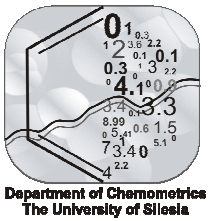
Challenges in handling complex metabolomic data

M. Daszykowski

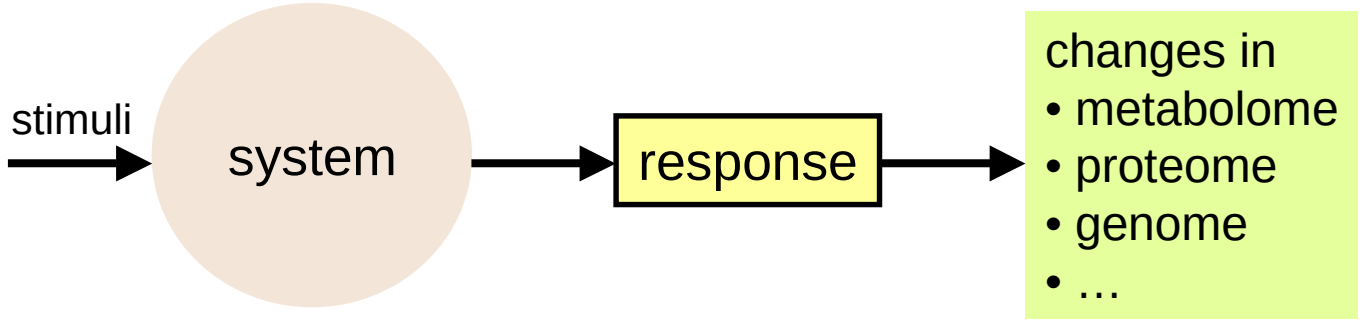


Department of Chemometrics
The Univeristy of Silesia
9 Szkolna Street
40-006 Katowice
Poland

http://www.chemometria.us.edu.pl

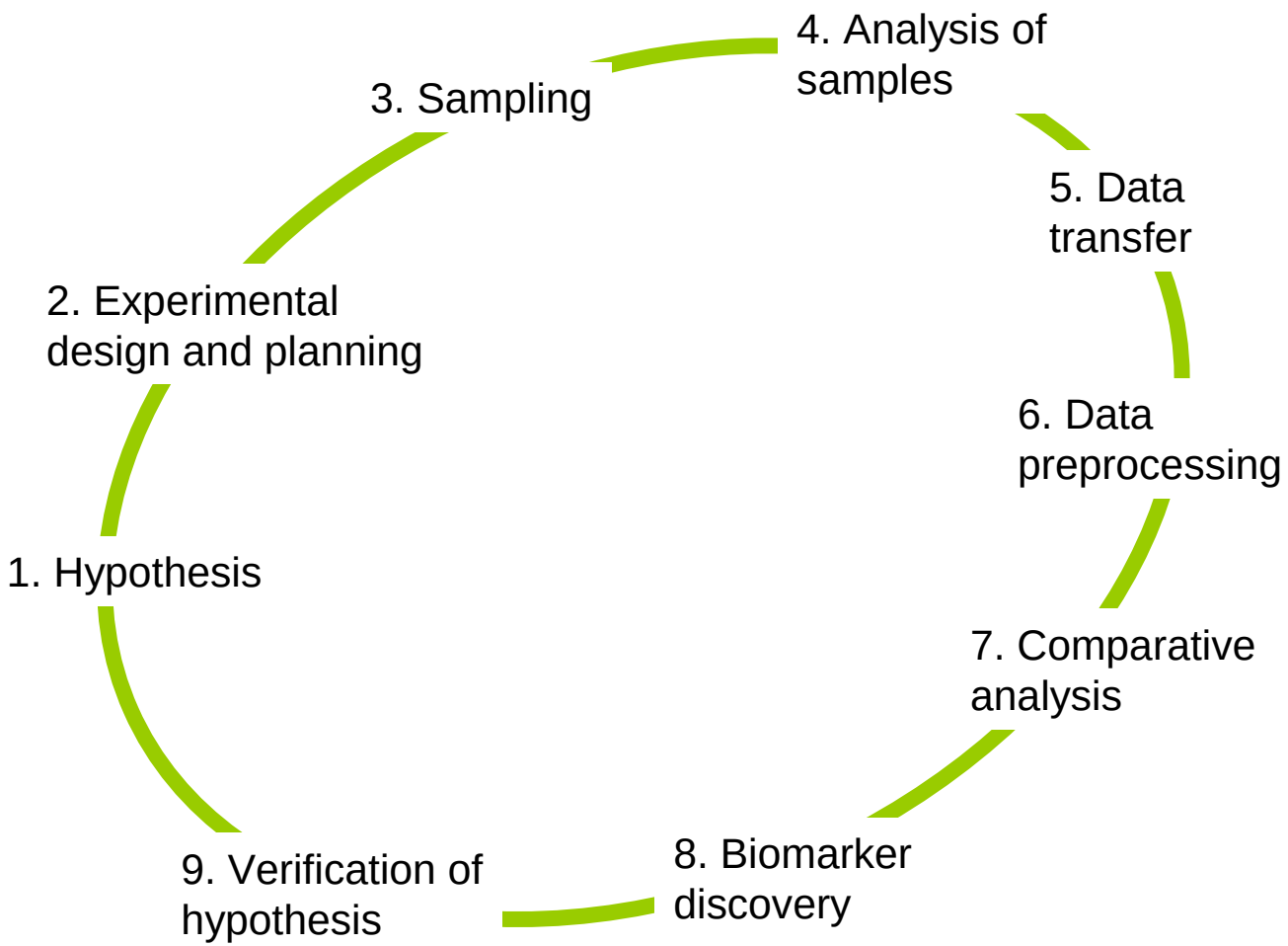
Metabolomics

- Metabolomics is the "systematic study of the unique chemical fingerprints that specific cellular processes leave behind" - specifically, the study of their small-molecule metabolite profiles [Wikipedia]



  - systematic changes in metabolome (qualitative and/or quantitative differences) – healthy/sick
  - monitoring changes in metabolome over time (a time course study)

Analytical process workflow



Non-targeted and targeted approach

- non-targeted approach – looking at all possible sample components (a fingerprint)
  - alignment problems
  - exploratory nature: at this stage identification of mixture components is not required

- targeted approach – monitoring a specific compound or a few compounds (so much desired in medical diagnostic …)
  - requires a relatively broad knowledge about system being studied
  - identification/quantification issue

Analytical technique - requirements

- a suitable analytical technique able to reveal as many compounds as possible and quantify them
  - NMR,
  - MS,
  - X-MS,
  - chromatography, …

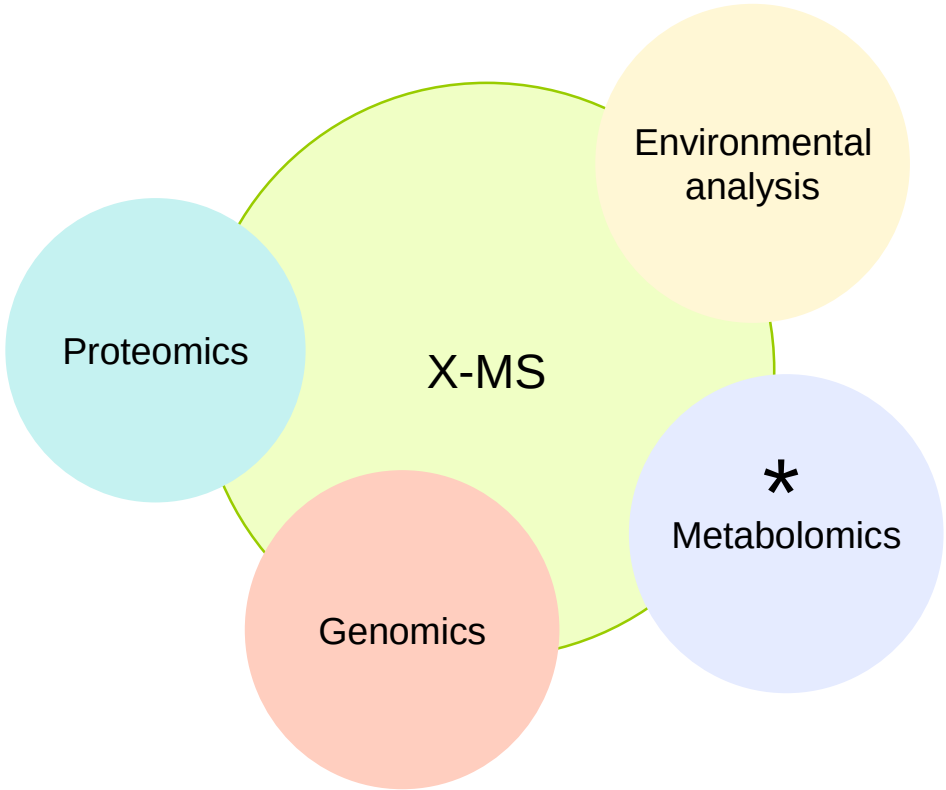
X-MS analytical techniques

- X-MS techniques become more and more popular
- they are very powerful (high resolution)
- different analytical techniques can be coupled successfully with MS

  - LC-MS
  - GC-MS
  - CE-MS
  - …

} X-MS

Applications of X-MS techniques



* G. Theodoridis, H..G. Gika, I.D. Wilson, LC-MS-based methodology for global metabolite profiling in metabonomics/metabolomics, **Trends in Analytical Chemistry, 27 (2008)** 251-260

J. Listgarten, A. Emili, Statistical and computational methods for comparative proteomic profiling using liquid chromatography – tandem mass spectrometry, Molecular & Cellular Proteomics, 4 (2005) 419-434

