La chimiométrie aux frontières du domaine proche infrarouge : explorer sans a priori les données spectroscopiques Raman, moyen infrarouge et térahertz.

Ludovic Duponchel

Laboratoire de Spectrochimie Infrarouge et Raman (UMR CNRS 8516)
Université Lille 1, 59655 Villeneuve d'Ascq, France.
ludovic.duponchel@univ-lille1.fr

HELIOSPIR 2014

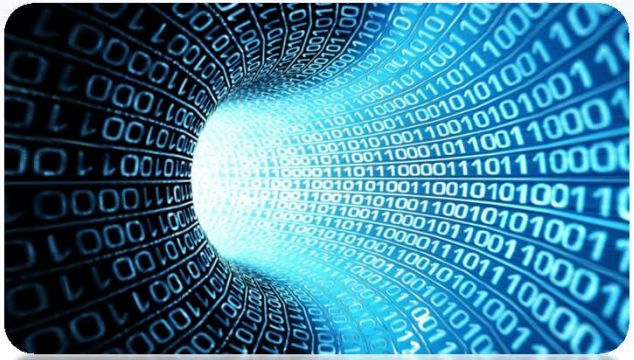
Université de Lille

cnrs

LASIR

We need tools !

Modern spectroscopic instrumentation: generation of an always larger amounts of data fairly automatically in a very short period of time.



The complexity of the data is always higher but what we believe to be an obstacle is a real opportunity to solve the given analytical problem.

⇨ development of chemometrics tools/methods to make sense of such data set and extract the maximum amount of useful information.

LASIR

What we can do with chemometrics ?

- Data pre-processing : filtering, denoising, baseline correction, signal amplification, artifact correction, data conditionning ...
- Design of Experiments (DOE) : a better understanding of the effects that different parameters have on a response (Screening, Saving time, Quantitative modelling, Optimisation)
- Exploratory Data Analysis : looking at relationships between samples / variables. Find structure, find outliers.
- Classification : check for groupings = unsupervised, existing groupings = supervised, visualize groupings, classify.
- Regression and multivariate calibration : relate instrumental measurements to characteristics of the sample of interest.
- Multivariate curve resolution : extract valuable profiles of pure species from multivariate dataset
- ...

LASIR

Outline

The potential of Multivariate Curve Resolution methods

- Probing the hydration shell of molecules with Terahertz Spectroscopy
- Characterizing a single cancer cell with synchrotron infrared microspectroscopy
- In situ monitoring of polymorphic transformations with multiset image analysis for hyperspectral Raman signal unmixing

LASIR

The potential of Multivariate Curve Resolution methods

Multivariate curve resolution has become one of the most important chemometrics method mainly due to its great potential for data analysis and its adaptability.

Objective : “Recovery of the response profile of pure components in an unresolved and unknown mixture obtained from evolutionary processes”.

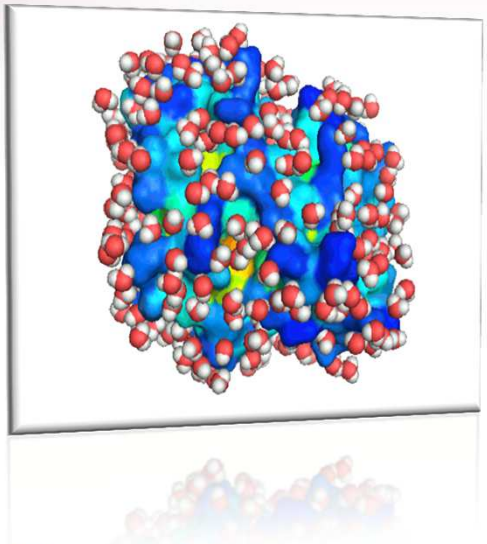
Roma Tauler

Response profile : spectra, pH profiles, time profiles, elution profiles …

(*) R. Tauler, *Chemomet. Intel. Lab. Systems*, 1995, 30, 133-146.

LASIR

The importance of hydration



Understanding the hydration mechanisms is crucial

→ Involved in many biological phenomena

Examples:

- Solvated-proteins dynamics is highly correlated to the solvent dynamics.
- Protein hydration plays an important role for the protein function expression

LASIR

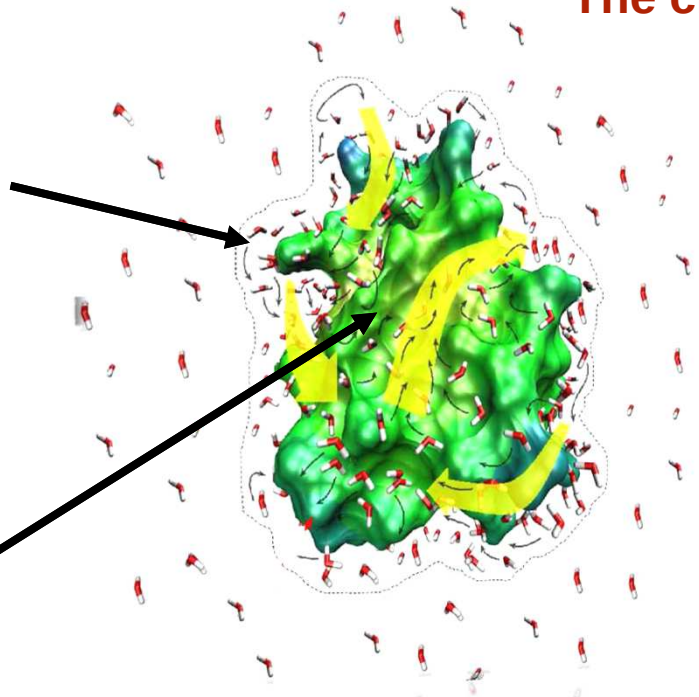
The classical hydration model

Non-charged molecule solvated in water : hydration is made through weak bonds such as hydrogen ones between the molecule and water molecules.

Consequence: the behavior of water molecules at the vicinity of the solvated molecule is changed.

The classical hydration model

« Bound water »
Forms the hydration shell of the molecule



« Bulk water »
far from the solvated molecule
(Same behavior as pure water).

Solvated molecule

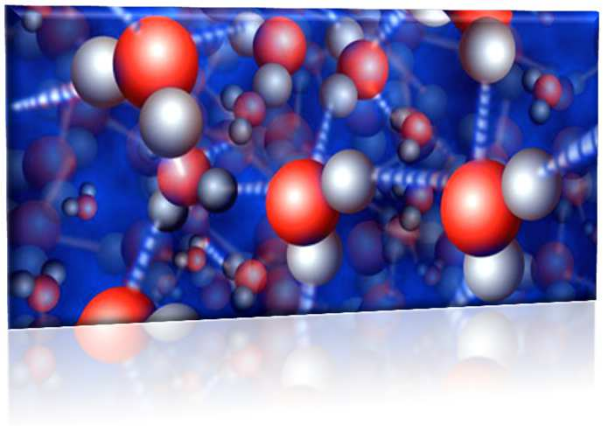
B. Born, M. Havenith (2009) J Infrared Milli Terahz Waves **30** 1245–1254

LASIR

Directly probing hydration shell of molecules

Observation: many studies on this topic but many questions remain (hydration-shell extent, hydration-process time scales or its thermodynamic-parameter dependence).

Main idea: probe directly the hydrogen bond network in water with Terahertz spectroscopy (100GHz-10THz) and obtain additional knowledge on hydration shell structure.



Overall objective: development of a global methodology (spectroscopic acquisition and chemometrics) in order to probe hydration shell of molecules.

LASIR

NIR Neighbors



THz spectroscopy

Raman spectroscopy

NIR spectroscopy Overtones and combinations of fundamental modes	**MIR spectroscopy** Fundamental modes	**Far spectroscopy** Fundamental modes

1 mm
10 cm^{-1}
0.3 THz

800 nm
0.8 µm
12500 cm^{-1}

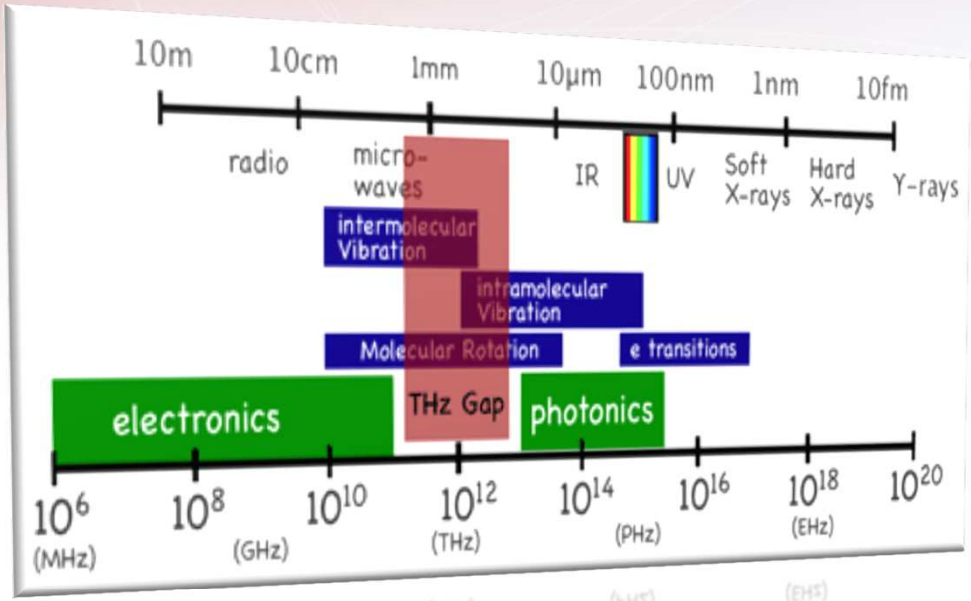
2500 nm
2.5 µm
4000 cm^{-1}

25 µm
400 cm^{-1}

100 µm
100 cm^{-1}
3 THz

LASIR

The Terahertz spectral domain



The Terahertz (Thz) spectral domain (teraherzt gap): a good way to probe low energy bonds such as hydrogen ones.

Problem : THz measurements on liquid samples are not easy due to huge water absorption.

To perform transmission measurements : powerful sources have to be used (with the risk of sample heating) or sample volumes have to be reduced and well-controlled.

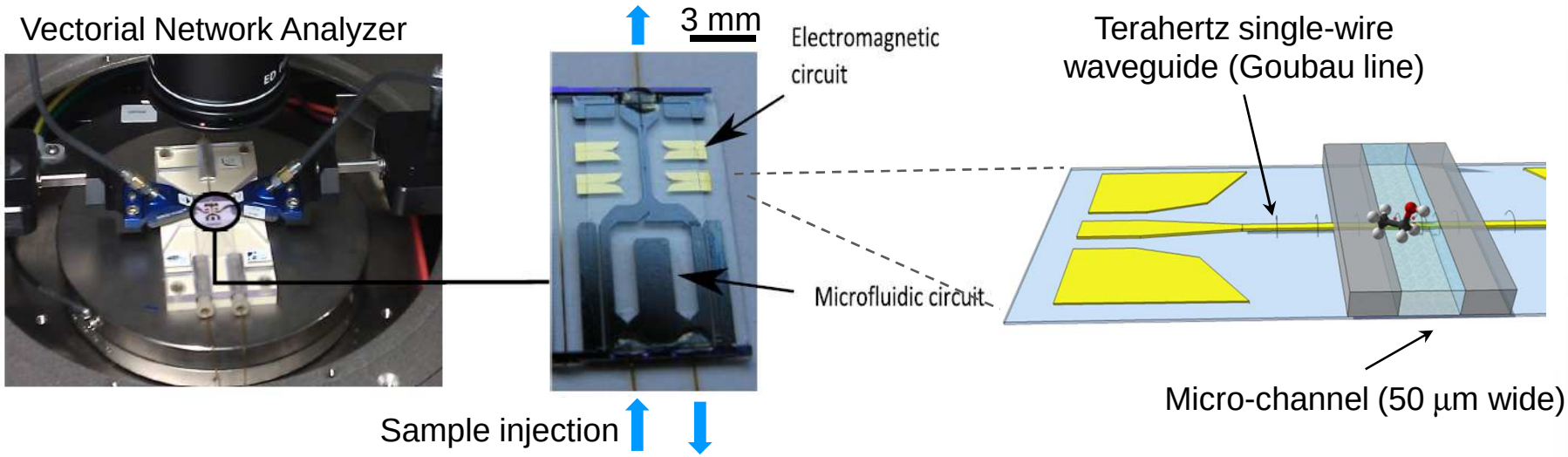
Our solution: development of an original microfluidic THz sensor to characterize hydration of molecules in a chip.

LASIR

Thz spectral acquisition in a microfluidic system

Liquid-absorption measurements: development of a two-function microsystem

- a microfluidic circuit drives the samples towards the analysis area.
- an integrated electromagnetic element (Goubau line) guides the THz waves through the sample to analyze.



Vectorial Network Analyzer connected to the wave-guide: measurement of amplitude and phase of both reflected and transmitted signals through the sample ⟹ calculation of an absorbance spectrum on the whole spectral domain (like in the mid infrared spectral range).

