Institut de Statistique, UCL, Belgium)

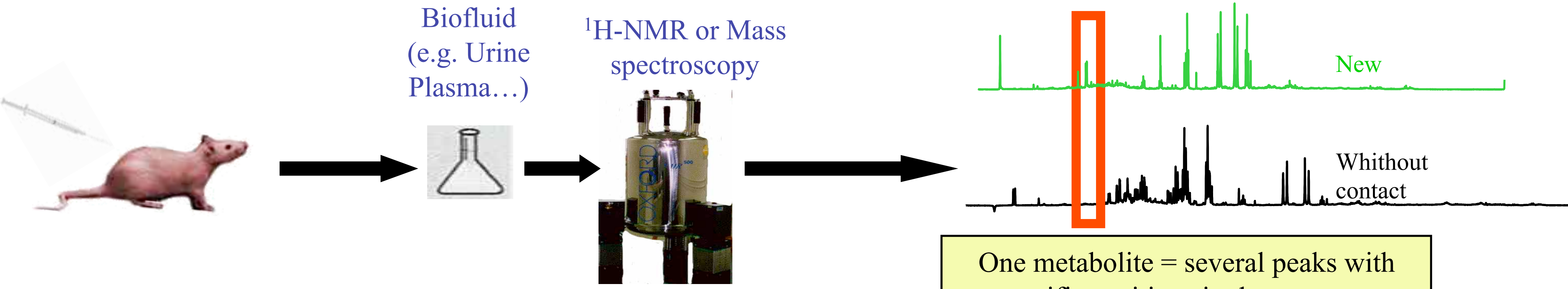
Joint work with Bernadette Govaerts and Michel Verleysen (UCL)

Metabonomics and biomarker identification

What is metabonomics ?

The study of biological responses to a stressor (ex: drug, disease) in the level of metabolites

Metabonomics in practice



Biofluid (e.g. Urine Plasma…)

^{1}H-NMR or Mass spectroscopy

One metabolite = several peaks with specific positions in the spectrum

Biomarker identification

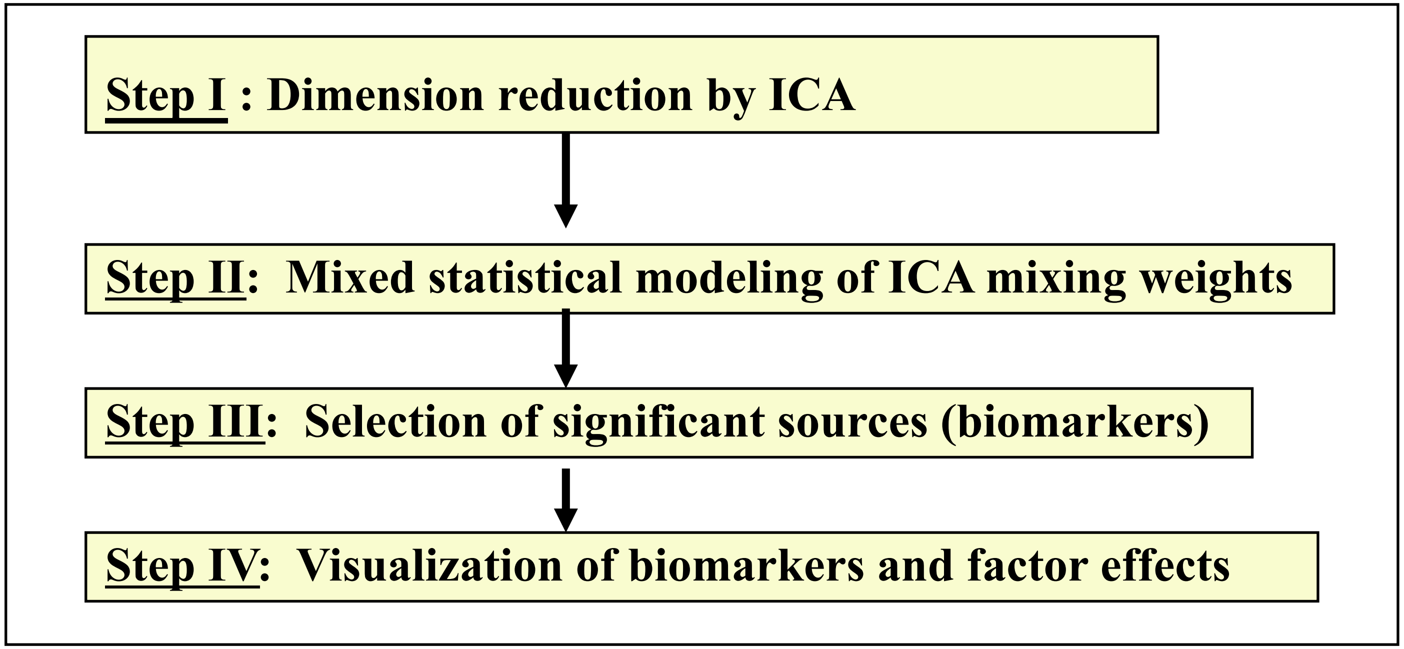
Find which metabolite or which part of the spectrum is alterated by a factor of interest (drug, disease…)

Objective of the talk:

to propose a methodology combining **ICA** and **statistical modeling for biomarker identification in ^{1}H-NMR spectroscopy.**

Outline of the talk

- Typical steps of a metabonomic study for the identification of biomarkers
- Overview of the methodology based on ICA and statistical modeling
- Data used in the talk
- Details of the methodology



Step I : Dimension reduction by ICA

Step II: Mixed statistical modeling of ICA mixing weights

Step III: Selection of significant sources (biomarkers)

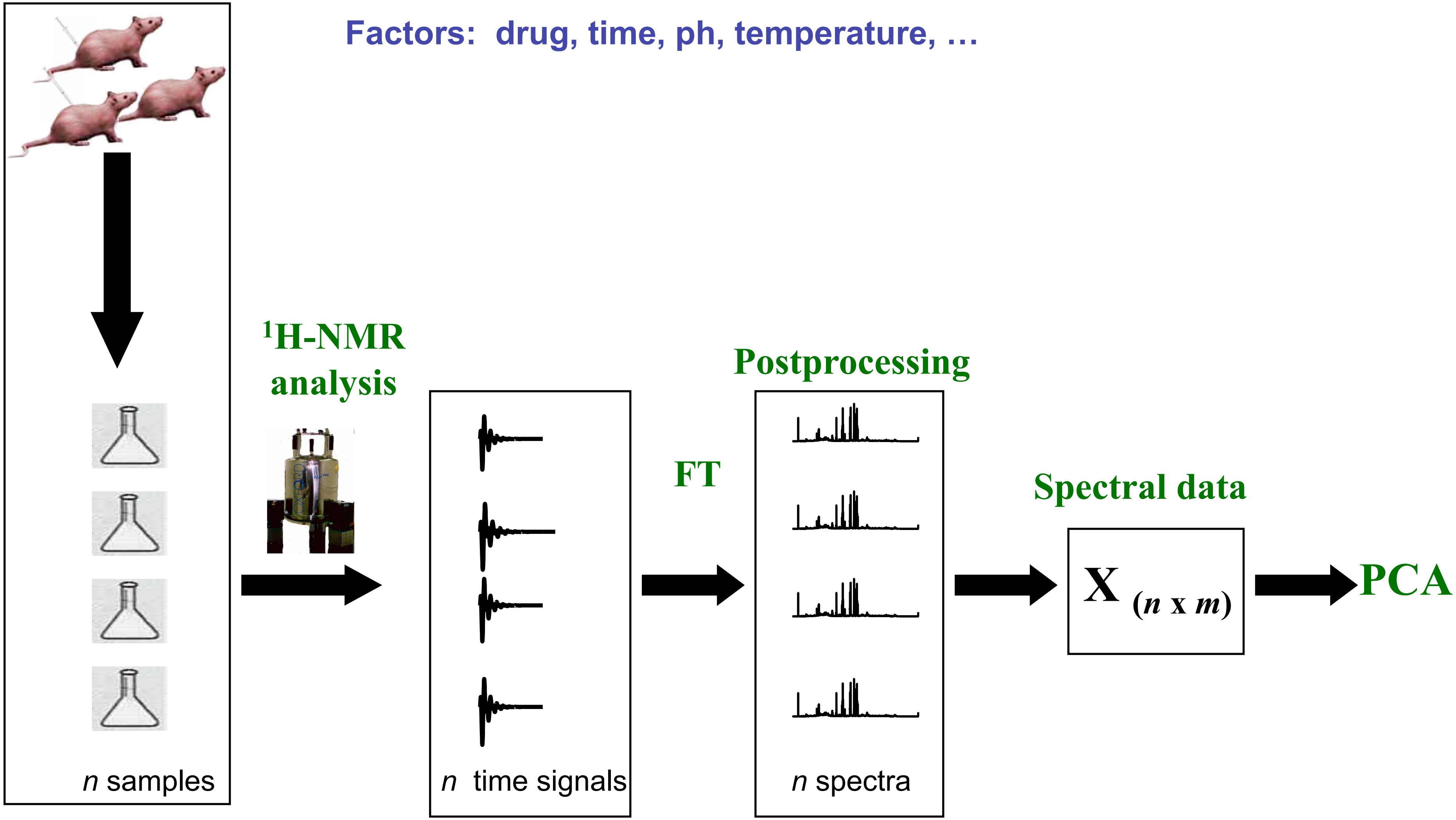
Step IV: Visualization of biomarkers and factor effects

- Conclusions.

Typical steps of a metabonomic study

Collection of biofluid samples under different conditions

Factors: drug, time, ph, temperature, …



n samples

^{1}H-NMR analysis

n time signals

FT

Postprocessing

n spectra

Spectral data

$\mathbf{X}_{(n \times m)}$

PCA

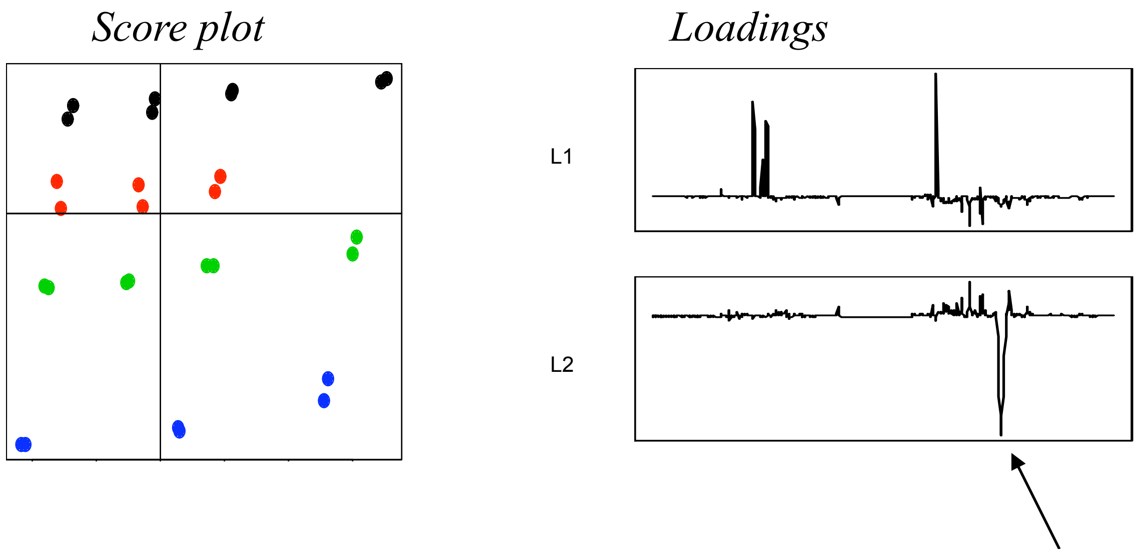
Typical steps of a metabonomic study

Spectral data

X (nxm)