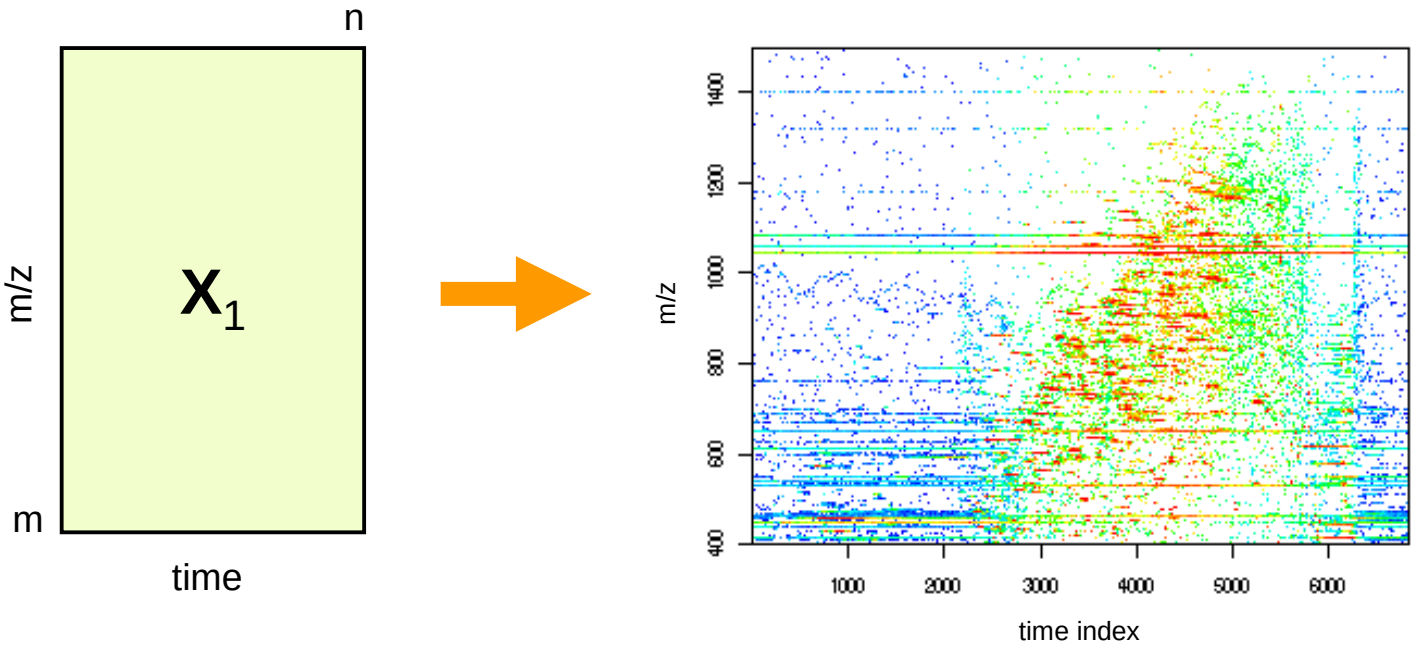
X-MS – the 2nd order advantage

- X-MS – two analytical techniques potentially providing an ‘orthogonal’ information
  - a typical example: HPLC-DAD
- studying peak purity
- resolving mixture components (e.g. MCR)

A. de Juan, R. Tauler, Chemometrics applied to unravel multicomponent processes and mixtures. Revisiting latest trends in multivariate resolution, Analytica Chimica Acta 500 (2003) 195-210

Example: the LC-MS data

- a list: m/z index, retention time index and ion abundance



Typical problems with the X-MS data

Data acquisition:

- **massive** analytical data (>300 Mb of data per sample)
- different data formats (netCDF, wiff, mzXML, …)
- different data forms – centroided vs. non-centroided
- issues related to mass spectra acquisition (ghost peaks, bleeding channels)

Data handling:

- not reproducible retention times
- signals quality (noise and background)

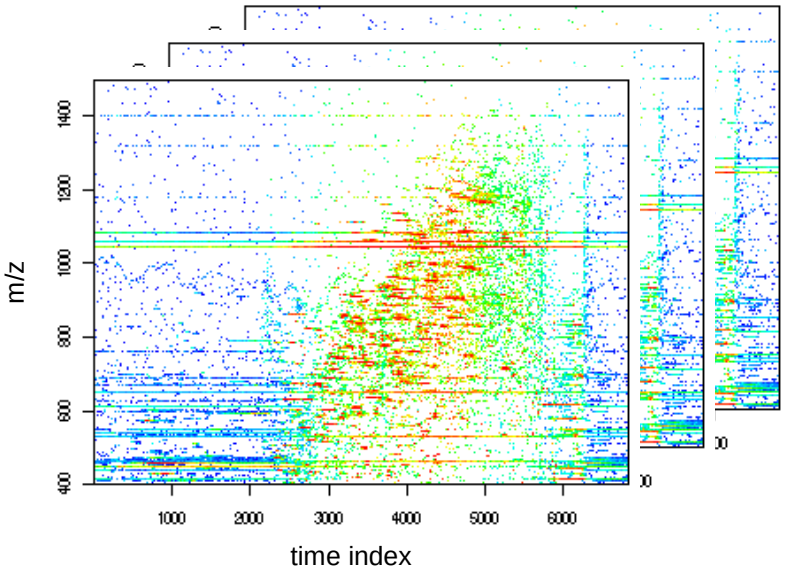
A real example: CE-MS

- D. Bäckström et al., Comparing CE-MS fingerprints of urine samples obtained after intake of coffee, tea or water, Anal. Chem., 80 (2008) 8946-8955

- 230 CE-MS signals

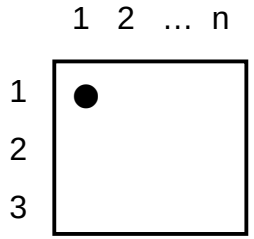
  - ca. **350** Mb of data per sample (only one mode of detection, either positive or negative)
  - total data size 80500 Mb = **115** CDs 700 Mb each

The X-MS data: peak table or fingerprint?



Collection of 2D LC-MS signals
(3 samples)

a peak table



1D, or 2D fingerprint

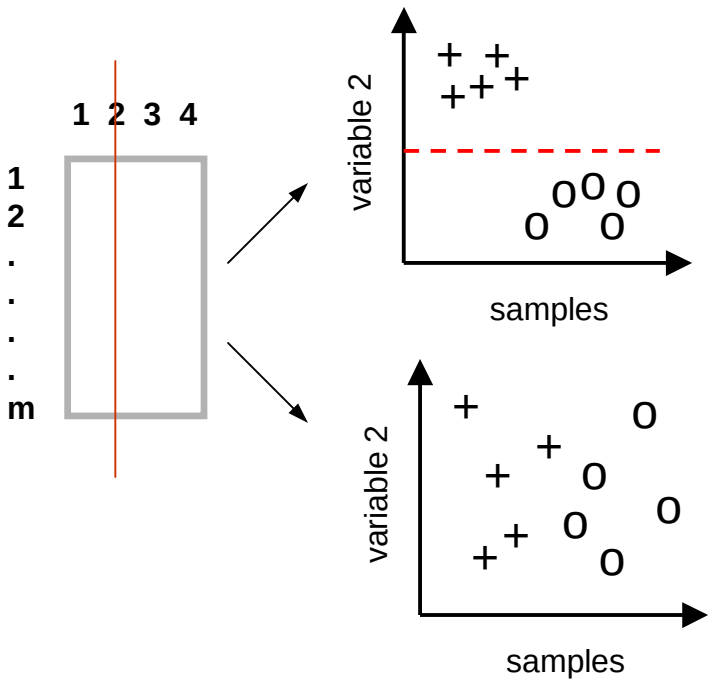
A need for chemometrics in metabolomics

- principles of experimental design
- elegant way of processing massive data
- making efficient use of available data
- dealing with correlation among data variables
- correcting certain deficiencies of individual signals and set of signals

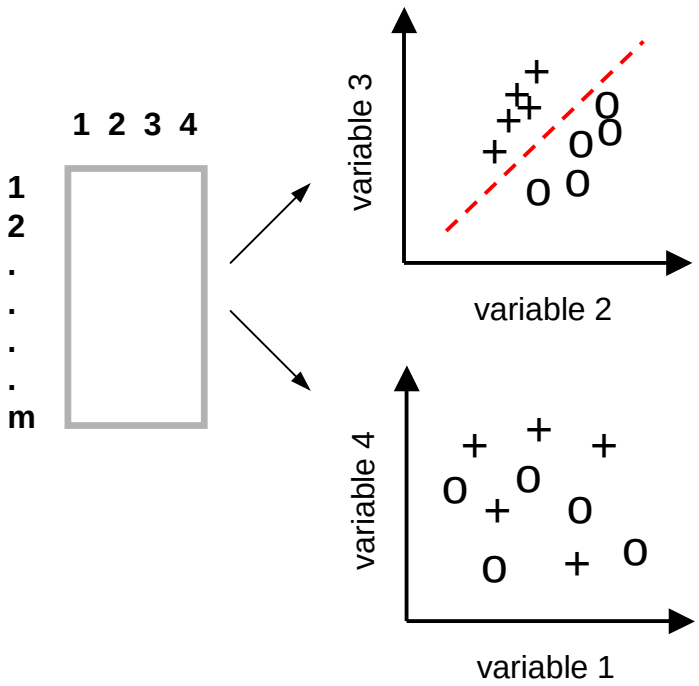
- exploring, visualizing and modeling of the X-MS data

Uni- vs. multivariate discrimination problem

- univariate = one variable

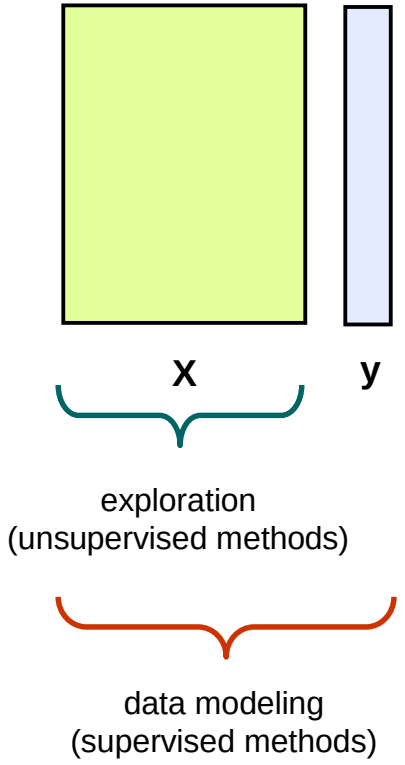


- multivariate = several variables



Multivariate data analysis

- Exploratory data analysis (unsupervised methods)
  - exploring similarities among samples and data variables (e.g. PCA, clustering techniques)
- Data modeling (supervised methods)
  - construction of classification / discrimination models (e.g. calibration techniques, PLS-DA, etc.)

