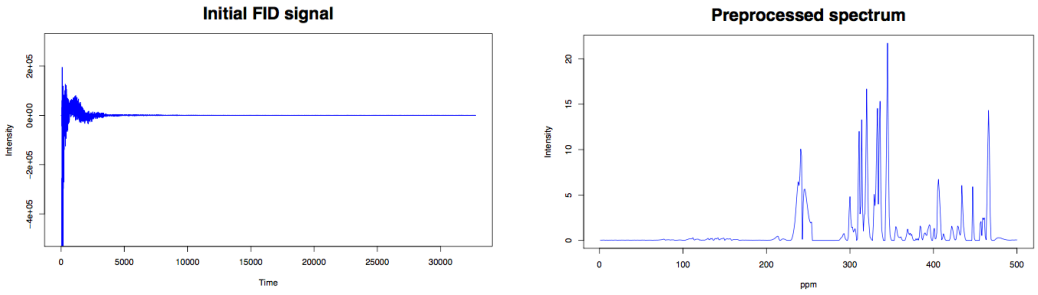
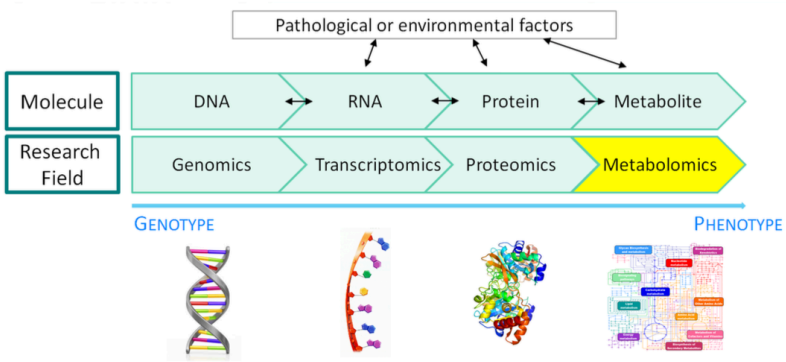
Atelier pré-conférence
RFMF 2018

Présentation et application de PepsNMR, un package R pour le prétraitement de spectres ^{1}H NMR



Metabolomics in the world of "omic's"



Metabolomics studies the link between a question of interest (disease…) and the metabolic images of the molecular content of biological samples (blood, urine, tissue…)

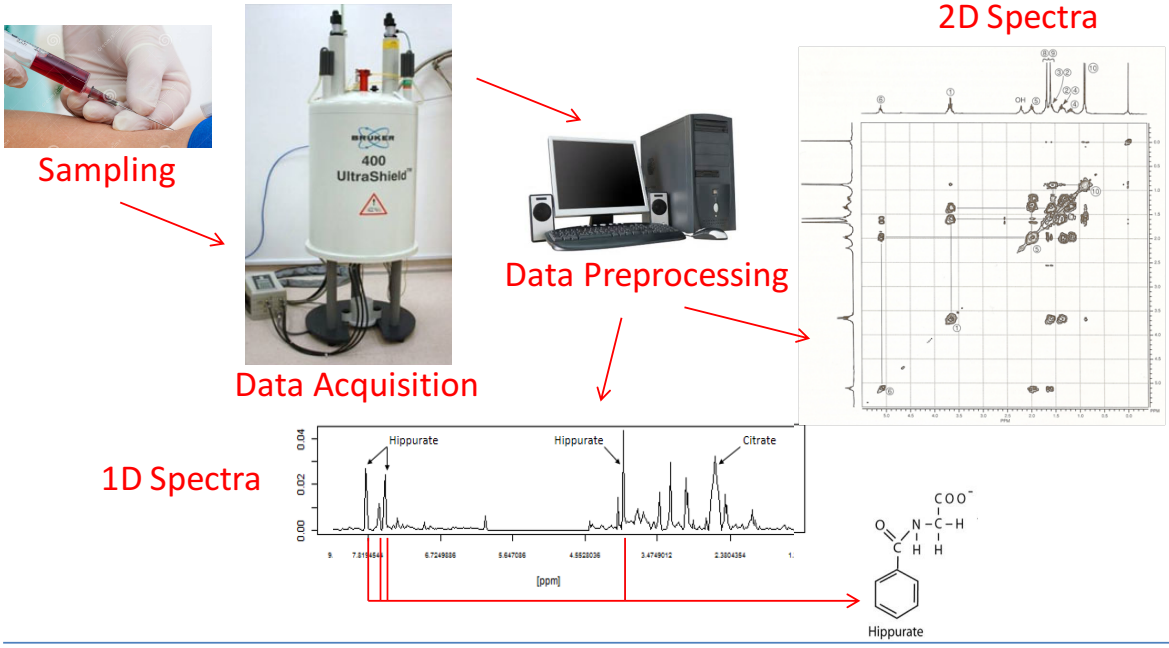
- Advantages : non invasive, quick, provide rich and dynamic information
- Disadvantages : very complex : you have ONE genome but your metabolic profile changes continuously.

Metabolomic platforms and data

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

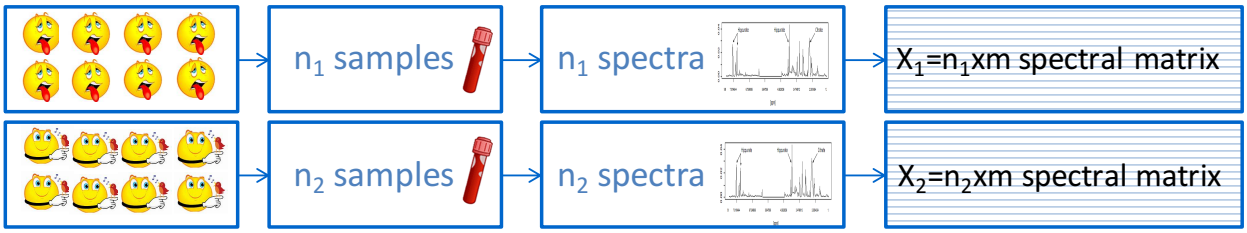
Final data : 1D or 2D spectra. Basic 1D H^1-NMR spectra = 45 000 points

One molecule : multiple peaks in the complex image



Typical metabolomic study and questions (1D spectra)

Two groups of subjects : ill and healthy.



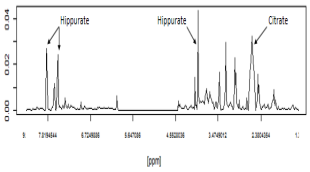
2 Main Goals

- 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.

Problems

- Nb of variables m >>> Nb of observations $n = n_1+n_2$
- Data are affected by noise of multiple sources : subject, time, sampling, data acquisition artifacts, …
- One molecule ≠ one fixed spectral variable => peak shifts, overlap, …

Where can the statisticians help ?



DOE : Design of complex multifactor metabolomic experiments – sample size estimation

Evaluate and compare the quality of spectral data sets (before anf after pretreatment)

Spectral data pre-processing based on complex mathematical and statistical algorithms

Data analysis

- Find biomarkers ("feature selection")
- Prediction models building
- Build models that integrate data of different sources (metabo, geno, clinical…)

Why spectral data pretreatment

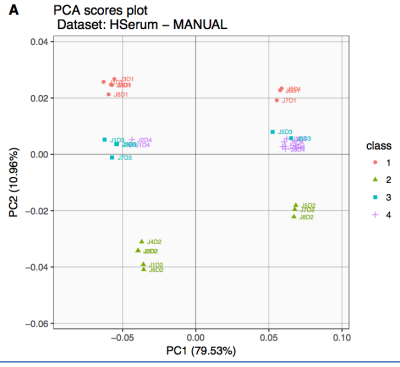
Spectral biological data are affected by many sources of variability

- Subject variability (between subjects and within over time)
- Nature of the biological sample (blood, urine…)
- Biais and noise of the measurement device, instrumental artifacts

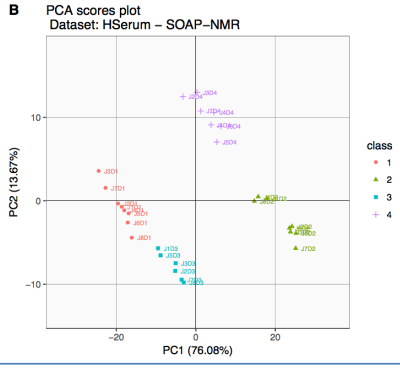
A good preparation (pre-treatment) of the spectra is crucial to maximize the information content of the data before statistical analysis.