S. Laurette, A. Treizebre and B. Bocquet, *Journal of Micromechanics and Microengineering*, 2011, 21, 065029.

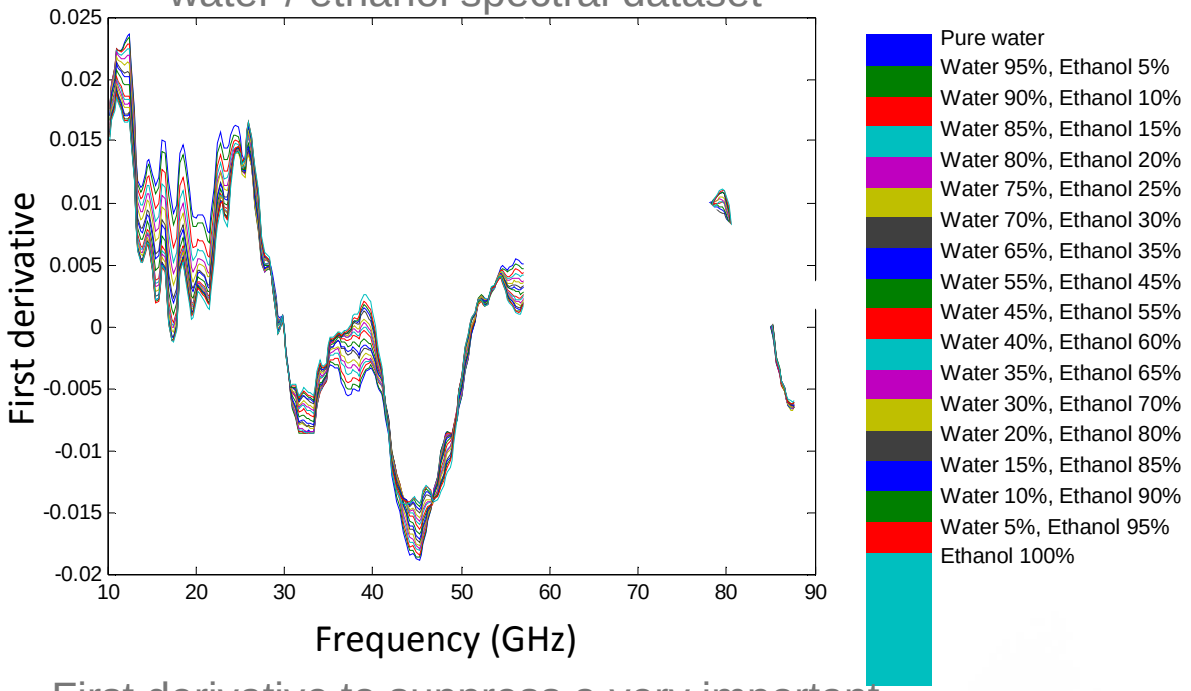
LASIR

Probing hydration shell of ethanol

Feasibility study of the concept: analyze of the ethanol / water system (thought to be a simple one).

Preparation of aqueous solutions of ethanol (alcohol volume ratio from 0 to 1)

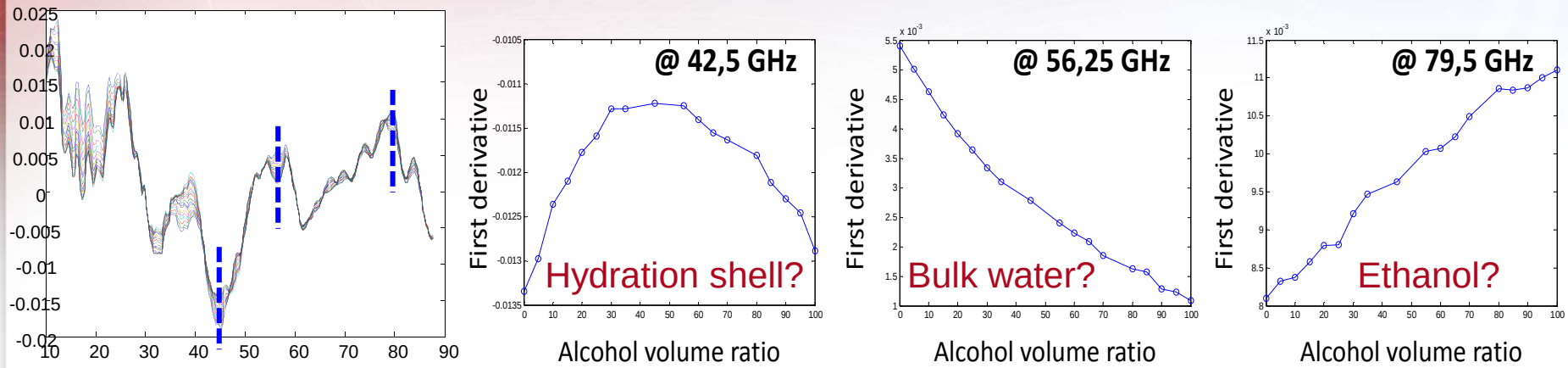
A Savitsky-Golay first derivative applied to the water / ethanol spectral dataset



First derivative to suppress a very important baseline shift completely masking the small spectral features.

LASIR

Using univariate spectral data observation



Possible to observe various signal shapes. Can be even reassuring compared to the classical hydration model.

Univariate observations can't give us an idea about the real number of contributions present in the chemical system. No guarantee that the used frequency is effectively specific to only one compound in the system (responses are often the addition of several contributions).

Univariate spectral data analysis : a partial vision of the chemical system and very often a biased one.

Our goal: use Multivariate Curve Resolution in order to retrieve information about the hydration shell in the ethanol / water system using the whole spectral domain with no *a priori*.

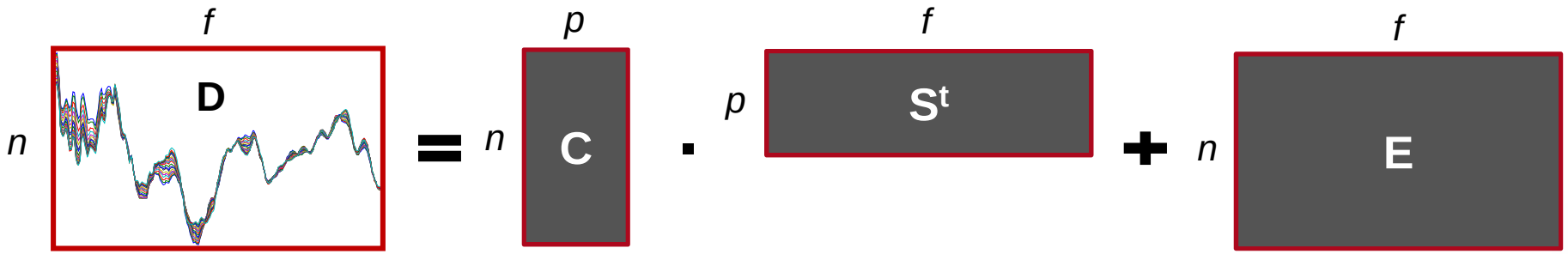
LASIR

Applying multivariate curve resolution method

The MCR-ALS(*) method:

- A model-free approach
- Unravels the pure contributions of all species in the spectroscopic dataset **D** without requiring chemical information about the underlying physicochemical system.

Constrained bilinear decomposition of the total instrumental response of a system **D** (*n*×*f*) into the product of two simpler matrices **C** (*n*×*p*) and **S**^{t} (*p*×*f*) i.e. the concentration profiles and corresponding pure spectra



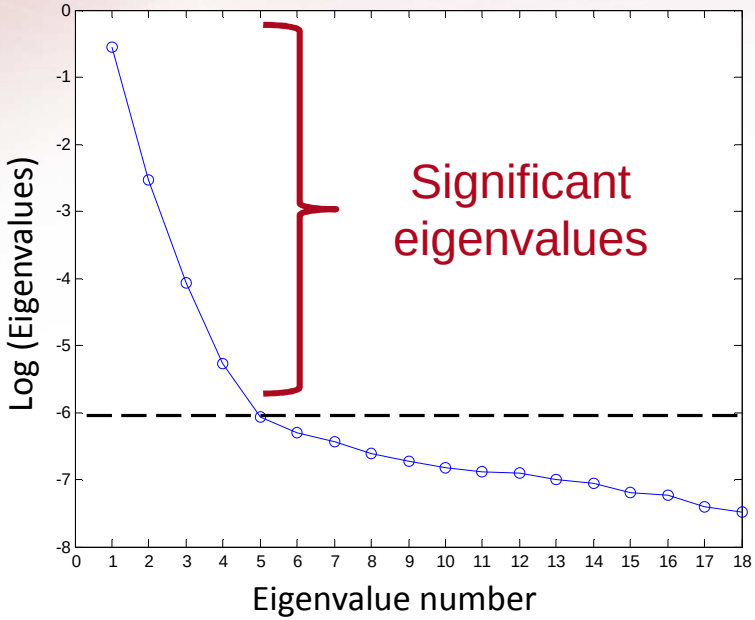
D(*n x f*) : Spectral data matrix, *n* = number of ethanol / water mixtures, *f* = frequencies in the spectral domain.
C(*n x p*) : Matrix of pure concentration profiles, *n* = number of ethanol / water mixtures, *p* = mathematical rank of D matrix = number of pure contributions in the chemical system.
S^t(*p x f*) : Matrix of pure spectral profiles.
E(*n x f*) : Residual matrix (unmodelled variations)

(*) R. Tauler, *Chemomet. Intel. Lab. Systems*, 1995, 30, 133-146.

LASIR

Evaluate the number of pure contributions

Singular Value Decomposition (SVD) of **D**



First step of the MCR-ALS methodology :

Evaluate *p* the mathematical rank of matrix **D** i.e. the number of pure contributions in the system (not an easy task even for a mathematician ☺).

Solution: try to separate significant singular values corresponding to real spectral contributions from non-significant ones corresponding to noise.
Due to a change of slope on the singular value curve, it is possible to estimate a threshold (dotted line) above which the contributions are considered significant.

As an initial assessment, we can say that **four species** are present in our **D** matrix. This first result is very important since it indicates that our microsystem is able to see more than water and ethanol.

LASIR

The MCR-ALS algorithm (some linear algebra)

Initial estimates of pure spectral profiles

S^t_{ini}

$$n\,D\,(f) = n\,C\,(p) \times p\,S^t\,(f) + n\,E\,(f)$$

Constrains: a way to decrease the total number of feasable solutions for the extracted profiles (non-negativity, unimodality, closure …)

$$D = C_{est} \cdot S^t_{ini}$$
$$C_{est} = D \cdot S_{ini} \cdot (S^t_{ini} \cdot S_{ini})^{-1}$$

C_{est}

C_{cstr}

$$D = C_{cstr} \cdot S^t_{est}$$
$$S^t_{est} = (C^t_{cstr} \cdot C_{cstr})^{-1} \cdot C^t_{cstr} \cdot D$$

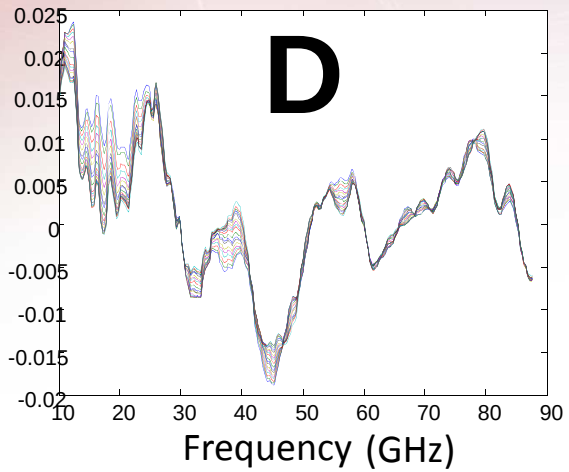
S^t_{est}

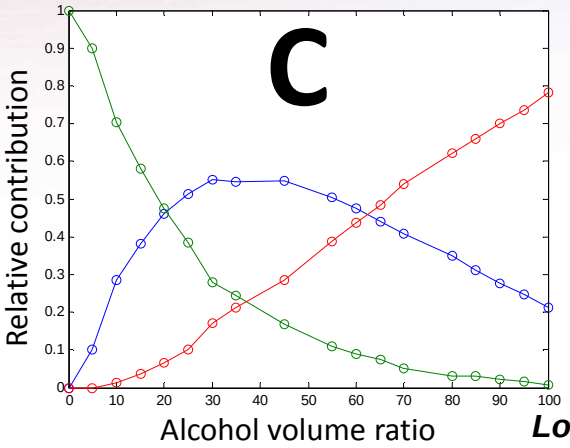
S^t_{cstr}

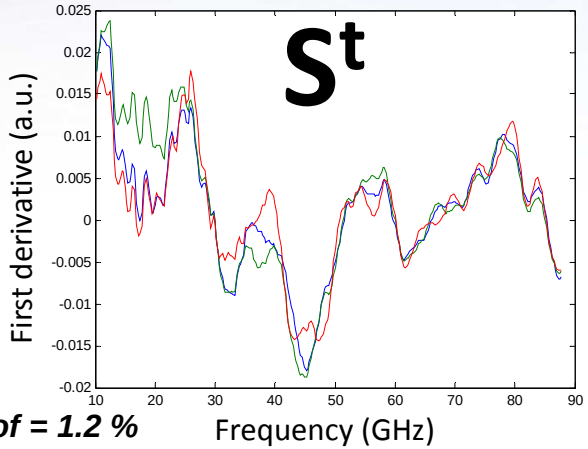
Loop until convergence

LASIR

MCR-ALS Results (3 contributions)