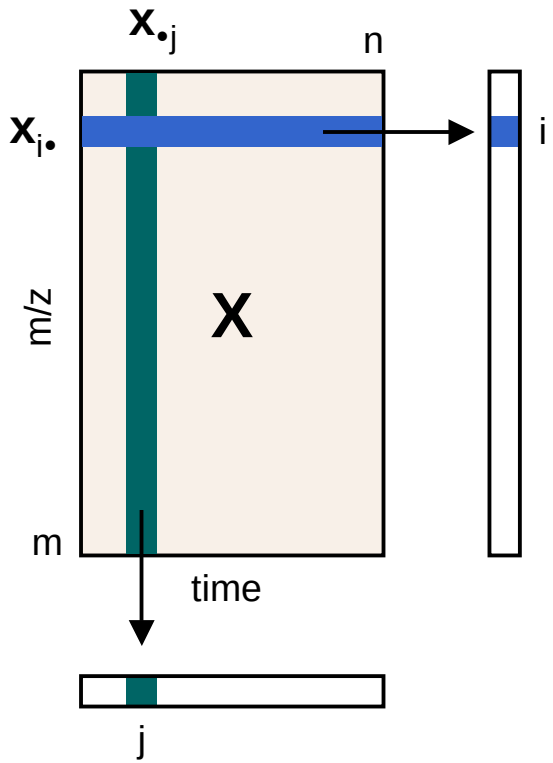


Strategies of the LC-MS data organization

- Strategy 1: guided by the available software (e.g., MarkerLynx)
- Strategy 2: Multivariate curve resolution (Resolver™)
- Strategy 3: by Johnsson *et al*.
- Strategy 4: the N-way approach
- Strategy 5: data reduction

M. Daszykowski, W. Wu, A.W. Nicholls, R.J. Ball, T. Czekaj, B. Walczak, Identifying potential biomarkers in LC-MS data, Journal of Chemometrics, 21 (2007) 292-302

Strategy 5: data reduction



$TIC_{\bullet j} = sum(\mathbf{x}_{\bullet j})$
$BPC_{\bullet j} = max(\mathbf{x}_{\bullet j})$

$TMS_{i\bullet} = sum(\mathbf{x}_{i\bullet})$
$BPP_{i\bullet} = max(\mathbf{x}_{i\bullet})$

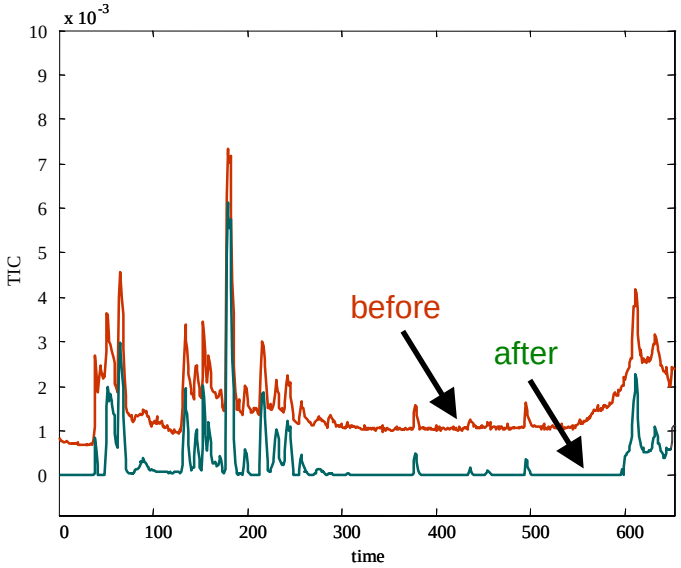
- **TIC** (Total Ion Chromatogram represents the sum of intensities over all m/z channels)
- **TMS** (Total Mass Spectrum represents the sum of intensities over all time channels)
- **BPC** (Base Peak Chromatogram represents the maxima of intensities over all time channels)
- **BPP** (Base Peaks Profile represents the maxima of intensities over all m/z channels)

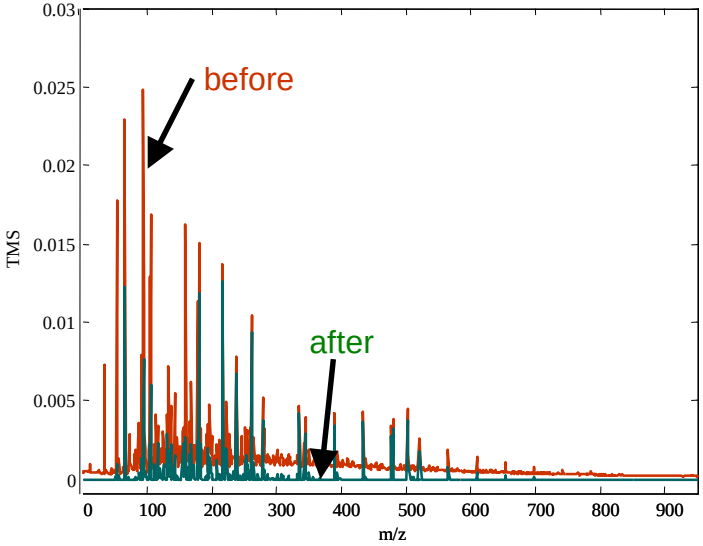
Data preprocessing

- enhancement of signal to noise ratio (S/N):
  - background removal
  - noise reduction
- synchronization of time axis across samples:
  - techniques for alignment of analytical signals
- specific transformations:
  - logarithmic transformation
  - standard normal variate transformation

Enhancement of signal to noise ratio

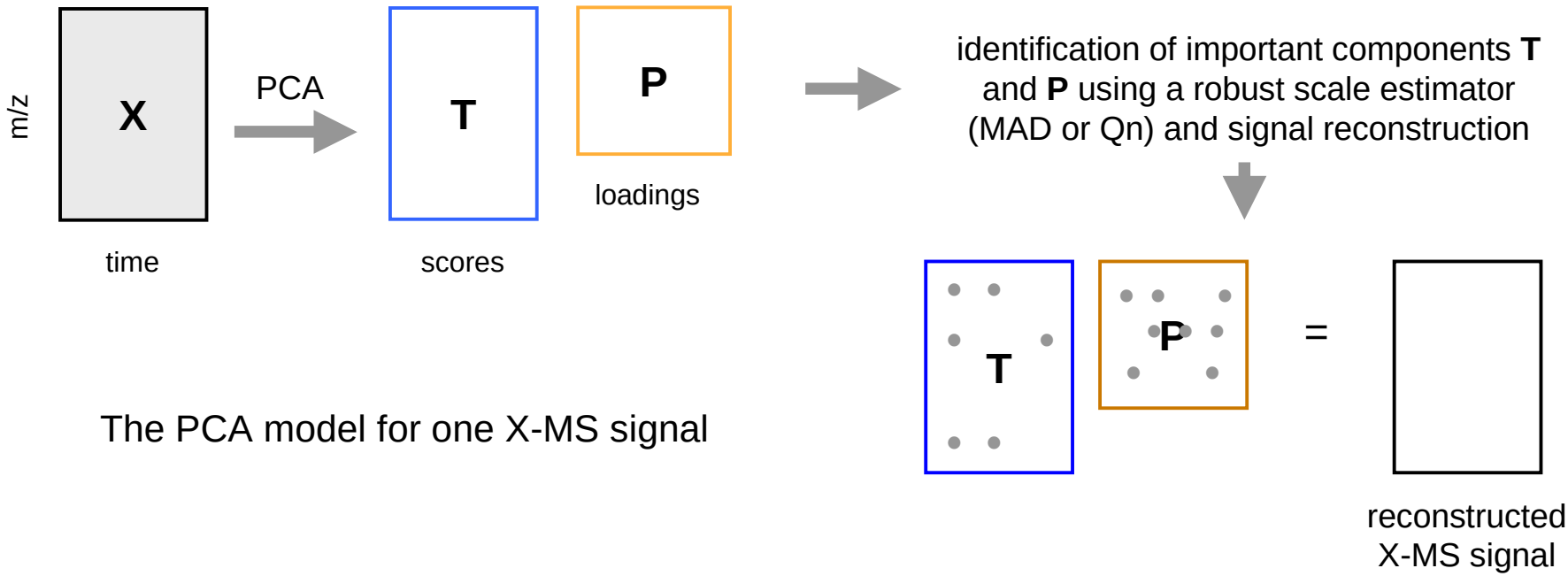
- elimination of noise (wavelets)
- background removal (e.g. PALS)