Example : 4 donors – 8 samples by donor

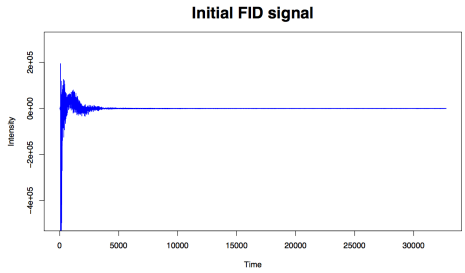
With manual pretreatment



With advanced PepsNMR pretreatment



Nature of ^{1}H-NMR data

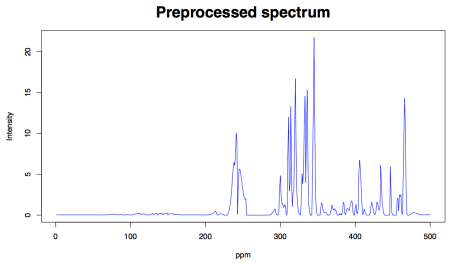


Raw data

- FID signal intensity over time
- 32768 variables (2^{15}) - Complex numbers

Transformed/final data

- Spectral intensity (descriptors) over part per million
- 250 to 1000 variables / Real numbers
- One metabolite = several descriptors and 1 or several peak(s)



Pretreatements

- Many steps including Fourier transform
- Usually done manually (spectrum by spectrum) with the help of an adapted software provided with the spectrometer (e.g. Topsin)
- Here we propose an "automatic" approach in R developed at ISBA - UCL

FID : Free Induction Decay = represents the net sum of distinct signal components from different proton nuclei

Ppm = chemical shift : frequency of the signal in a unit independent from the spectrometer and related to the position of an internal standard for which the ppm is set (usually the TMSP at 0).

Steps of the pre-processing process in PepsNMR

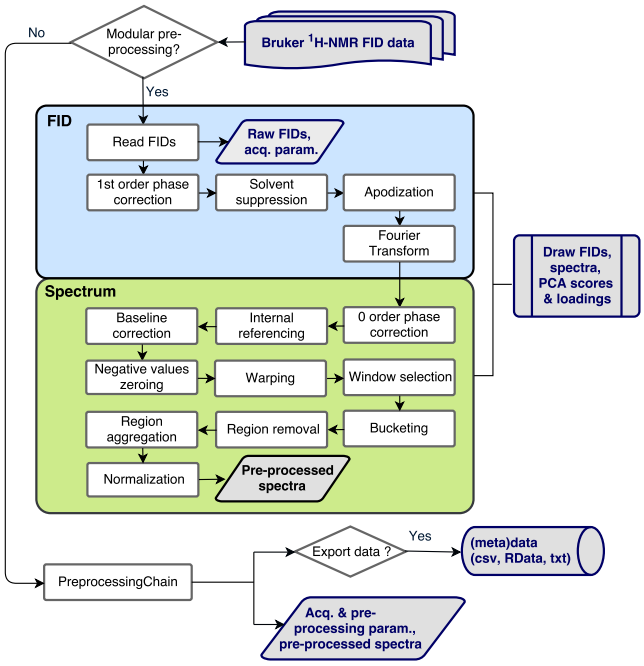
PepsNMR or **P**ackaged **E**xtensive **P**reprocessing **S**trategy for **NMR** data

Origins: Matlab library (BUBBLE) developped by Eli Lilly & Paul Eilers in collaboration with UCL [Vanwinsberghe, 2005; Rousseau, 2011]

Developments of PepsNMR:

- Semi-automated & flexible pre-processing scheme
- Recognized methods and in-house algorithms
- Classic and advanced pre-processing steps
- Free to use (R package) and documented

Article available online:
M. Martin et al., 2018



PepsNMR flowchart with available functions, inputs and outputs

Details of PepsNMR Steps

The NMR pre-processing shows:

- an example of the complexity of bioinformatic data pretreatment. The expertise necessary to understand these steps is issued from analytical chemistry.
- that the statisticians may help a lot in this context
  - Several pre-processing steps need statistical expertise e.g. in non parametric statistics.
  - Several steps include tuning parameters which may be optimized by experimental design methods.
  - The real efficiency of each step in never guarantied and may be checked after pre-processing with adequate statistical quality criteria.

A bad pre-processing will never be compensated by a good data analysis!

Pre-processing steps overview

Steps	Description
Group Delay Correction	Correct for the Bruker Group Delay.
Solvent Suppression	Remove the solvent signal from the FID.
Apodization	Increase the Signal-to-Noise ratio of the FID.
Fourier Transform	Transform the FID into a spectrum and convert the frequency scale (Hz -> ppm).
Zero Order Phase Correction	Correct for the zero order phase dephasing.
Internal Referencing	Calibrate the spectra with an internal compound referencing.
Baseline Correction	Remove the spectral baseline.
Negative Values Zeroing	Set negatives values to 0.
Warping	Warp the spectra according to a reference spectrum.
Window Selection	Select the informative part of the spectrum.
Bucketing	Data reduction.
Region Removal	Set a desired spectral region to 0.
Zone Aggregation	Aggregate a spectral region into a single peak.
Normalization	Normalize the spectra.

Online package repository: Github

- Link: https://github.com/ManonMartin/PepsNMR
- install the package in R:

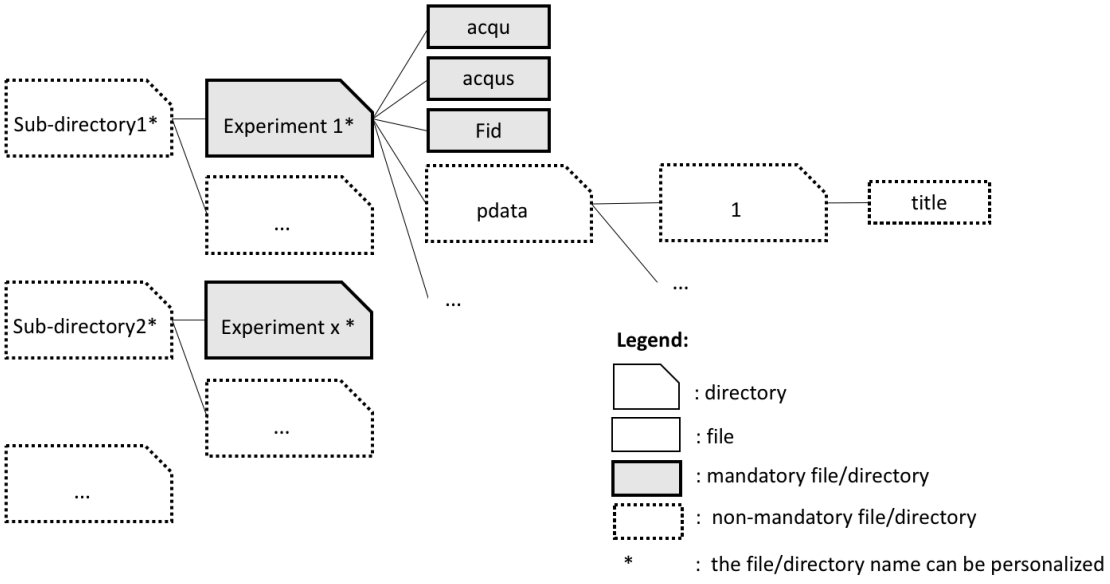
```
install.packages("devtools", dependencies = TRUE)
require(devtools)
install_github("manonmartin/pepsnmr", build_vignettes = TRUE,
               dependencies = TRUE)
```