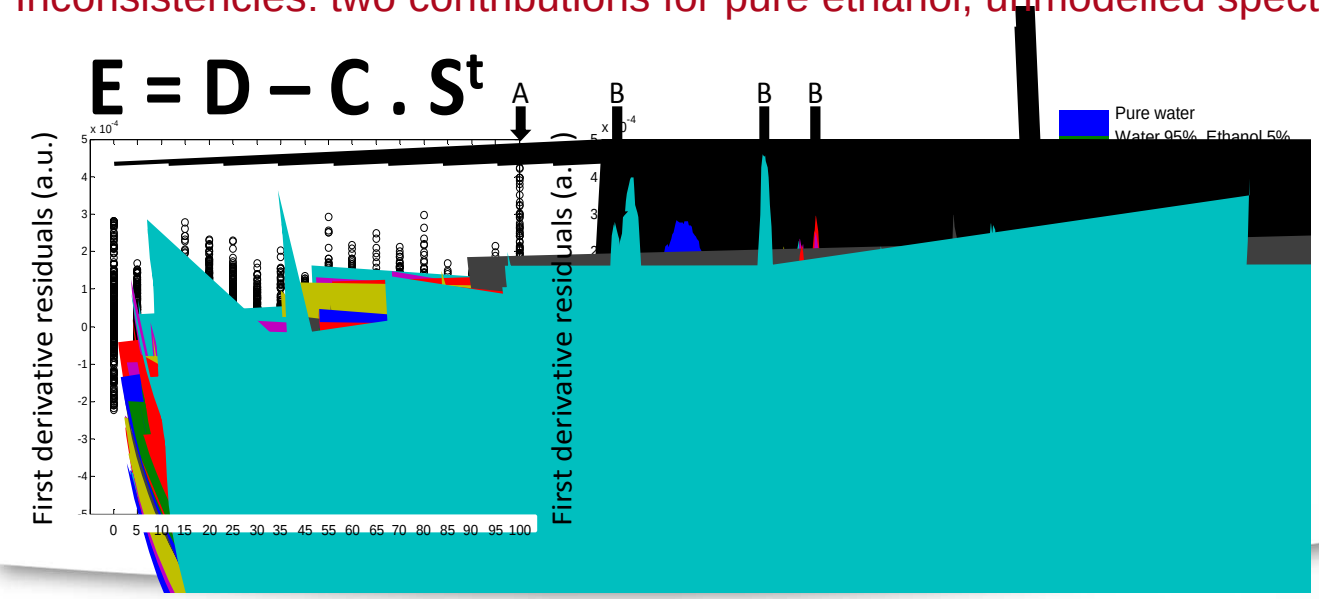




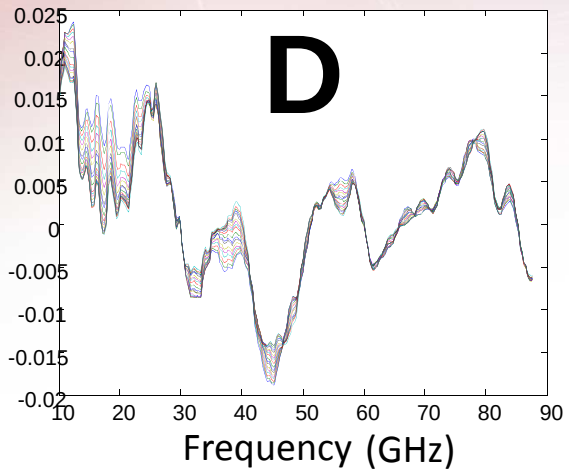


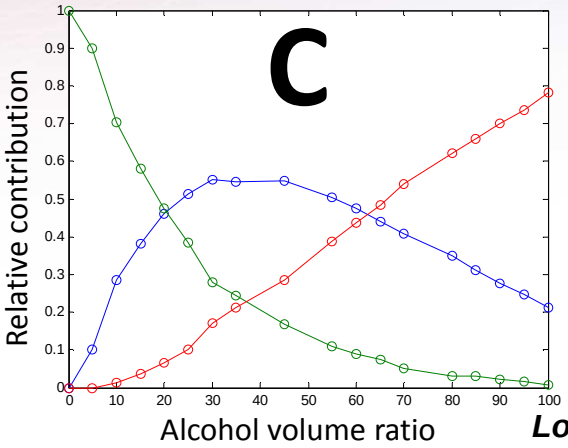
Lof = 1.2 %

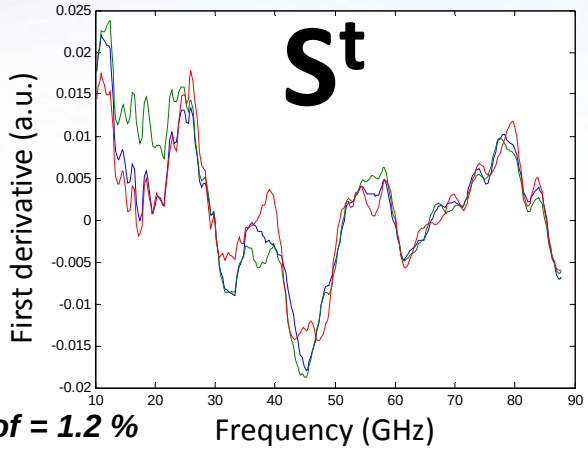
Observations: green curve may correspond to « bulk water » and the red one to ethanol.
First assumption: blue curve can be the hydration shell contribution.
Inconsistencies: two contributions for pure ethanol, unmodelled spectral variations remain on E



MCR-ALS Results (3 contributions)

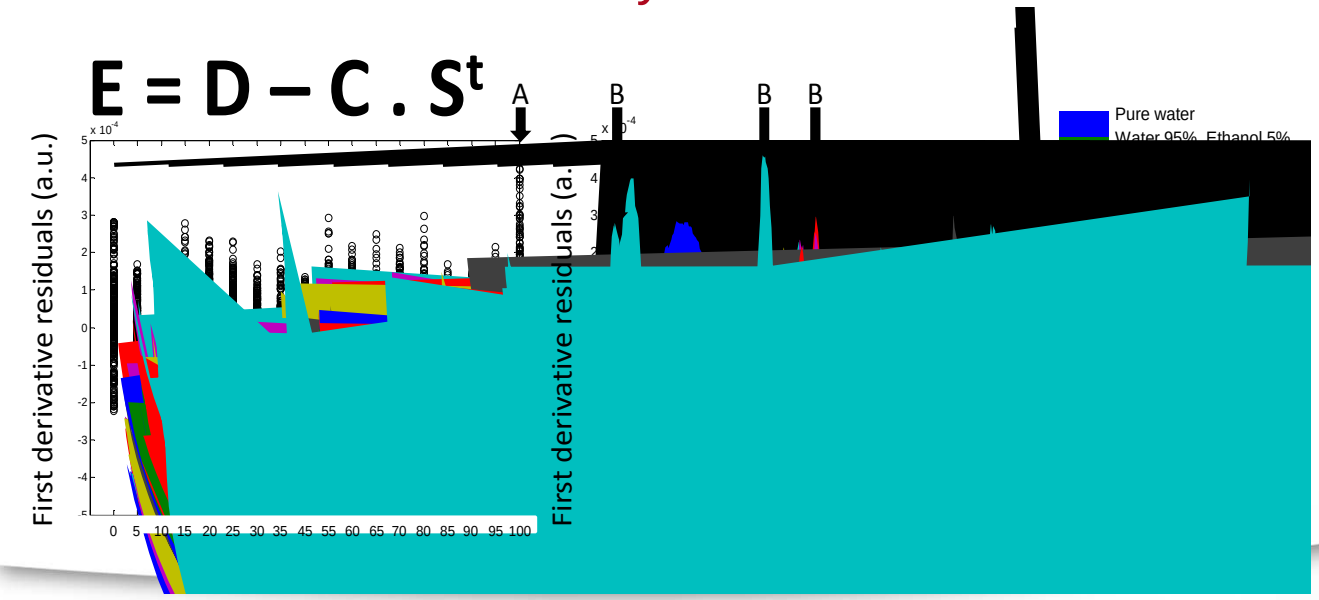




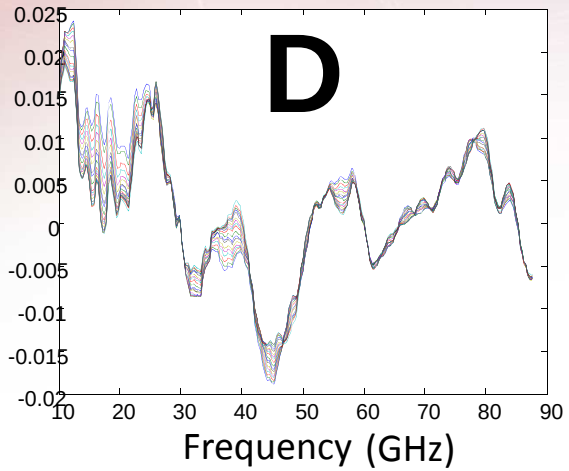


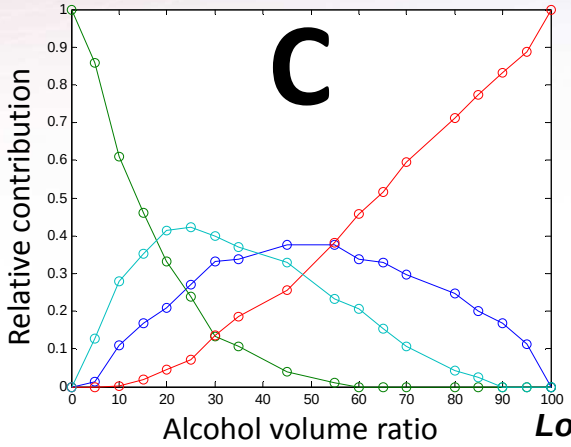
Lof = 1.2 %

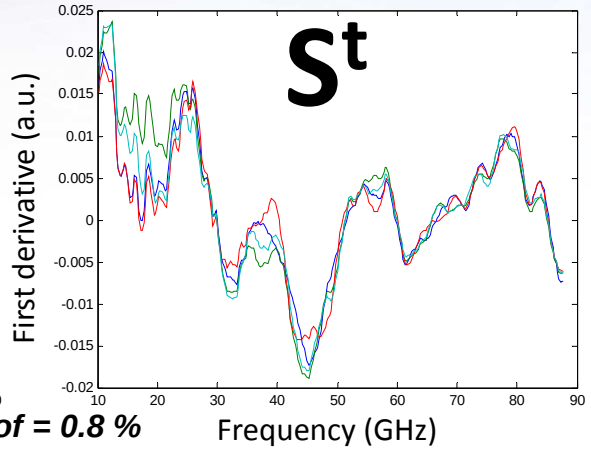
Impossible to express all spectral variations with only 3 contributions.
The well established hydration model seems too simple.



MCR-ALS Results (4 contributions)





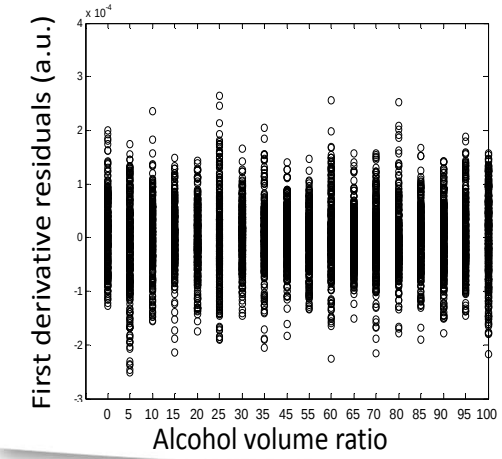


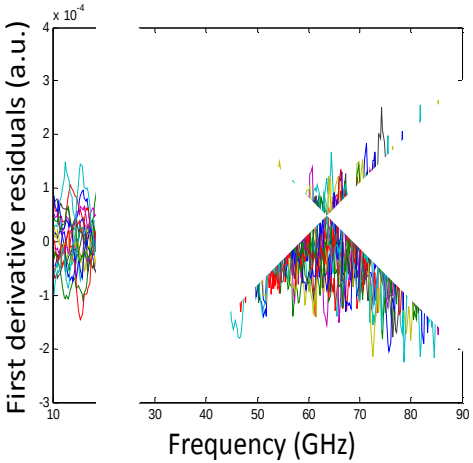
Lof = 0.8 %

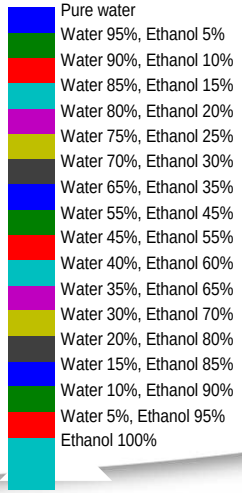
Observations:

- More consistent results with four contributions (also given by SVD).
- Two species (in light and dark blue) appear to be the ”non-bulk” water or the hydration shell.

$$E = D - C \cdot S^t$$

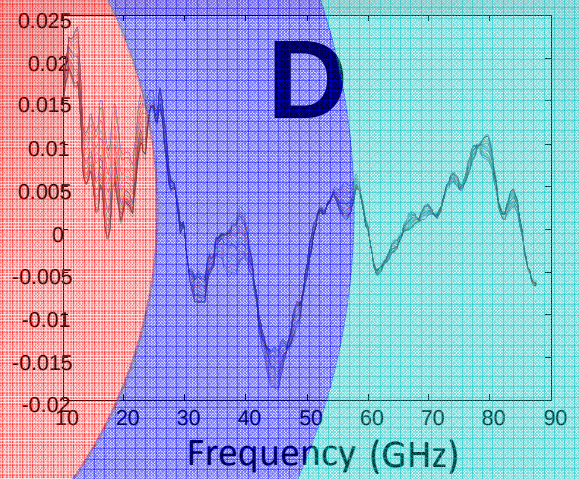


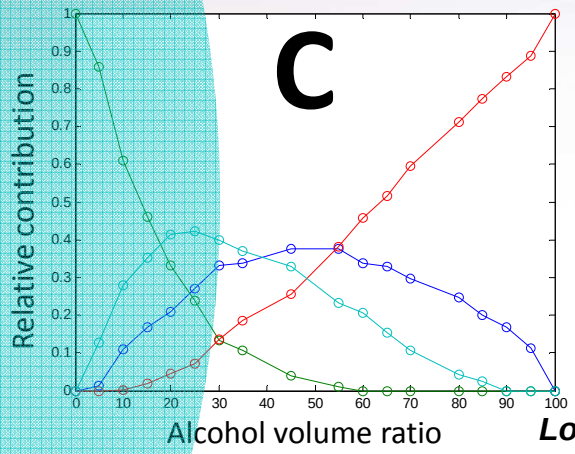


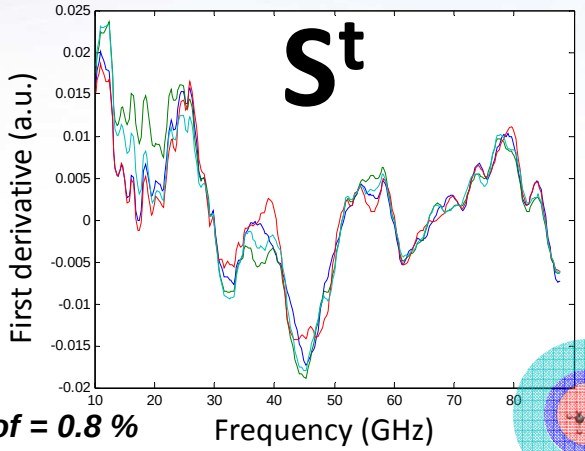


L. Duponchel *et al.*, *Chemometrics and Intelligent Laboratory Systems* 123(5), p28 (2013).

MCR-ALS Results (4 contributions)