PCA:

- ➢ Reduction of the dimension to obtain uncorrelated principal components
- ➢ Examination of the 2 first components to identify biomarkers



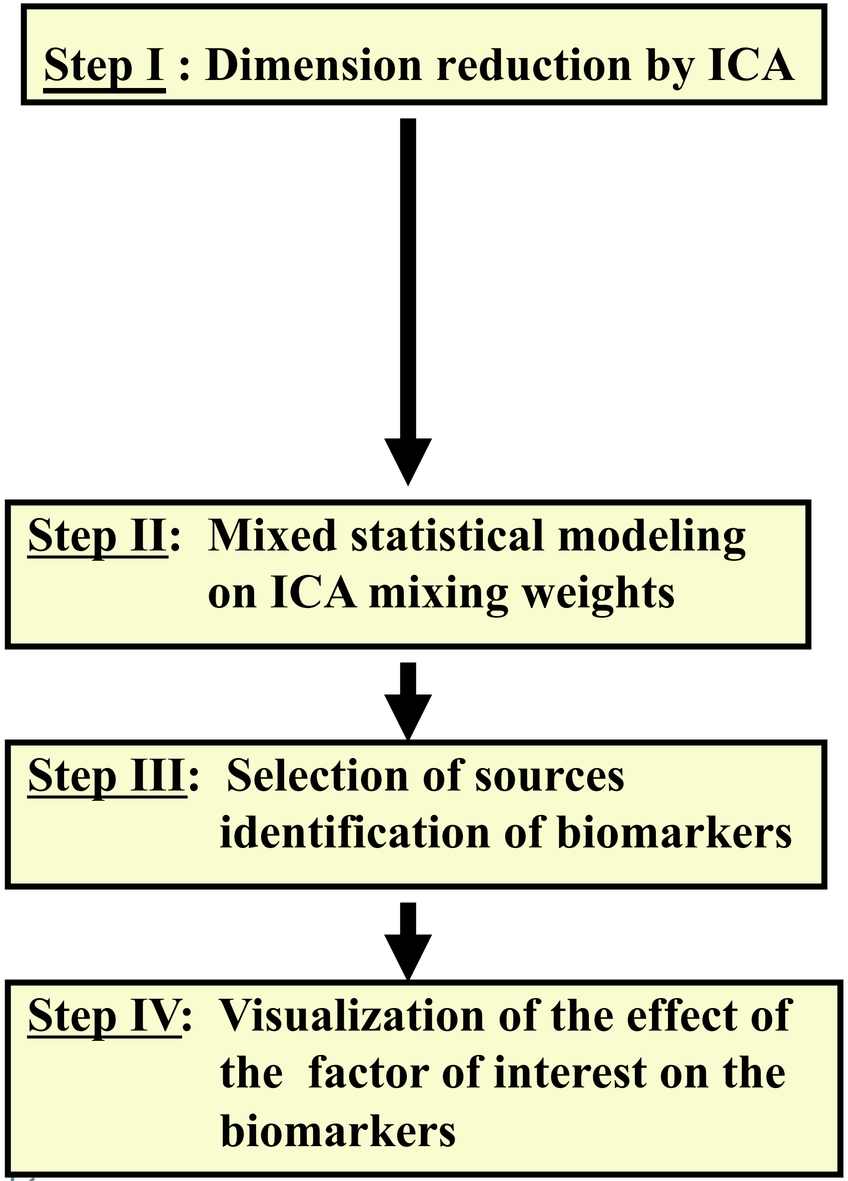
ex: this peak plays an important role

ex: colors = 4 groups of disease

Identification of biomarker

This is only powerful if the biological question is related to the highest variance in the dataset!

Methodology based on ICA and statistical modeling



Step I : Dimension reduction by ICA

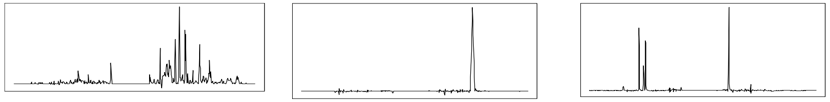
Step II: Mixed statistical modeling on ICA mixing weights

Step III: Selection of sources identification of biomarkers

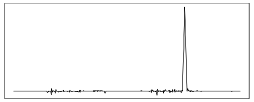
Step IV: Visualization of the effect of the factor of interest on the biomarkers

$$X^{TC} = \underset{\text{Components}}{S} \, . \, \underset{\text{Weights} \approx \text{quantity}}{A^{T}}$$

Examination of the ALL components:
to **visualize unconnected molecules in samples**



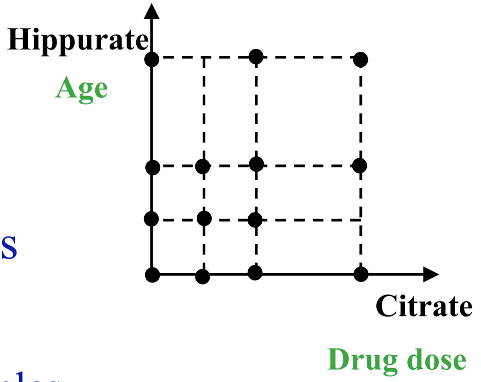
$$A^{T} = Z^{1}\beta + Z^{2}\gamma + \varepsilon$$



$$S^{*} \subset S$$

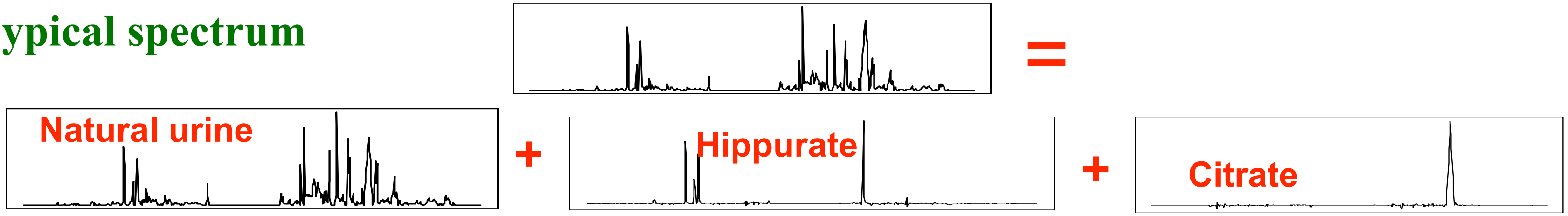
Data used in this talk

- **Prepared samples**
  - ➢ to know the spectral regions that should be identified as biomarkers
  - ➢ Mixtures of urine with citrate and hippurate
  - ➢ 14 experimental conditions – 2 replicates per condition = 28 samples



- **Spectra postprocessing**
  - ➢ Using Bubble a tool developped by Eli Lilly optimised for urine samples
  - ➢ Normalisation : unit sum - Resolution : 600ppms

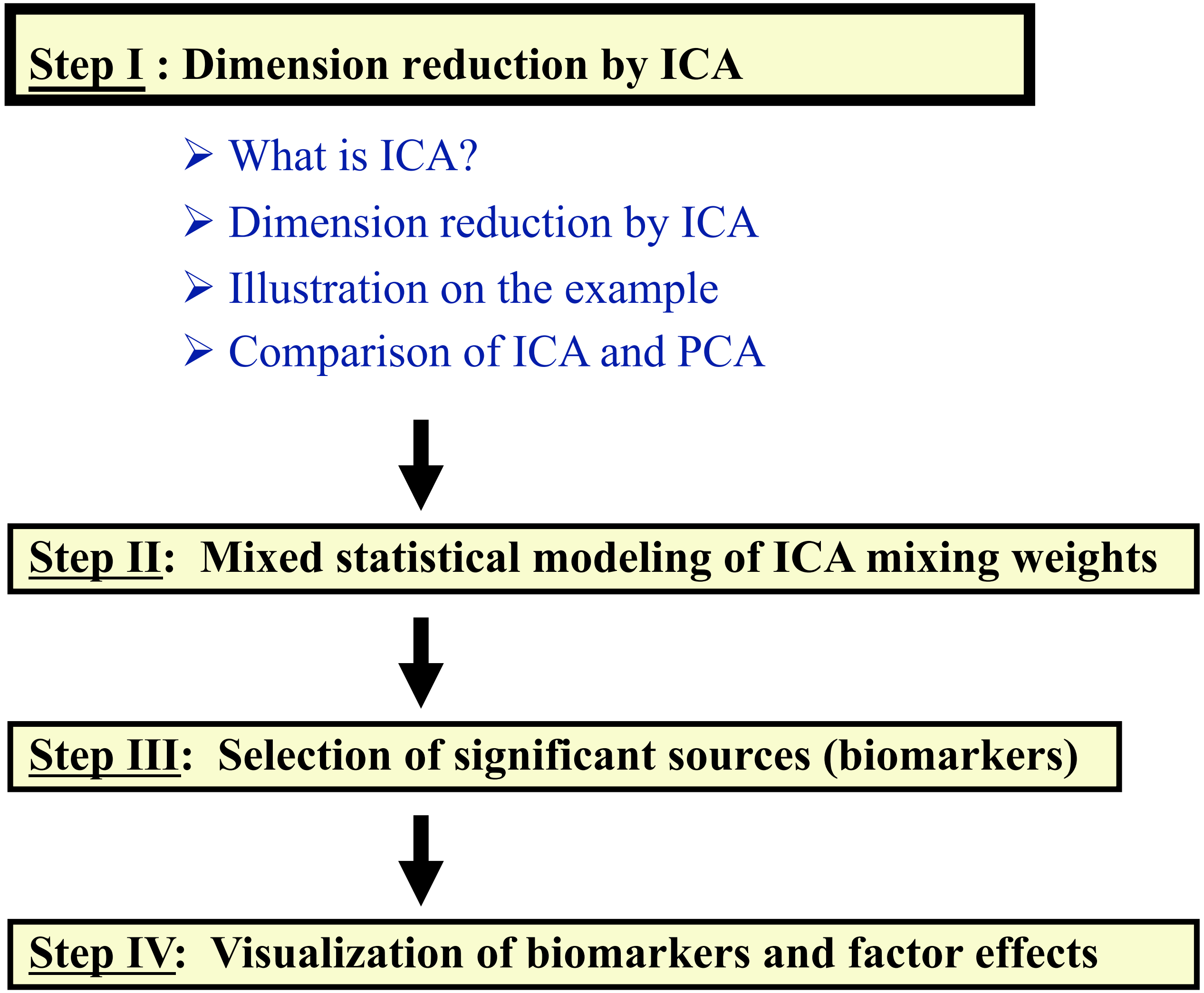
- **Typical spectrum**



<u>Hypothetical question</u>

- ➢ Assimilate the concentration of citrate as a **drug dose** received by the subject
  of hippurate as the **age** of the subject
- ➢ **Goal = to find a biomarker for the drug dose**
  i.e. discover « automatically » the citrate peak from the 28 spectra.

Methodology based on ICA and statistical modeling



Step I : Dimension reduction by ICA

- What is ICA?
- Dimension reduction by ICA
- Illustration on the example
- Comparison of ICA and PCA

Step II: Mixed statistical modeling of ICA mixing weights

Step III: Selection of significant sources (biomarkers)

Step IV: Visualization of biomarkers and factor effects

$$X^{TC} = S.A^{T}$$

Step I : What is Independent component analysis (ICA)?