- On Github: README.md with code that can be copy-pasted for spectral pre-processing
- Vignette (application of the package on a real data case) in R:

```
require(PepsNMR)
vignette("PepsNMR_minimal_example")
```

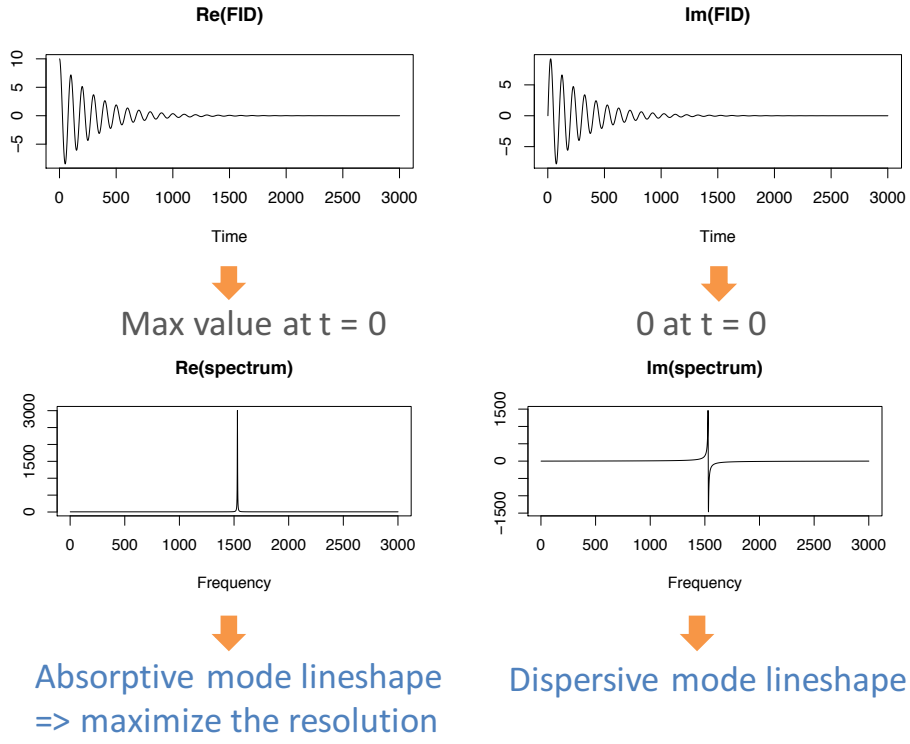

Step 0 : Read Bruker Fids

- Need to have well organized data
- Different ways to organise and read the data
- Experiment names: from the *title file* or the *"Experiment …" directories* names



Phase shifts (1)

1 perfectly phased FID/spectrum

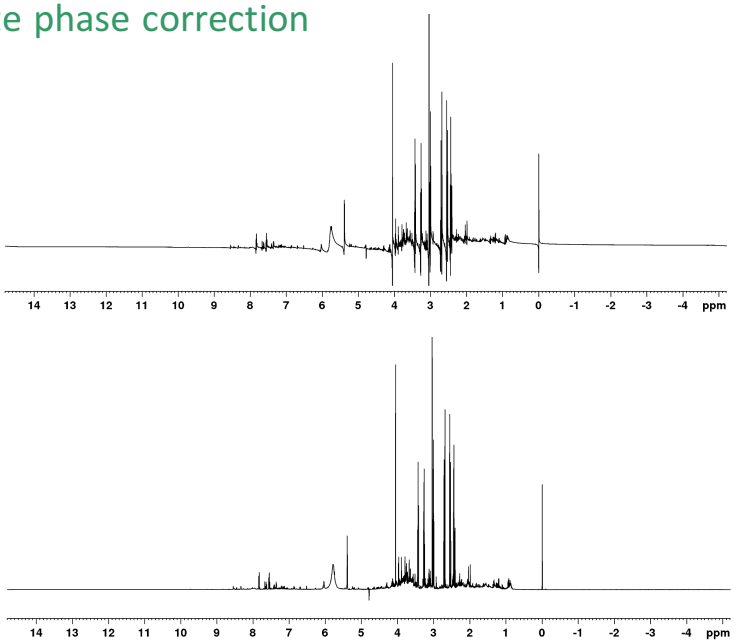


Phase shifts (2)

- Spectra have a mix of absorptive and dispersive signals

  ⟹ The phase must be corrected

- Phase correction is a crucial step: the integration of signals depends on an adequate phase correction



Step 1 : Group delay correction (1)

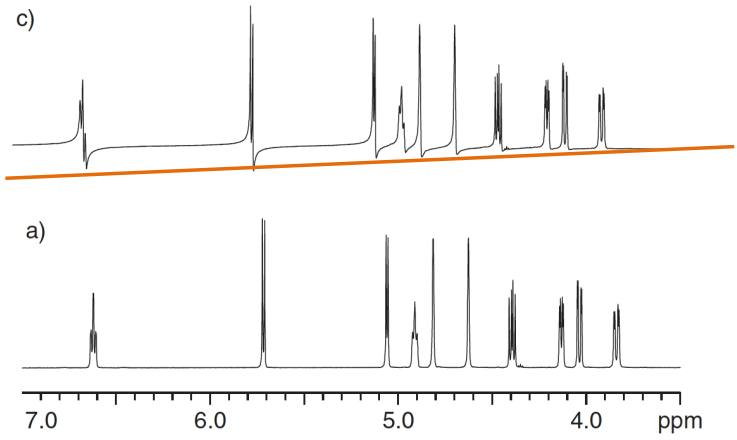
Zero (φ_0) and first order ($\varphi_1(\nu)$) phase shifts: $S = S_{phased}\, e^{i(\varphi_0+\varphi_1(\nu))}$

Presence of a **digital filter**

- The first tens of points in the FID are not part of the recorded signal and are called the **group delay**
- Since the phase shift differs across signals, it introduces a first order phase shift linearly related to ν (—)

First order phase shift

Correctly phased spectrum



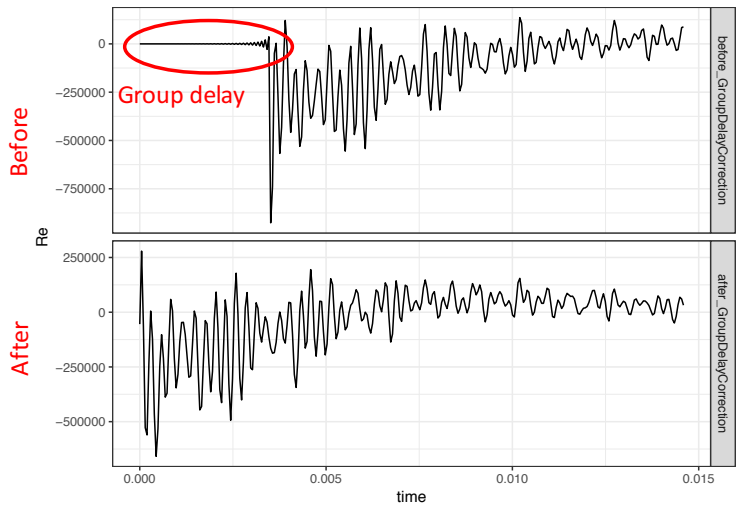
Claridge, T. D. (2016). High-resolution NMR techniques in organic chemistry (Vol. 27). Elsevier.