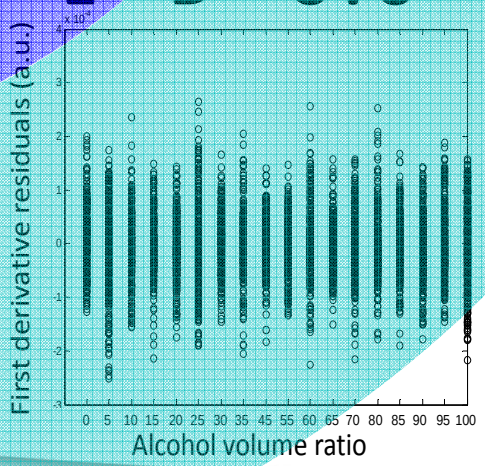


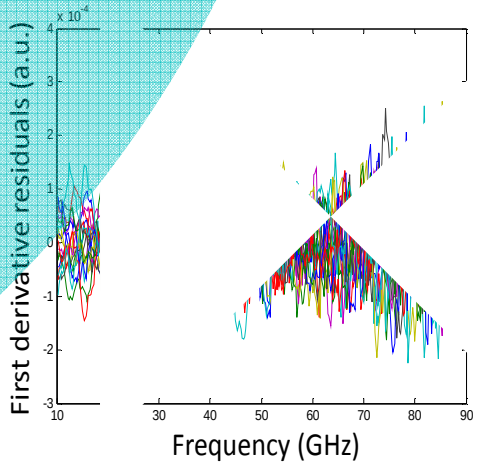
Lof = 0.8 %

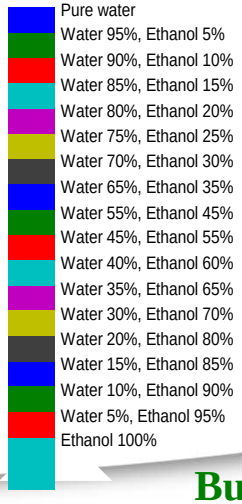
Our hypothesis: the two hydration-water contributions is the result of a two-layer hydration-shell structure:

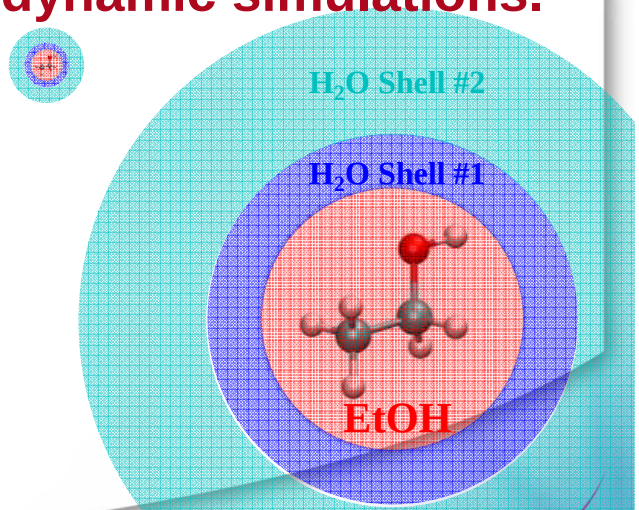
These results were validated with molecular dynamic simulations.

$$E = D - C \cdot S^t$$









LASIR

Multivariate Curve Resolution in imaging spectroscopy

Imaging Spectroscopy

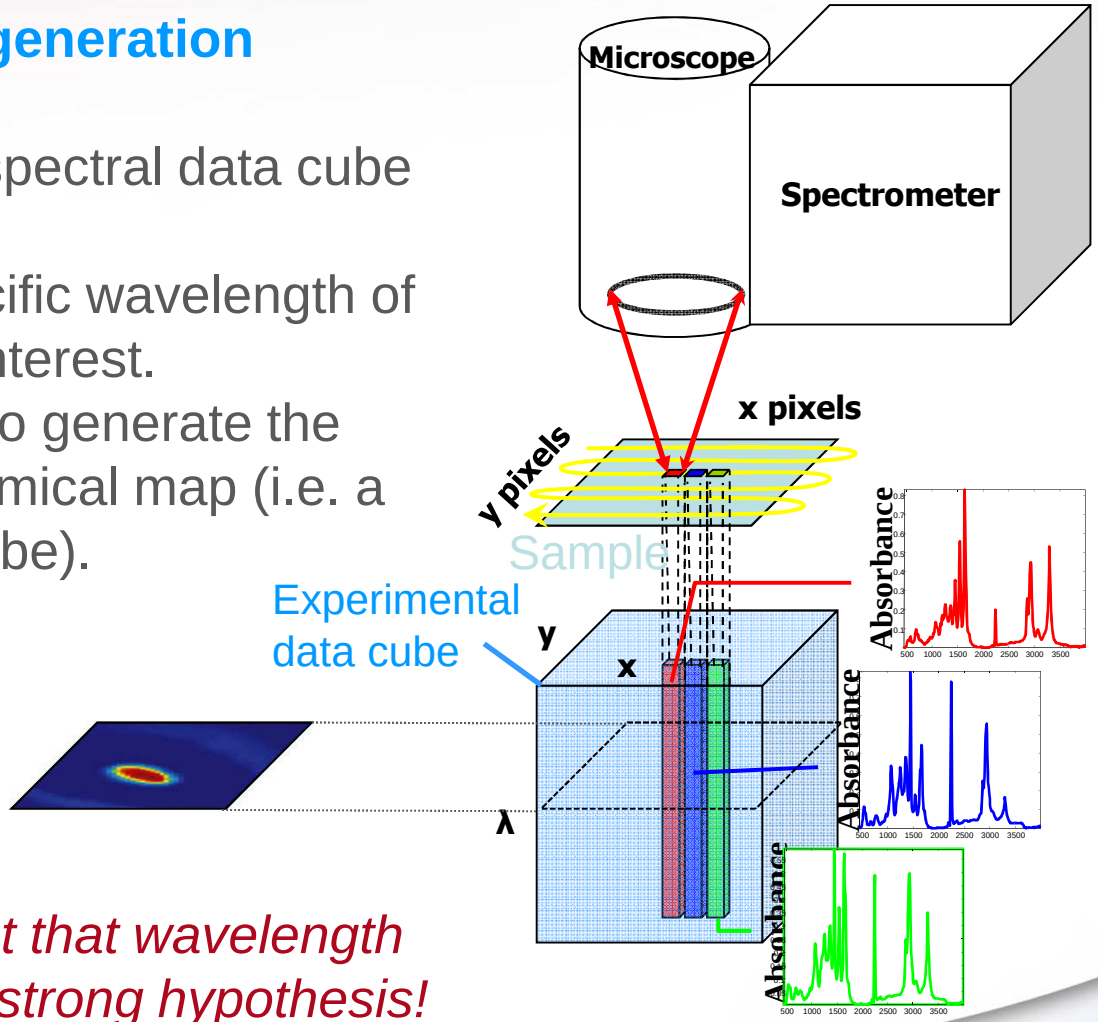
- A key field of analytical chemistry.
- Powerful tool for characterizing the molecular distribution of different chemical compounds in heterogeneous materials.
- Development of many chemometrics methods in order to extract more hidden information from such experiments.

LASIR

Multivariate Curve Resolution in imaging spectroscopy

The classical method for concentration maps generation

1) Acquisition of the spectral data cube (ex: mapping).
2) Selection of a specific wavelength of the compound of interest.
3) Signal integration to generate the corresponding chemical map (i.e. a slice of the data cube).



Often we forget that wavelength specificity is a strong hypothesis!

Multivariate Curve Resolution in imaging spectroscopy

Wavelength specificity is a strong hypothesis

Some drawbacks:

• Usually necessary to know all compounds present in the analyzed sample.

➢If a compound is forgotten = possible interferences = overestimations of concentration = false distribution maps = biased vision of analytical reality.

➢impossible to select a specific spectral range for an unexpected compound (no image).

• Sometimes impossible to find a specific spectral range for each compound due to the complexity of the sample (high number of species) and/or the high bandwidth of the considered spectroscopy.

What can be done with Multivariate Curve Resolution ?

LASIR

Multivariate Curve Resolution in imaging spectroscopy

MCR-ALS Methodology

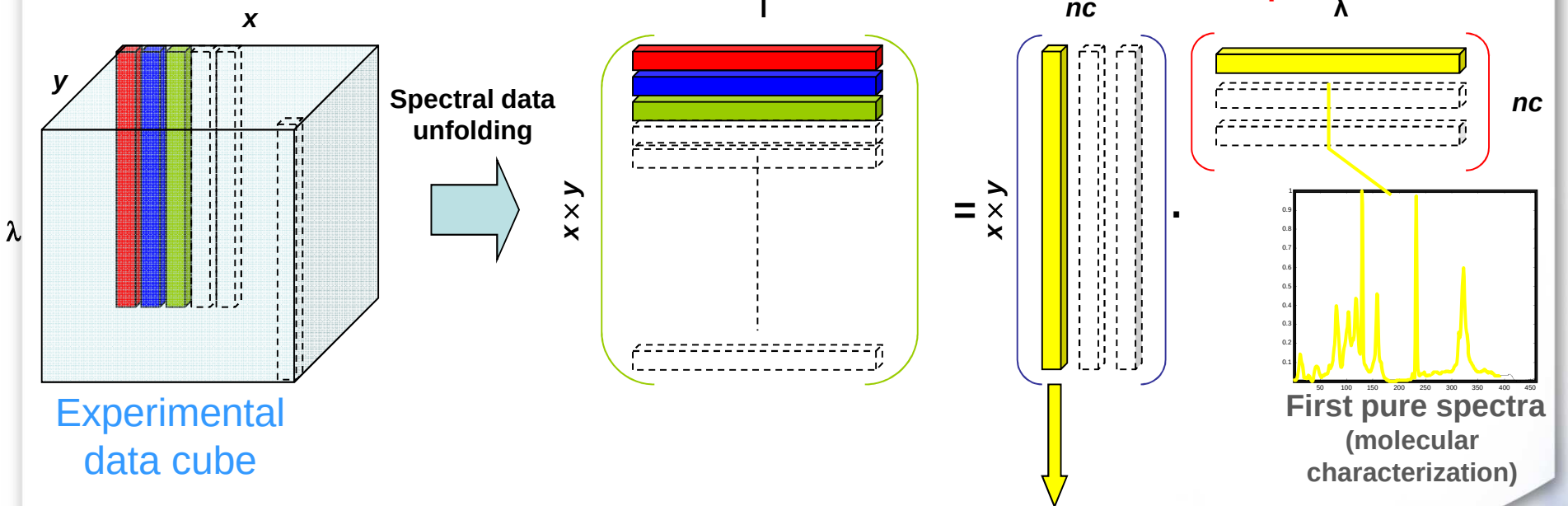
- Unfolding of the experimental data cube.
- Rank evaluation of the **D** matrix (*nc*: number of spectral contributions in the dataset).
- Simultaneous extraction of **C** and **St**
- Refolding **C** in order to generate chemical maps.

Multivariate curve resolution (bilinear decomposition)
Simultaneous extraction with no prior knowledge

D
Spectral data matrix

C
Pure compounds concentration matrix

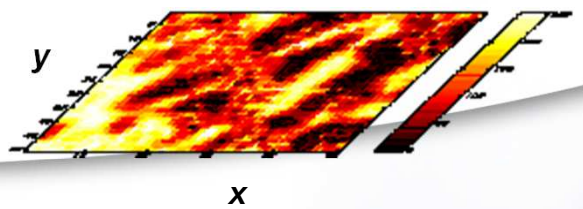
S^T
Pure compounds Spectral matrix



Experimental data cube

First pure spectra
(molecular characterization)

Refolding of the first pure compound concentration distribution (pixels space)



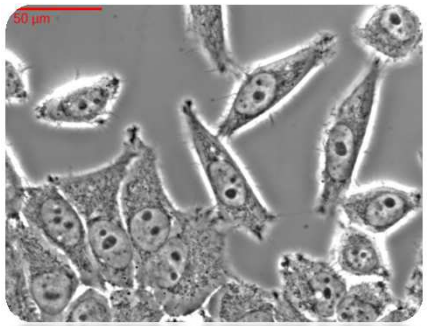
LASIR

MIR Imaging on a single HeLa cancer cell

Why do we need Infrared Synchrotron radiation ?

- **1000** times brighter than our IR spectrometer sources : highest signal to noise ratio -> possible to observe small absorbance variations.

- Possible to obtain the highest spatial resolution : necessary if we want to observe details for very small sample.



Is it really enough?

