Multivariate Analysis in LabSpec 6 (MVA)

Vincent Larat
Raman Applications Scientist
HORIBA Scientific

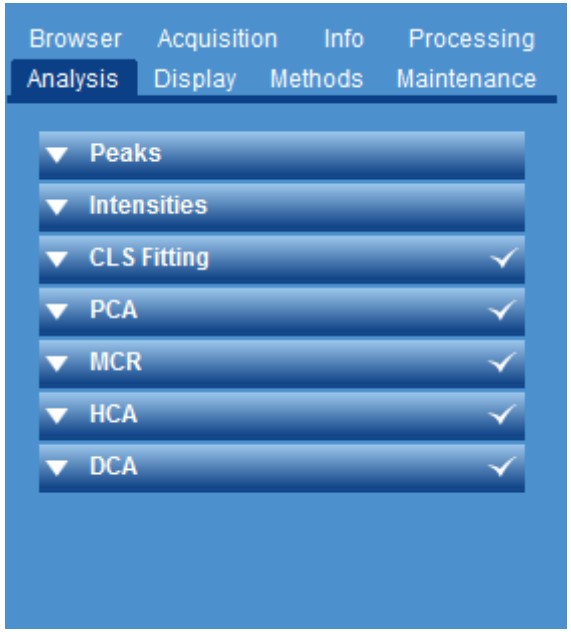
Outline

- Introduction

- 1- Decomposition
  - Decomposition Principle
  - Principal Component Analysis (PCA)
  - Multivariate Curve Resolution (MCR)

- 2- Clustering
  - Hierarchical Cluster Analysis (HCA)
  - Divisive Cluster Analysis (DCA)

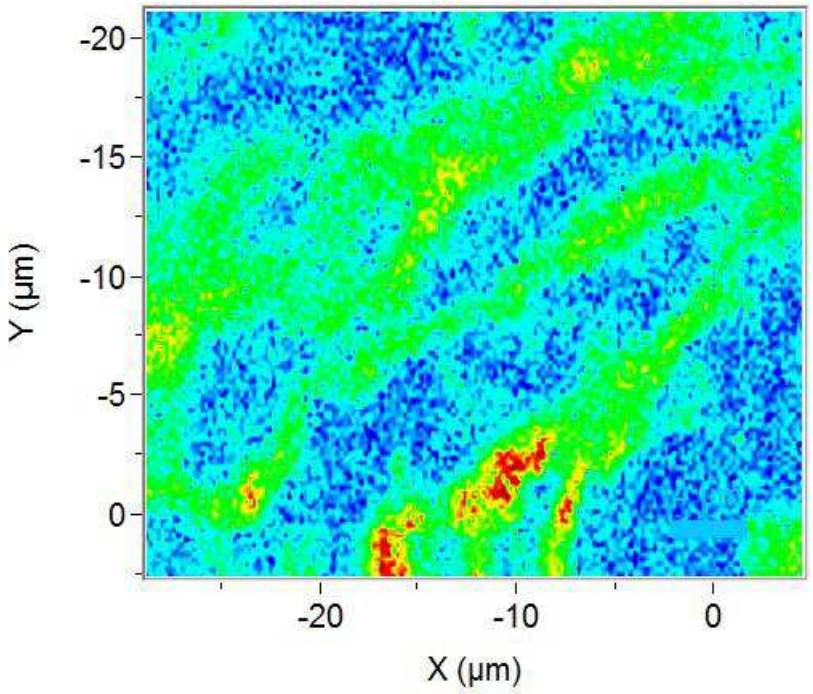
- 3- Data preprocessing



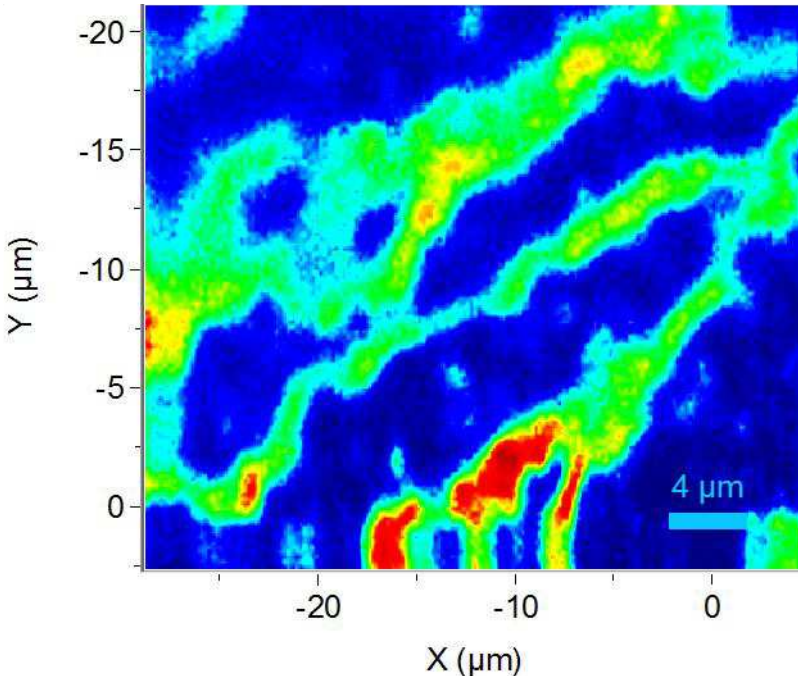
Introduction

Which one do you prefer?

Cursors (univariate)



MVA

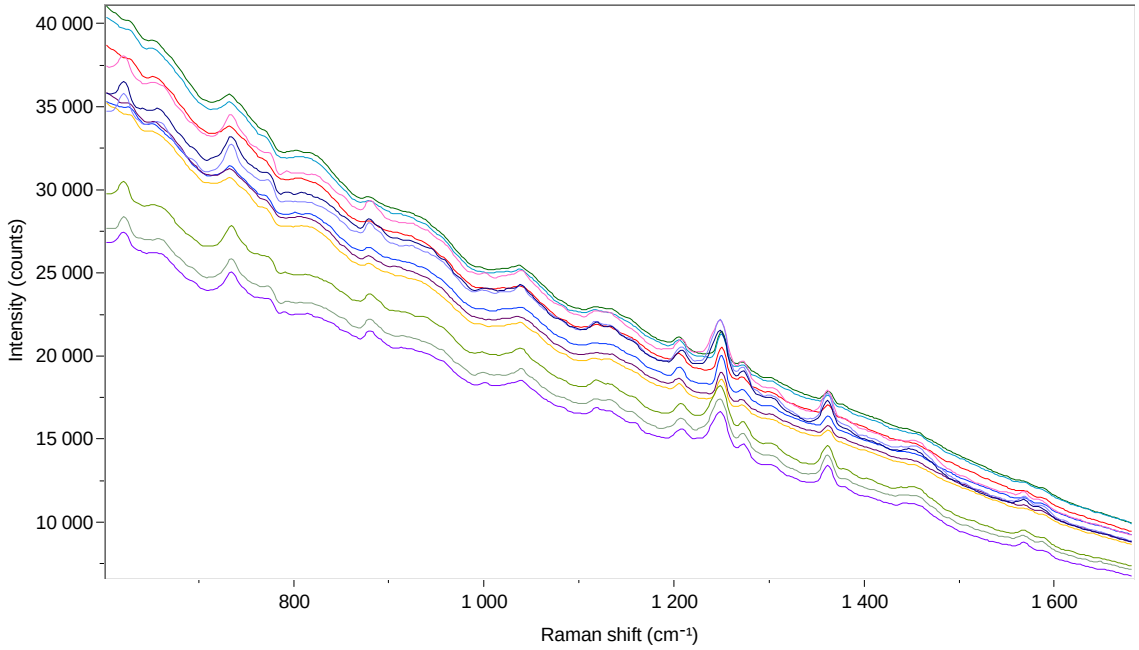


Introduction

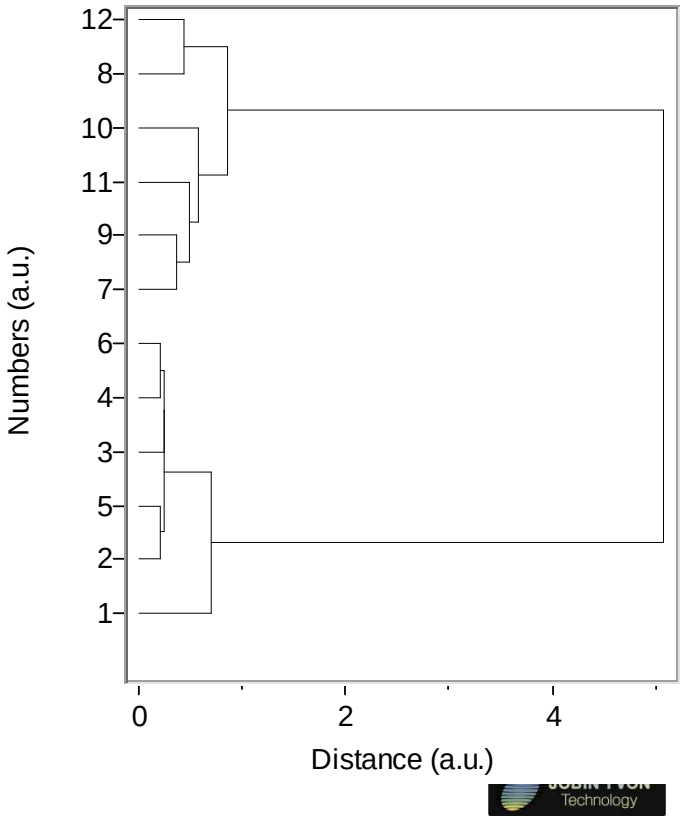
Are there any spectral difference in this dataset?