Step 1 : Group delay correction (2)

The Bruker pre-acquisition delay τ is known from the acquisition parameters (either GRPDLY or DECIM/DSPFVS)

Procedure:

- Apply Fourier transform on the FID (allows for a non-integer delay)
- Remove the delay τ
- Apply inverse Fourier transform on the spectrum to recover the FID



Step 2 : Solvent signal suppression

Solvent residuals signal is problematic:

- Highly variable
- Not of interest & can mask useful information
- Affects other pre-processing steps (baseline correction, phase corrections, etc.)

Hypothesis:

Solvent is the main component in the analysed biological samples

Principle [Eilers, 2003]:

A statistical smoother is used to estimate and remove the useless signal ***Z***

Min (Least Squares criterion + lambda*roughness)

deviations of Z from the FID roughness penalty

Meta-parameter: smoothing parameter lambda

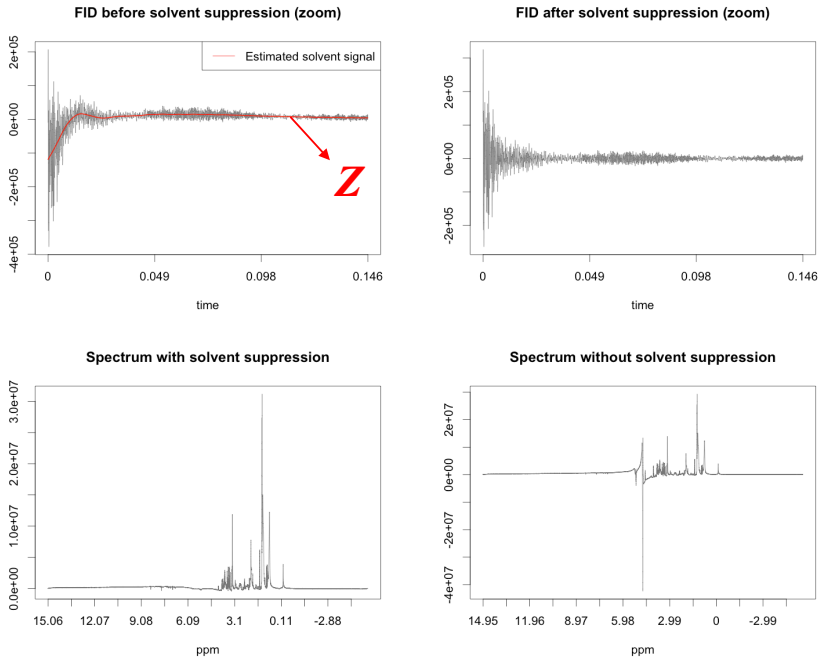
Determine how smooth is the solvent signal : λ ↑ => smoother signal

Statistical expertise is necessary ☺

Illustration of solvent residuals suppression

$$\min Q = \text{Least Squares criterion} + \text{Smoothness Penalty}$$

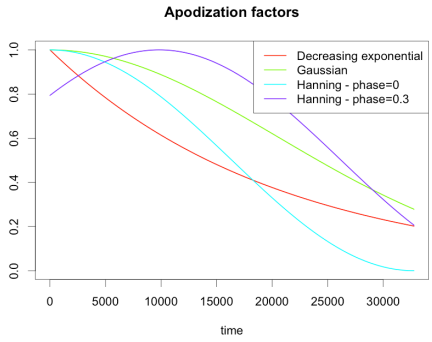
$$= \sum_{j=1}^{m} w_j (F^{old}(\nu_j) - Z(\nu_j))^2 + \lambda \sum_{j=1}^{m} (\Delta^2 Z(\nu_j))^2$$



Step 3 : Apodization

NMR signal intensity decays through time whereas the registered noise has a random noise with a constant amplitude.

- Apodisation purpose: improve the sensitivity and/or the resolution
- Procedure: multiply the FID by a factor called a **weighting function**
- Several classes of factors in Peps:
  - Negative exponential
  - Gaussian
  - Hanning
  - Hamming
  - Cos2



- Example: the negative exponential function to emphasise spectral zones with high signal to noise ratio.

$$W(t) = \exp(-LB * t) \text{ with } LB : \text{line broadening factor}$$

Common transformation in spectroscopy. No real need of statisticians

Step 3 : Sensitivity/resolution improvement

Sensitivity (SNR) enhancement

- over time:
  - Noise is random with a constant amplitude
  - FID signal is decaying
- The SNR is higher at the beginning of the FID

⇒ Decreasing factor needed to emphasize spectral areas with high SNR

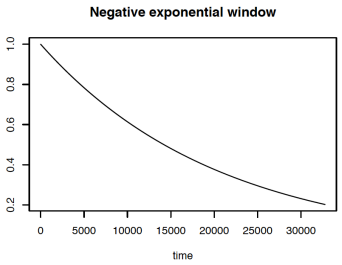
Resolution enhancement

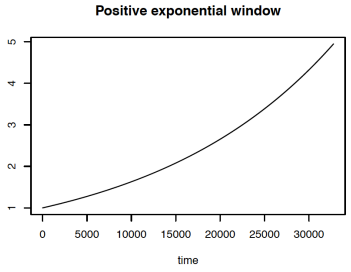
- A faster decay implies broader lines and decreased peaks height

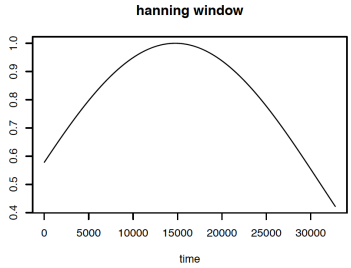
⇒ Increasing factor needed to delay the decay

- A decaying factor used for SNR improvement reduces the resolution

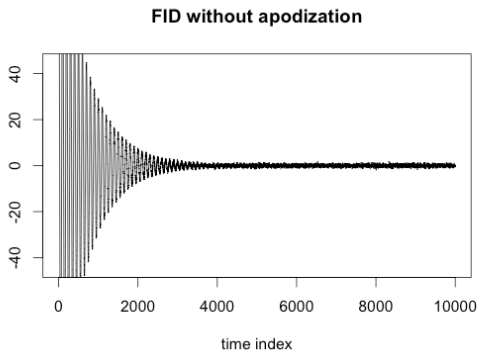
⇒ A **trade-off** solution is to use a concave factor