Background removal

- using robust criterion to detect relevant signal components

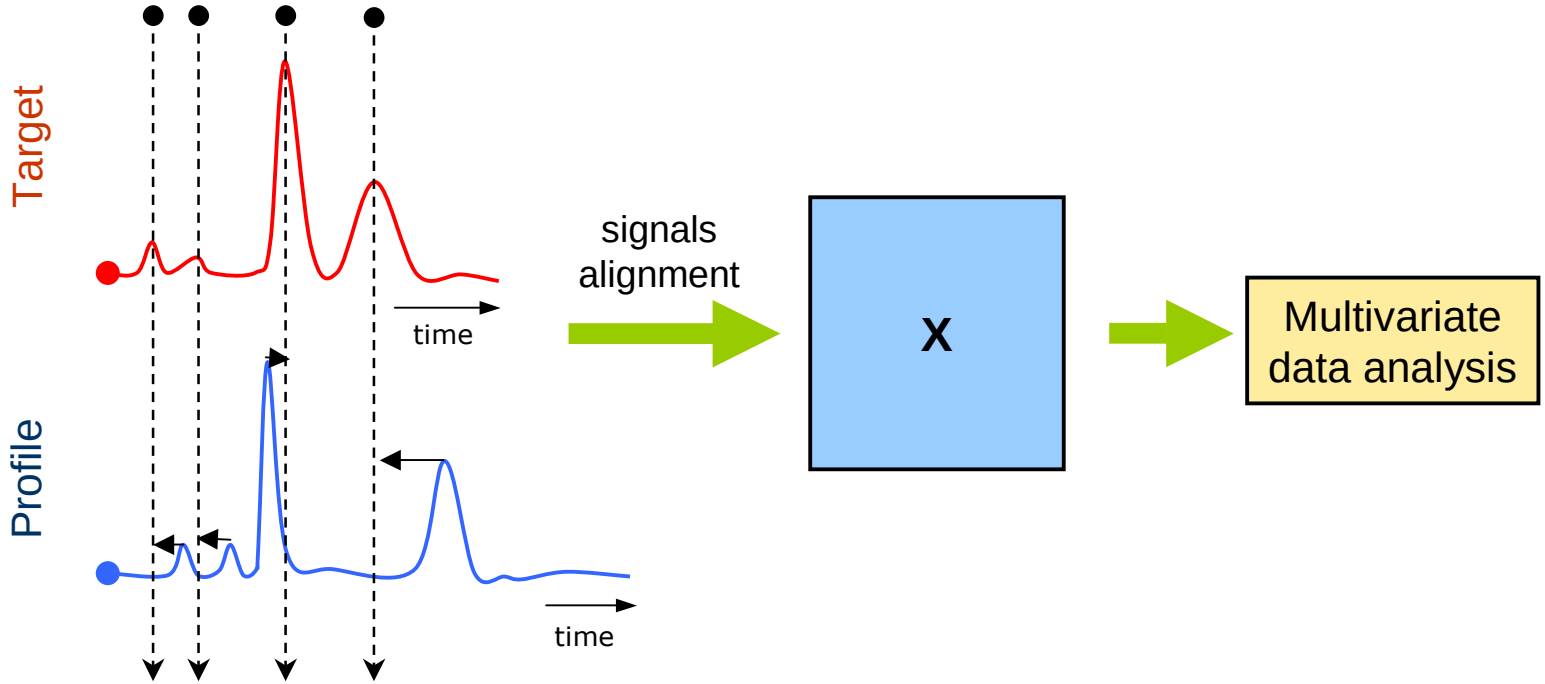


The PCA model for one X-MS signal

M. Daszykowski, W. Wu, A.W. Nicholls, R.J. Ball, T. Czekaj, B. Walczak, Identifying potential biomarkers in LC-MS data, Journal of Chemometrics, 21 (2007) 292-302

Synchronization of time axes

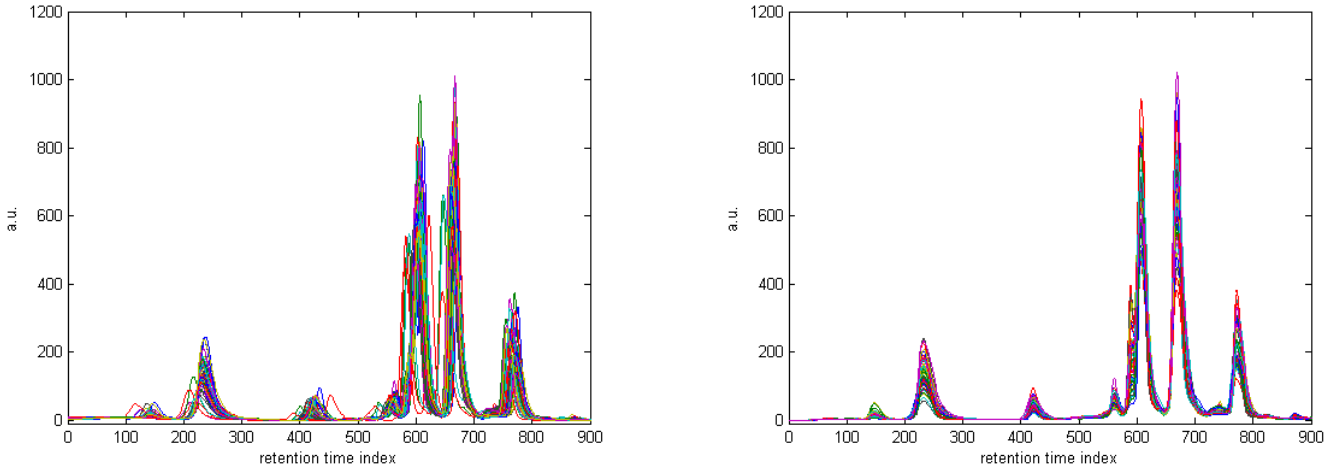
- Alignment techniques
  - e.g., correlation optimized warping (COW) of 1D signal



N.-P. V. Nielsen, J.M. Carstensen, J. Smedsgaard, Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping, Journal of Chromatography A 805 (1998) 17-35

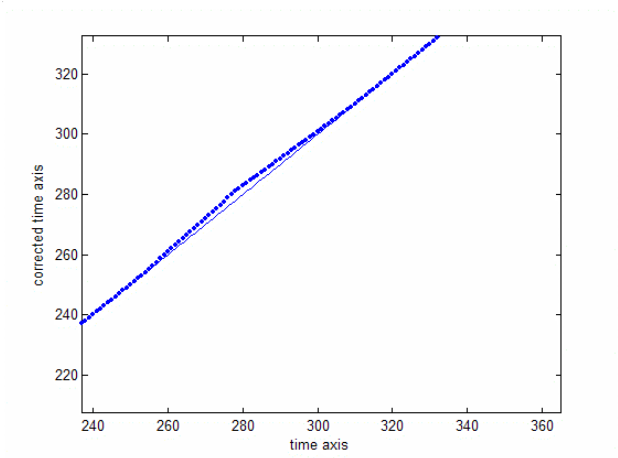
Synchronization of time axes

- HPLC chromatograms of green tea extracts



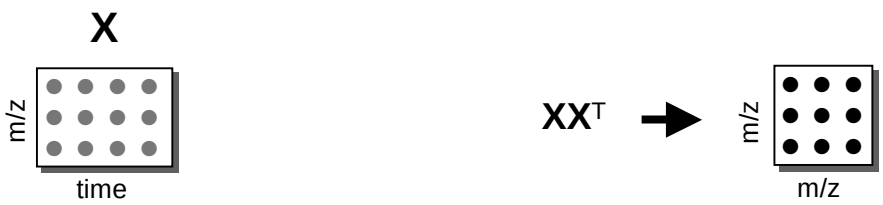
Synchronization of time axes in LC-MS

- Two steps of the LC-MS signals alignment:
  - finding a suitable transformation of the time axis for TIC target profile and TIC profile to be aligned
  - transformation of individual m/z channels of a profile according to determined transformation



Approaches for peak shifts handling

- markers
- correction of peak shifts using alignment methods
  - 1D alignment
  - 2D – taking a second order advantage
- a no-alignment approach – creating a Gram matrix by excluding a problematic data dimension



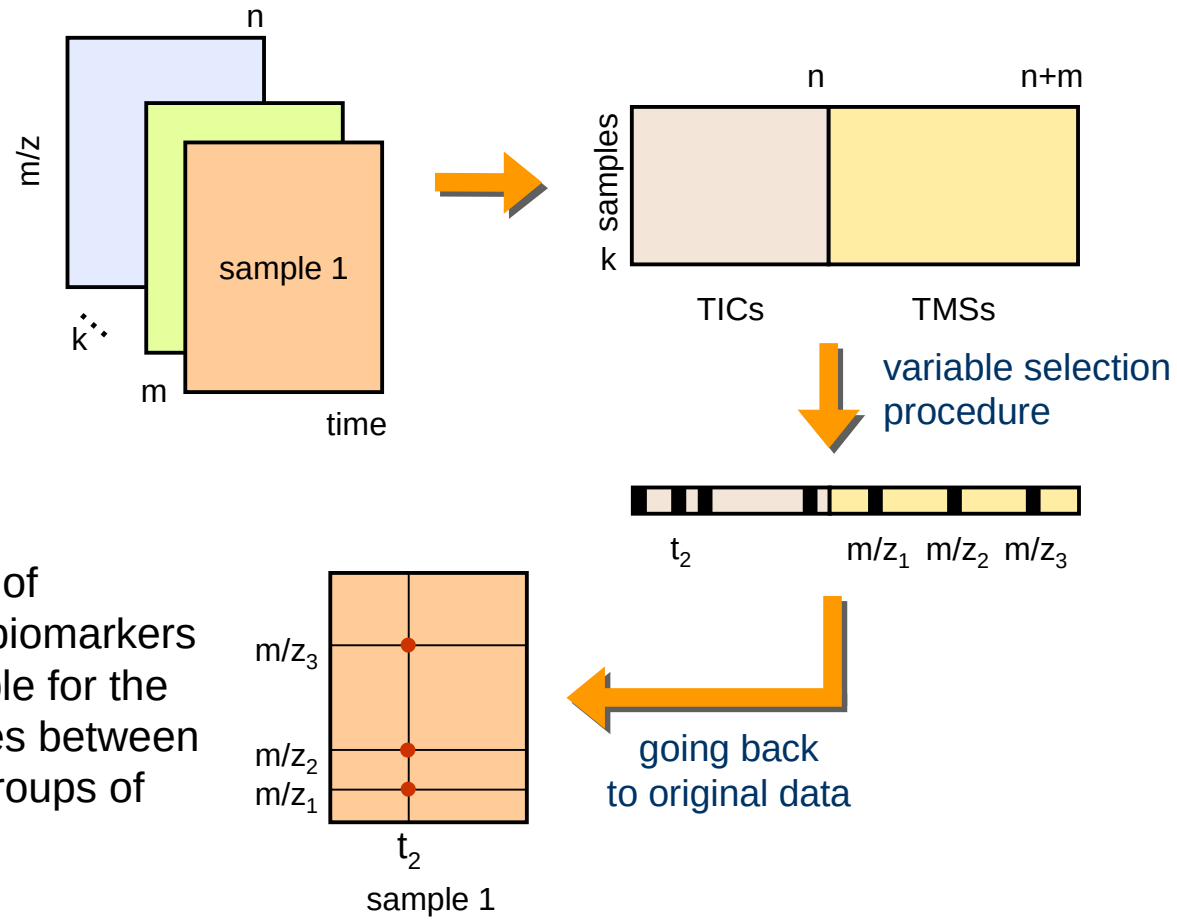
Examples of applications

Goal and data set studied

- goal: identification of possible biomarkers
- studying effect of metabolome of fasting in comparison with a strict diet in healthy volunteers (a two-class discrimination problem) with the LC-MS technique
  - 90 urine samples (44 samples of class 1 and 46 samples of class 2)
  - to validate the discrimination models the available data were divided into model and test sets with the aid of Kennard and Stone algorithm