Is it possible to show clusters?

Raw spectra



MVA



Multivariate vs Univariate Analysis

- Univariate analysis
  - Only one variable: e.g. monitoring the peak intensity (or area) in a dataset

- Multivariate analysis
  - Several variables observed at the same time: e.g. use of several/all of the wavelengths of the spectra
  - Captures the main variance from a dataset, expressed by ‘meaningful’ and interpretable factors (loadings)
  - Multiple algorithms are available
  - Used for :
    - Decomposition and clustering purposes
    - Quantitative purposes

Outline

- Introduction
- 1- Decomposition
  - Decomposition Principle
  - Principal Component Analysis (PCA) and Multivariate Curve Resolution (MCR)
- 2- Clustering (quantitative analysis)
  - Hierarchical Cluster Analysis (HCA)
  - Divisive Cluster Analysis (DCA)
- 3- Data preprocessing

1- Decomposition

Decomposition principle

- It is a factorization technique. It helps identifying the key components within a dataset
- For a dataset of n spectra, each individual spectrum is decomposed into a linear combination of components (also known as factors or loadings):

- Spectrum 1 = Score $_{1,1}$ * Loading 1 + Score $_{1,2}$ * Loading 2 +
- Spectrum 2 = Score $_{2,1}$ * Loading 1 + Score $_{2,2}$ * Loading 2 +
- .
- .
- .
- Spectrum n = Score $_{n,1}$ * Loading 1 + Score $_{n,2}$ * Loading 2 +

Loadings kept

Loadings discarded

- In most cases, only a limited number of factors (or loadings) will be sufficient to describe the dataset. The remaining ones will express mostly noise (residuals)

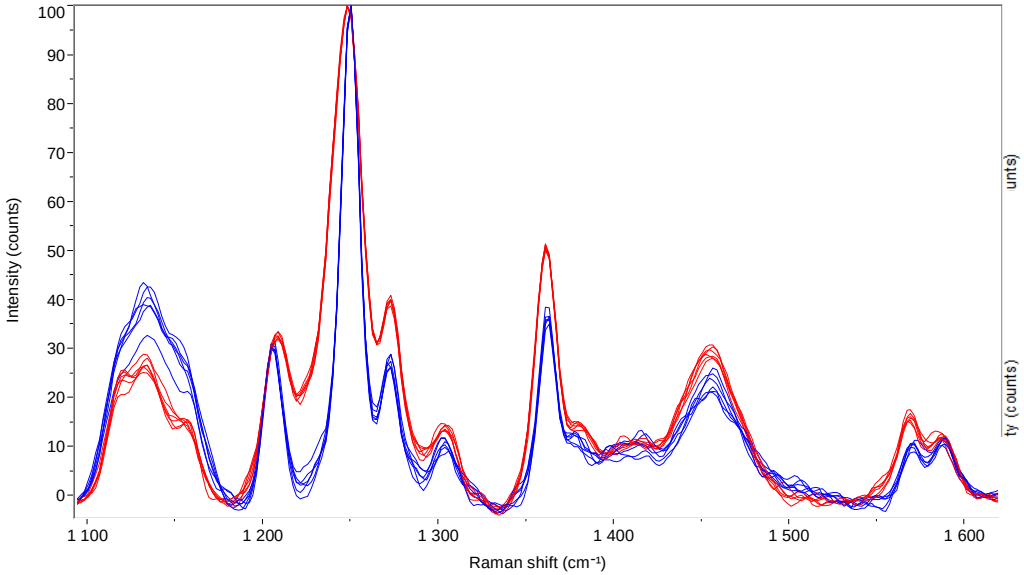
$$S_i = \sum_j s_{i,j} \times L_j + E$$

residuals

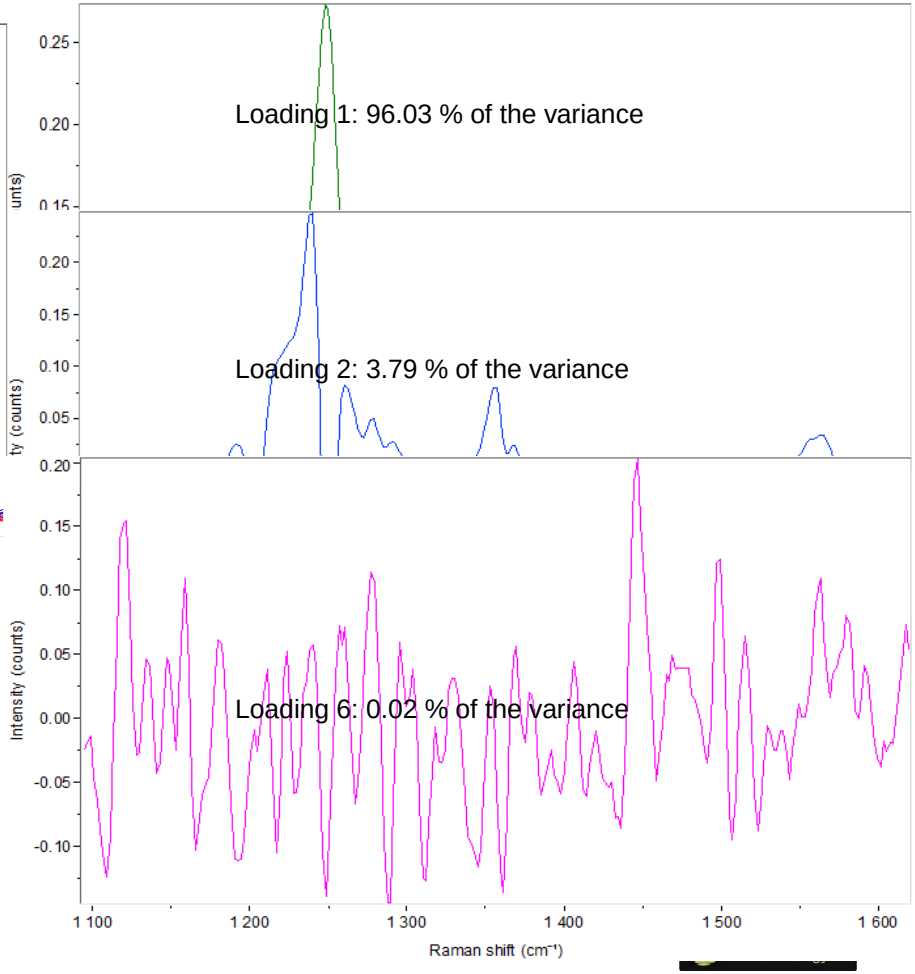
1- Decomposition

Decomposition principle

- Example: 6 capsules filled with form 1 and 6 capsules filled with form 2



	Components	% variations	accum. % variation
1	PCA_1	96.03	96.03
2	PCA_2	3.79	99.82
3	PCA_3	0.04	99.86
4	PCA_4	0.04	99.90
5	PCA_5	0.02	99.92
6	PCA_6	0.02	99.94



1- Decomposition

Decomposition principle

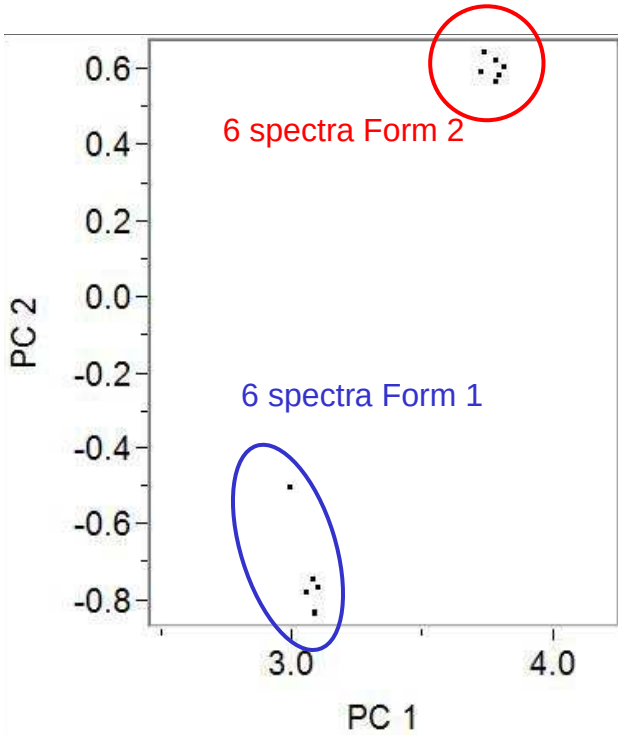
- Example: 6 capsules filled with form 1 and 6 capsules filled with form 2