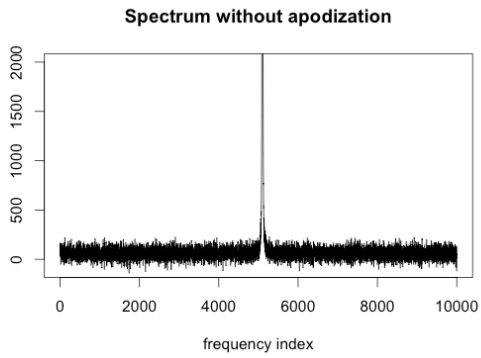


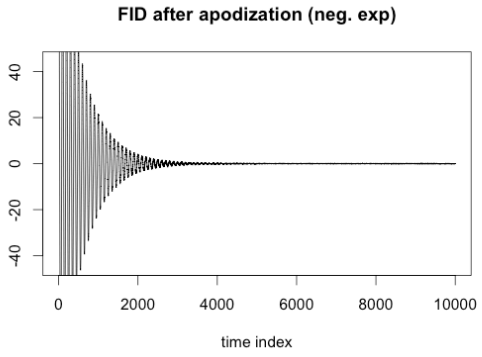


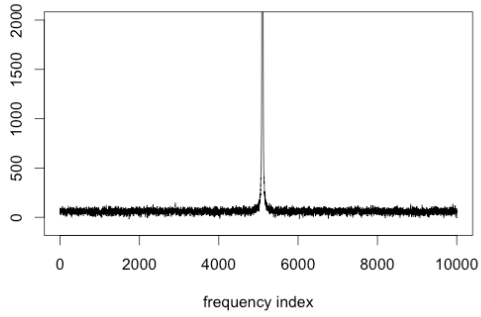


Step 3 : Example of apodization with a negative exponential









Step 4 : Fourier transform

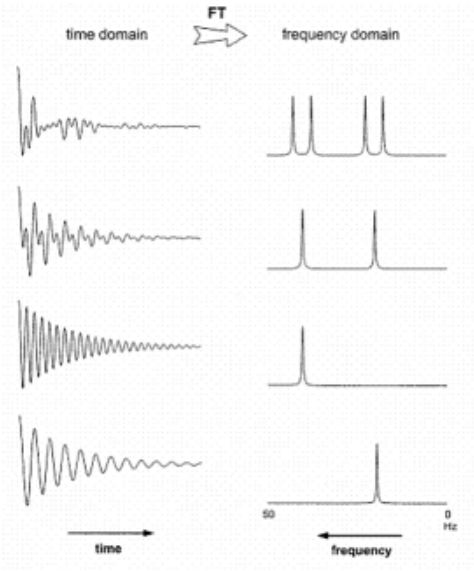
The Fourier Transform

- recovers intensities of individual signals from the mixed time signal.
- is a well known transformation used in many areas.
- Converts the signal in the time domain into a spectrum in the frequency domain (Hertz).

For each frequency ν of the spectra, the intensity F is calculated from the discrete FID time signal S_t with the Discrete Fourier transform (Fast Fourier algorithm):

$$F(\nu) = \sum_{t=0}^{m-1} S_t \exp(-2i\pi\nu t/m)$$

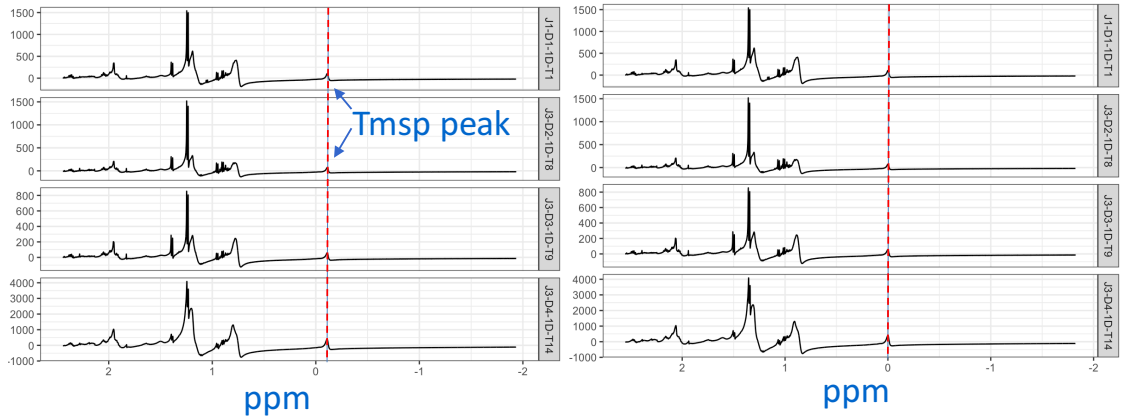
This is a mathematical transformation. No Stat is needed but it is good to know about FT as a statistician ☺…



Step 5 : ppm conversion and referencing to an internal standard

After Fourier transform, spectral coordinates are frequencies that depends on the external magnetic field .

1. Frequencies in Hertz are re-xpressed as **chemical shifts δ** in parts-per-million (ppm), a unit independent from the device.
2. All spectra are the aligned on the peak of a molecule of reference added to the sample before analysis (usually TMSP). TMSP = 0ppm
3. ppm are presented in decreasing order on the x axis.



Step 6 : Zero order phase correction

The axis along which the signal is recorded cannot be predicted

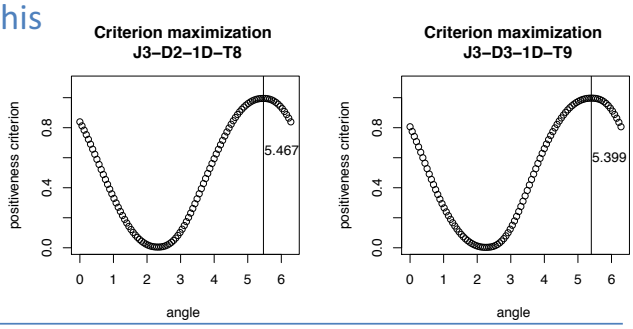
⇒ arbitrary phase φ_0 That can be expressed as $F = F^* \exp(i\phi_0)$ where F^* is the phased spectrum.

Hypothesis:

The real part of the spectrum should be in pure absorptive mode (with strictly (+) intensities)

Correction principle:

- Rotate the spectrum with a series of angles
- Measure a positiveness criterion on the real part of the spectrum
- Select the angle that maximises this criterion



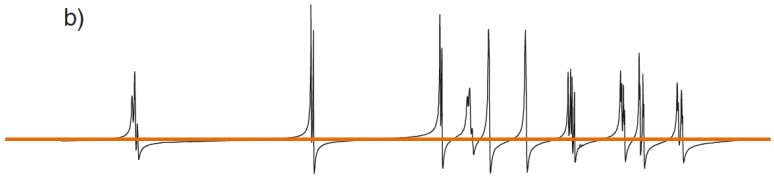
Positiveness criteria:

- rms: $\frac{\sum([positive\ intensities]^2)}{\sum([intensities]^2)}$
- max: maximum positive intensity

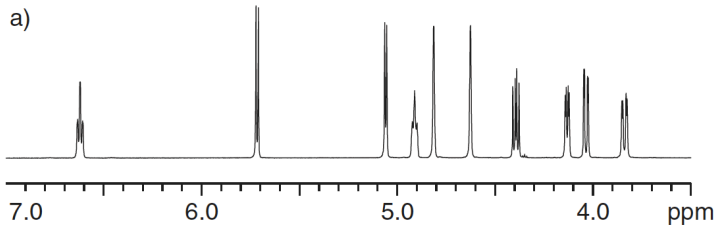
Step6 : Zero phase shift illustration

Equal phase shift across all signals (–)

Zero order phase shift



Correctly phased spectrum



Claridge, T. D. (2016). High-resolution NMR techniques in organic chemistry (Vol. 27). Elsevier.

Step 7 and 8 : Baseline correction and zeroing of neg values

In an ideal H-NMR spectra, the baseline is flat but several acquisitions and pre-processing artefacts can create fluctuation in the baseline

Principle: Estimation of a baseline Z with AsLS for each spectra and calculation of the "corrected" spectra F* = F – Z

Methods: simple line, polynomial, non parametric smoother.