The idea:

- Each observed vector of data (spectrum) is a linear combination of unknown independent (not only linearly independent) components

$$x_i = \sum_{k=1}^{l} s_k a_{ki} = s_1 a_{1i} + s_2 a_{2i} + \ldots + s_l a_{li}$$

- The ICA provides the independent components (sources, s_k) which have created a vector of data and the corresponding mixing weights a_{ki}.

How do we estimate the sources?

with linear transformations of observed signals that maximize the **independence** of the sources.

How do we evaluate this property of independence?

Using the **Central Limit Theorem** (*), the independence of sources components can be reflect by non-gaussianity.

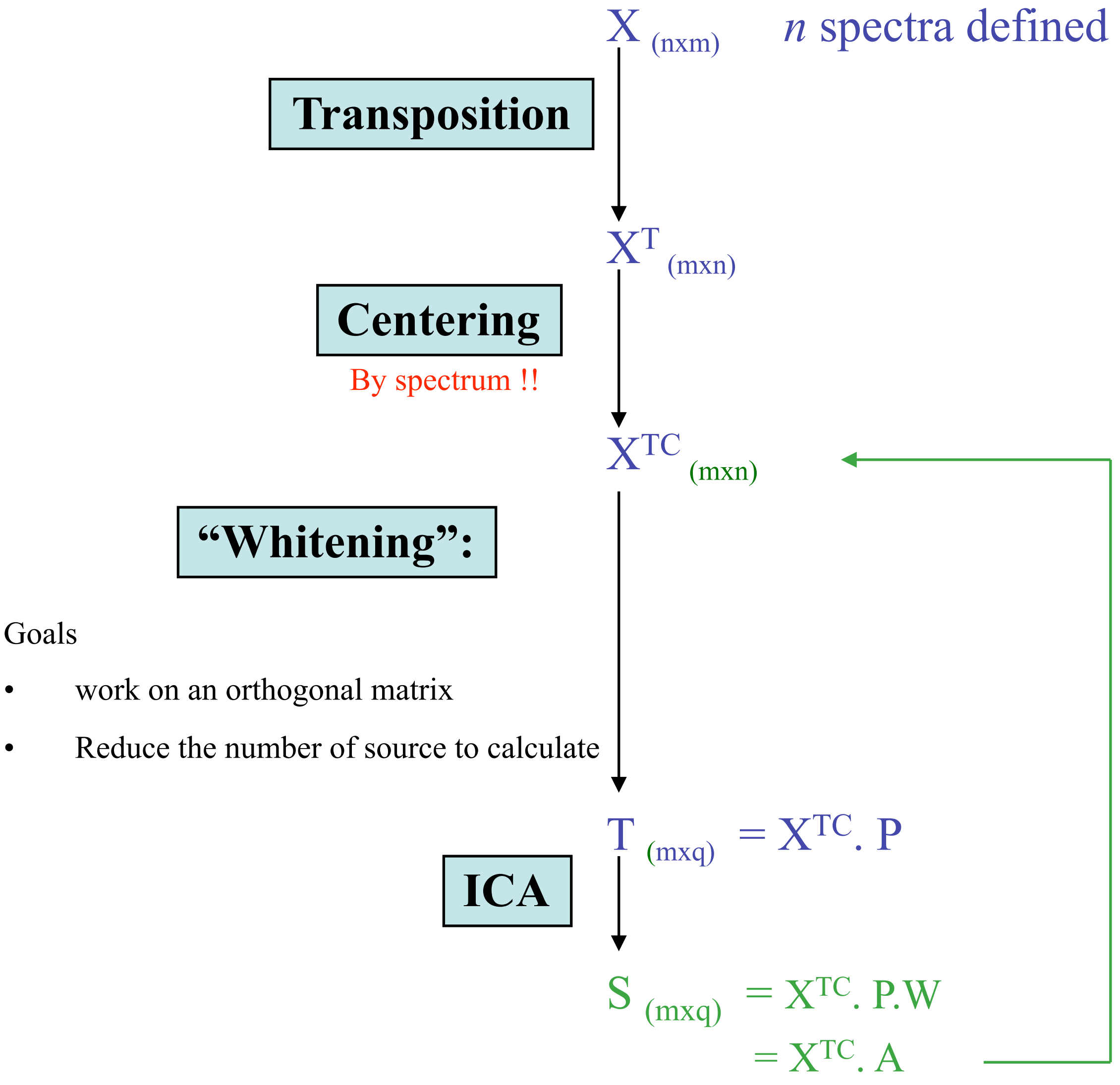
Solving the ICA problem consists of finding a **demixing matrix which maximises the non-gaussianity** of the estimated sources under the constraint that their variances are constant.

Fast-ICA algorithm:

- uses an objective function related to **negentropy**
- uses **fixed-point iteration scheme**.

** almost any measured quantity which depends on several underlying independent factors has a Gaussian PDF*

Step I : dimension reduction by ICA :



$X_{(n \times m)}$ *n* spectra defined by *m* variables ex: (28x600)

Transposition

$X^{T}_{(m \times n)}$

Centering

By spectrum !!

$X^{TC}_{(m \times n)}$

“Whitening”:

Goals

- work on an orthogonal matrix
- Reduce the number of source to calculate

$T_{(m \times q)} = X^{TC}.\,P$

ICA

$S_{(m \times q)} = X^{TC}.\,P.W$

$= X^{TC}.\,A$

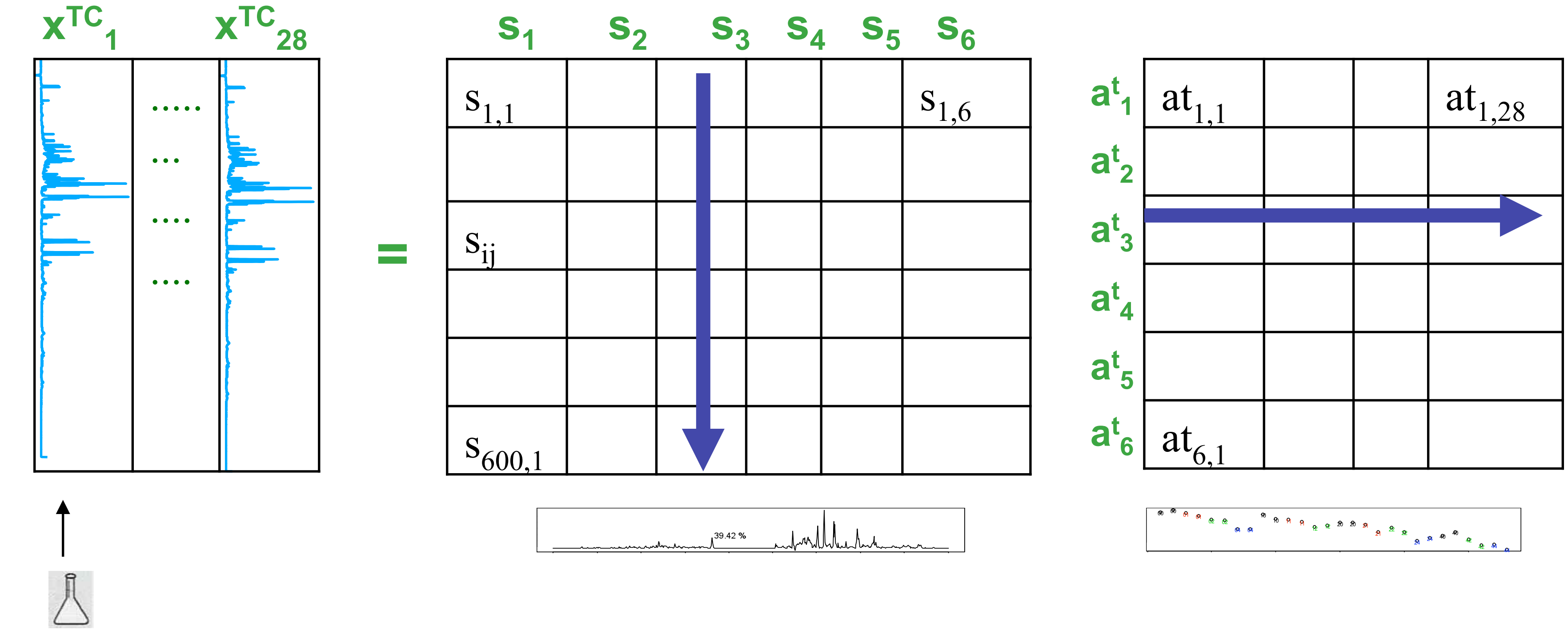
$$\boxed{X^{TC} = S.A^{T} + E}$$

Each spectrum is a weighted sum of the independent spectral expressions which each one can correspond to an independent (composite) metabolite contained in the studied sample.

(a^{T} , weight ≈ quantity)

Step I : Example

$$\mathbf{X}^{TC}{}_{(600 \text{ x } 28)} = \mathbf{S}_{\ (600 \text{ x } 6)} \quad \mathbf{A}^{T}{}_{\ (6\text{x}28)}$$



$$\begin{array}{c|c|c|c|c|c}
s_1 & s_2 & s_3 & s_4 & s_5 & s_6 \\
\hline
s_{1,1} & & & & & s_{1,6} \\
 & & & & & \\
s_{ij} & & & & & \\
 & & & & & \\
 & & & & & \\
s_{600,1} & & & & &
\end{array}$$

$$\begin{array}{c|cccc}
 & & & & \\
\hline
at_1 & at_{1,1} & & & at_{1,28} \\
at_2 & & & & \\
at_3 & & & & \\
at_4 & & & & \\
at_5 & & & & \\
at_6 & at_{6,1} & & &
\end{array}$$