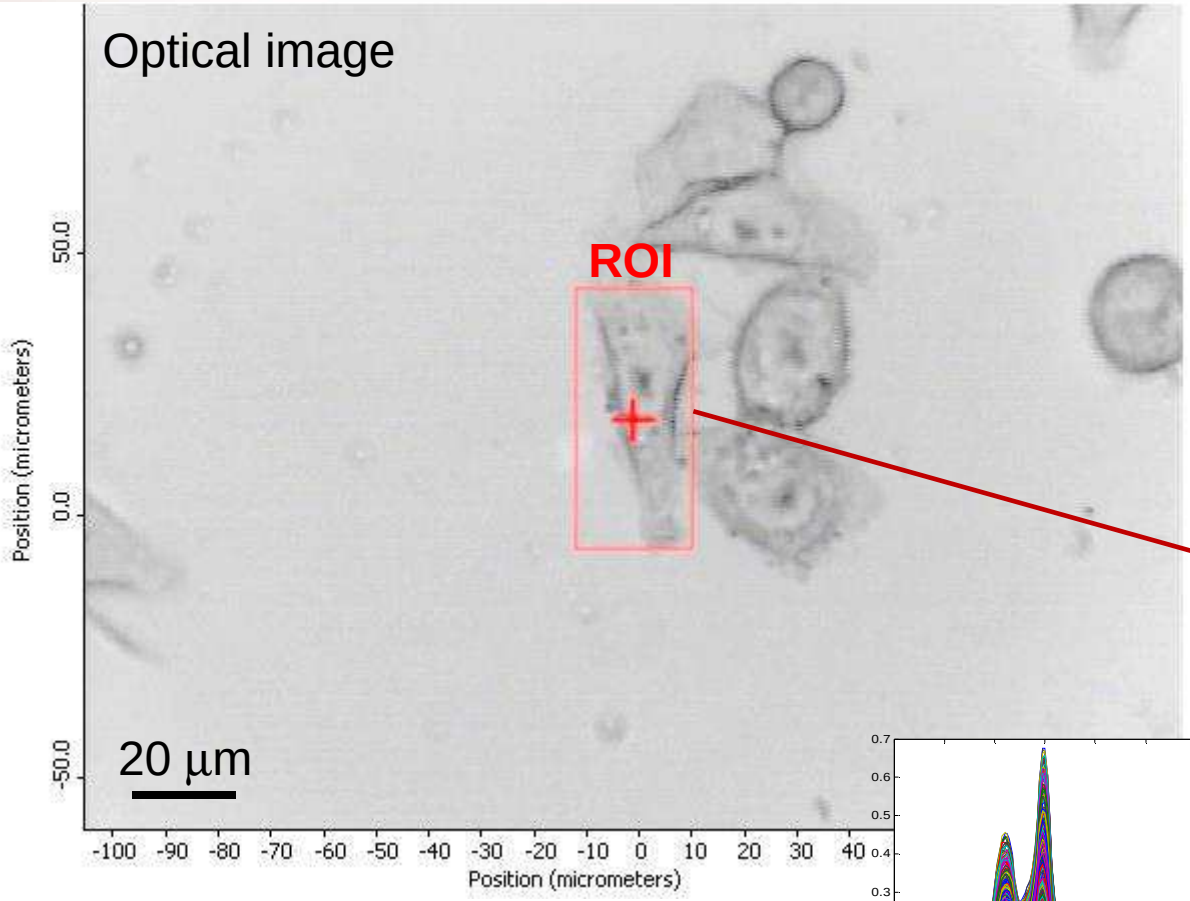
Rayleigh criterion: $R \sim \lambda$
IR spectral range : $2.5\ \mu m < \lambda < 25\ \mu m$
Cell dimensions : few tens micron

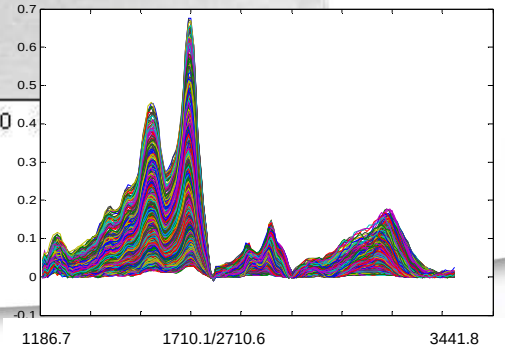
What can be done with chemometrics (MCR-ALS and Super-resolution)

LASIR

A first spectral exploration

Characterizing a single HeLa cell with IR Synchrotron micro-spectrometry

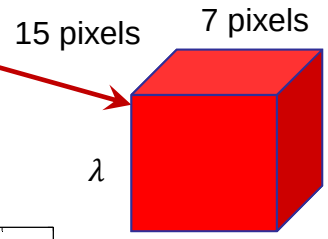




Region of interest: 52.5 µm x 24.5 µm

Pixel size: 3.5 µm x 3.5 µm

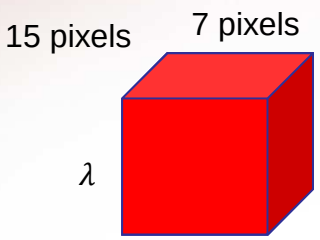
Spectral range :
1000-4000 cm^{-1}
4 cm^{-1} spectral resolution



105 spectra in the spectral data cube

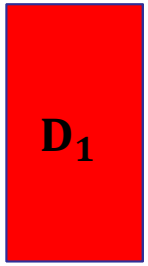
A first spectral exploration

Chemical map extraction with MCR-ALS

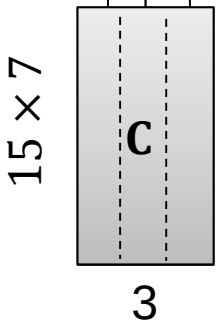


105 spectra in the spectral data cube

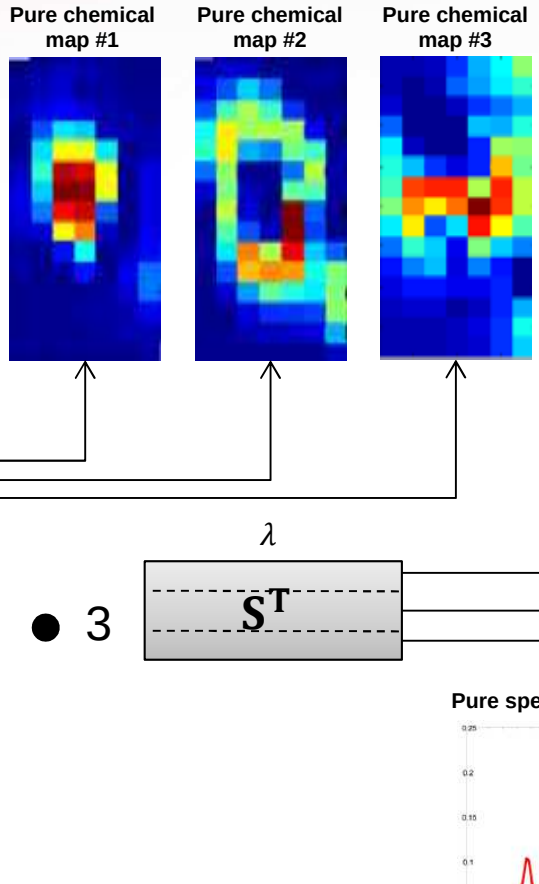
Unfolding

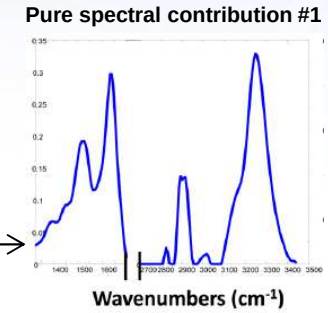


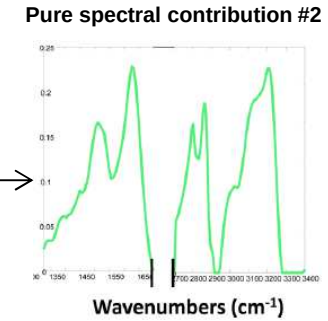
=

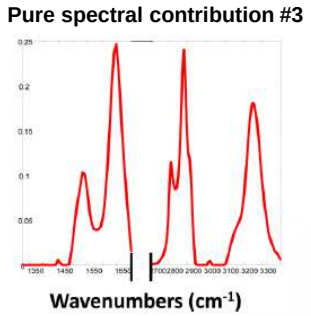


•









It was not really possible to find a specific wavenumber!

A possible molecular identification but a « poor spatial resolution ». Is there any thing we can do concerning spatial resolution?

LASIR

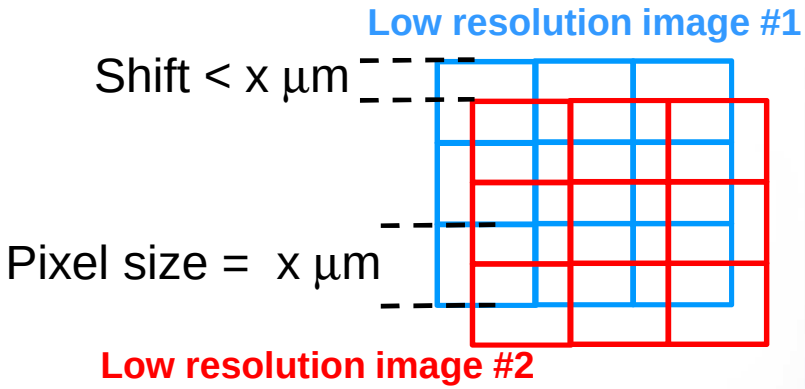
The super-resolution concept in spectroscopic imaging

The super-resolution concept

Background: Real discipline of the signal processing community (early works in 1980, a real emerging scientific research in 2000).

Concept: Simultaneous exploitation of several low resolution images of the same object (observed from “different angles”) in order to obtain one higher resolution image.

Condition of application: sub-pixellic shift between low resolution images (image shifts lower than the pixel size of low resolution image).



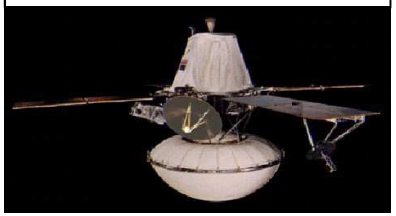
LASIR

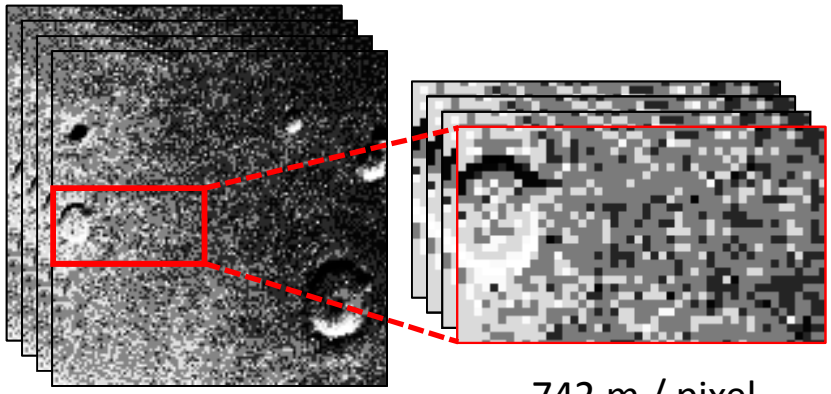
A super-resolution example in space observation

NASA's Viking Mission: the mission objectives were to obtain high resolution images of the Martian surface, characterize the structure and composition of the atmosphere and surface, and search for evidence of life.