PepsNMR:

Uses an asymmetric least squares smoothing algorithm (AsLS) to estimate the baseline function. (Eilers & Boelens, 2005):

Min ⌊Asymmetric Least Squares criterion⌋ + ⌊lambda*roughness⌋

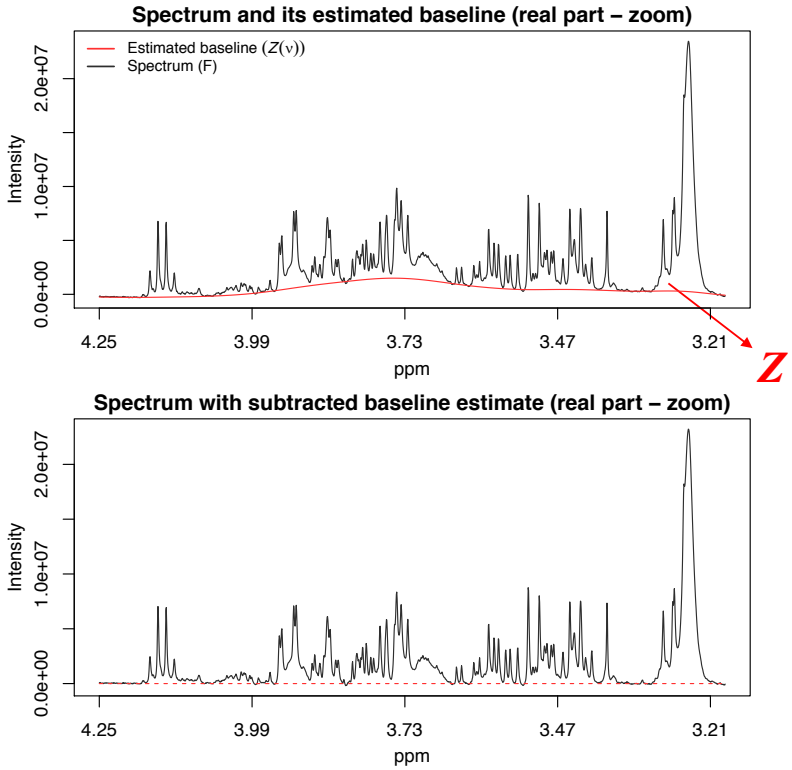
uneven weights *p* for (+) and (-) deviations of *Z* from the spectrum will favour positive corrected intensities

roughness penalty

Meta-parameters:

- **Asymmetry parameter** (**p**): = weight for (-) deviations, if $p \uparrow$ => the estimated baseline will more frequently be above the spectrum. Default value = 0.05
- **Smoothing parameter** (**lambda**): $\lambda \uparrow$ => smoother signal

Step 7 and 8: Illustration of baseline estimation and removal



After baseline suppression, remaining negative values are replaced by 0

Step 9 : Warping

Big problem in RMN

Peak misalignment due to variations in the acquisition conditions and samples properties (e.g. pH).

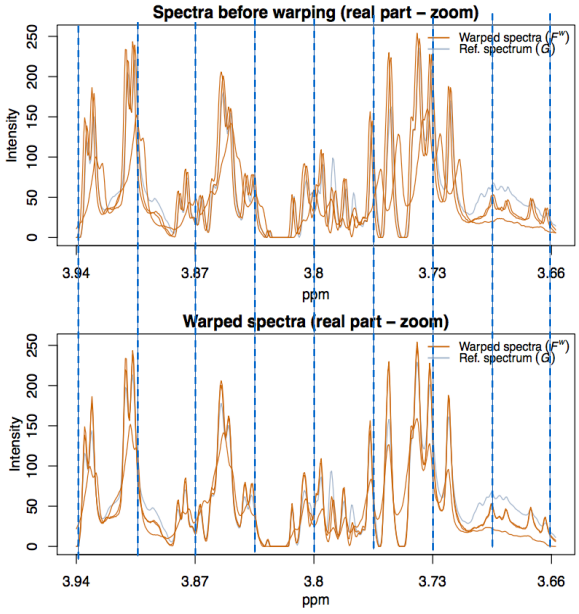
Warping

Transformation/deformation of the ppm axis in order to realign the unaligned peaks.

Method: $F^{new}(\nu) = F^{old}(w(\nu))$

Estimation of a non parametric optimal warping function $w(\nu)$ combining polynomials and B-splines based on a reference spectrum

$$w(\nu) = \underbrace{\sum_{k=0}^{K} \beta_k \nu^k}_{\text{polynomials}} + \underbrace{\sum_{l=1}^{L} \alpha_l B_l}_{\text{B-splines}}$$



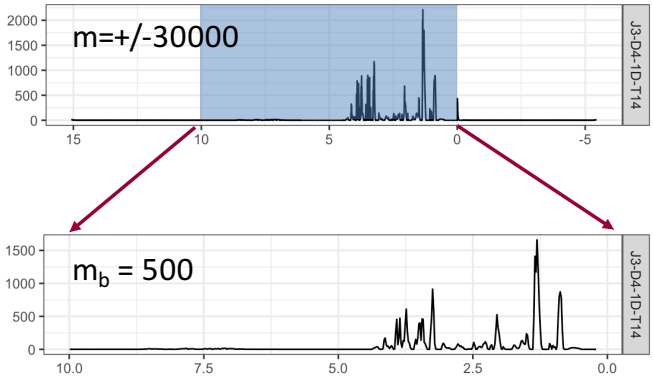
Steps 10, 11 : Window selection and Bucketing

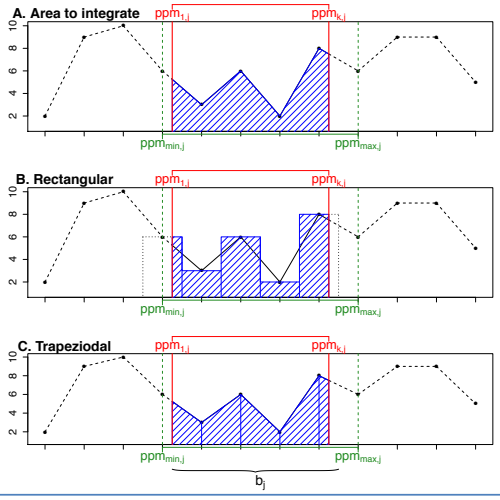
Window selection

Crop of the part of the ppm domain used for data analysis.

Bucketing

- Integration of the *m* original spectral intensities into $m_b << m$ predefined intervals to reduce the dimensionality.
- Bucketing renders data analysis easier and reduces the residual alignment problems (soft alignment).

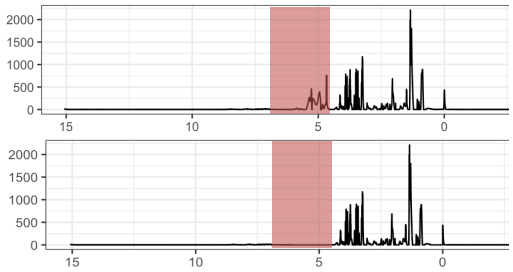




Steps 12, 13 : Region removal and region aggregation

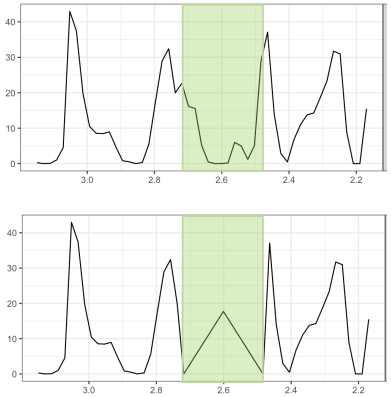
Region removal

Replaces by 0's or Na's regions of the spectra which are specially unstable and/or are not of interest in the analysis. e.g. : water area



Region aggregation

Aggregates, in one peak, zones of the spectra which have big misalignment problems that cannot be solved by warping. e.g. Citrate peak for urine spectra.