M. Daszykowski, W. Wu, A.W. Nicholls, R.J. Ball, T. Czekaj, B. Walczak, Identifying potential biomarkers in LC-MS data, Journal of Chemometrics, 21 (2007) 292-302

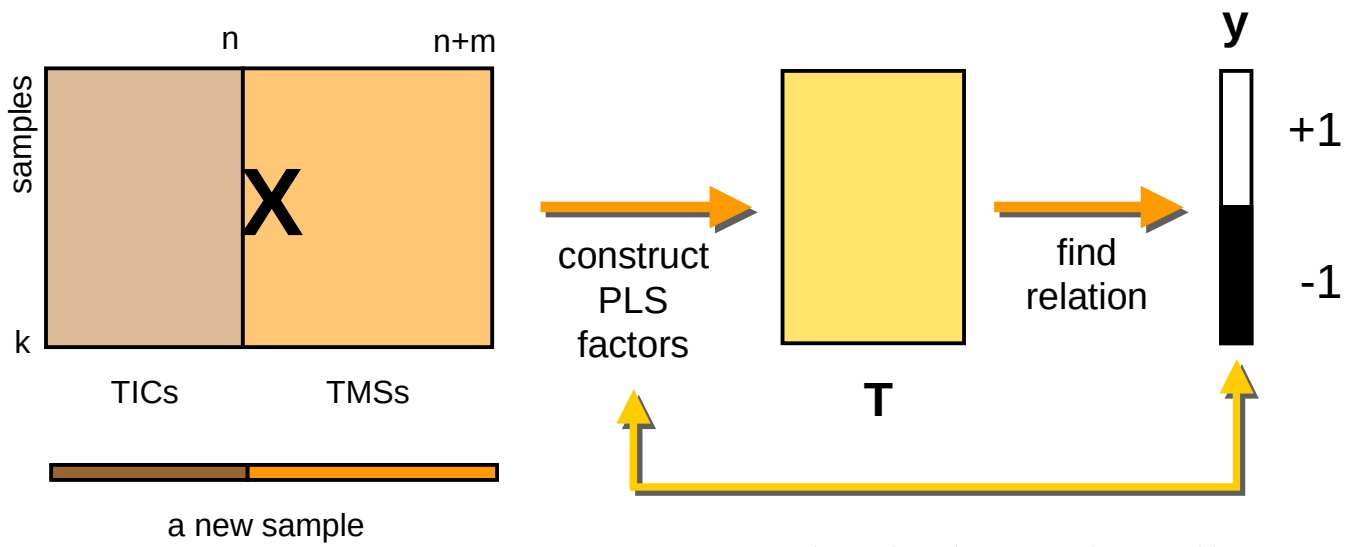
Strategy 5: data reduction



detection of possible biomarkers responsible for the differences between studied groups of samples

Linear discrimination problem

- discriminant Partial Least Squares approach (PLS-DA)



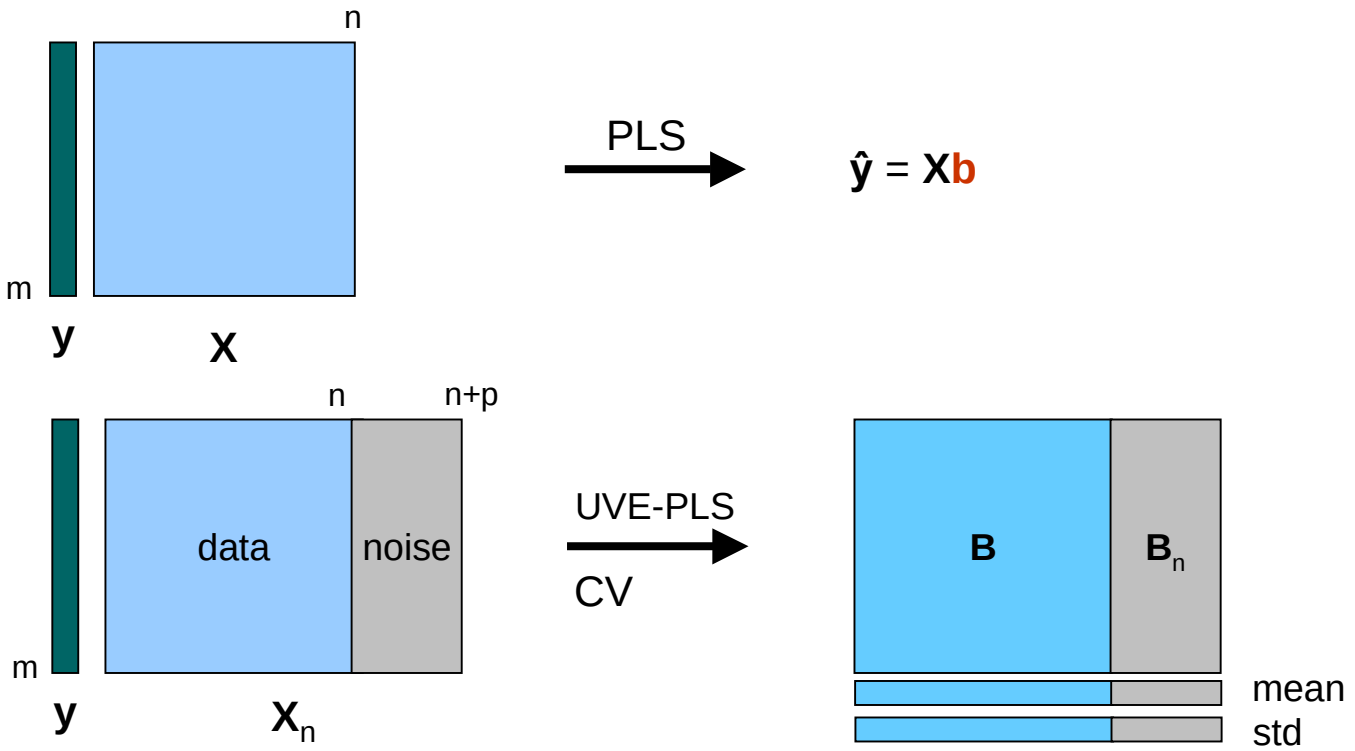
$$Q = \max\left(\mathrm{var}(\mathbf{Xc}) + \mathrm{cov}(\mathbf{Xc}, \mathbf{y})\right)$$

y = Xb

Xc – projection of original data onto a given PLS factor

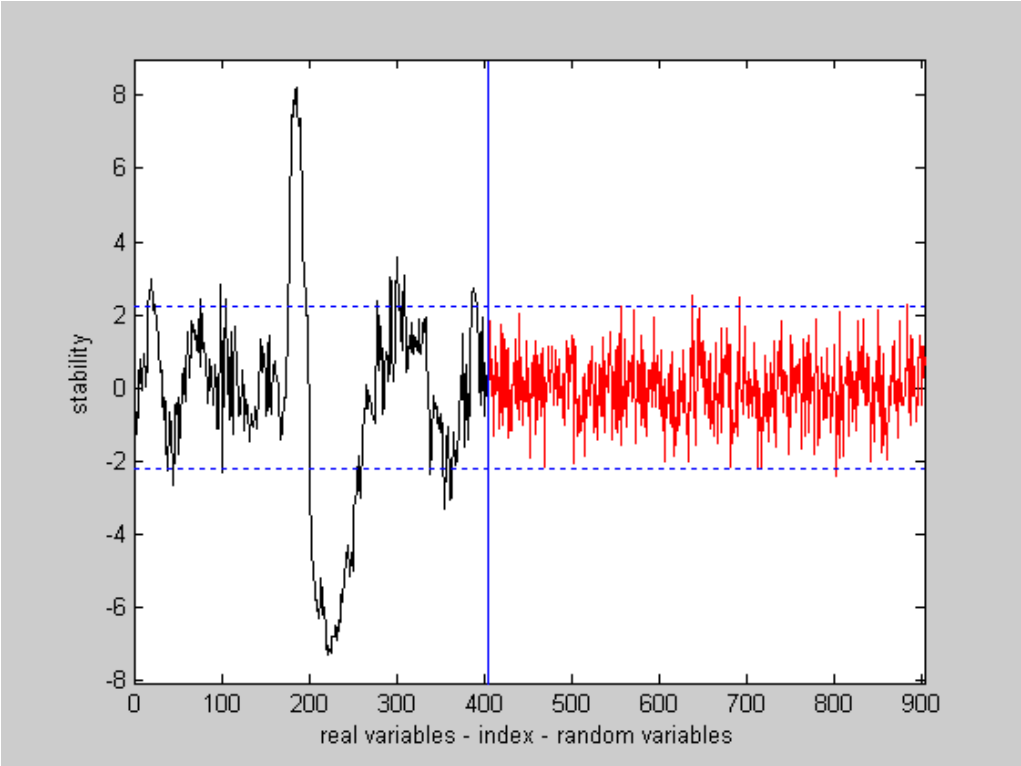
UVE-DPLS

- uninformative variable elimination PLS (detection of stable variables)