Source :
INTELLIGENT SYSTEMS DIVISION
at Ames Research Center

Simultaneous use of 24 shifted images of mars acquired by vicking orbiter 1





Visible images

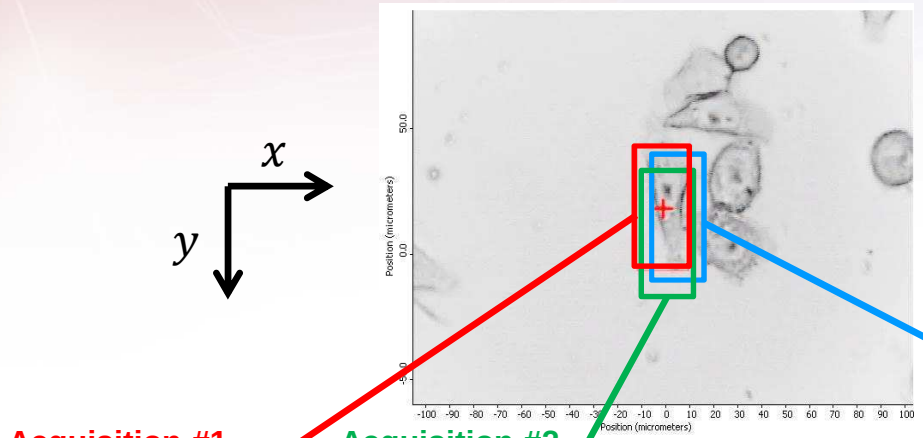
742 m / pixel

Super-resolution

186 m / pixel

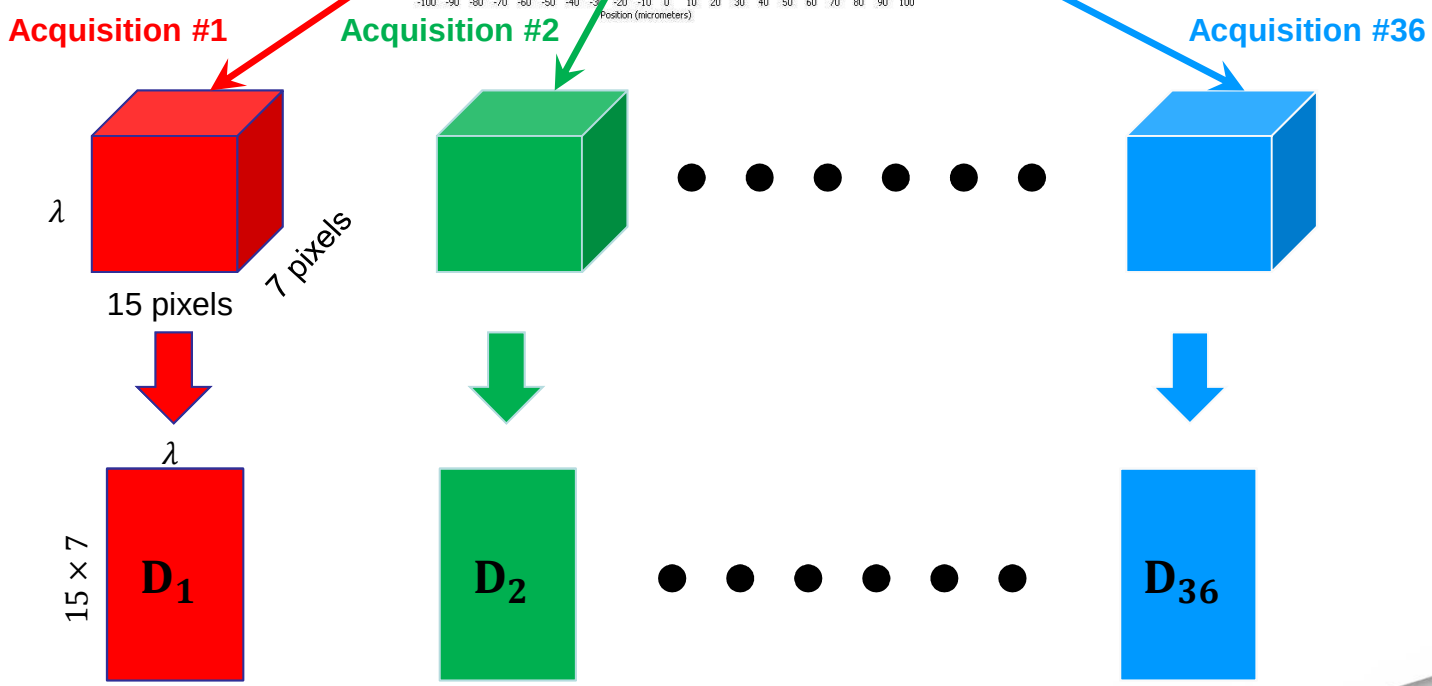
LASIR

Super-resolution acquition setup

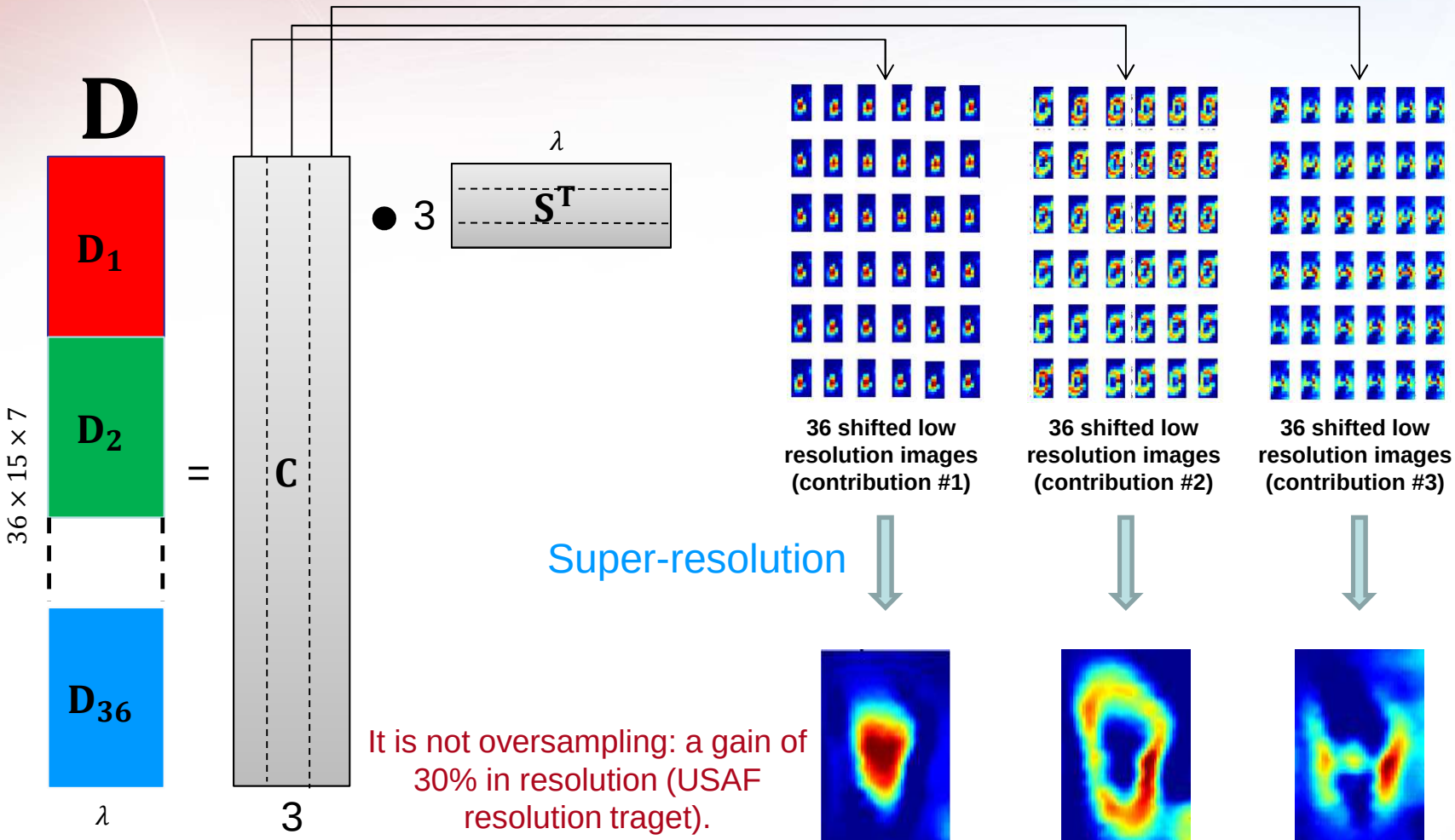


Sub-pixel shifts between consecutive acquisition (multiple of 0.6 μm in x and y directions.

⇒ Acquisition of 36 spectral data cubes



Multiset MCR-ALS analysis + Super-resolution



Column-wise matrix augmentation

It is not oversampling: a gain of 30% in resolution (USAF resolution traget).

S. Piqueras, L. Duponchel, M. Offroy, F. Jamme, R. Tauler, A. de Juan, *Analytical Chemistry*, 85(13), p6303 (2013).

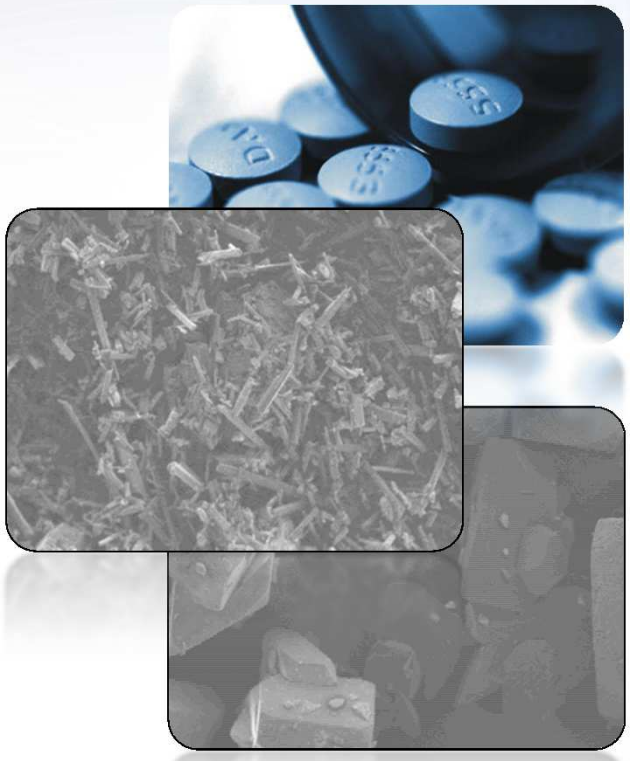
LASIR

Polymorphism study with in situ hyperspectral Raman imaging

Polymorphism is defined by the existence of several crystalline forms for the same chemical compound.

Differences in crystal structure = variations in chemical and physical properties = different behavior of the molecule.

In pharmaceutical industry: relevant to know the potential presence of polymorph transformations induced by different parameters, such as light exposure, pressure, temperature changes…



Aim of the study: use of multiset analysis on the series of in situ Raman hyperspectral images acquired during a thermal induced transformation of carbamazepine as the optimal way to extract useful information about polymorphic transformations.

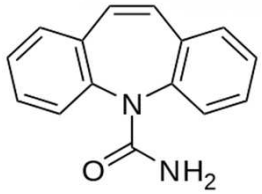
S. Piqueras, L. Duponchel, R. Tauler, A. de Juan, *Analytica Chimica Acta*, 819, p15 (2014)

LASIR

Polymorphism study with in situ hyperspectral Raman imaging

Experimental setup

Sample: anhydrous commercial carbamazepine (CBZ) obtained from Sigma–Aldrich (≥99% r.a). Powder slightly compacted (2 mm thickness) in order to obtain a flat surface.

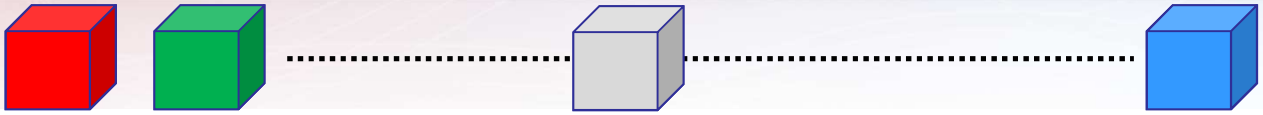


Raman imaging: Labram HR spectrometer, Horiba JY (@632.81nm, from 100 to 1650cm−1). Integration time 0.5 s, pixel size 10 μ, mapping size 20 x 20 pixels. Total acquisition time / cube: 40 min.

Nine Raman data cubes were collected along the thermal degradation at different temperatures (from T=25◦C to T = 160 ◦C with THMS600 accessory from Linkam).

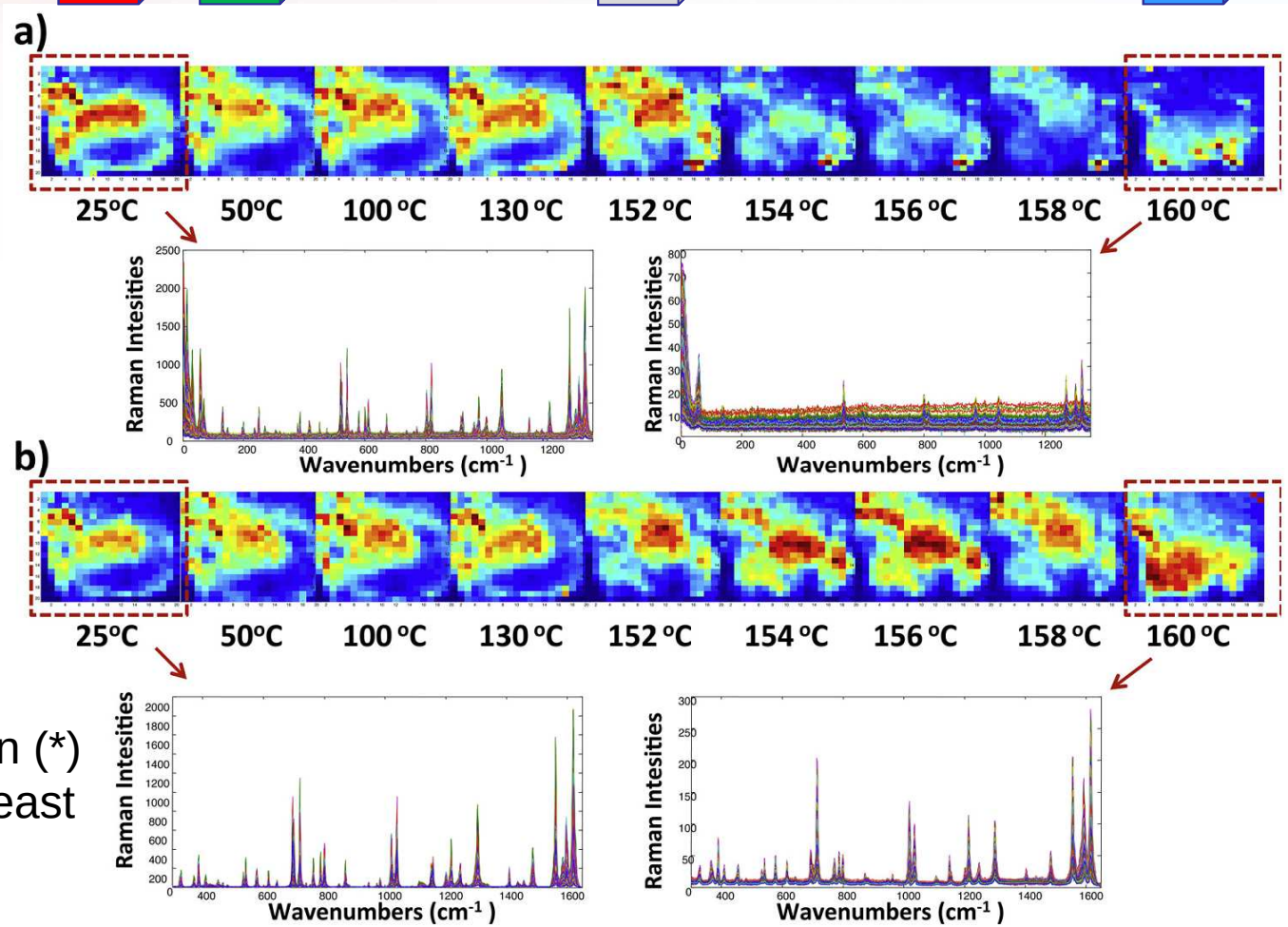
LASIR

A first exploratory analysis with global intensity maps



Raw data

AsLs correction (*) (Assymetric Least Squares)