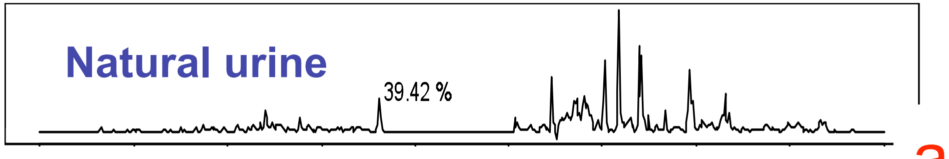
Urine

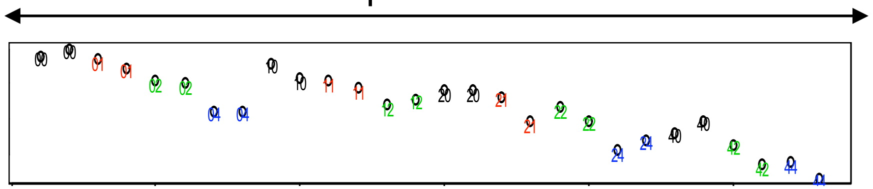
+ citrate

+ hippurate

Sources : S $_{(600 \times 6)}$

Mixing weigthsA^T





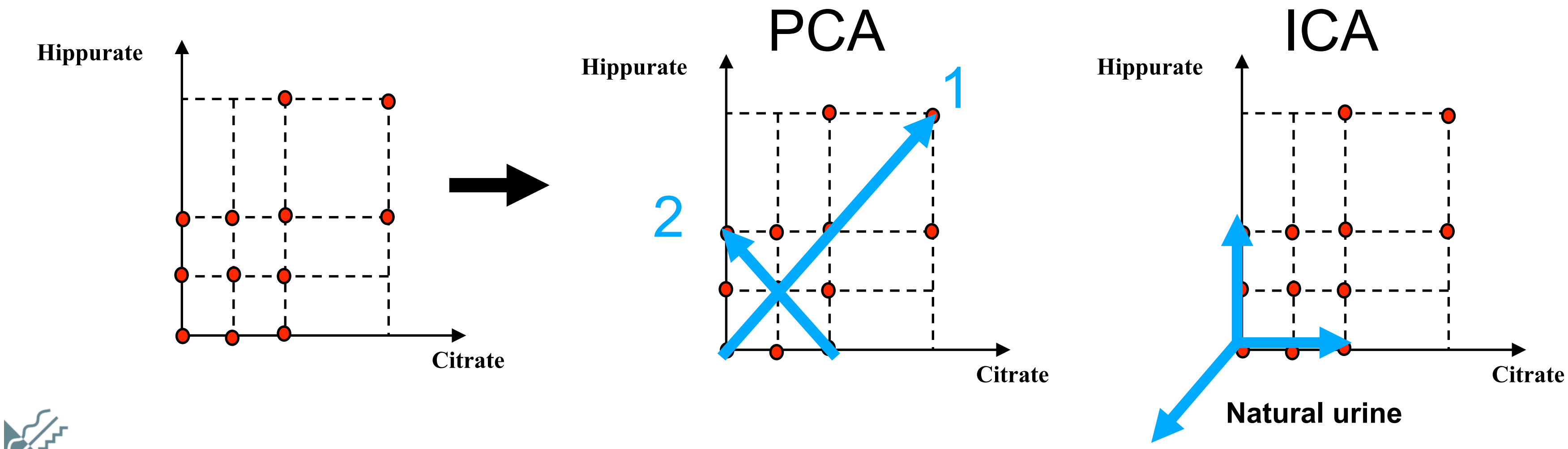
$a^T_{2,8}$

Citrate

Hippurate

Step I: Comparison with the usual PCA

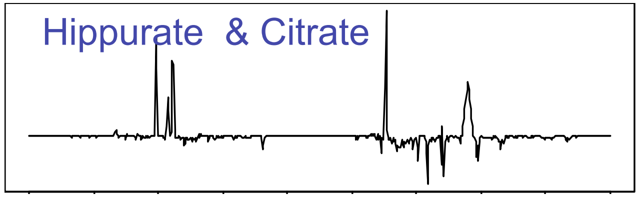
- **Similarities**: projection methods linearly decomposing multi-dimensional data into components.
- **Differences:**
  - ➢ ICA uses $X^T_{(mxn)}$ (PCA uses $X_{(nxm)}$)
  - ➢ The number of sources, *q*, has to be fixed in ICA
  - ➢ Sources are not naturally sorted according to their importance in ICA
  - ➢ The **independence condition** = the biggest advantage of the ICA:
    - independent components (ICA) are more meaningful than uncorrelated components (PCA)
    - more suitable for our question in which the component of interest are not always in the direction with the maximum variance.



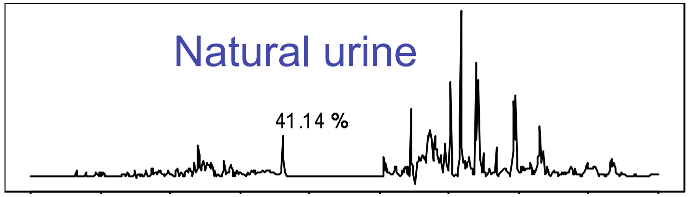
PCA

ICA

Loading 1



Hippurate & Citrate



Natural urine

s_1

Loading 2

Citrate

s_2

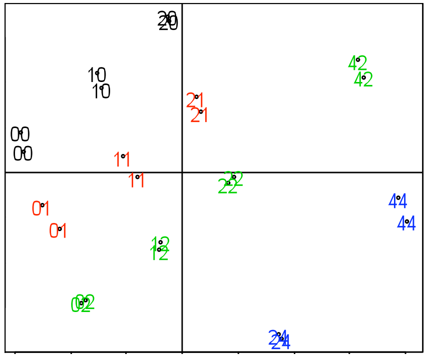
Hippurate & Citrate

Loading 3

Hippurate

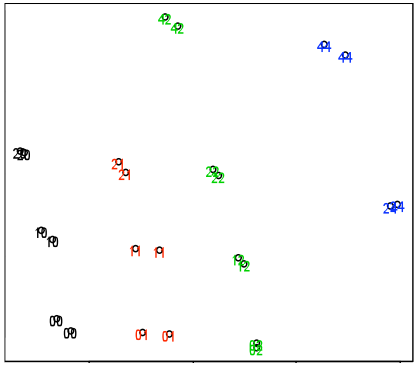
s_3

PC2



PC1

$a^T{}_3$

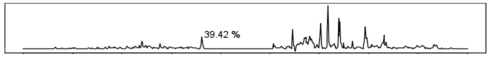


$a^T{}_2$

Methodology based on ICA and statistical modeling

<u>Step I</u> : Dimension reduction by ICA

$$X^{TC}_{(600x28)} = S \; . \; A^T$$



Some of these sources present the biomarkers.

Which ones?

<u>Step II</u>: Mixed statistical modeling on ICA mixing weights

$$A^T = Z^1\beta + Z^2\gamma + \varepsilon$$

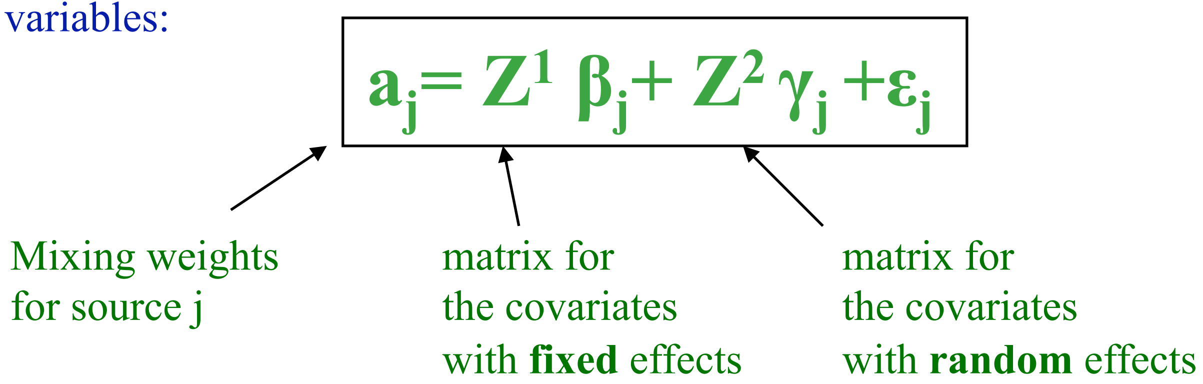
<u>Step III</u>: Selection of significant sources (biomarkers)

$$S^* \subset S$$

<u>Step IV</u>: Visualization of biomarkers and factor effects

Step II: statistical modeling of ICA mixing weights

- **For each of the *q* sources** s_j, we assume a linear relation between its vector of weights and the design variables:



$$\boxed{a_j = Z^1 \beta_j + Z^2 \gamma_j + \varepsilon_j}$$

Mixing weights for source j

matrix for the covariates with **fixed** effects

matrix for the covariates with **random** effects

- **Models with fixed and random effects covariates : Mixed model:** $a_j = Z^1 \beta_j + Z^2 \gamma_j + \varepsilon_j$
- **Models with only random effects covariates :** $a_j = Z^2 \gamma_j + \varepsilon_j$

  → ex: biomarker to explore variance component (machines, subjects, laboratories)

- **Models with only fixed effects covariates :** $a_j = Z^1 \beta_j + \varepsilon_j$
  - Case 1: **categorical covariates: ANOVA**

    → ex: biomarker to discriminate 3 groups of subjects: disease1, disease2 & sane

  - Case 2: **quantitative covariates : linear regression**

    → ex: biomarker to explore the severity of an illness, the concentration of a drug

Step II: Fit a model: example

- **For each of the q = 6 recovered s_j,** we construct a **multiple linear regression model** with 2 fixed quantitative covariates and no interaction:

$$\boxed{a_j = \beta_{j0} + \beta_{j1} y_1 + \beta_{j2}\, y_2 + \varepsilon_j}$$

Mixing weights for source j

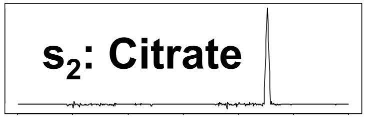
Drug dose (covariate of interest)

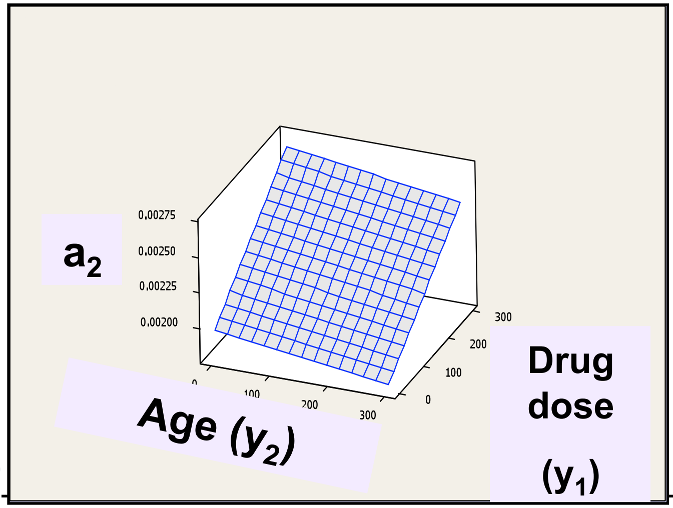
Age

- **For each of the 6 sources s_j, the fitted model** by least square technique is :

$$\hat{a}_j = b_{j0} + b_{j1}\, y_1 + b_{j2}\, y_2$$

Ex:



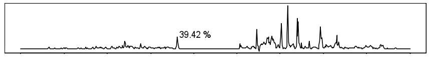


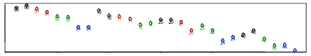
Methodology based on ICA and statistical modeling