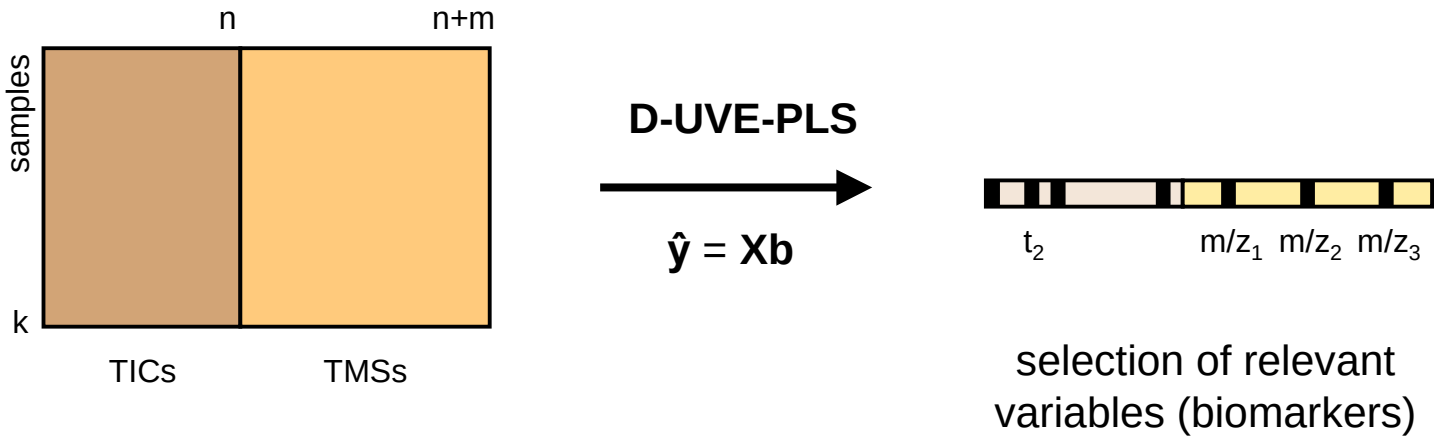


UVE-PLS: stability concept

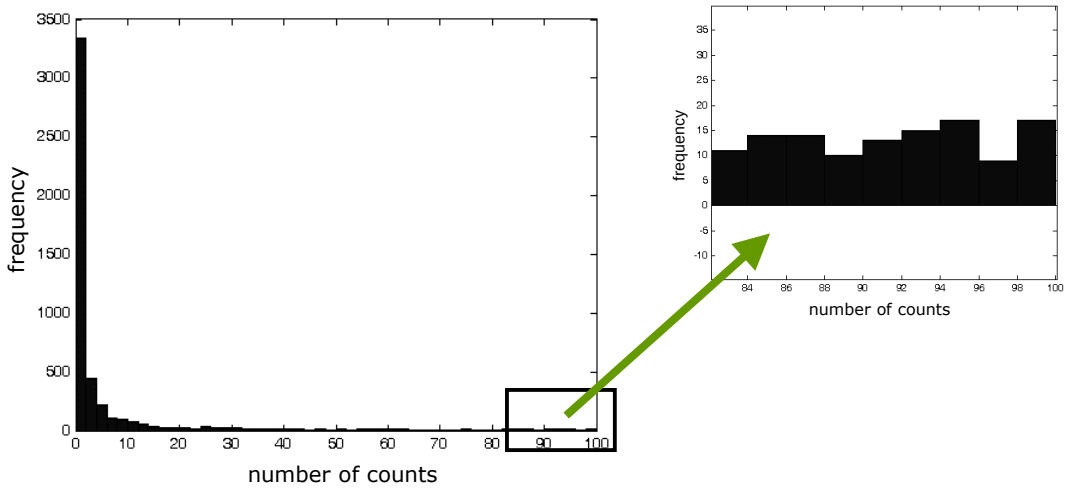
$$stab_j = \frac{mean(\mathbf{b}_j)}{std(\mathbf{b}_j)}$$

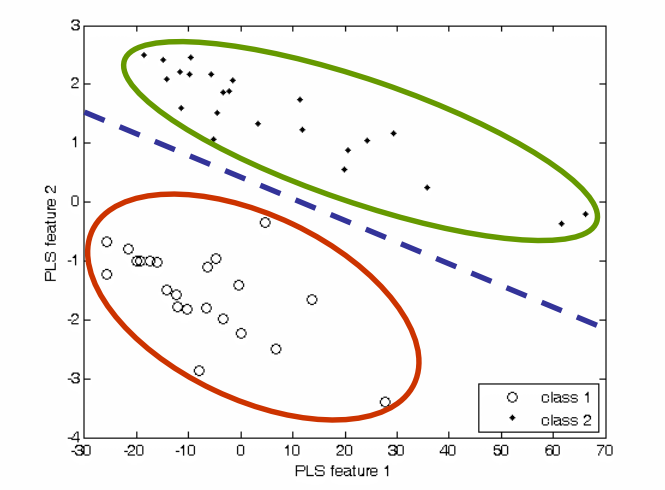


Variable selection D-UVE-PLS



Monte Carlo D-UVE-PLS





Projection of samples on the plane defined by the two first PLS features
(TMS profiles after the SNV transformation)

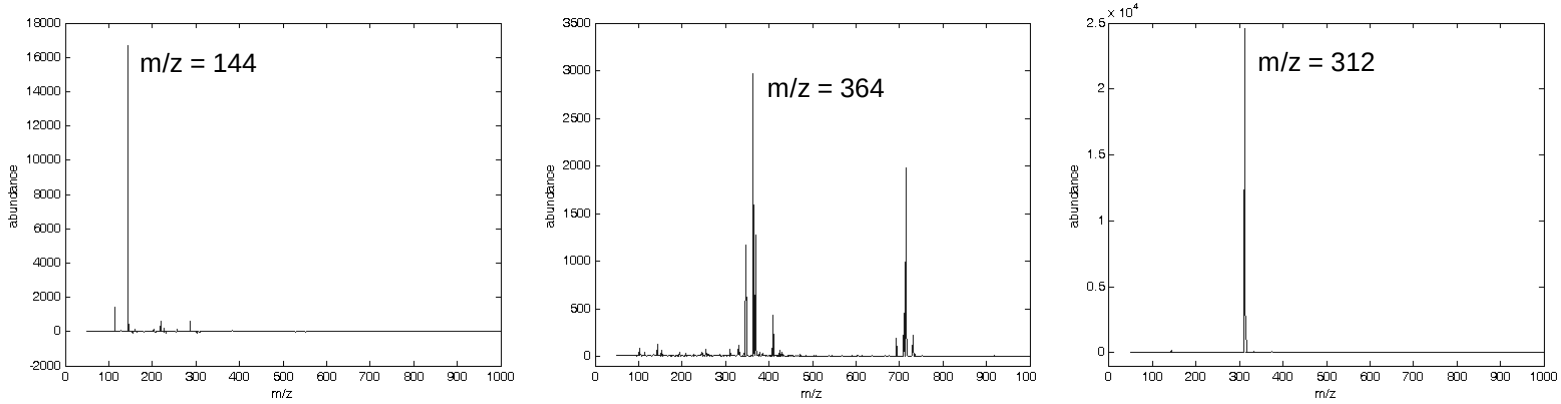
Different Monte Carlo UVE-PLS models

Table 1 Results of Monte Carlo UVE-PLS for the SNV transformed profiles

Profile(s)	Model complexity	RMSEP	Number of selected variables
TIC	5	0.205	**20** (80%)
TMS	3	0.162	**6** (95%)
TIC+TMS	3	**0.163**	**10** (95%)
BPC	4	0.203	**6** (95%)
BPP	1	0.219	**2** (95%)
BPC+BPP	4	0.175	**9** (95%)

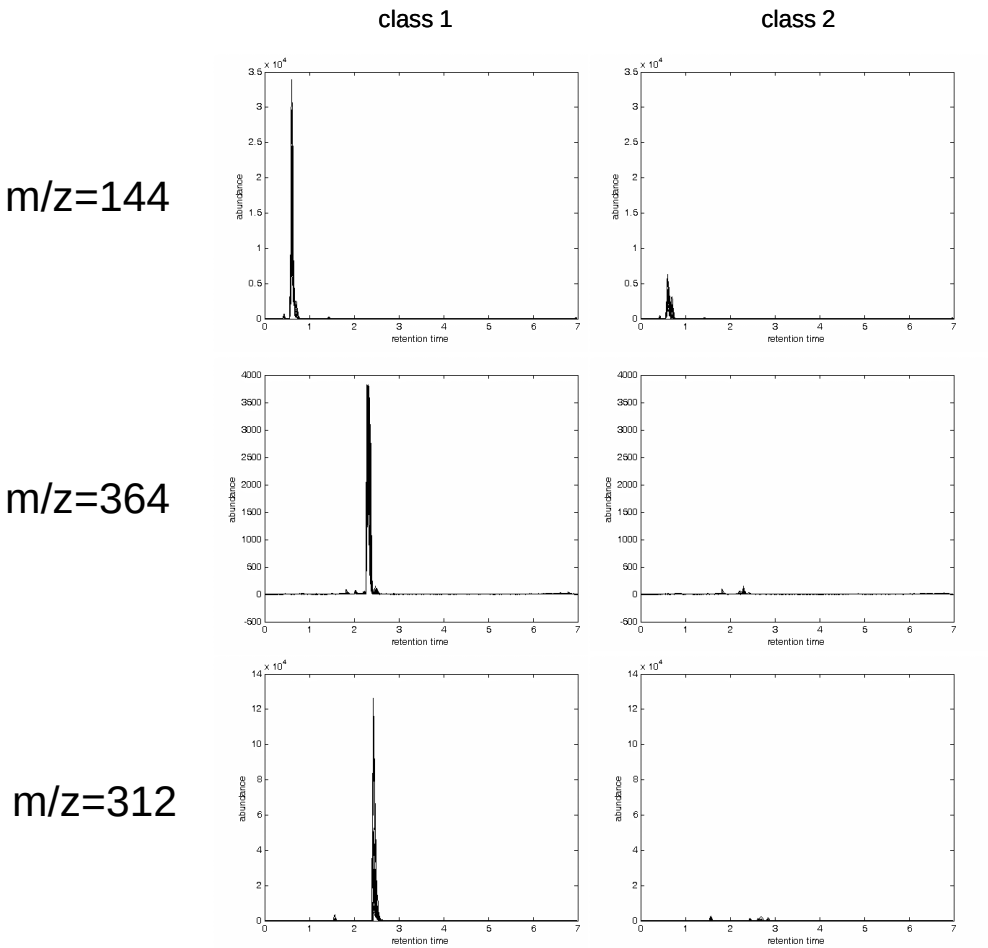
TIC (Total Ion Chromatogram represents the sum of intensities over all time channels)
TMS (Total Mass Spectrum represents the sum of intensities over all m/z channels)
BPC (Base Peak Chromatogram represents the maxima of intensities over all time channels)
BPP (Base Peaks Profile represents the maxima of intensities over all m/z channels)