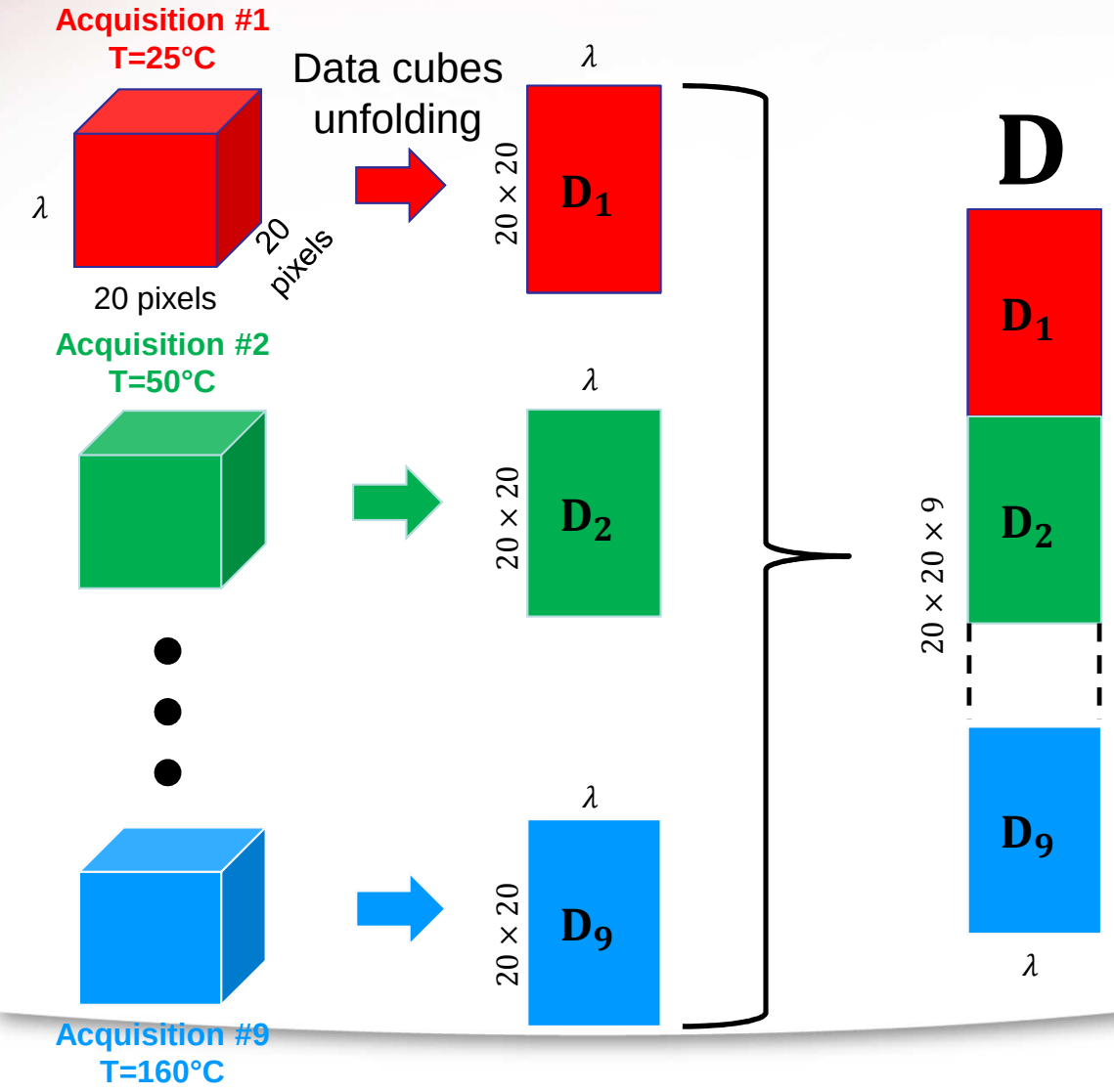


(*) P. Eilers, *Analytical Chemistry*, 75(14), 3631-3636 (2003).

LASIR

Next step: global rank estimation

The global rank is potentially an estimation of the total number of compound in the system



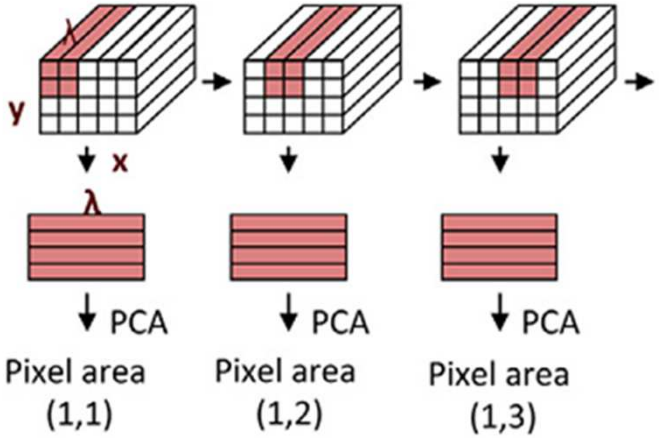
Singular value decomposition (SVD) applied on the column-wise augmented matrix **D**

Detection of three significant singular values = three spectral contributions present in the system.

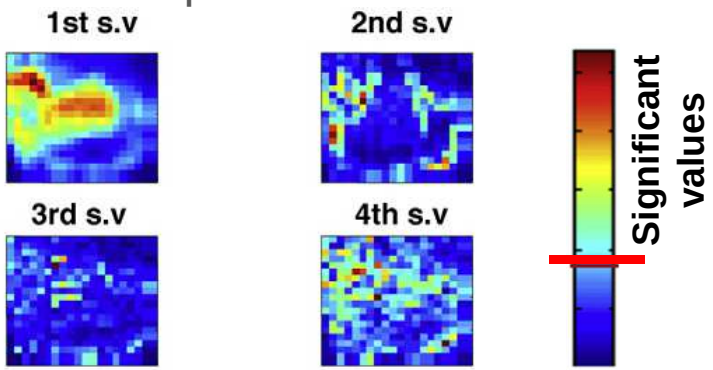
Exploratory local rank analysis

Provide information on the local complexity (compound overlap) at the pixel level of the sample analyzed

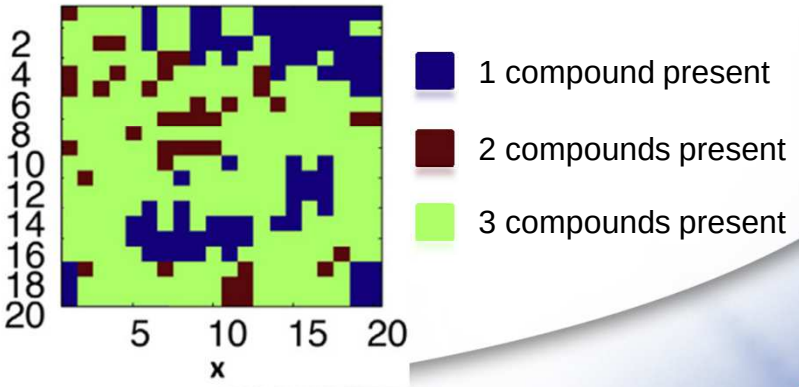
Given an experimental data cube, the FSIW-EFA algorithm (Fixed-Size image Window - Evolving Factor Analysis) performs PCA analyses by moving small windows around each individual pixel area across the full image (*).



4 singular values per pixel => generation of 4 singular value maps



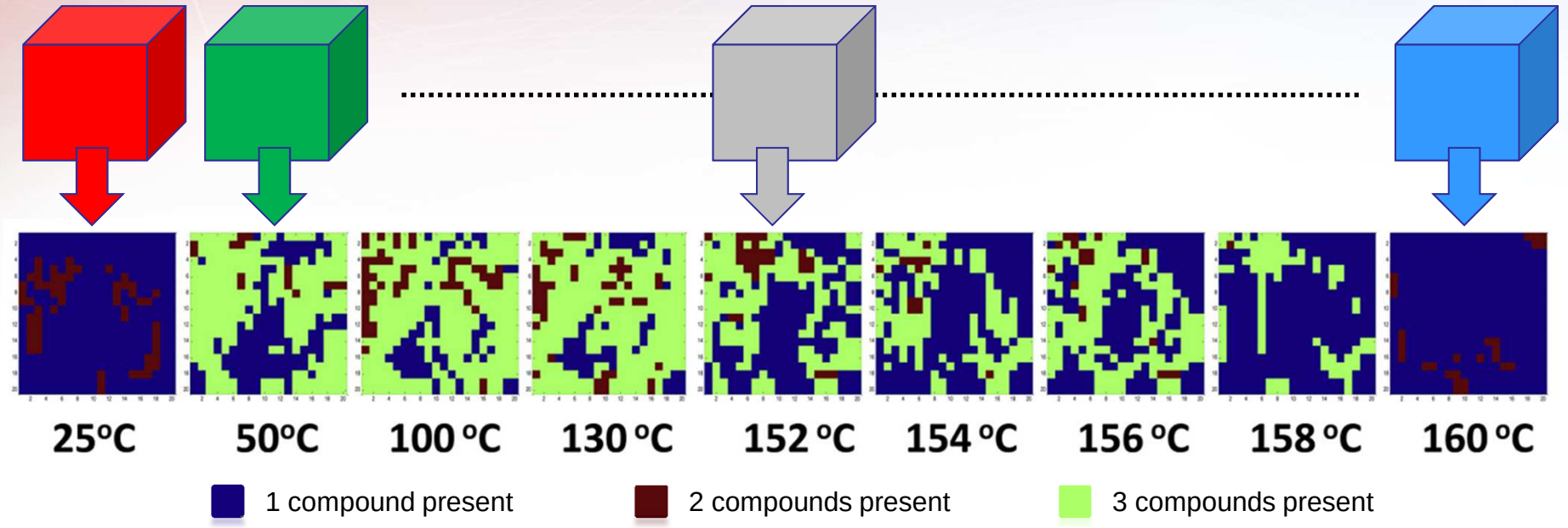
Local rank map: representation of the number of significant singular values per pixel (local number of compounds)



(*) A. De Juan, M. Maeder, T. Hancewicz, R. Tauler, *Chemometrics and Intelligent Laboratory Systems* 77 (2005) 64–74.

LASIR

Local rank analysis along thermal degradation



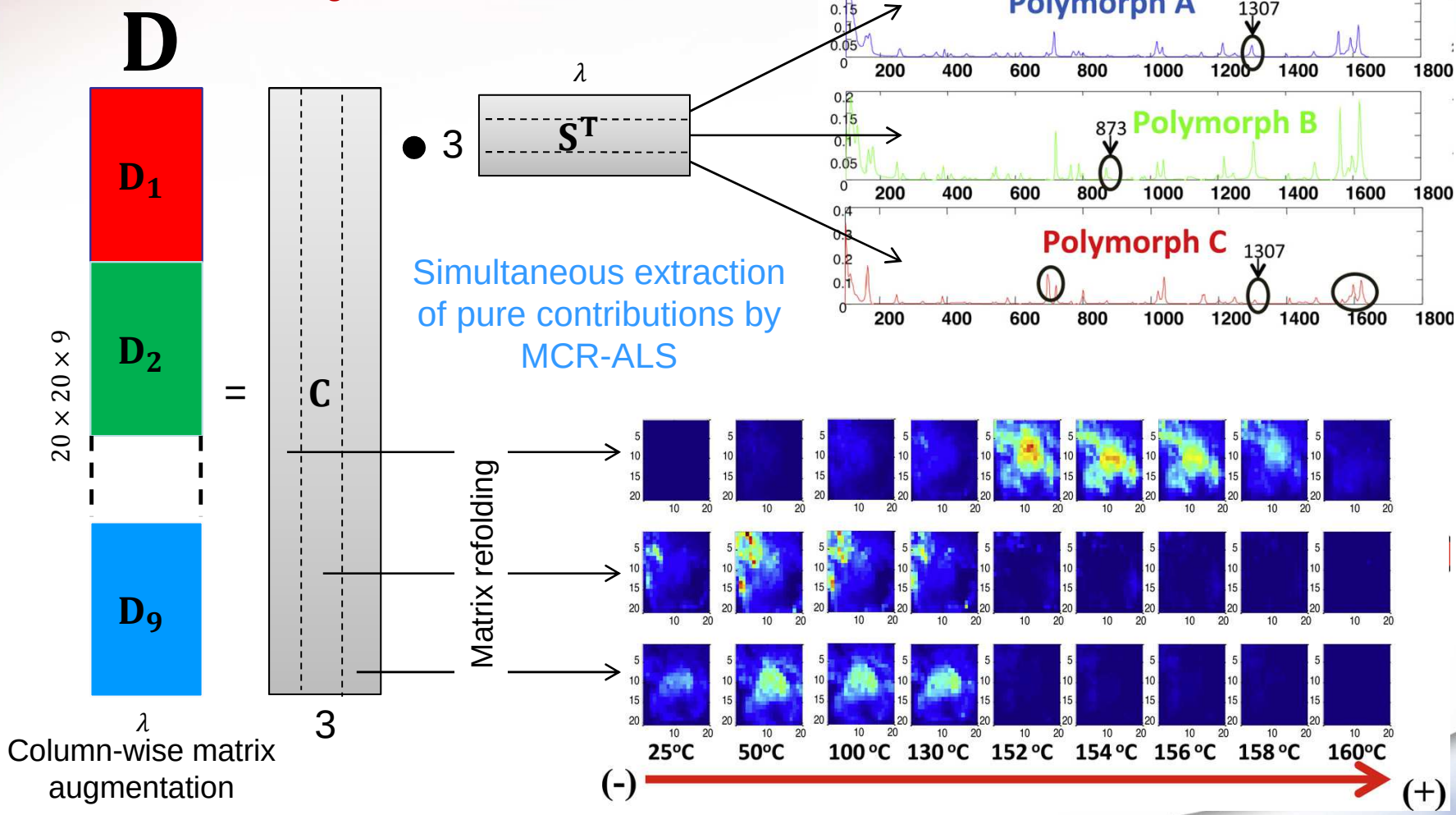
- @ 25 °C: zones with 2 compounds are observed (rather unexpected).
- increase of the complexity up to 130 °C (more green regions).
- Image complexity decreases until 160 °C (where only some pixels with 2 compounds).

The more important: observing pixels in the image where the number of overlapping constituents is smaller than the total (i.e., where some constituents are absent) is crucial to obtain a unique solution in the image multiset resolution analysis.

LASIR

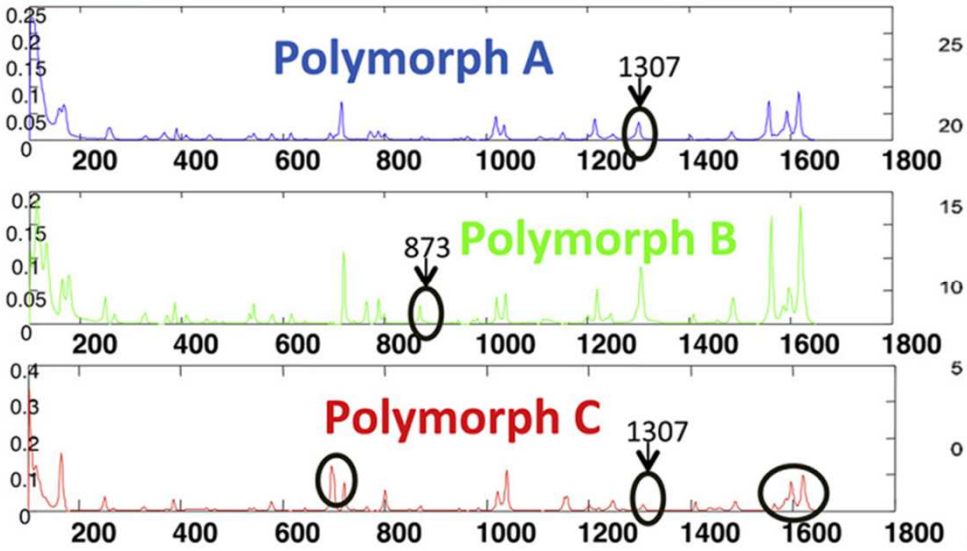
Image multiset MCR-ALS analysis

The nine different data cubes share the same spectral range => MCR can be applied on the column-wise augmented matrix **D**.