1- Decomposition

Decomposition principle

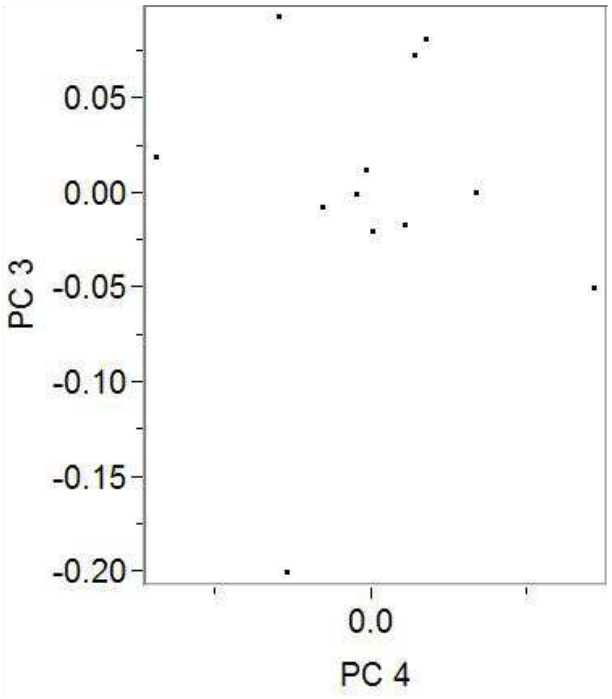
- Example: 6 capsules filled with form 1 and 6 capsules filled with form 2

  - Score plot: PC1 vs PC2



  - Score plot: PC4 vs PC3

Form 2? Form 1?



Outline

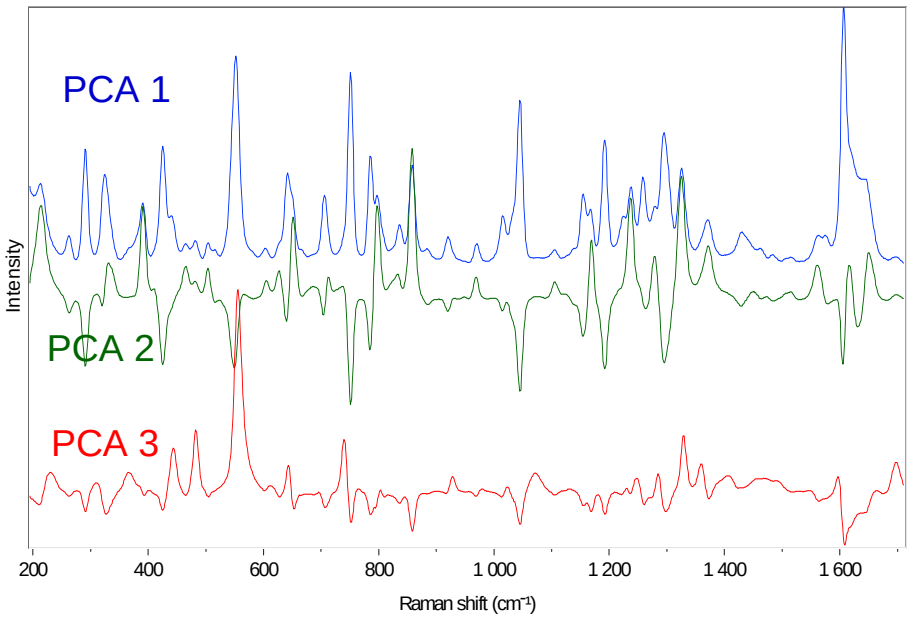
- Introduction
- 1- Decomposition
  - Decomposition Principle
  - Principal Component Analysis (PCA) and Multivariate Curve Resolution (MCR)
- 2- Clustering (quantitative analysis)
  - Hierarchical Cluster Analysis (HCA)
  - Divisive Cluster Analysis (DCA)
- 3- Data preprocessing

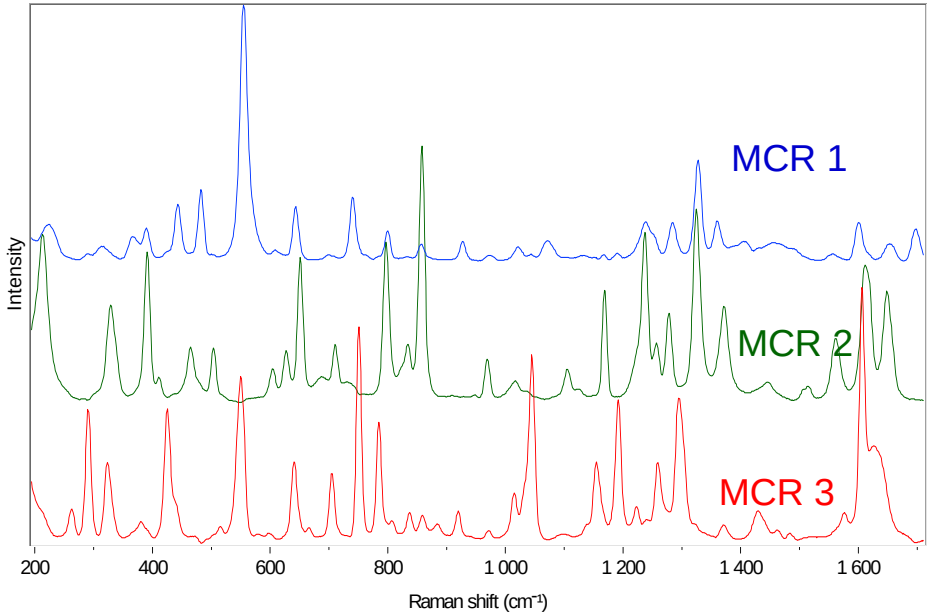
1- Decomposition

PCA and MCR

- Example: Pharmceutical tablet
  - 3 main constituents: aspirin, caffeine, acetaminophen
  - Mapping: dataset of 9595 spectra (101 x 95)
  - PCA and MCR applied to this dataset