Identifying potential biomarkers



MS spectra

Identification of biomarkers



Exploratory analysis of 2D X-MS signals

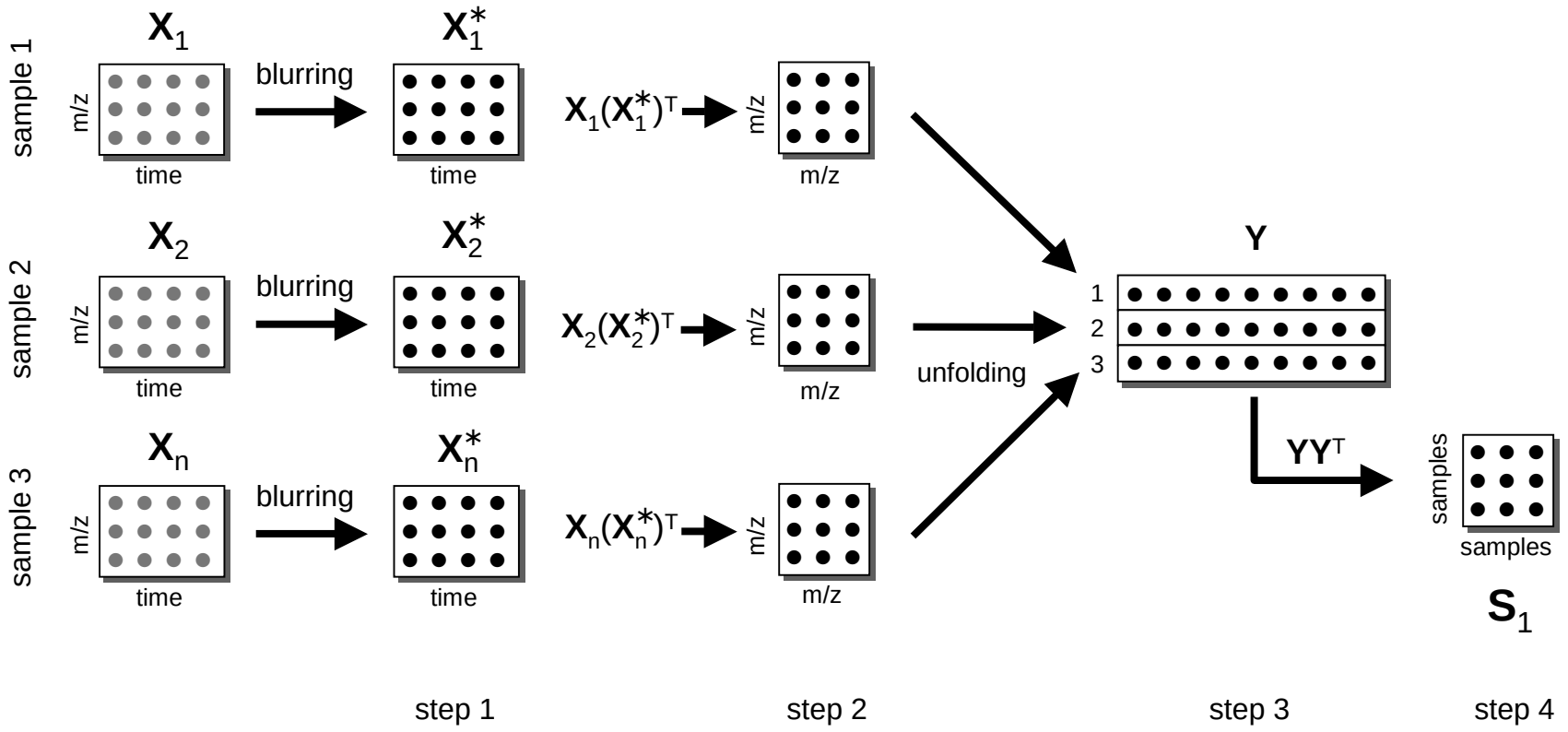
- studying urine profiles of patients before and after paracetamol intake with the use of CE-MS

- a comparative analysis of 2D signals - Rv

$$\mathbf{Rv}(\mathrm{X},\mathrm{Y}) = \frac{\mathrm{trace}\left\{\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{Y}\mathbf{Y}^{\mathrm{T}}\right\}}{\sqrt{\mathrm{trace}\left\{\mathbf{X}\mathbf{X}^{\mathrm{T}}\right\}\cdot\mathrm{trace}\left\{\mathbf{Y}\mathbf{Y}^{\mathrm{T}}\right\}}} = \frac{\mathrm{vec}\left(\mathbf{X}\mathbf{X}^{\mathrm{T}}\right)^{\mathrm{T}}\cdot\mathrm{vec}\left(\mathbf{Y}\mathbf{Y}^{\mathrm{T}}\right)}{\sqrt{\left(\mathrm{vec}\left(\mathbf{X}\mathbf{X}^{\mathrm{T}}\right)^{\mathrm{T}}\cdot\mathrm{vec}\left(\mathbf{X}\mathbf{X}^{\mathrm{T}}\right)\right)\cdot\left(\mathrm{vec}\left(\mathbf{Y}\mathbf{Y}^{\mathrm{T}}\right)^{\mathrm{T}}\cdot\mathrm{vec}\left(\mathbf{Y}\mathbf{Y}^{\mathrm{T}}\right)\right)}}$$

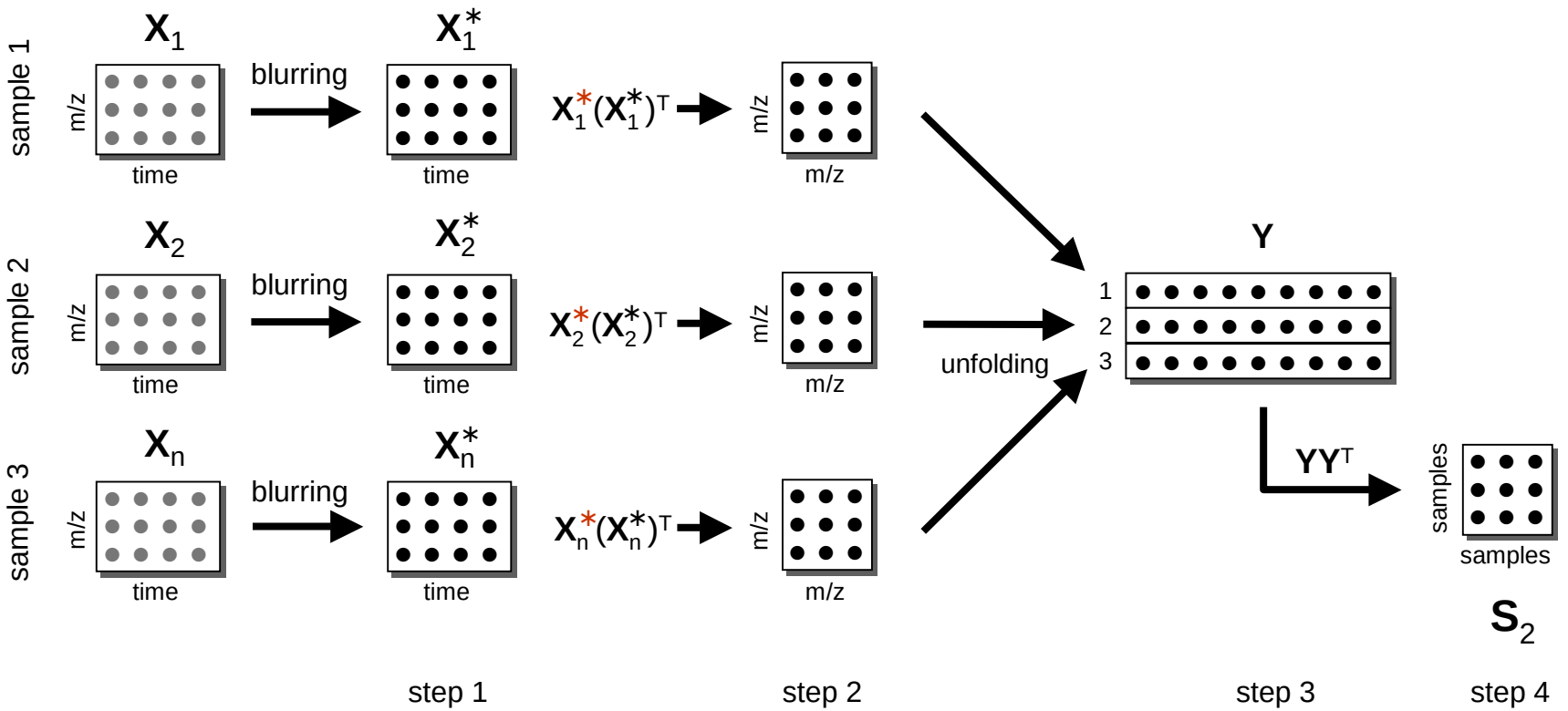
M. Daszykowski, R. Danielsson, B. Walczak, No-alignment-strategies for exploring a set of two-way data tables obtained from capillary electrophoresis-mass spectrometry, Journal of Chromatography A, 1192 (2008) 157-165