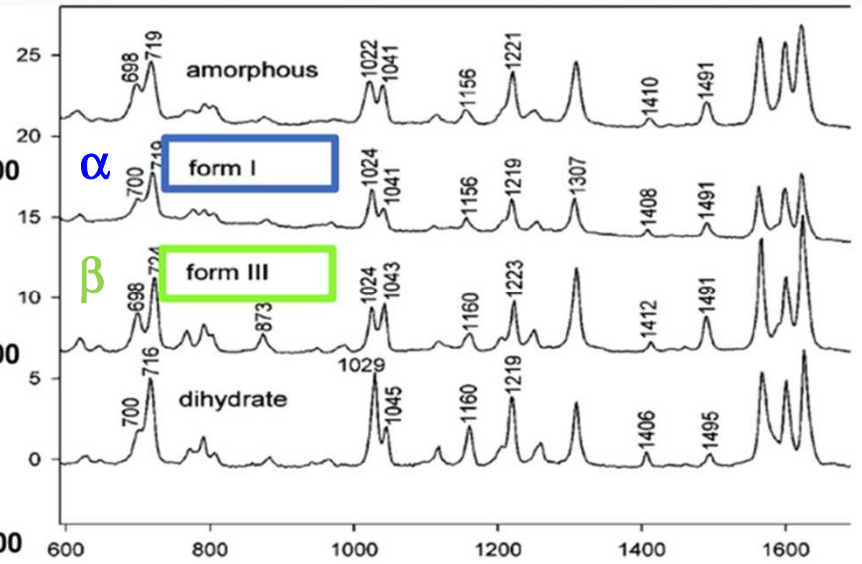


Spectral interpretation

Pure spectra extracted with MCR

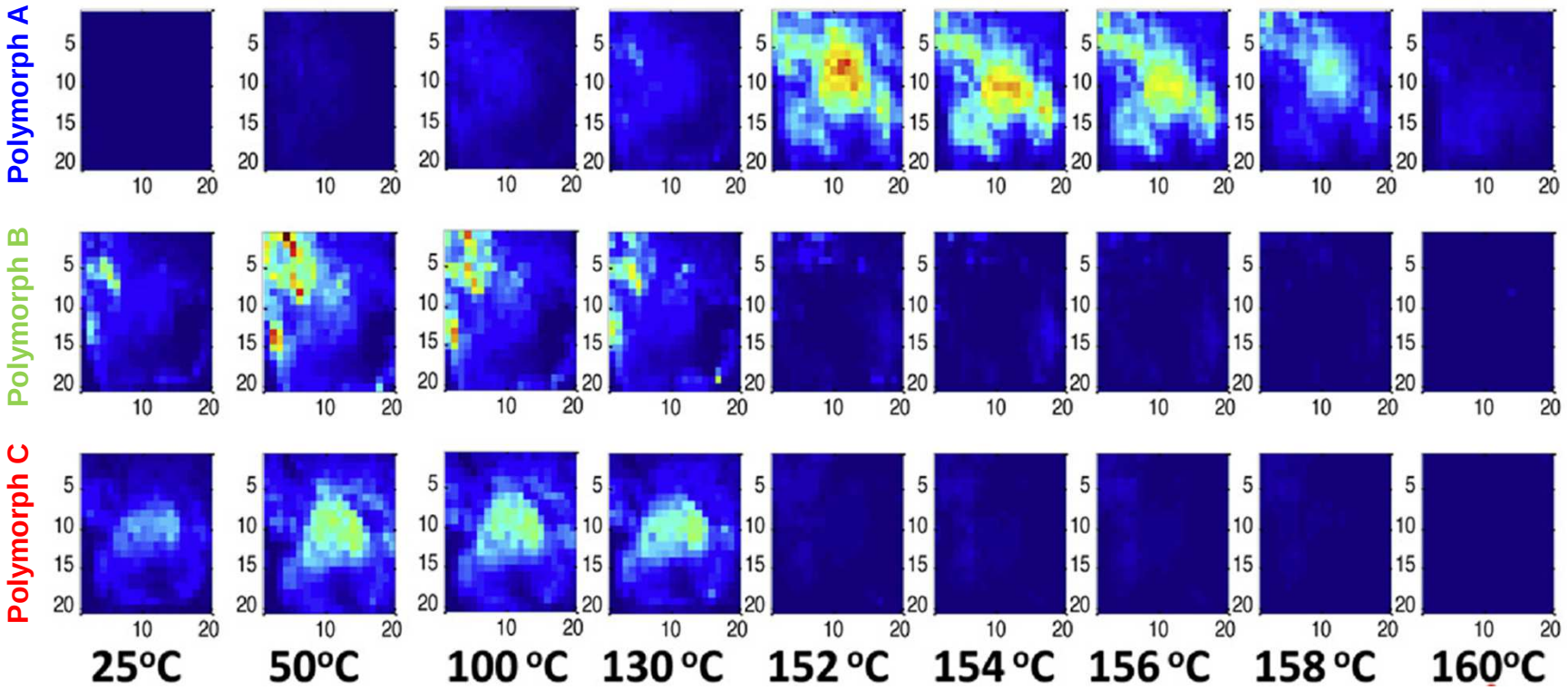


Reference spectra from litterature



- Possible to retrieve the two well-known polymorph forms 1 and 3 also called alpha form and beta form respectively.
- Extraction of a unknown polymorph C with specific spectral features.
- Difficult to have a more important spectral overlap between pure compounds: more or less impossible to generate unbiased chemical maps with the classical integration method.

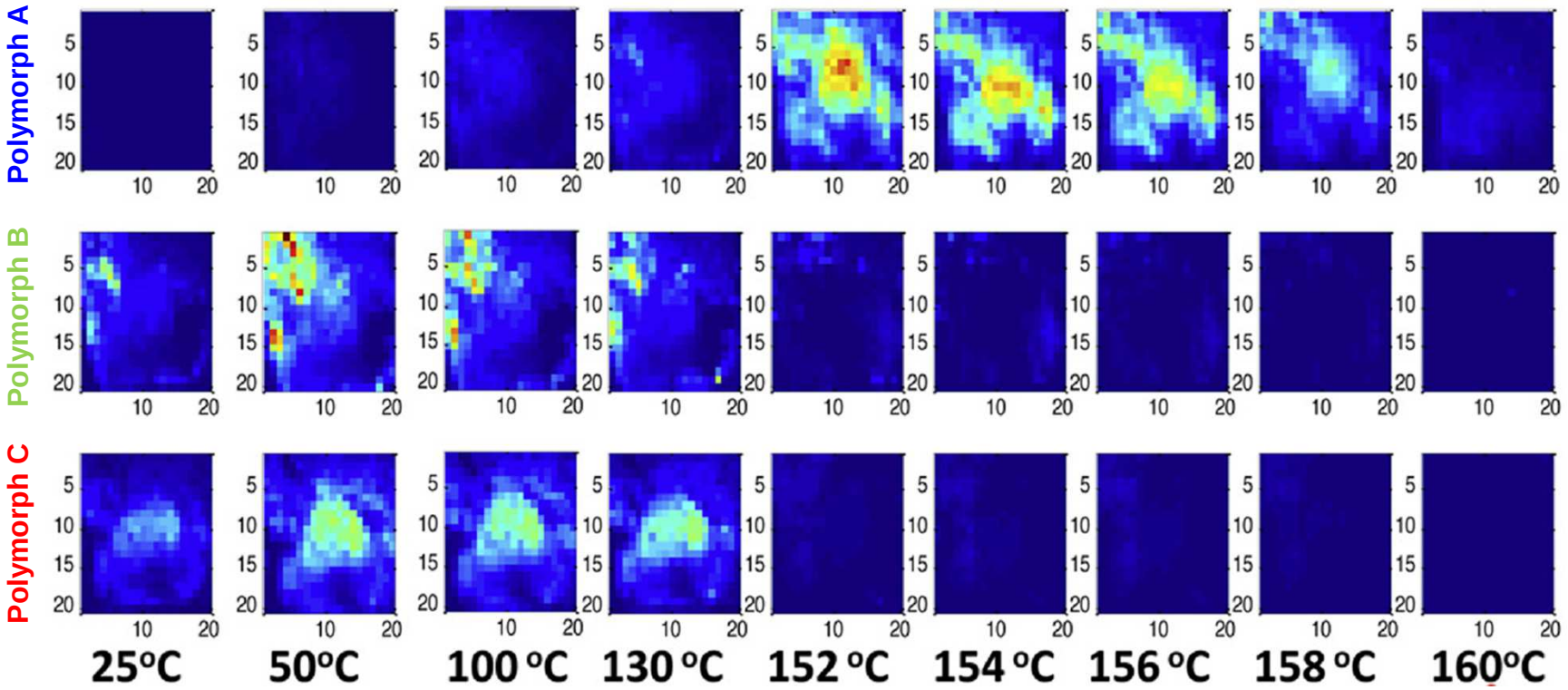
Chemical maps interpretation



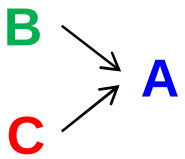
- Distribution maps allow following the process evolution at a pixel (local) level and interpret spatial process evolution.
- Polymorphs B and C are present at 25°C.
- Polymorph A emerges and B and C decay as temperature increases along the process.
- All compounds melt at the end of the process with the consequent decrease in Raman signal and loss of intensity in the distribution maps.

LASIR

Chemical maps interpretation

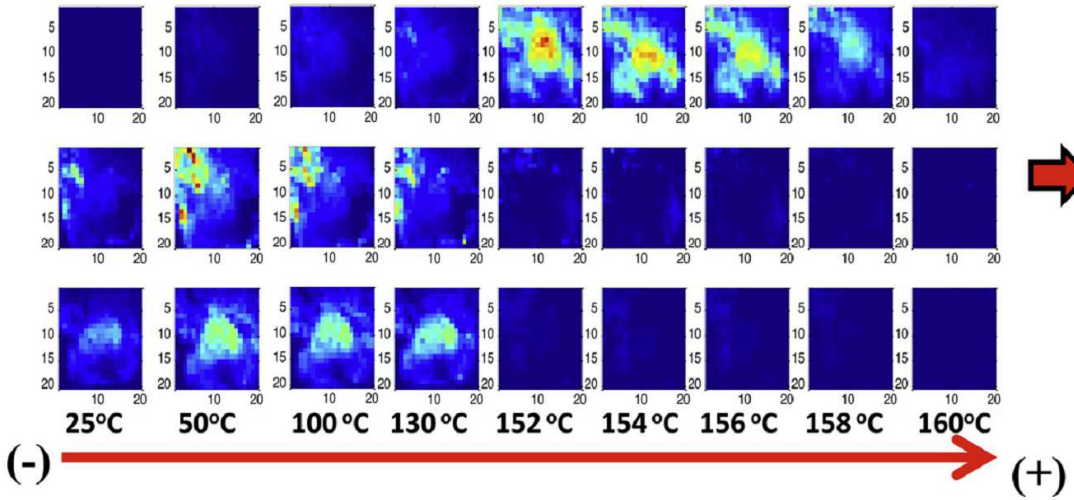


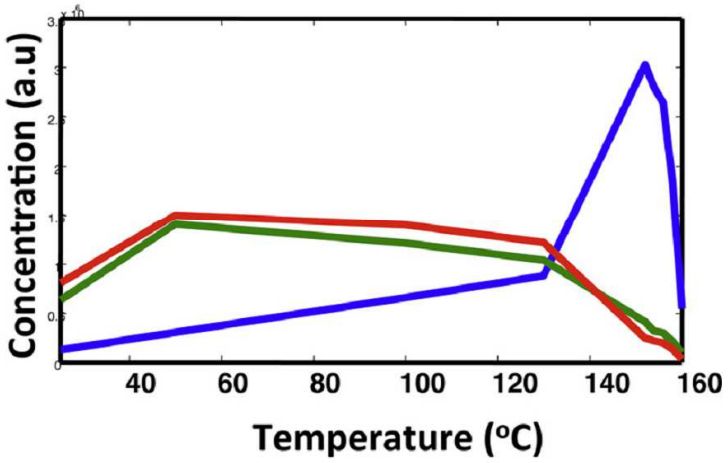
- Observation of different dominant spatial zones of initial compounds B and C.
- Polymorphic product A extends on the spatial areas of B and C
  => the single final product can be obtained from the two initial ones following the parallel pathways :



LASIR

Chemical maps interpretation





Mean values of the distribution maps of each compound

Mean concentrations of **B** and **C** are correlated along the process

⇒ if we don't use imaging spectroscopy but a classical one for bulk analysis, MCR would extract only one contribution for the initial product (probably a mix of B and C).

⇒ Impossible to discover such the parallel pathways.

LASIR

Conclusions

- Richness of vibrational spectrscopy
- Chemometrics is a good way to push the limits of scientific instrumentation (high level or routine basis)
- Chemometrics and more precisely Multivariate Curve Resolution method offers significant potential for the analysis of complex dataset with no prior knowledge about the underlying physico-chemical system. A High degree of adaptability.
- Chemometrics is accessible to all (I'm just a physical-chemist ☺) as long as we take the time to discover it.
- Chemometrics is not a push button concept!

LASIR

Acknowledgments

Dr Anna De Juan
Sara Piqueras (PhD student)

Marc Offroy (PhD student)

Dr Bertrand Bocquet
Simon Laurette (PhD student)

Dr Paul Dumas
Dr Frederic Jamme

Dr Peyman Milanfar

The next GFC symposium

http://chimio2015.sciencesconf.org

SOCIETE FRANÇAISE
DE STATISTIQUE

LASIR

Thank you for your attention!