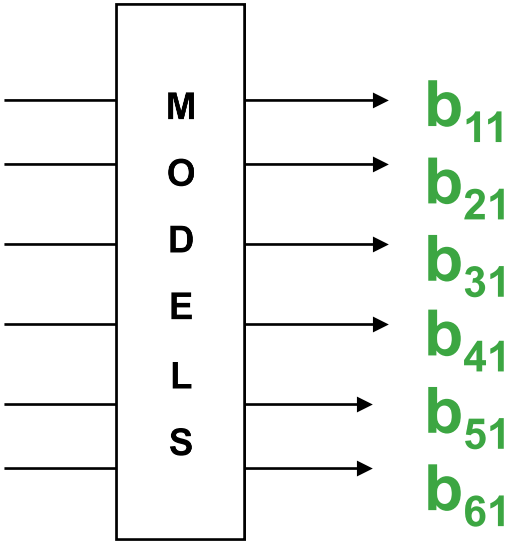
Step I : Dimension reduction by ICA

Step II: Mixed statistical modeling on ICA mixing weights

$X^{TC}_{(600x28)} = S \quad . \quad A^T$







MODELS

b_{11}

b_{21}

b_{31}

b_{41}

b_{51}

b_{61}

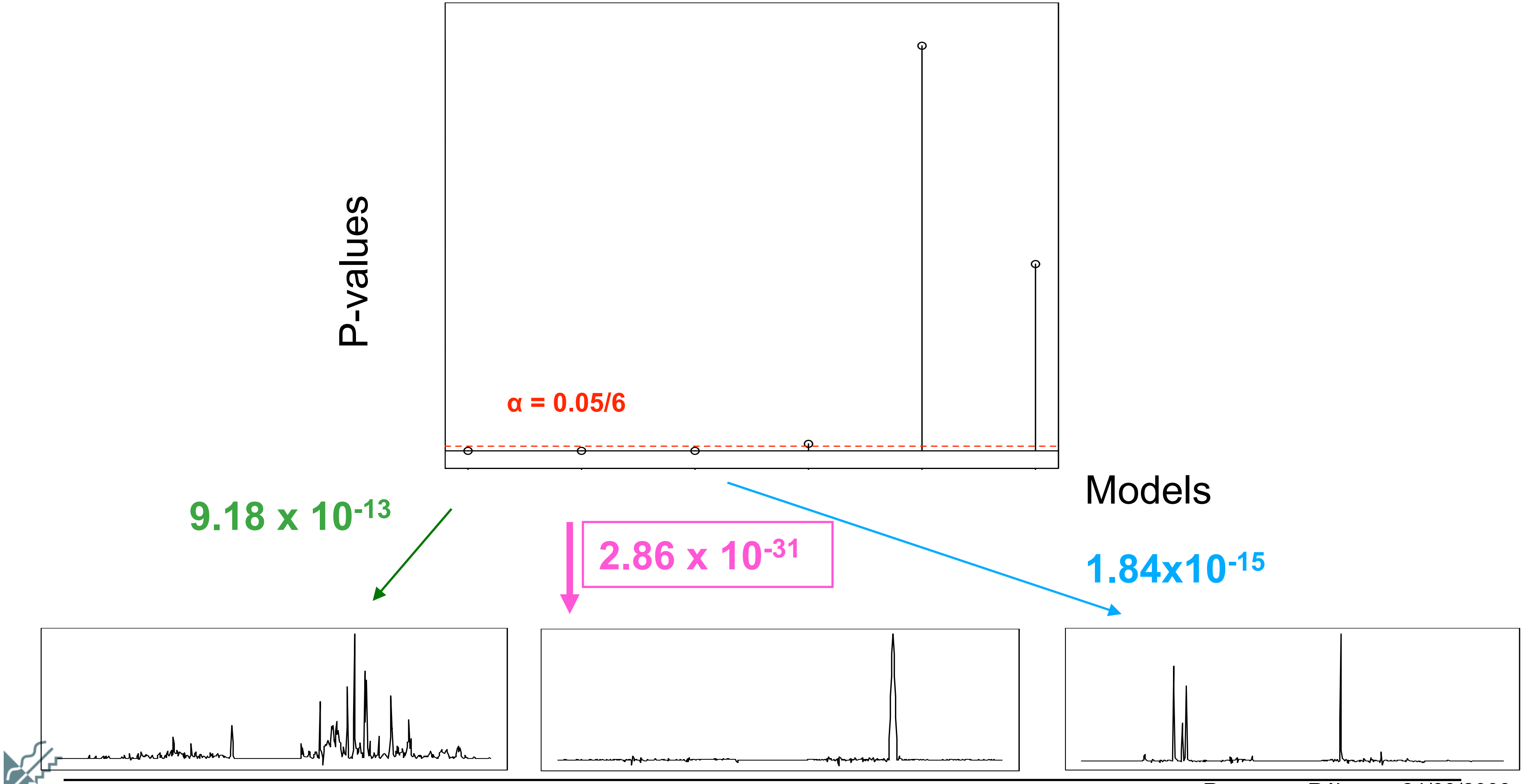
Step III: Selection of significant sources (biomarkers)

$S^* \subset S$

Step IV: Visualization of biomarkers and factor effects

Step III: Selection of significant sources, biomarker identification

- **Goal: we want to select the sources presenting a significant effect** of the covariate of interest on their weights.
- **For each source,** F or t **test of hypothesis and Bonferroni correction of the level of significance.**



9.18×10^{-13}

2.86×10^{-31}

1.84×10^{-15}

Methodology based on ICA and statistical modeling

<u>Step I</u> : Dimension reduction by ICA

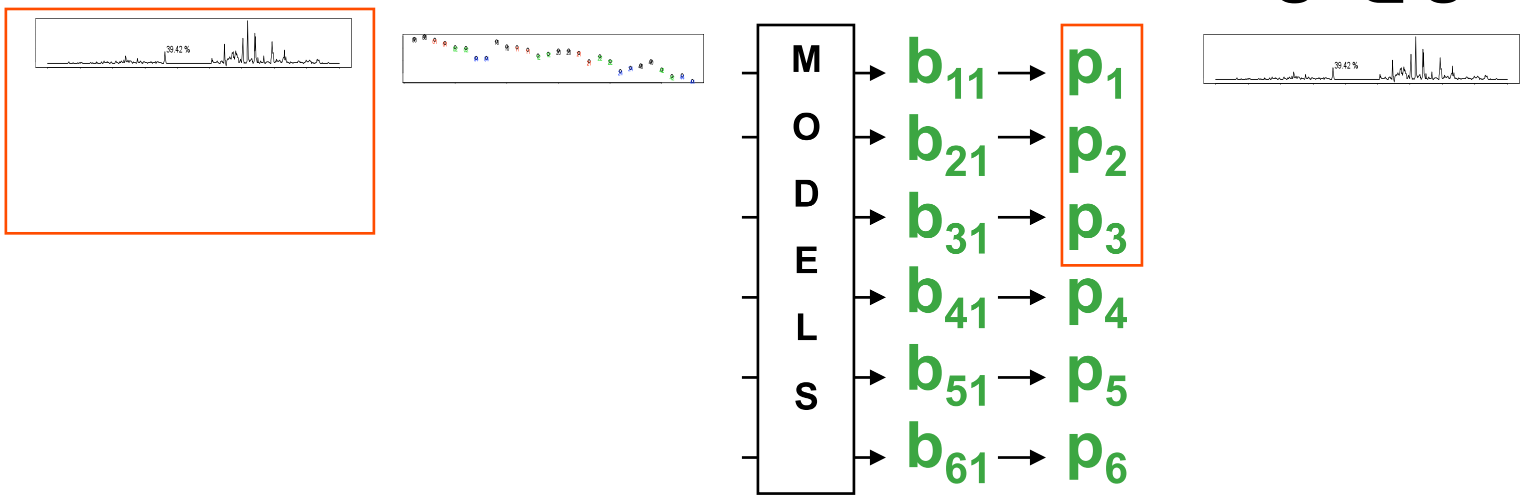
↓

<u>Step II</u>: Mixed statistical modeling on ICA mixing weights

↓

<u>Step III</u>: Selection of significant sources (biomarkers)

$X^{TC}_{(600x28)} = S \; . \; A^T$



$S^* \subset S$

↓

<u>Step IV</u>: Visualization of biomarkers and factor effects

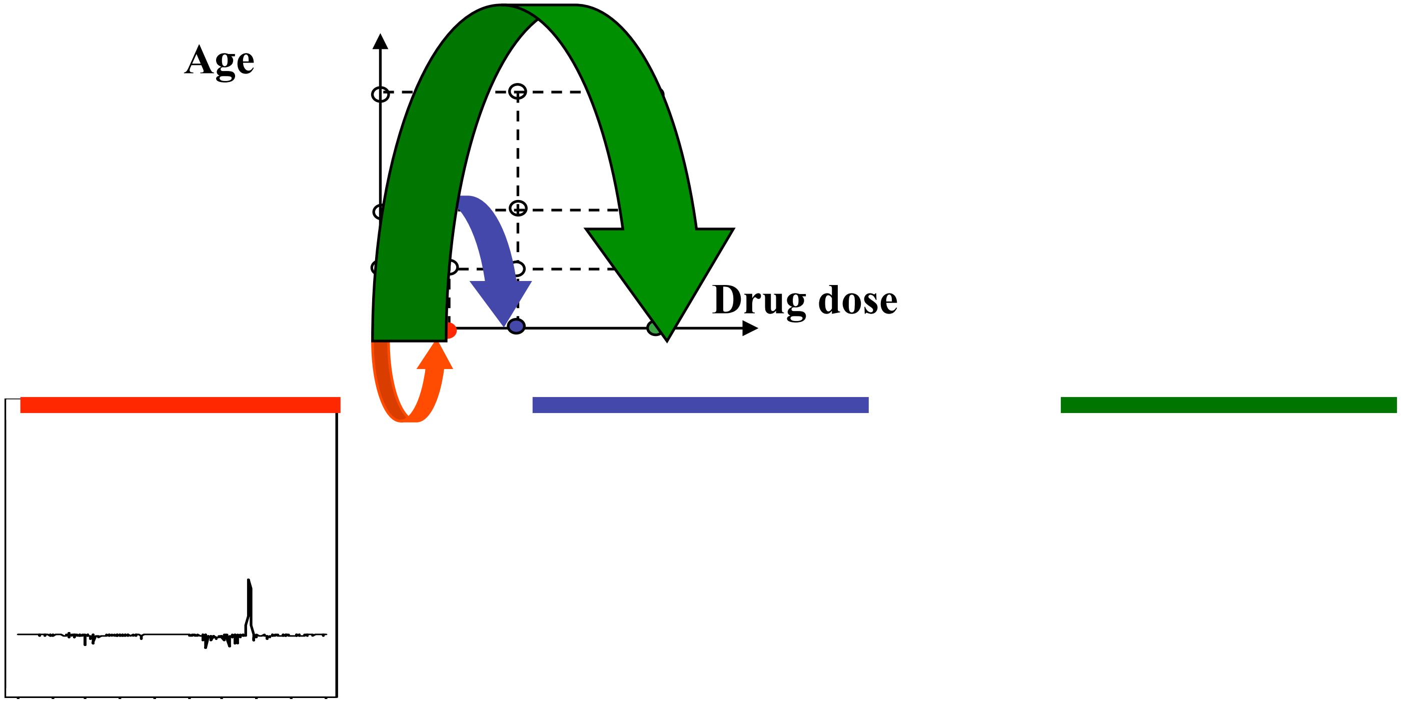
Step IV : Comparison of the intensities in biomarkers

- **Goal:** visualize the **effects** on the biomarker caused by ≠ changes in the variable of interest.
- **Choose values of the variable of interest:**

  ex: y_1= drug dose

  y_1^1 : a first value of reference y_1^2 : a new value of interest of y_k
- **Compute contrast:** ex: the effect on the biomarker of the change of y_1 from $\mathbf{y_1^1}$ to $\mathbf{y_1^2}$:

$$C_1 = S * \beta_k^{*} (y_k^2 - y_k^1)$$

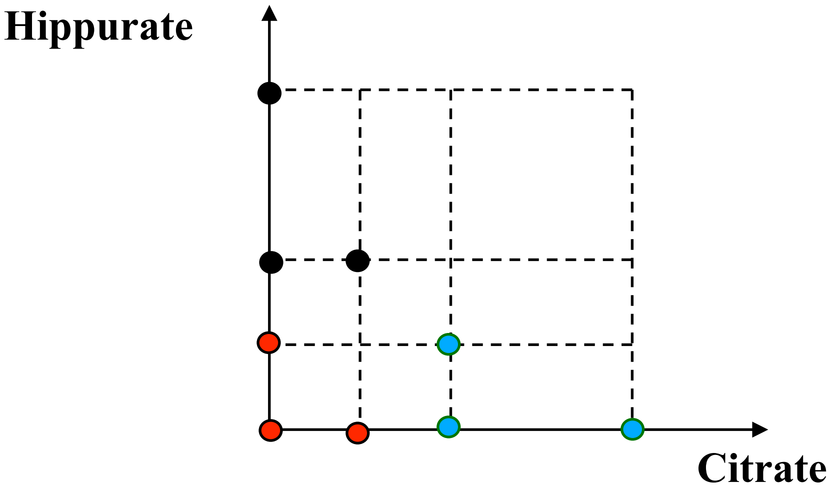


Conclusions:

- With the presented methodology combining ICA with statistical modeling,
  - we visualize the independent metabolites contained in the studied biofluid (through the sources) and their quantity (through the mixing weights)
  - we identify biomarkers or spectral regions changing significantly according to the factor of interest by a selection of source.
  - we compare the effects on these spectral biomarkers caused by different changes of the factor of interest.
- In comparison with the PCA, ICA:
  - gives more biologically meaningful and natural representations of this data.