- Loadings

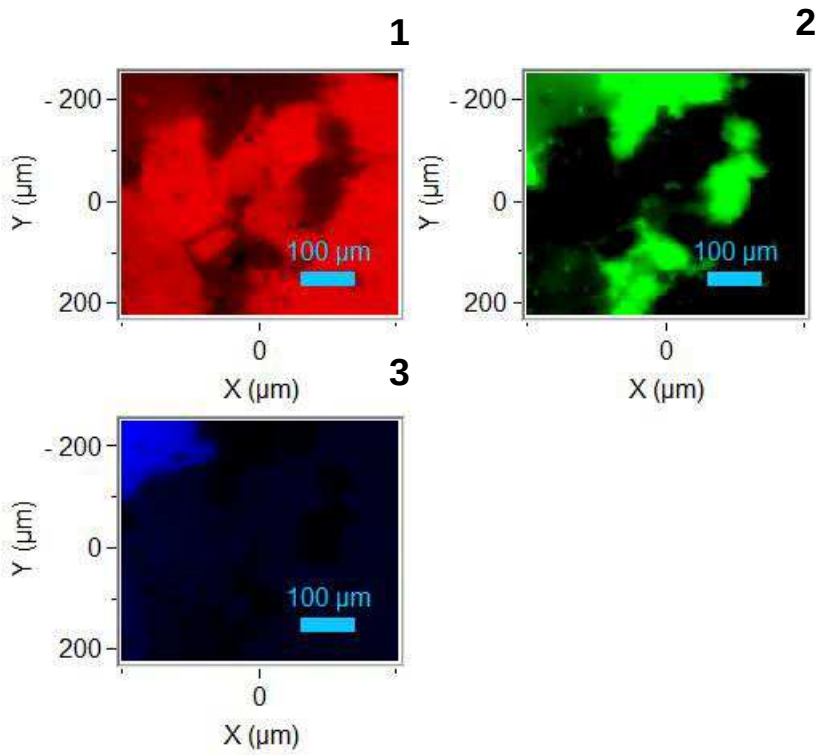




1- Decomposition

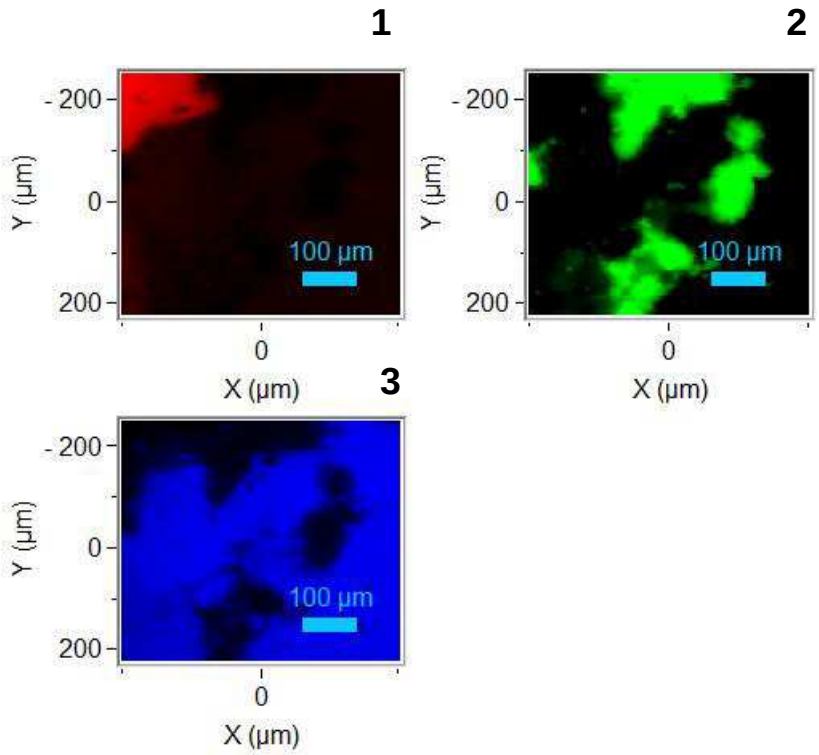
PCA and MCR

PCA: Score plots



Components	% variations	accum. % variation
PCA_1	85,26	85,26
PCA_2	8,31	93,56
PCA_3	4,08	97,64

MCR: Score plots



Loadings	% variations	accum, % variation
MCR_1	9,80	9,80
MCR_2	19,71	29,51
MCR_3	68,13	97,64

PCA and MCR

- PCA: Principal Component Analysis

  - Main characteristics:
    - Orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of uncorrelated components
    - Ranking of the principal components (the first one always captures the most variance)
    - Components are not always easy to interpret as they may have positive and negative peaks

  - When to use it?
    - Exploratory data analysis
      - ✓ Identification of groups
      - ✓ Outliers in the dataset
      - ✓ Help finding out the number of constituents
    - Applicable to any datasets including maps

PCA and MCR

- MCR: Multivariate Curve Resolution

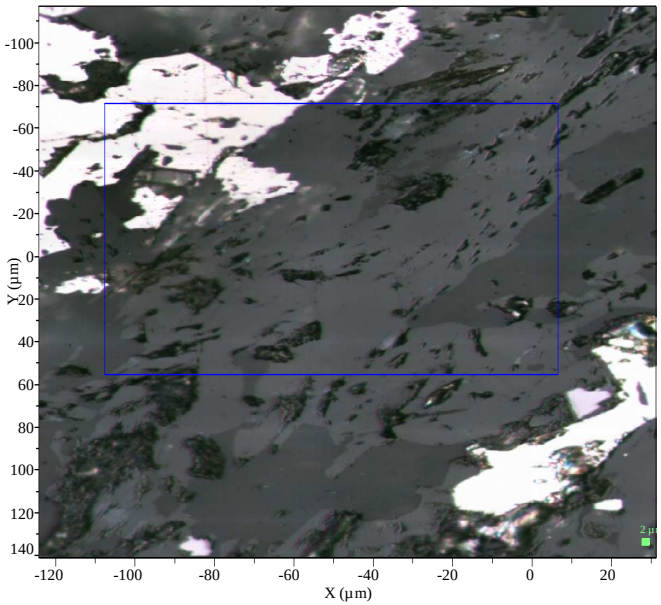
  - Main characteristics:
    - Components are no longer ranked
    - Components can be readily interpreted: the algorithm includes a non-negativity constraint explaining why the spectra do not show negative peaks as in PCA.

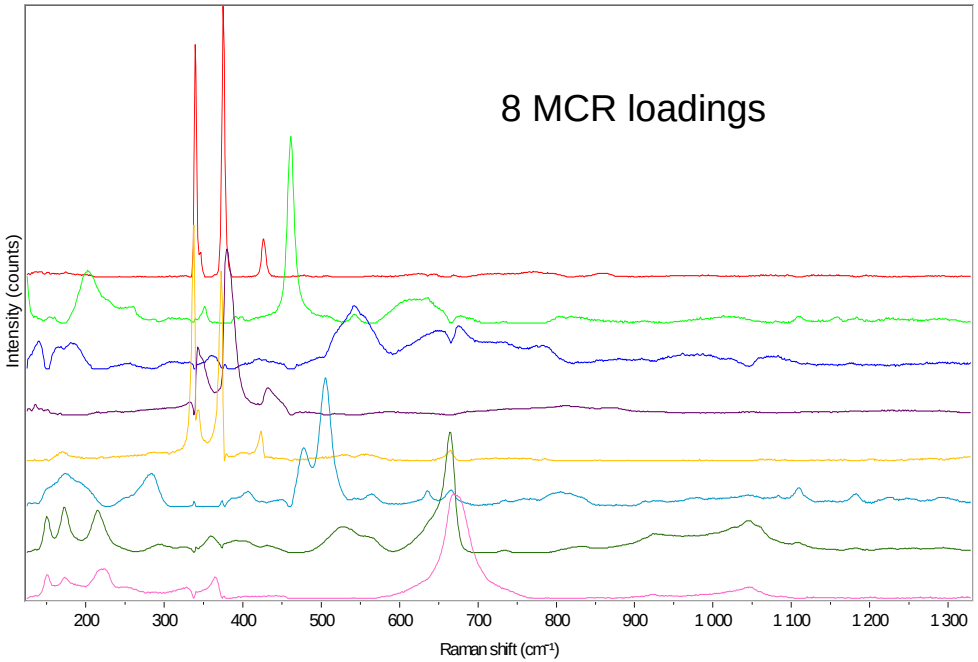
  - When to use it?
    - Applicable to any datasets including maps
    - When interpretable components are required
    - MCR component can be searched in libraries.

1- Decomposition

Example of MCR analysis

- Geological sample
  - Mapping: 5460 spectra (70 x 78)

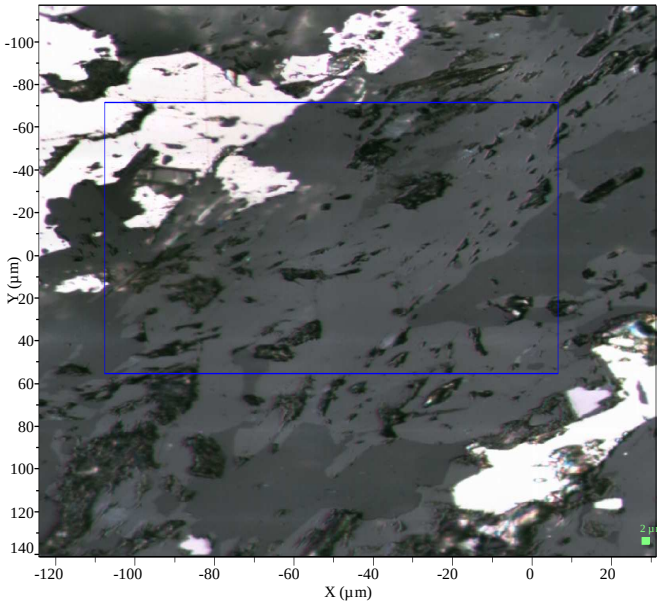




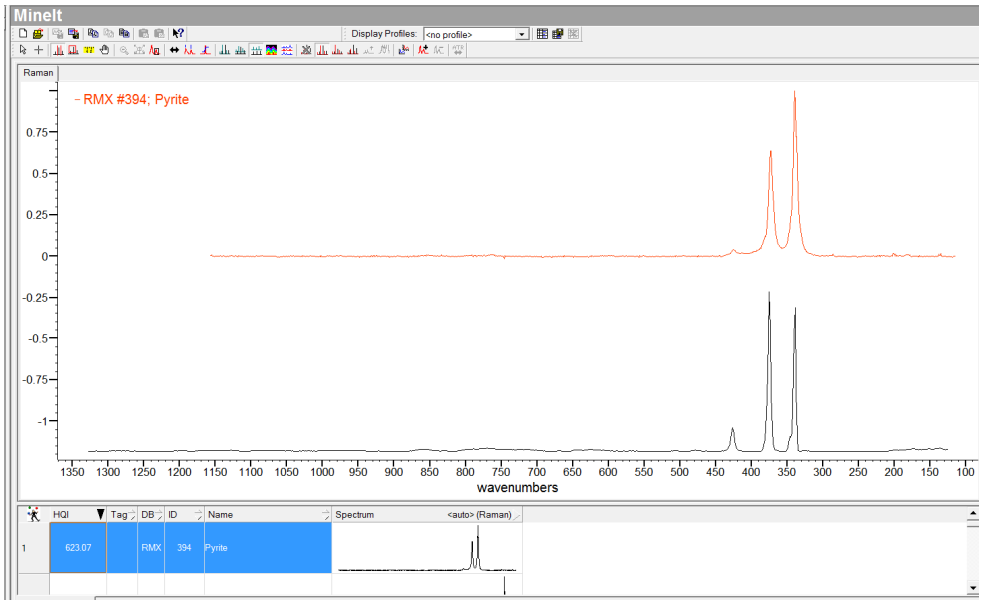
1- Decomposition

Example of MCR analysis

- Geological sample
  - Search of the MCR loadings in KnowItAll® HORIBA Edition with HORIBA spectral libraries