Step 14 : Normalization

Normalization definition

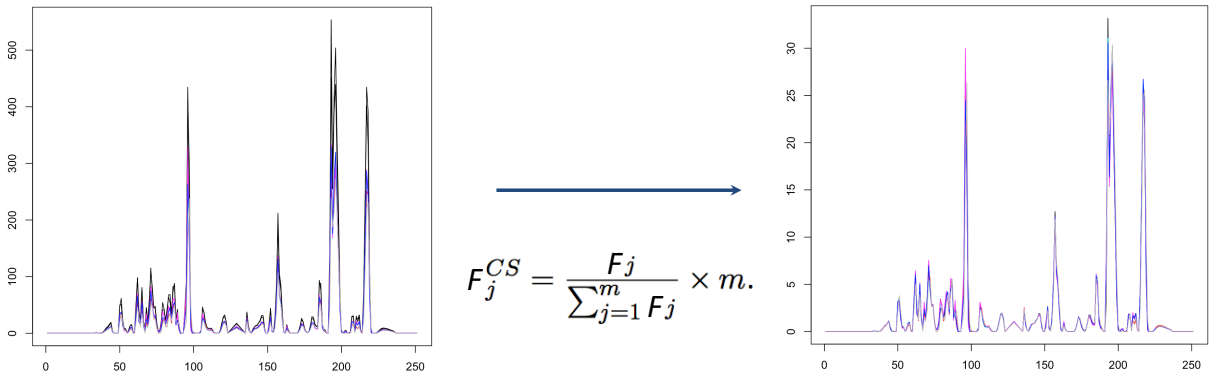
Row operation: $F_i^* = F_i \times c_i$ where each spectrum *i* is multiplied by a constant term c_i.

Goals : get data with the same scale, reduce dilution problems, …

Most used normalization methods (available in PepsNMR)

Mean (or constant sum), pqn , reference peak, median, ….

Example of the Constant Sum Normalization



PepsNMR in Galaxy NMR processing tools

Workflow4Metabolomics 3.0

https://galaxy.workflow4metabolomics.org/

" Welcome to the collaborative portal dedicated to metabolomics data processing, analysis and annotation for Metabolomics community."

Tools for NMR data in W4M: read Bruker raw data, pre-processing, alignment, and bucketing

PepsNMR in NMR processing tools:

- Toolboxes **NMR_Read** and **NMR_Preprocessing**
- Steps included: Read fids and Group Delay Correction => Negative values Zeroing

Inputs for **NMR_Read**: Bruker raw data as a zip file uploaded with "Upload File from your computer" in W4M

Inputs and outputs for the NMR_Preprocessing Toolbox

Input/output Format: *.tabular* (as outputted by NMR_Read)

with field separator "tab" and decimal separator "."

Data matrix

	A	B	C	D	E	F	G	H	I
1		0	3.57E-05	7.14E-05	0.0001071	0.0001428	0.0001785	0.0002142	0.0002499	0.0002856	0.0
2	ADG10003u_007	0+0i	0+0i	0+0i	0+0i	0+0i	0+0i	0+0i	0+0i	0+0
3	ADG10003u_009	0+0i	0+0i	0+0i	0+0i	0+0i	0+0i	0+0i	0+0i	0+0
4	ADG10003u_015	0+0i	0+0i	0+0i	0+0i	0+0i	0+0i	0+0i	0+0i	0+0
5	ADG10003u_017	0+0i	0+0i	0+0i	0+0i	0+0i	0+0i	0+0i	0+0i	0+0
6	ADG10003u_021	0+0i	0+0i	0+0i	0+0i	0+0i	0+0i	0+0i	0+0i	0+0

Sample metadata with the acquisition parameters:

	A	B	C	D	E	F	G	H	I	J
1	TD	BYTORDA	DIGMOD	DECIM	DSPFVS	SW_h	SW	O1	DT	
2	ADG10003u_007	65536	1	1	12	12	14005.60224	20.01162551	3290.5	3.57E-05
3	ADG10003u_009	65536	1	1	12	12	14005.60224	20.01162551	3290.5	3.57E-05
4	ADG10003u_015	65536	1	1	12	12	14005.60224	20.01162551	3290.5	3.57E-05
5	ADG10003u_017	65536	1	1	12	12	14005.60224	20.01162551	3290.5	3.57E-05
6	ADG10003u_021	65536	1	1	12	12	14005.60224	20.01162551	3290.5	3.57E-05

- TD: Time domain size
- BYTORDA: Determine the endianness of stored data. If 0 -> Little Endian; if 1 -> Big Endian
- DIGMOD: Digitization mode
- DECIM: Decimation rate of digital filter
- DSPFVS: DSP firmware version
- SW_h: Sweep width in Hz
- SW: Sweep width in ppm
- O1: Spectrometer frequency offset
- DT: Dwell time in microseconds

Workflow4Metabolomics

Evaluation of the pre-processing quality

Since:

- A long and complicated pretreatment does not guarantee that the spectra will be "good" at the end.
- A good statistical data analysis can not compensate a bad pre-processing.
- Spectral quality is usually based on subjective visual inspection

⇒ A methodology and criteria are then necessary to evaluate and compare the efficiency of several pretreatments methods.

Some tools are developed at our institute ISBA/UCL for this purpose and called **MIC** indices

MIC : motivation and definition

Concept : **MIC** for Metabolomic Informative Content

For ONE sample (blood, urine…), SEVERAL metabolomic spectra can

- Result from different sample preparations.
- Be produced by different measurement devices MS, H-NMR, ≠Mhz,…
- Be measured by different acquisition methods (1D, 2D, COSY, HSQC).
- Result from different data pretreatments.

Need

A methodology to quantify et compare the quality of 2 (or more) spectral matrices generated from a SAME set of bio-samples

Principle

If the goal is to classify data in 2 to k groups, the dataset should ideally

- maximize the structured group signal and
- minimize the noise and other unnecessary information

We want to quantify this by adequate indices

MIC approaches used at UCL

Data used to calculate MIC indices

- A set of *n* spectral organized in groups.
- The experiments are organized such that spectra of a group should be similar (e.g. repeated sampling on a same subject, sample repeated measures, …)

MIC tools and indices

- Inertia measures
- PCA
- Unsupervised clustering
- PLS-DA
- ASCA+ and ANOVA-PCA+ (Not seen in this course)

  MIC quantification and source of variation understanding in complex designs.

R package with the MIC tools: https://github.com/ManonMartin/MBXUCL

Case study

Human serum : 4 donors, 8 measures of each donor sample over 8 days.
Compare two pre-processing with PepsNMR (basic and full)

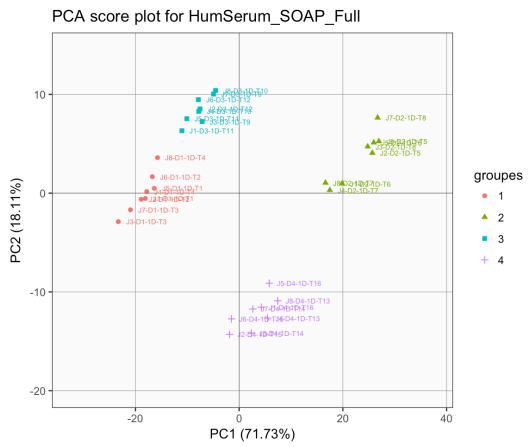
PCA

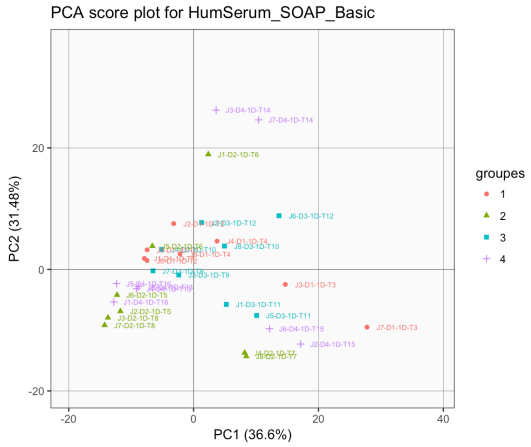
Method

Apply PCA to the data set and identify the groups with different markers or colors. We expect good group clusters in the 2 first Principal Components.