Thank you for your attention

Example2: the data



Group 1= disease 1

Group 2= disease 2

Group 3= no disease

- 18 spectra of 600 values
- 1 characteristic in Y

X(18x600) **Y (18x1)**

y_1= disease group of the rat (qualitative)

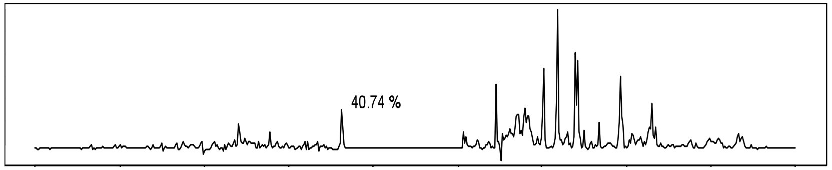
- **We want biomarkers for group of disease described in y_1.**

→ a model with qualitative covariates

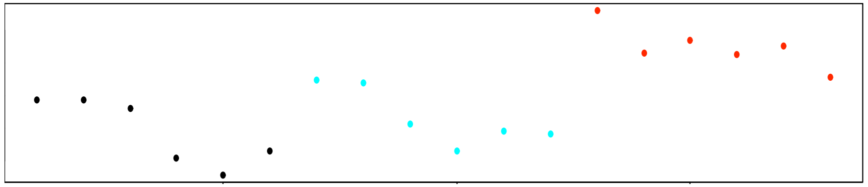
Example2 :Part I. Dimension reduction by ICA

$X^{TC} = S.A^{T}$

$S_{(600 \times 5)}$



$A^{T}_{(5x18)}$



Example 2: Part II: biomarkers discovery through statistical modeling

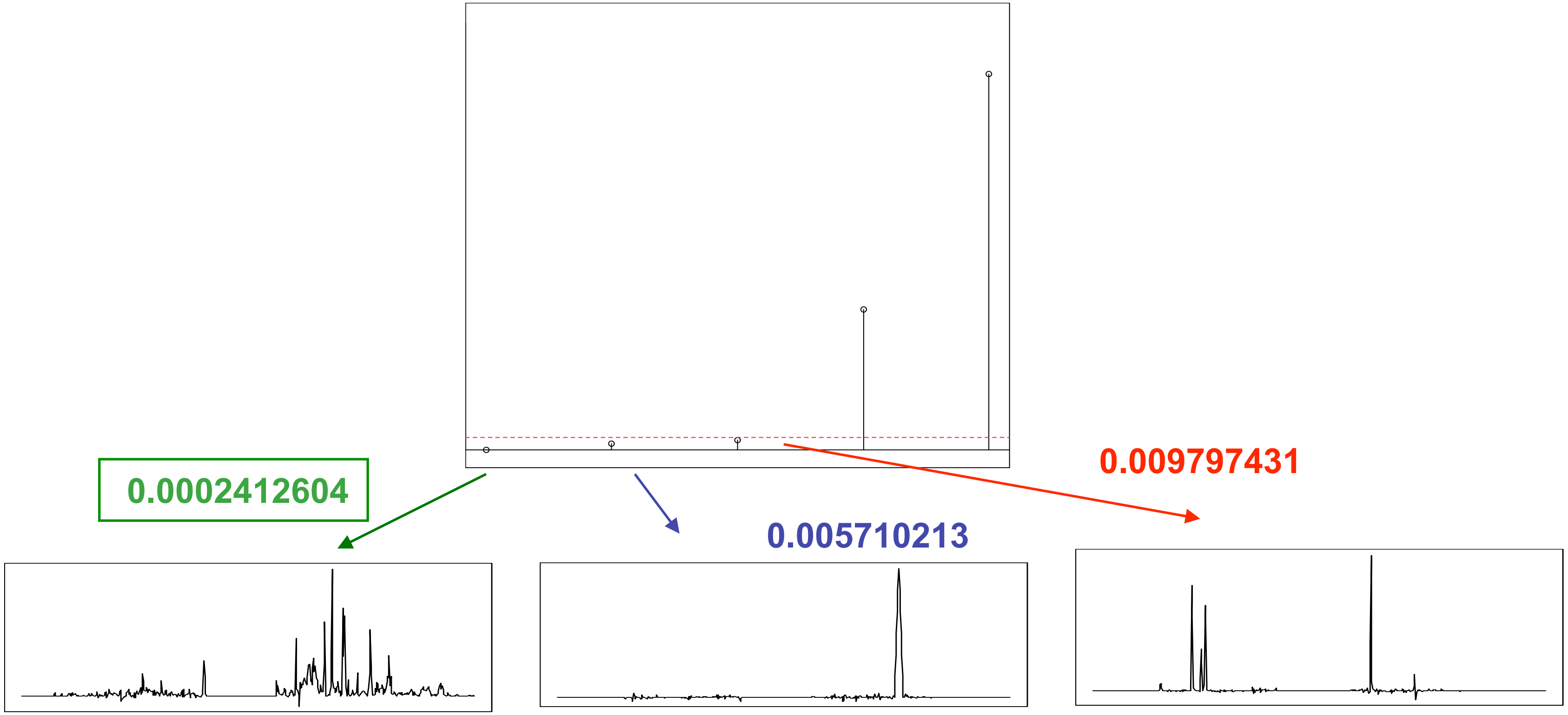
Step 1: Fit a model on A^T

Models with only a categorical covariate with fixed effects: ANOVA I

$$a_j = Z^1 \beta_j + \varepsilon_j$$

Step 2: Biomarker identification:

- For each of the *q* recovered s_j, test the effect of y_1 → F_j statistics→ $\mathbf{p_j}$
- **Bonferroni correction:** select, in a *(m x r)* matrix **S*, the *r* sources with $p_j < 0.05/q$**



0.0002412604

0.005710213

0.009797431

Step 3: Comparison of the intensities in biomarkers

- **Goal:** comparison of the **effects** on the biomarker caused by ≠ changes in y_k.
- **Choose 3 or more values of y_k:**
  - $\mathbf{y_k^1}$: a first value of reference of y_k
  - $\mathbf{y_k^2}$: a new value of interest of y_k
  - $\mathbf{y_k^3}$: a second new value of interest of y_k
- **Compute:**
  - The effect on the biomarker of the change of y_k from $\mathbf{y_k^1}$ to $\mathbf{y_k^2}$:

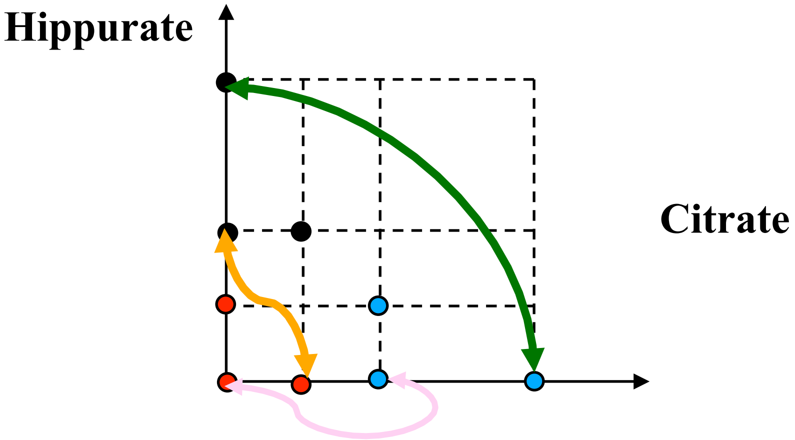
$$C_1 = S^* \beta_k^* (y_k^2 - y_k^1)$$

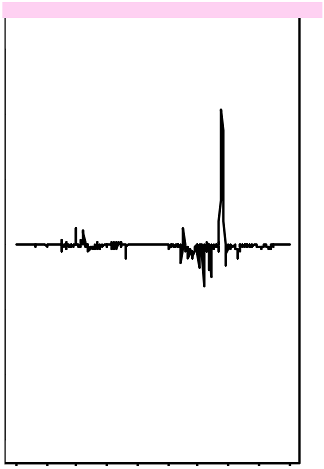
  - The effect on the biomarker of the change of y_k from $\mathbf{y_k^1}$ to $\mathbf{y_k^3}$:

$$C_2 = S^* \beta_k^* (y_k^3 - y_k^1)$$

Step 3: Comparison of the intensities in biomarkers

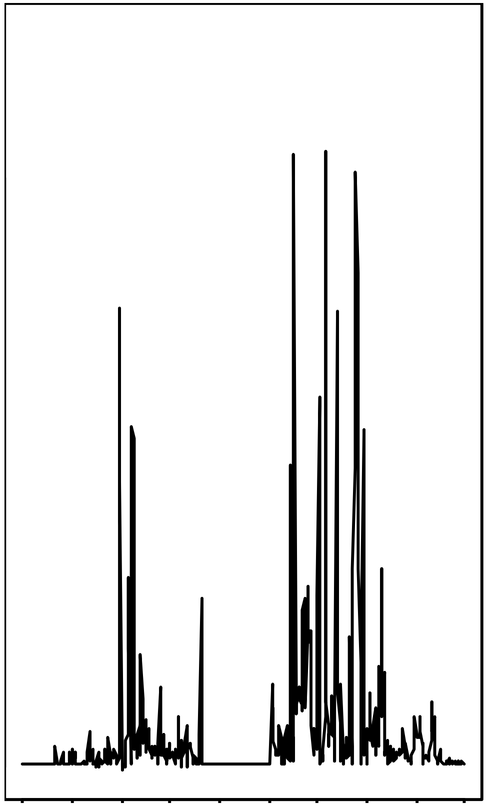
Goal: comparison of the **effects** on the biomarker caused by the changes of group.