MCR Loading 1 = Pyrite

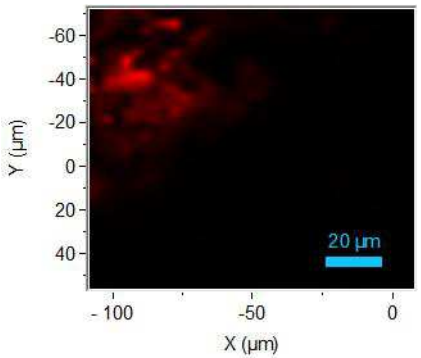


1- Decomposition

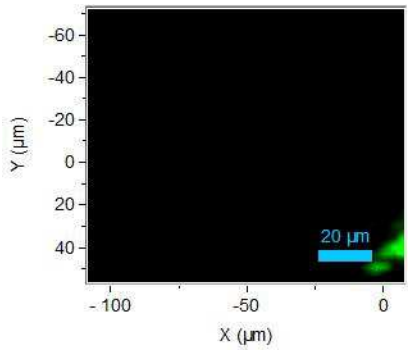
Example of MCR analysis

- Geological sample
  - Identification of the MCR loadings in KnowItAll® HORIBA Edition

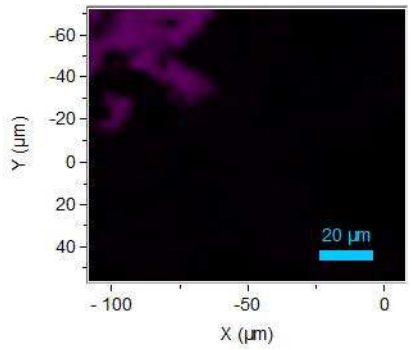
MCR 1: Pyrite

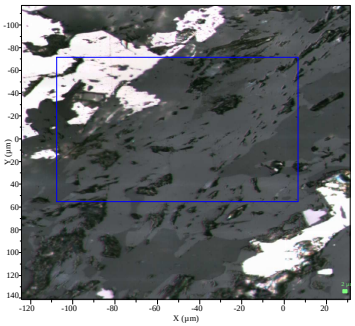


MCR 2: Quartz

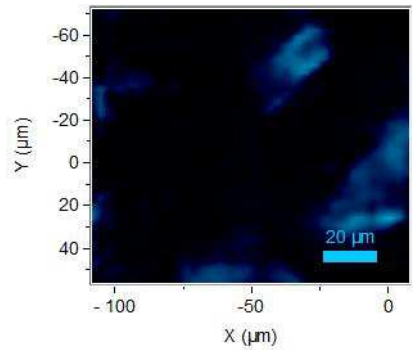


MCR 4: Pyrite with different crystallinity

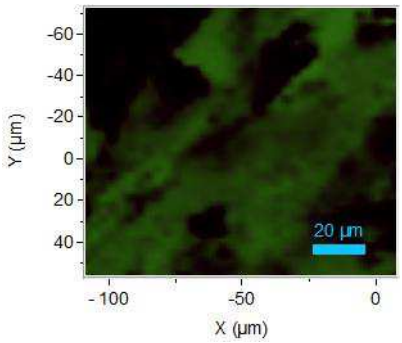




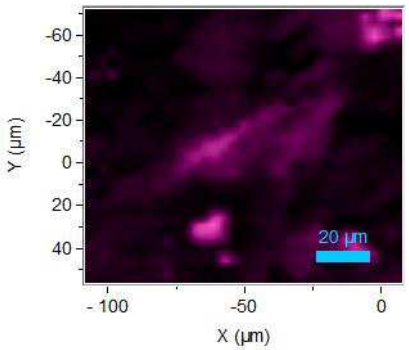
MCR 6: Bytownite (tectosilicate)



MCR 7: Tremolite (inosilicate)



MCR 8: Ilmenite



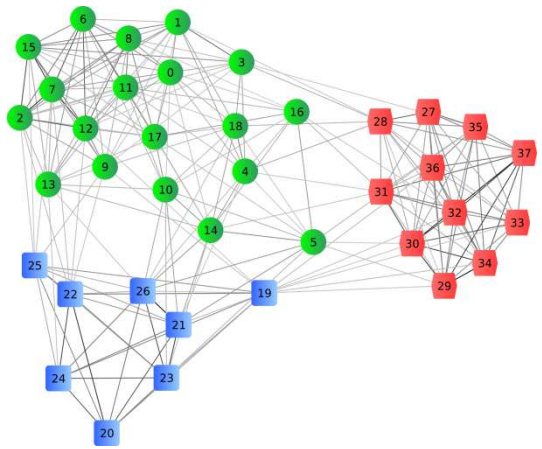
Outline

- Introduction
- 1- Decomposition
  - Decomposition Principle
  - Principal Component Analysis (PCA) and Multivariate Curve Resolution (MCR)
- 2- Clustering
  - Hierarchical Cluster Analysis (HCA)
  - Divisive Cluster Analysis (DCA)
- 3- Data preprocessing

2- Clustering

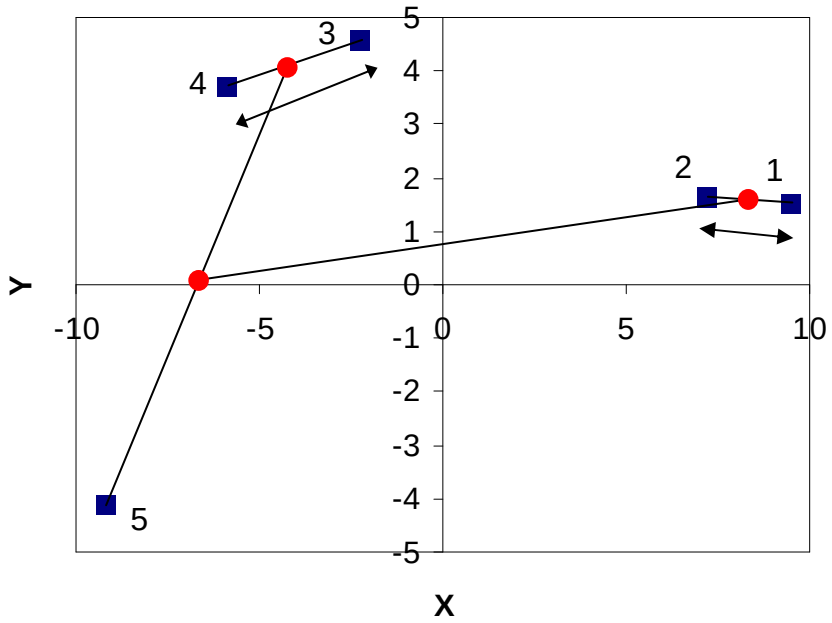
Clustering principle

- Clustering techniques are used to build groups of objects in which each object of the group presents more similarities with the other objects of the same group than with the objects of the other groups.
- It can be applied in many different application fields, such as biology, medicine, economics, marketing, computer science, etc…
- In spectroscopy it is used to classify spectra of a dataset into groups (or clusters) having similar spectral properties.
- There are several classes of clustering:
  - Hierarchical clustering analysis (HCA)
  - Divisive clustering analysis (DCA)
  - ….