Variance-covariance approach*



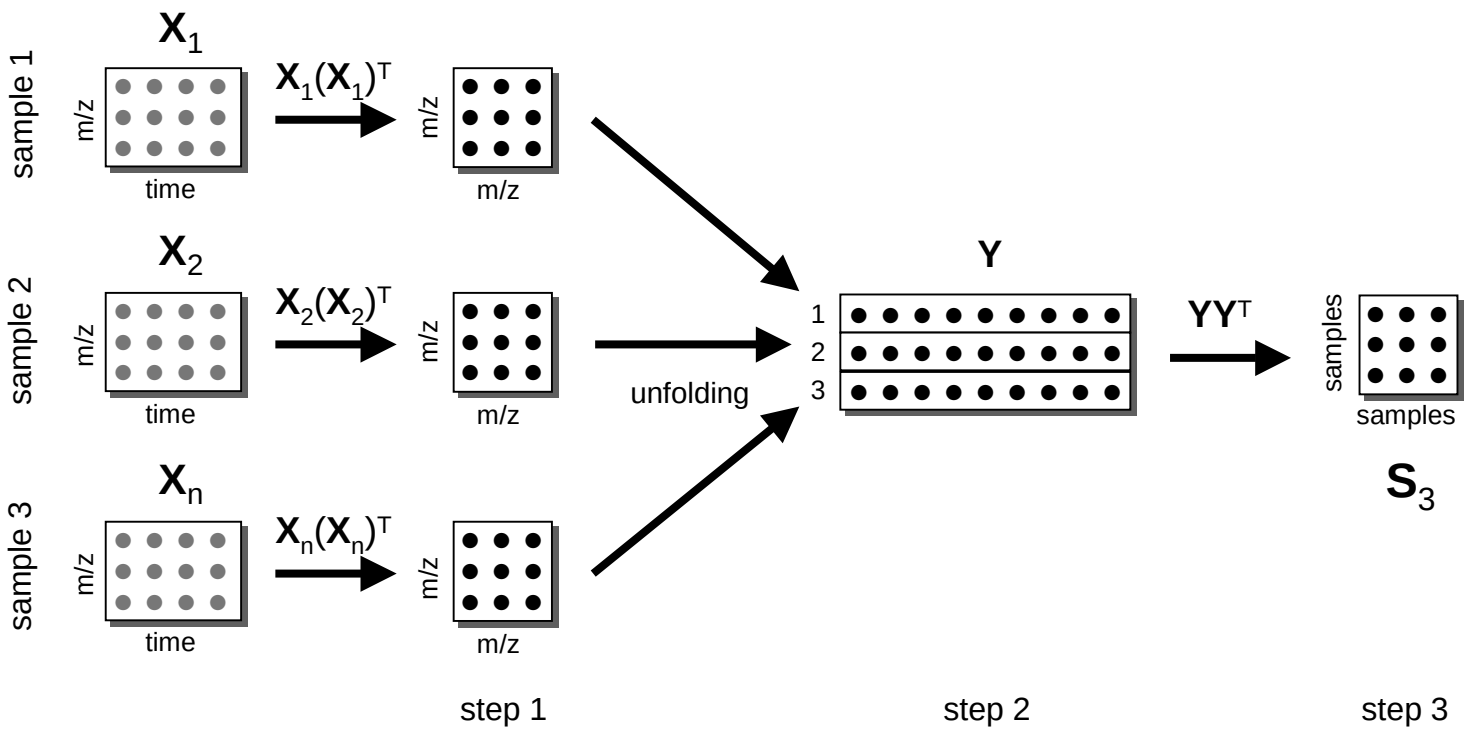
R. Danielsson, D. Bäckström, S. Ullsten, Rapid multivariate analysis of LC/GC/CE data (single or multiple channel detection) without prior peak alignment, Chemometrics and Intelligent Laboratory Systems 84 (2006) 33-39

Variance-covariance approach**



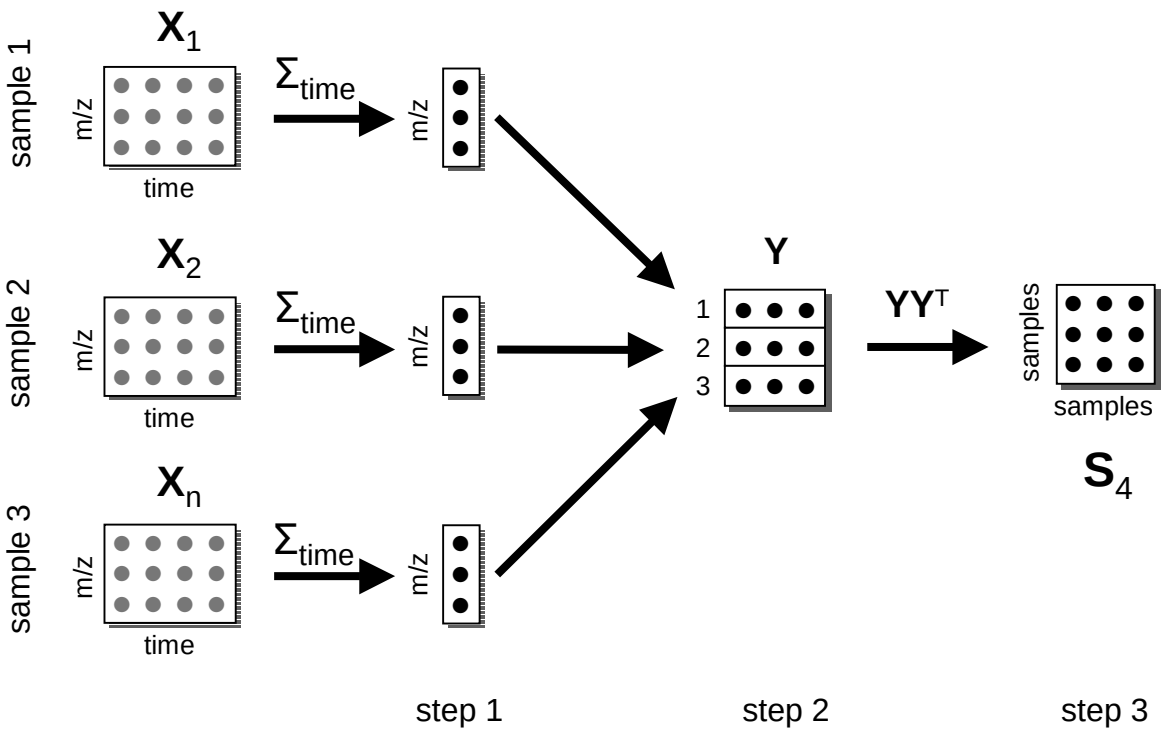
R. Danielsson, D. Bäckström, S. Ullsten, Rapid multivariate analysis of LC/GC/CE data (single or multiple channel detection) without prior peak alignment, Chemometrics and Intelligent Laboratory Systems 84 (2006) 33-39

Gram matrix approach



M. Daszykowski, R. Danielsson, B. Walczak, No-alignment-strategies for exploring a set of two-way data tables obtained from capillary electrophoresis-mass spectrometry, Journal of Chromatography A, 1192 (2008) 157-165

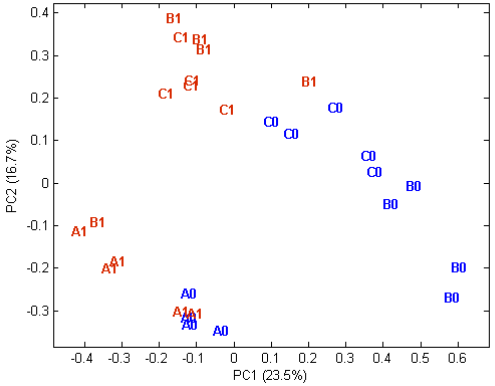
Complete profiles approach

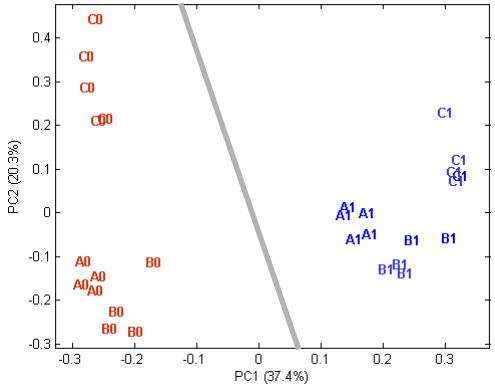


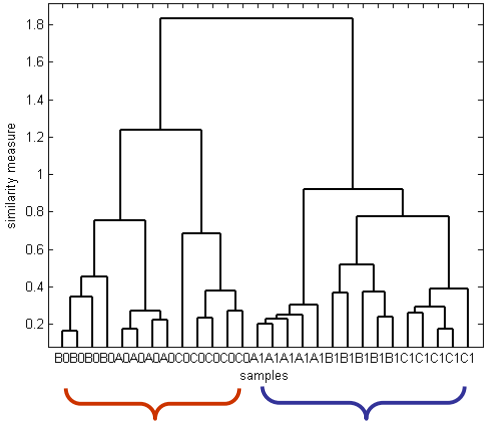
M. Daszykowski, R. Danielsson, B. Walczak, No-alignment-strategies for exploring a set of two-way data tables obtained from capillary electrophoresis-mass spectrometry, Journal of Chromatography A, 1192 (2008) 157-165

Exploring CE-MS data

- a set of CE-MS signals of human urine, before and after paracetamol intake
- Gram approach (Rv)







A, B, C – sample collection ocasion

0 – before

1 – after paracetamol intake

M. Daszykowski, R. Danielsson, B. Walczak, No-alignment-strategies for exploring a set of two-way data tables obtained from capillary electrophoresis-mass spectrometry, Journal of Chromatography A, 1192 (2008) 157-165

Discriminant model: PLS-DA

- discrimination of samples before and after paracetamol intake
- construction of similarity matrix ($\mathbf{S}_1$, $\mathbf{S}_2$, $\mathbf{S}_3$ and $\mathbf{S}_4$)
- 6-fold cross-validation

	R^2	Q^2
$\mathbf{S}_1$	98.8%	79.7%
$\mathbf{S}_2$	98.8%	100%
$\mathbf{S}_3$	98.8%	100%
$\mathbf{S}_4$	98.4%	100%

SSy = (y-mean(y))'*(y-mean(y))

R^2 = 100*(1-[(Ycal-y)'*(Ycal-y)]/SSy)

Q^2 = 100*(1-[(Yval-y)'*(Yval-y)]/SSy)

Conclusions

- preprocessing of the X-MS data is essential for further chemometric data analysis
- dimensionality reduction strategy leads to:
  - easier LC-MS data handling
  - satisfactory discrimination model for data studied
- Monte Carlo UVE-PLS approach enables construction of a reliable discrimination model based on a few selected features
- the selected variables offers a possibility of biomarkers identification

Further challenges

- data fusion
  (combining data from different sources, e.g., GC-MS, LC-MS, etc.)
- efficient handling of X-MS data
- variable selection and good validation of selected variables
- alignment of 2D, 3D (GCxGC-MS) signals
- can we validate somehow available software?
- will you find the same biomarkers if you move your method to another lab, use different instrument, software, etc?
- how to make use of multi channel detection efficient and wide?
- providing and making benchmark data available