Case Study

	PC 1	PC 2	PC 3	PC 4
HumSerum_PEPS_Full	0.7173	0.8984	0.9501	0.9653
HumSerum_PEPS_Basic	0.366	0.6808	0.8124	0.9064





Inertia criteria

Method

Quantification of the percentage of the total variation of the data (total inertia) due to the variability between groups and within groups.

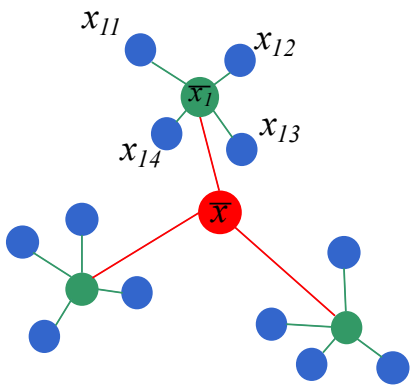
Formula

$$I_T = \frac{1}{n}\sum_{i=1}^{n} d^2(x_i, \bar{x})$$

$$= I_{\text{Within groups}} + I_{\text{Between groups}}$$

$$= \frac{1}{n}\sum_{j=1}^{g}\sum_{x_i \in G_j} d^2(x_i, \bar{x}_j) + \frac{1}{n}\sum_{j=1}^{g} n_j d^2(\bar{x}_j, \bar{x})$$

Where *d* is the euclidian distance



Case Study

	BI	WI	TI
HumSerum_ PEPS'_Full	88.01	11.99	100
HumSerum_ PEPS_Basic	20.54	79.46	100

Unsupervised clustering

Method

- Apply a non supervised clustering method to the spectra (hierarchical clustering, k-means, …)
- Use indices to quantify after classification if the clusters are
  - homogeneous (e.g. Dunn (DI) and Davies-Bouldin (DBI) indexes)
  - correspond to the true ones (e.g. Rand (RI) and Adjusted Rand (ARI) indexes)

Case Study

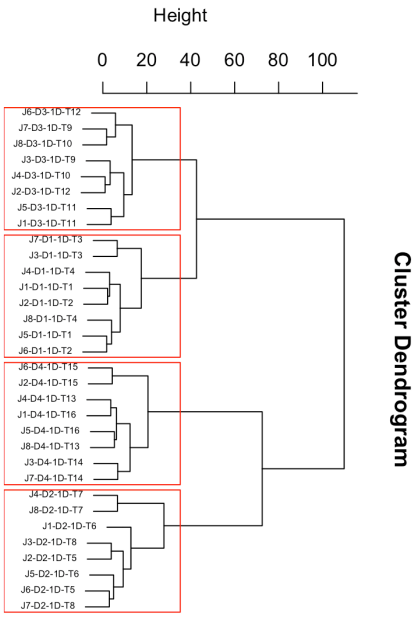
	DunnW	DunnKM	DBW	DBKM	RandW	RandKM	AdjRandW	AdjRandKM
HumSerum_ PEPS_Full	0.6406	0.6406	0.7385	0.7385	1	1	1	1
HumSerum_ PEPS_Basic	0.3055	0.3908	2.051	0.9944	0.6714	0.7077	0.09719	0.1969

Reference: B. Féraud et al., 2015

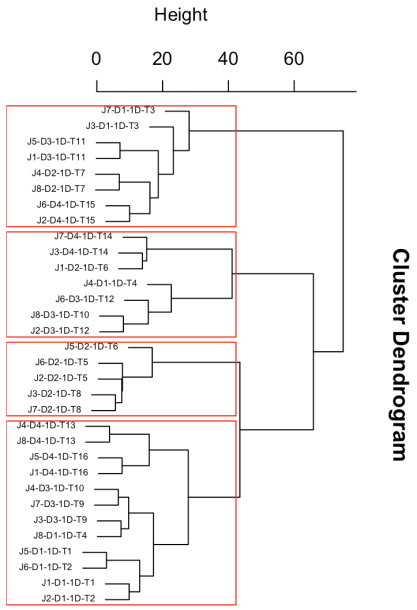
Unsupervised clustering

Example: dendograms

Full PepsNMR



Basic PepsNMR



PLS-DA

Method

Fit a PLS-DA model which predicts the group from the spectra

Compare the quality of the models in prediction and cross-validation.

- **RMSEP:**

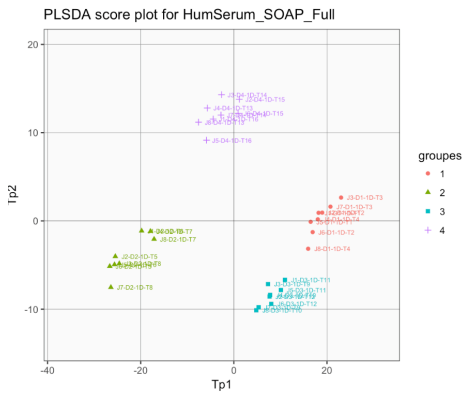
	Y1	Y2	Y3	Y4
HumSerum_PEPS_Full	0.1026	0.08399	0.1139	0.09619
HumSerum_PEPS_Basic	0.3821	0.2334	0.3288	0.31

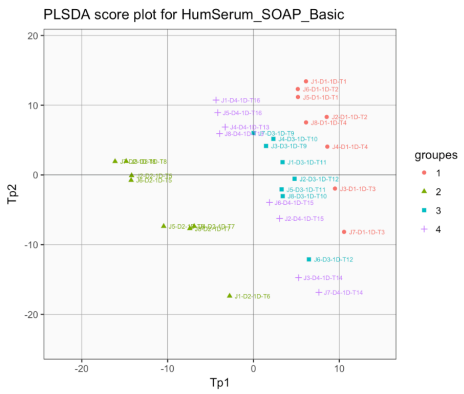
- **R2:**

	Y1	Y2	Y3	Y4
HumSerum_PEPS_Full	0.9716	0.978	0.9664	0.9694
HumSerum_PEPS_Basic	0.4906	0.8036	0.579	0.6976

- **Q2:**

	Y1	Y2	Y3	Y4
HumSerum_PEPS_Full	0.9438	0.9624	0.9308	0.9507
HumSerum_PEPS_Basic	0.2215	0.7094	0.4235	0.4875





Bibliographie

- Vanwinsberghe J. (2005),Bubble: development of a matlab tool for automated 1H-NMR data processing in metabonomics. Master's thesis, Strasbourg University, France.
- Rousseau R. (2011), Statistical contribution to the analysis of metabonomics data in 1H NMR spectroscopy. PhD thesis, Institut de statistique, biostatistique et sciences actuarielles, Universit ́e catholique de Louvain, Belgium.
- Martin M. et al. (2018), "PepsNMR for 1H NMR metabolomic data pre-processing," Anal. Chim. Acta, vol. 1019, pp. 1–13.
- Eilers, P. H. (2003). A perfect smoother. Analytical chemistry, 75(14), 3631-3636.
- Eilers P.H., Boelens H.F. (2005), Baseline Correction with Asymmetric Least Squares Smoothing, Medical Centre Report (unpublished results), Leiden University.
- Féraud, B.; Govaerts, B.; Verleysen, M.; de Tullio, P. (2015) Statistical treatment of 2D NMR COSY spectra in metabolomics: data preparation, clustering-based evaluation of the Metabolomic Informative Content and comparison with 1H-NMR. In: Metabolomics, Vol. 11, no.6, p. 1756-1768. doi:10.1007/s11306-015-0830-7