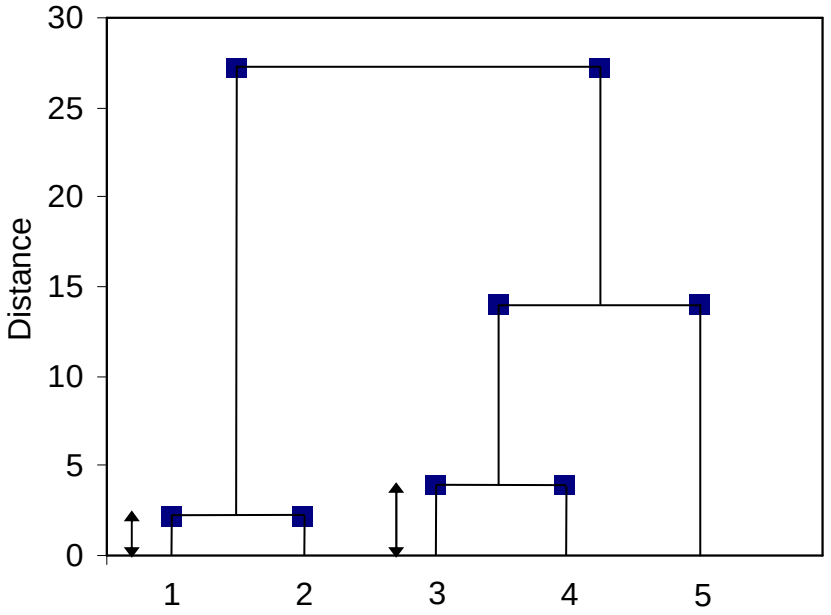


2- Clustering

Hierarchical Cluster Analysis (HCA)



Dendrogram



At start, each object is its own cluster

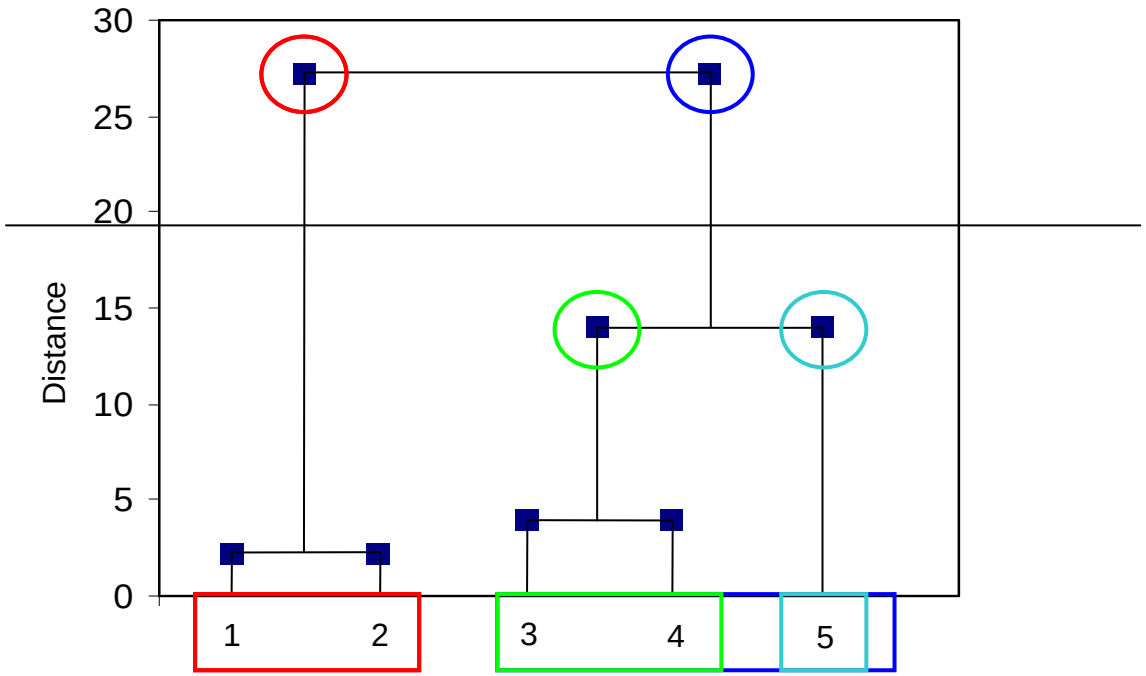
Find the 2 closest objects and define a new point at their center. They are paired to form a cluster.

Find the next 2 closest objects and define a new point at their center. They are paired to form another cluster

Find the next 2 closest objects and define a new point at their center (repeated for each cycle).

Hierarchical Cluster Analysis (HCA)

- The number of classes will depend on where you set the cursor



- 2 classes:
  - Class 1: sample 1 and 2
  - Class 2: sample 3, 4 and 5
- 3 classes:
  - Class 1: sample 1 and 2
  - Class 2: sample 3 and 4
  - Class 3: sample 5

Hierarchical Cluster Analysis (HCA)

- Use of Ward's algorithm
  - Known as 'agglomerative' clustering method.
  - Unsupervised method.
  - Ward's algorithm involves minimization of within-cluster variance.
  - At each cycle, the distances between each clusters are calculated; clusters are then joined in order to minimize the variance within the newly formed clusters.
  - Choice of the number of classes is critical

- Advantages
  - Display of the results with dendrograms

- Limitations
  - This method involves a lot of calculation and should therefore be limited to relatively small datasets, comprising less than 1000 spectra.
  - Has the tendancy to split data in groups of roughly equal size

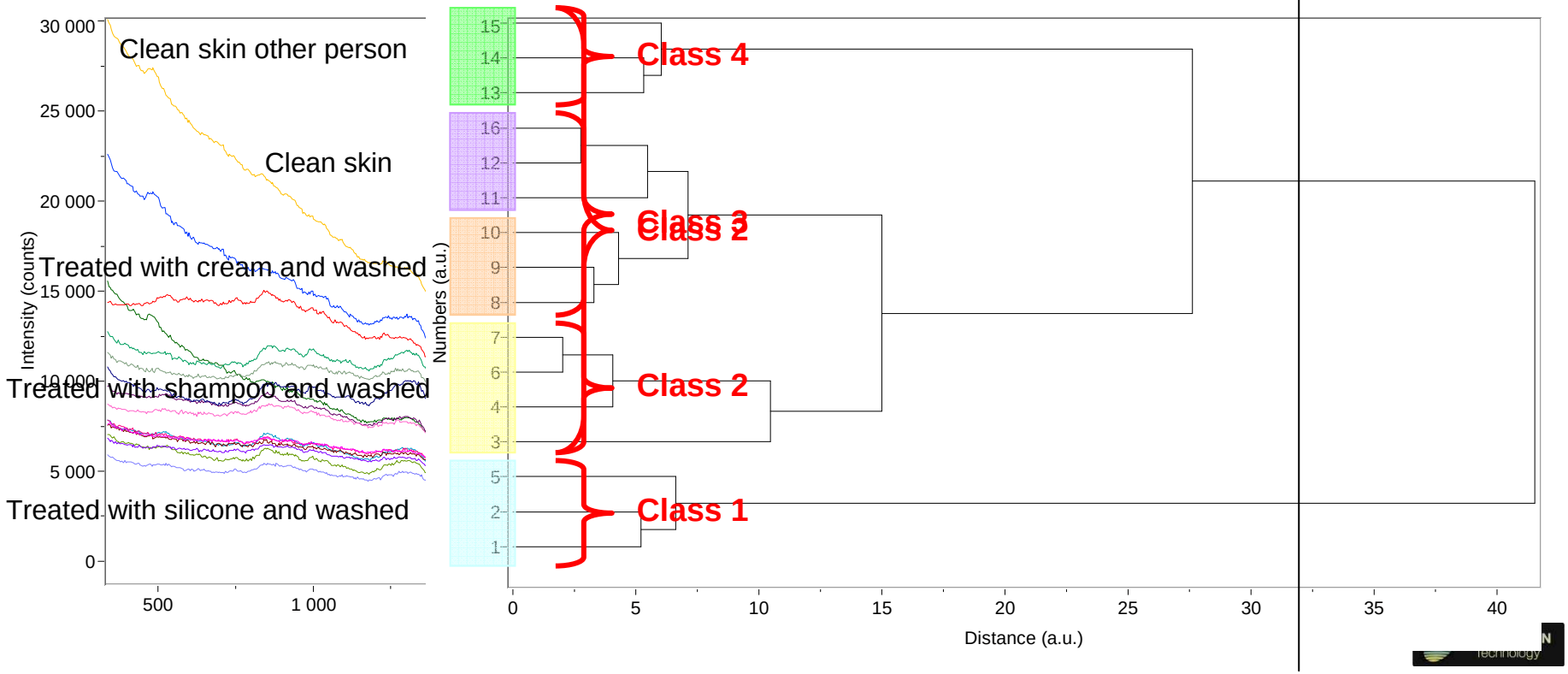
2- Clustering

Hierarchical Cluster Analysis (HCA)