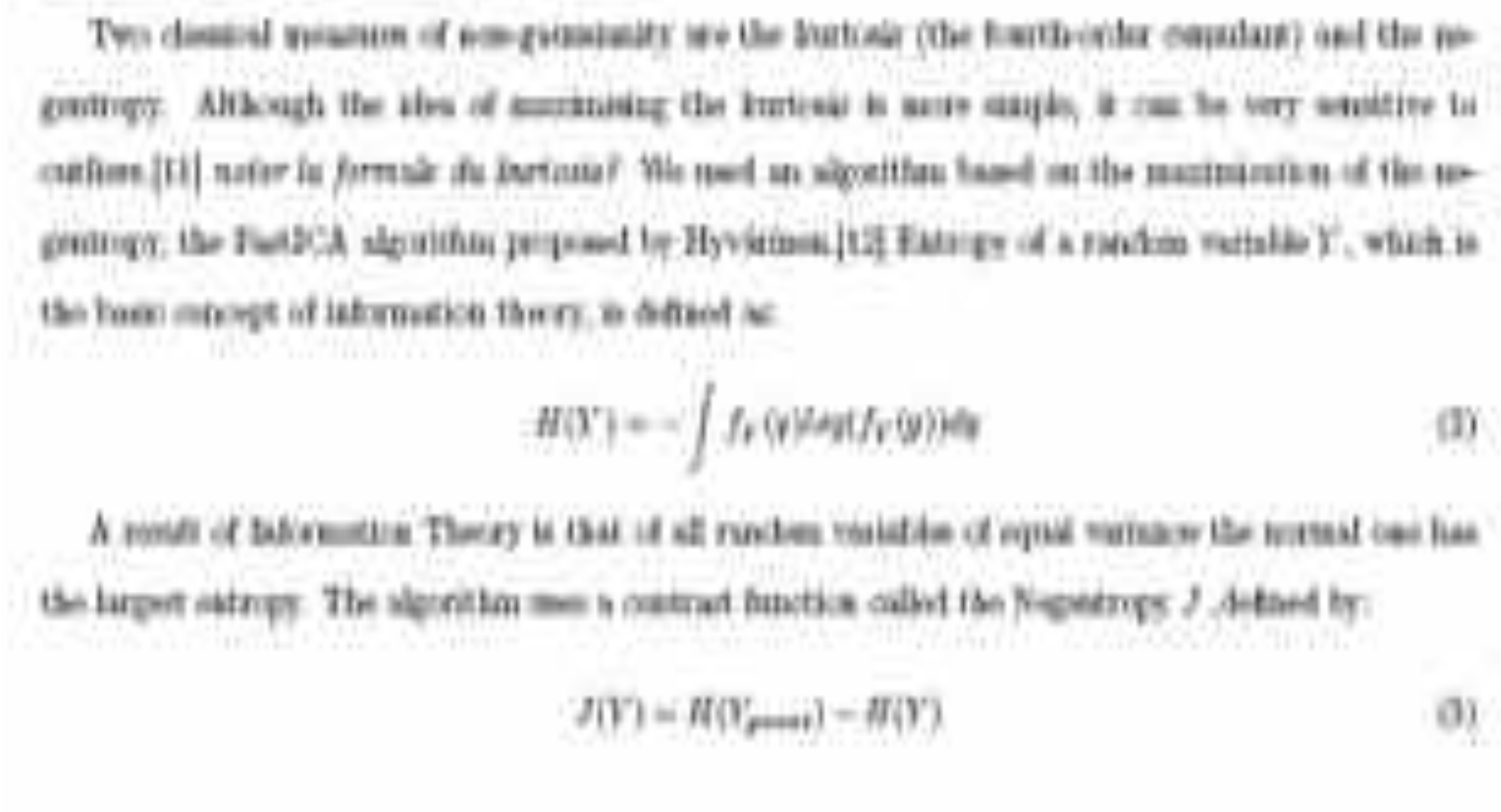




Others slides

Example 2: the reconstructed spectra





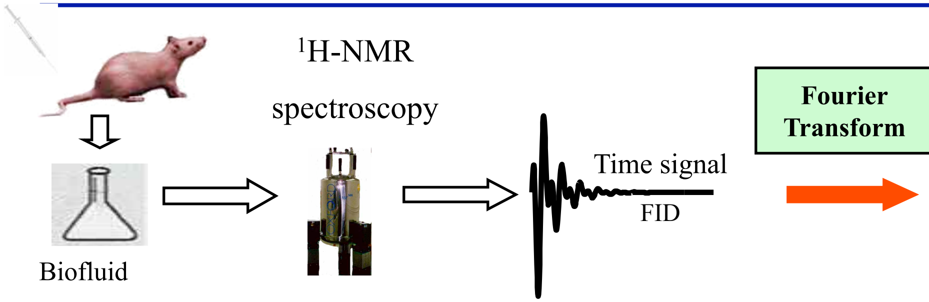
Two classical measures of non-gaussianity are the kurtosis (the fourth-order cumulant) and the negentropy. Although the idea of maximizing the kurtosis is more simple, it can be very sensitive to outliers [11] *noter la formule du kurtosis?* We used an algorithm based on the maximization of the negentropy, the FastICA algorithm proposed by Hyvärinen [12]. Entropy of a random variable Y, which is the basic concept of information theory, is defined as:

$$H(Y) = -\int f_Y(y) \log(f_Y(y)) dy \qquad (1)$$

A result of Information Theory is that of all random variables of equal variance the normal one has the largest entropy. The algorithm uses a contrast function called the Negentropy J, defined by:

$$J(Y) = H(Y_{gauss}) - H(Y) \qquad (2)$$

Pre-treatments of spectra



^{1}H-NMR spectroscopy

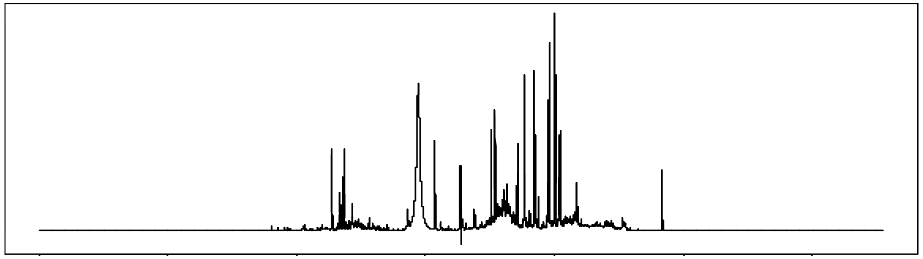
Biofluid

Time signal

FID

Fourier Transform

Initial spectrum: 65536 points



Solvent suppression:

Whittaker smoother

Apodization

Fourier Transform

Phase correction

Baseline correction:

Assymetric least square

Whittaker smoother

Normalization

Median method

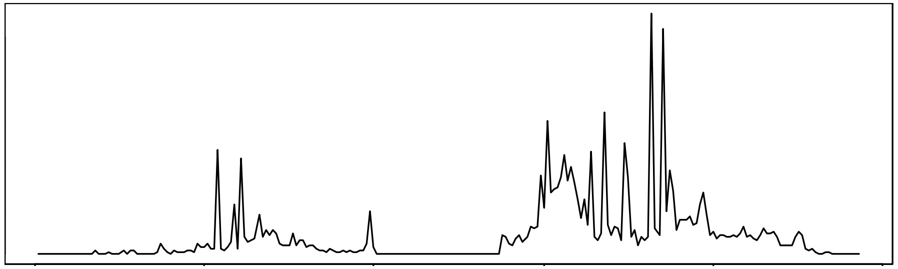
Peaks alignement

Parametric time warping

Data reduction

Pre-treatment of spectra

Spectrum ready for analysis: 600 points



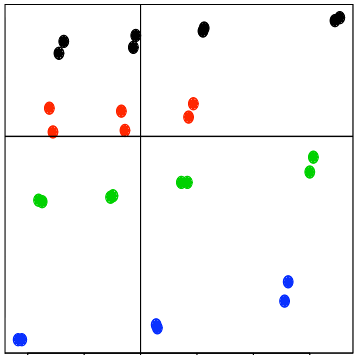
USUAL

I. Reduction of the dimension: PCA

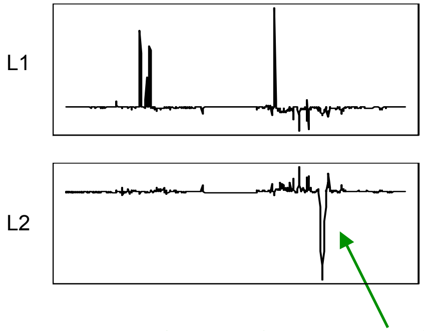
$X^C = TP$

- **Principal components** are :
  - uncorrelated
  - in the direction of maximum of variance
- **Examination of the 2 first components:**

Score plot



Loadings



ex: Citrate plays an important role

Identification of biomarker

This is only powerful if the biological question is related to the highest variance in the dataset!

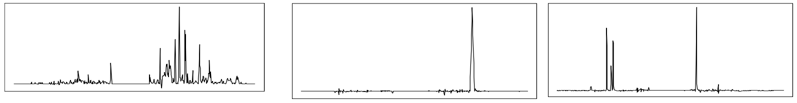
NEW

I. Reduction of the dimension: ICA

$X^{TC} = SA^T$

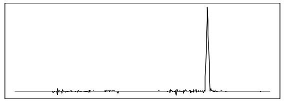
- **Components** :
  - are independent
  - with a biological meaning
- **Examination of the ALL components:**

to **visualize unconnected molecules in samples**



II. Biomarker discovery through Statistical modelling

Identification of biomarker



Comparison of the intensities of biomarkers between spectra from ≠ conditions

Example: controlled data

- **Advantage of controlled data**:

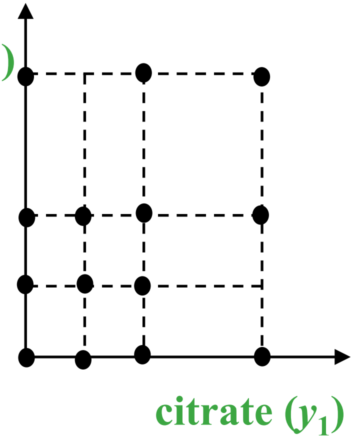
  we know the spectral regions that should be identified as biomarkers.

- **The controlled data :**

  - 28 spectra of 600 points**:** **X(28 x 600)**
  - Each spectrum = a sample of urine

    + a chosen concentration of Citrate

    + a chosen concentration of Hippurate



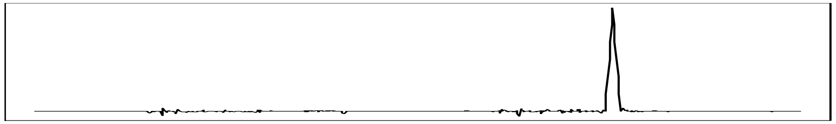
X(28x600) **Y (28x2)** **y_1= concentration of citrate**

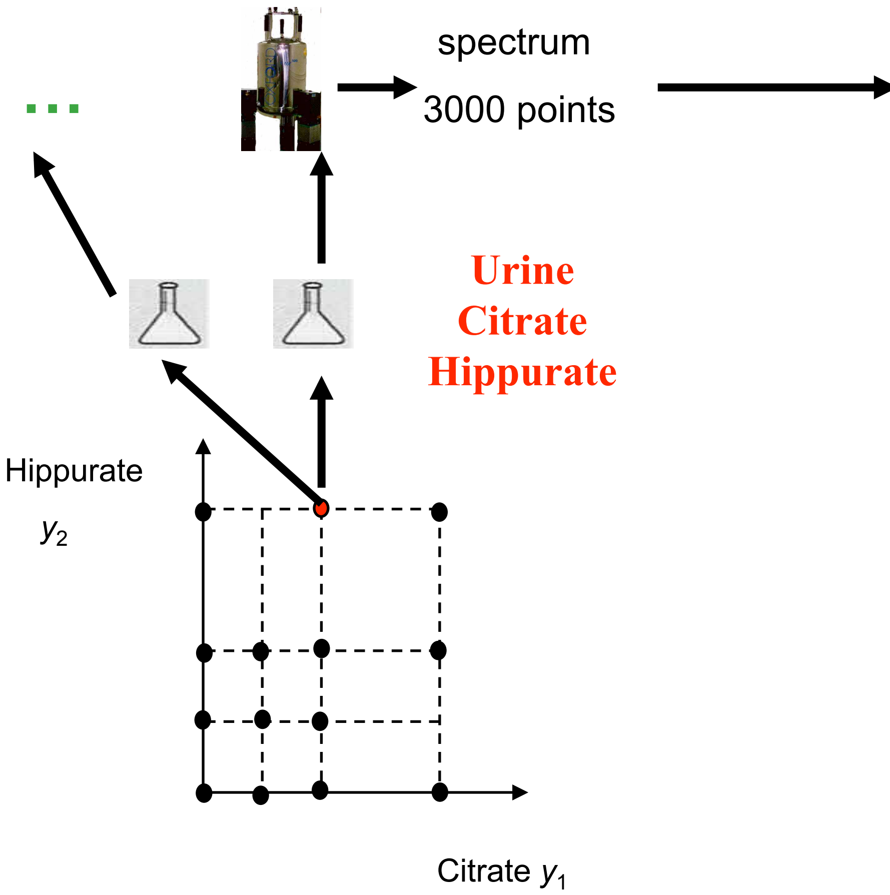
y_2= concentration of hippurate

- **We need a biomarker to detect changes of the level of citrate described by y_1**