- Spectra of skin
  - Clean skins
  - Treated skins with products and washed

4 classes
2 classes

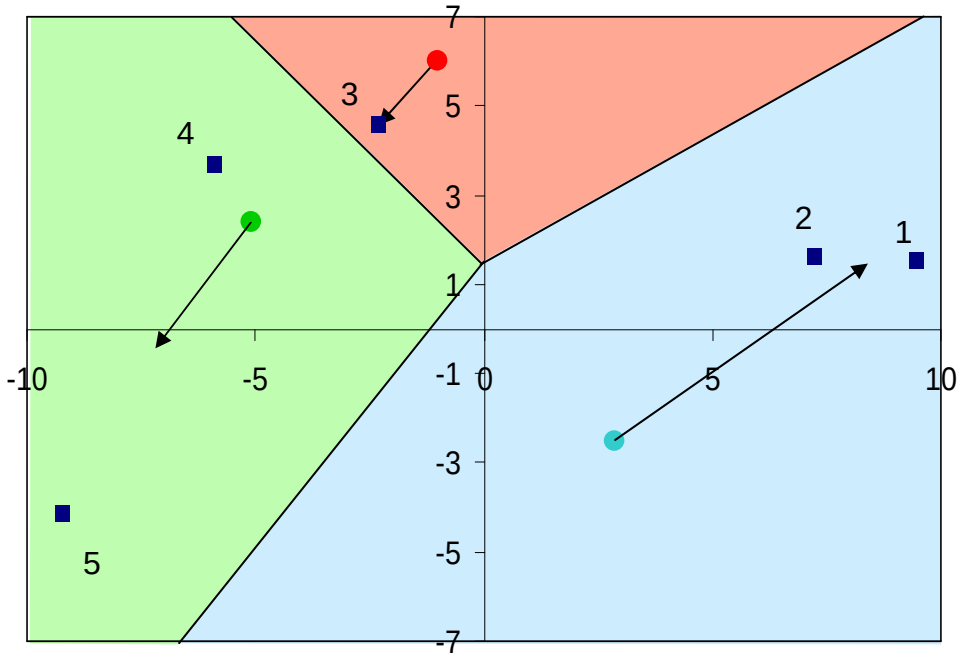
	Components	class	counts %
1	HCA_1	3	18.75
2	HCA_2	4	25.00
3	HCA_3	6	37.50
4	HCA_4	3	18.75



Outline

- Introduction
- 1- Decomposition
  - Decomposition Principle
  - Principal Component Analysis (PCA) and Multivariate Curve Resolution (MCR)
- 2- Clustering
  - Hierarchical Cluster Analysis (HCA)
  - Divisive Cluster Analysis (DCA)
- 3- Data preprocessing

Divisive Cluster Analysis (DCA)

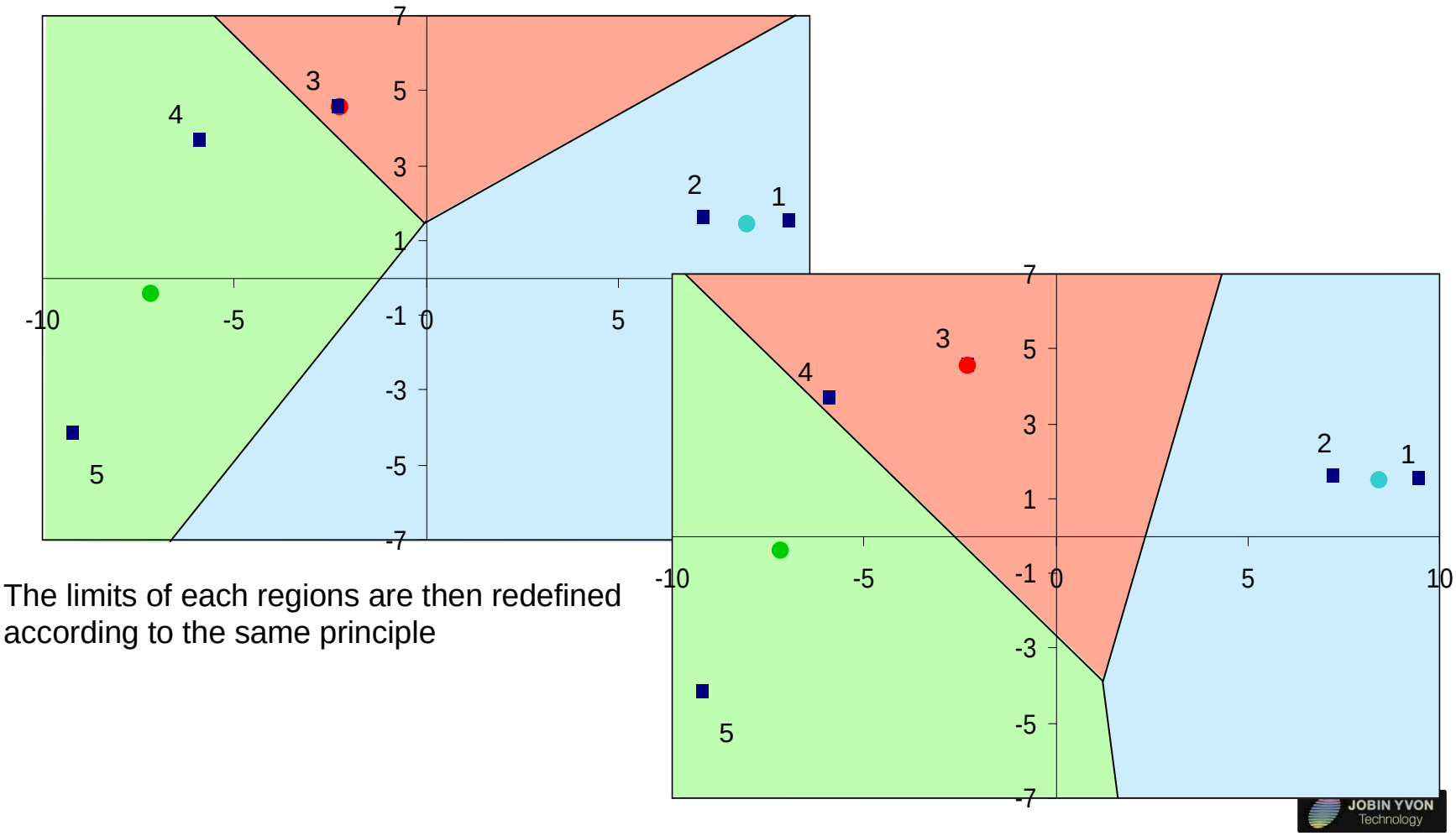


A number of points (centroids) corresponding to the number of classes are set at random positions.

The space is divided into regions, in such a way that all the points within each region are closer to its centroid than to any other one.

New centroids are then set, corresponding to the mean of the points in each region.

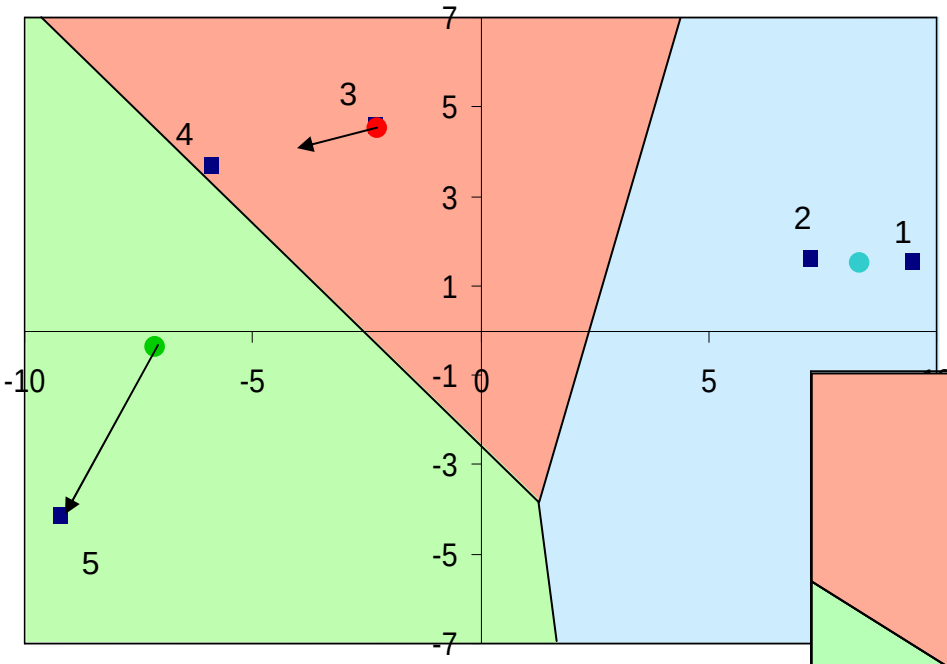
Divisive Cluster Analysis (DCA)