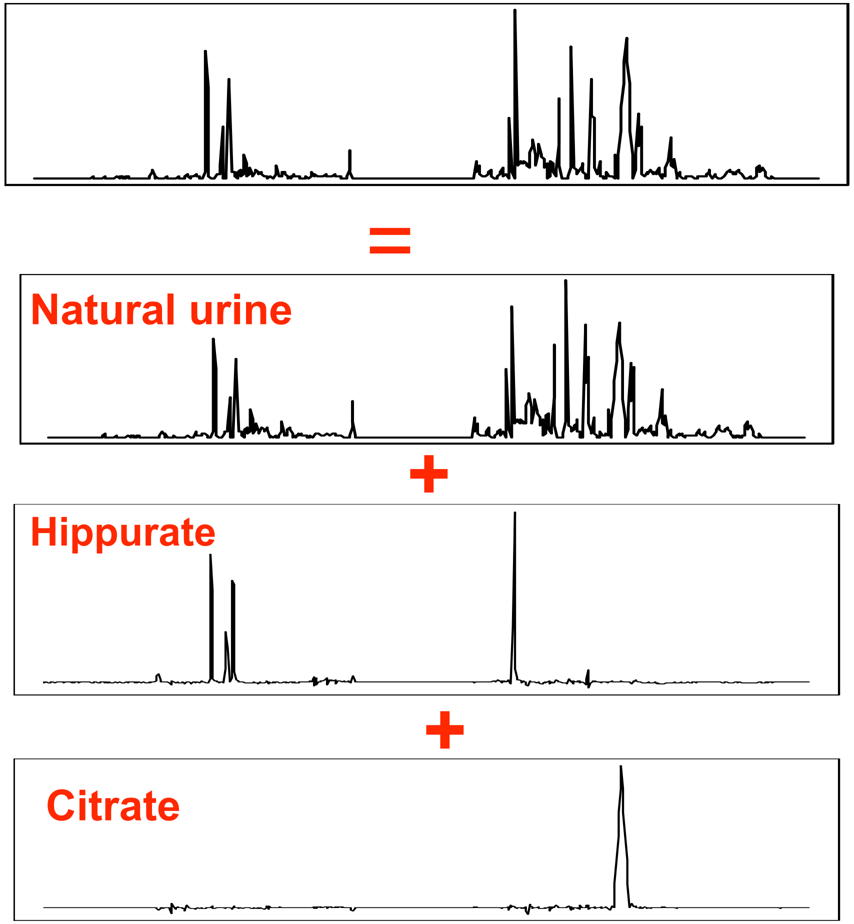
« Which are the spectral regions x_j the most altered when the y_1 changes?»

Spectral regions corresponding to Citrate = the biomarkers to identify.





A spectrum of 600 values x_j with $\sum x_j = 1$



14 mixtures in 2 replications

28 samples

The biomarkers to identify.= spectral regions corresponding to Citrate

Step II: Fit a model: example

- **For each of the $q = 6$ recovered $s_{j,}$** we construct a **multiple linear regression model** with 2 fixed quantitative covariates and no interaction:

$$\boxed{a_j = \beta_{j0} + \beta_{j1} y_1 + \beta_{j2}\, y_2 + \varepsilon_j}$$

Mixing weights
for source j

Drug dose

Age

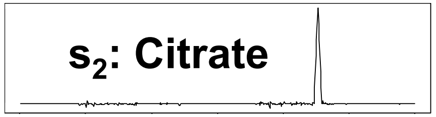
- **For each of the q recovered s_j, the fitted model** by least square technique is :

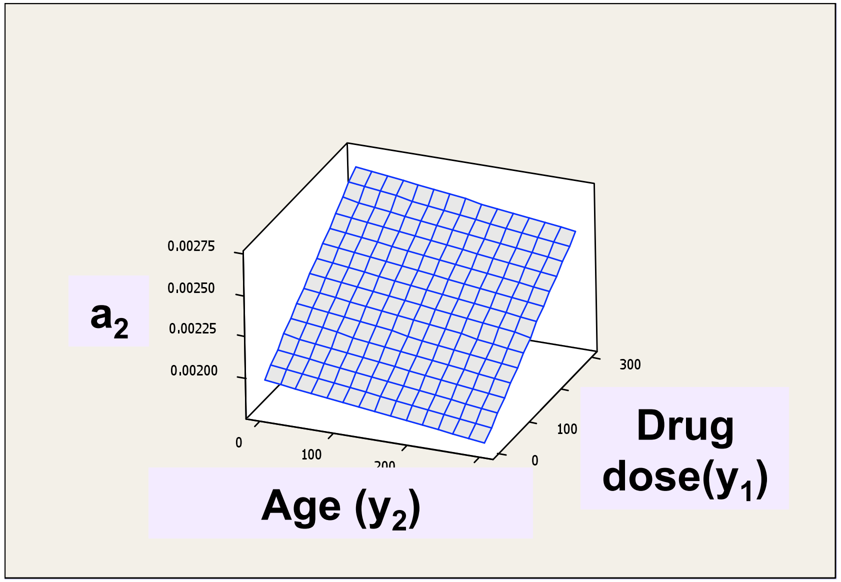
$$\hat{a}_j = b_{j0} + b_{j1}\, y_1 + b_{j2}\, y_2$$

- **In this example, we want to identify biomarkers for the concentration of a drug. The covariate of interest is y_1.**

- **Output: a vector b_1 giving the 6 values of the effect of the drug concentration on each of the 6 mixing weights**

Step II: Fit a model: example





(y_1)

Methodology based on ICA and statistical modeling

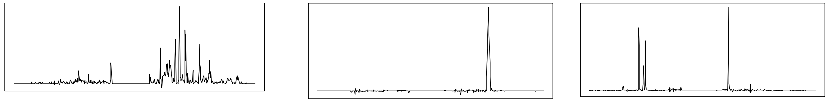
Step I : Dimension reduction by ICA

$$\boxed{X^{TC} = S \, . \, A^{T}}$$

Components : Weights ≈ quantity

Examination of the ALL components:

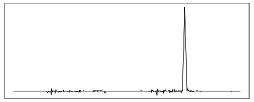
to **visualize unconnected molecules in samples**



Step II: Mixed statistical modeling on ICA mixing weights

$$\boxed{A = Z^{1}\beta + Z^{2}\gamma + \varepsilon}$$

Step III: Selection of sources identification of biomarkers

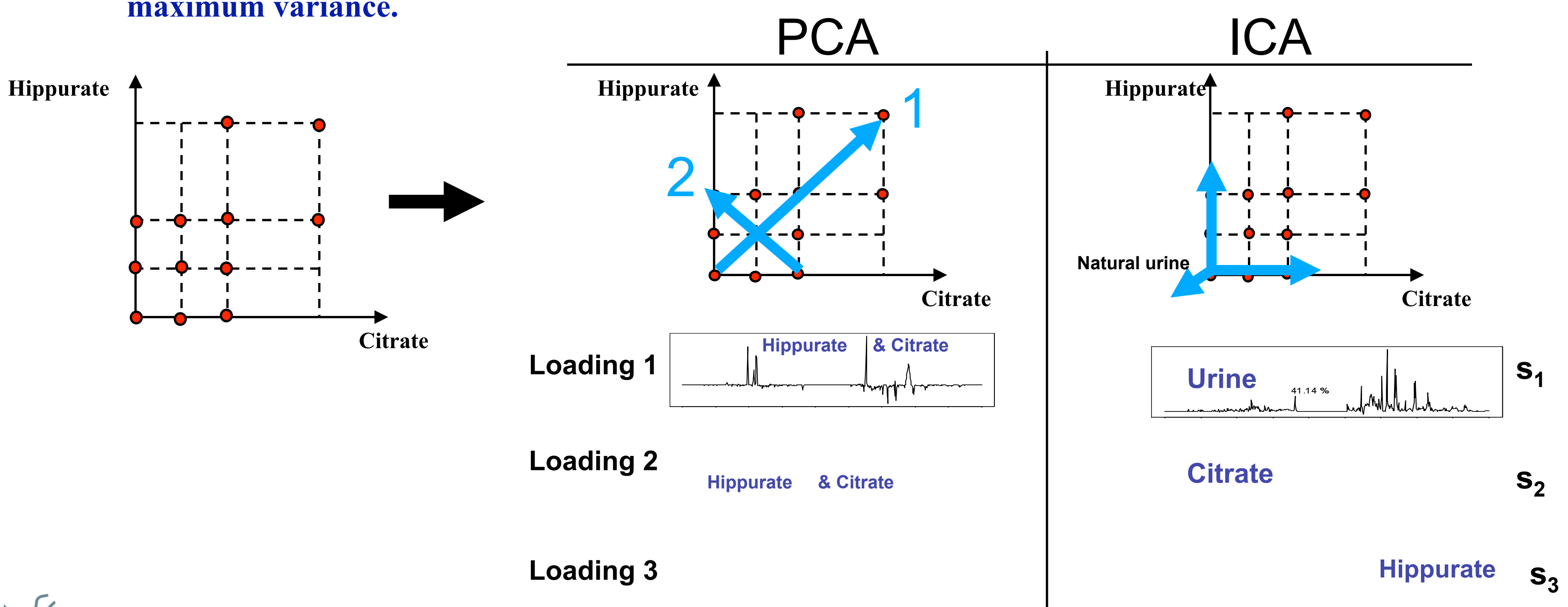


$$S^{*} \subset S$$

Step IV: Visualization of the effect of the factor of interest on the biomarkers

Step I: Comparison with the usual PCA

- **Similarities**: projection methods linearly decomposing multi-dimensional data into components.
- **Differences:**
  - ICA uses $X^T_{(mxn)}$ (PCA uses $X_{(nxm)}$)
  - The number of sources, *q*, has to be fixed in ICA
  - Sources are not naturally sorted according to their importance in ICA
  - The **independence condition** = the biggest advantage of the ICA:
    - independent components (ICA) are more meaningful than uncorrelated components (PCA)
    - **more suitable for our question in which the component of interest are not always in the direction with the maximum variance.**


Hippurate
Citrate
PCA
ICA
Hippurate
1
2
Citrate
Hippurate
Natural urine
Citrate
Loading 1
Hippurate
& Citrate
Loading 2
Hippurate
& Citrate
Loading 3
Urine
41.14 %
s_1
Citrate
s_2
Hippurate
s_3