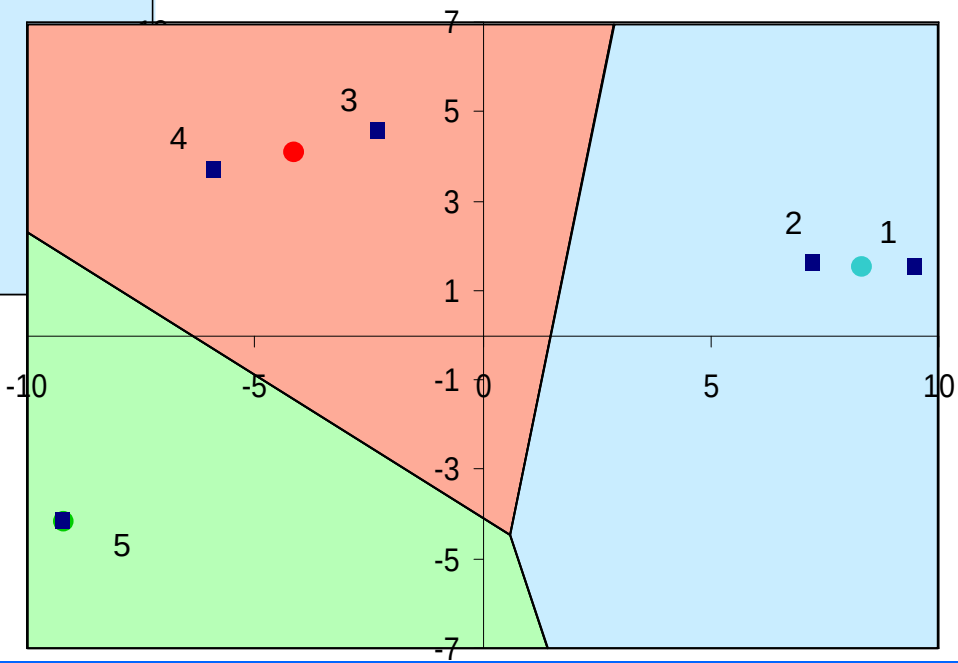
The limits of each regions are then redefined according to the same principle

2- Clustering

Divisive Cluster Analysis (DCA)



The whole process is repeated until convergence has been reached, i.e. when the assignments of each sample in the different groups no longer change.



- 3 classes are then set:
  - Class 1: sample 1 and 2
  - Class 2: sample 3 and 4
  - Class 3: sample 5

Divisive Cluster Analysis (DCA)

- Use of K-means
  - K-means: the objects (spectra) are partitionned into k clusters so that each object belongs to the cluster with the nearest mean.
  - Unsupervised method
  - Involves a number of cycles with recalculation of all means until no change in cluster assignment of all objects (spectra) is observed
  - Choice of the number of classes is critical

- Advantages
  - Usually well suited for large datasets (maps)

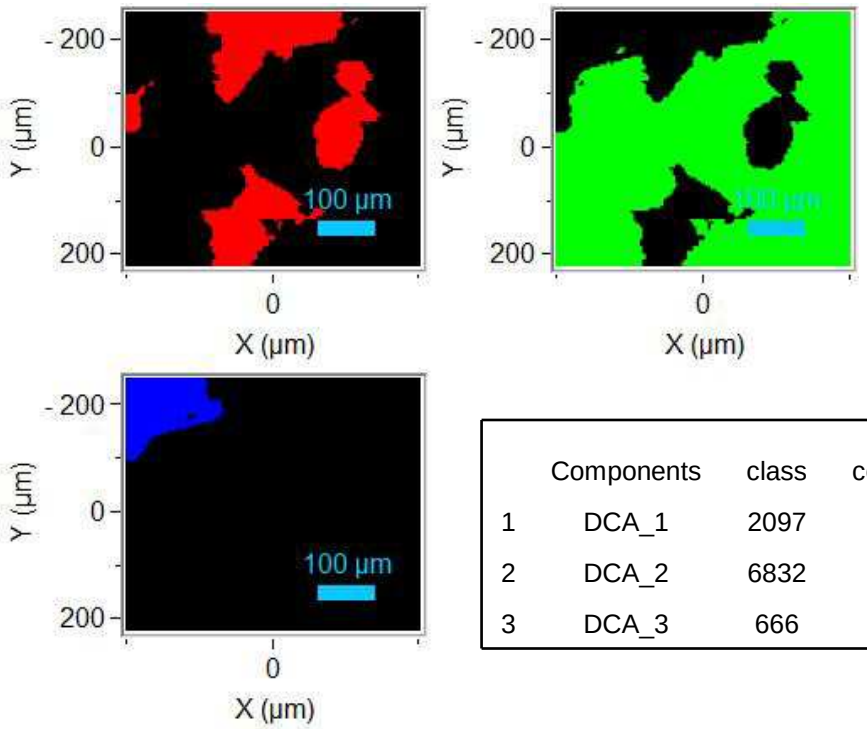
- Limitations
  - Does not produce dendrograms

2- Clustering

Divisive Cluster Analysis (DCA)

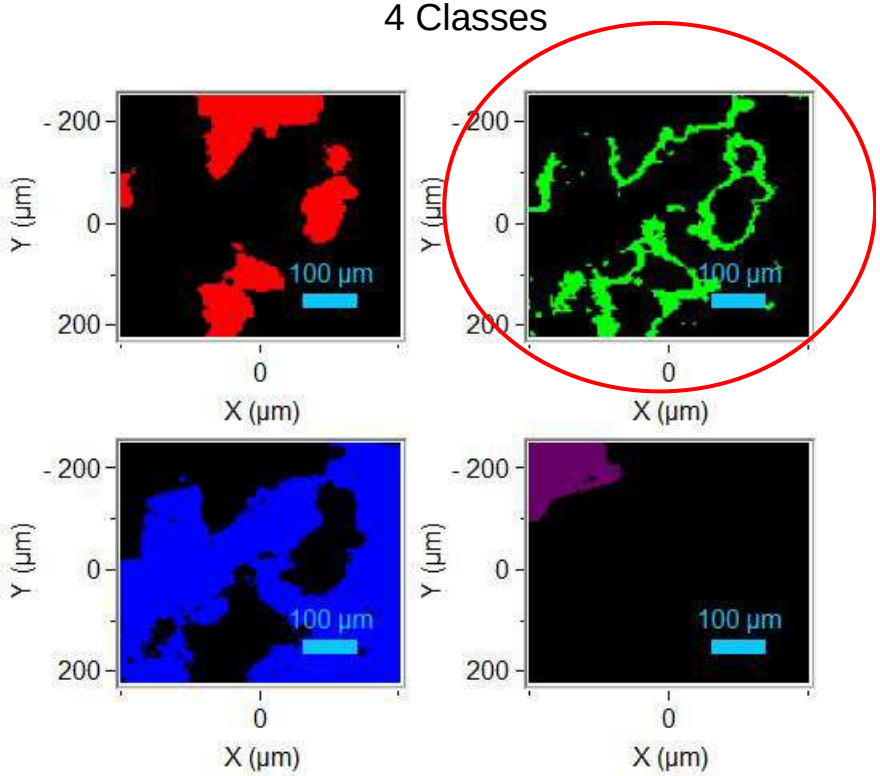
- Pharmaceutical tablet

3 Classes



	Components	class	counts %
1	DCA_1	2097	21.86
2	DCA_2	6832	71.20
3	DCA_3	666	6.94

4 Classes



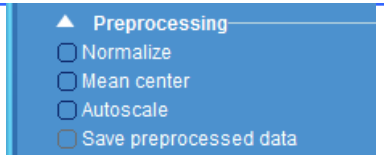
	Components	class	counts %
1	DCA_1	1760	18.34
2	DCA_2	1360	14.18
3	DCA_3	5811	60.56
4	DCA_4	664	6.92

Outline

- Introduction
- 1- Decomposition
  - Decomposition Principle
  - Principal Component Analysis (PCA) and Multivariate Curve Resolution (MCR)
- 2- Clustering
  - Hierarchical Cluster Analysis (HCA)
  - Divisive Cluster Analysis (DCA)
- 3- Data preprocessing

3- Data preprocessing

Prepocessings included in the MVA module



- Normalize
  - Each spectrum is normalized to unit area: the area under the spectral curve is then equal to 1
  - Compensates for the differences due to the measurement itself (drift in the laser power, differences of focus, etc…)
  - Applicable to most of the datasets
  - Should be preceded by baseline correction (or derivative)

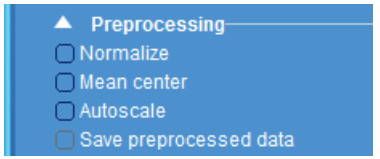
- Mean centering
  - Subtract the mean spectrum of the dataset from each single spectrum
  - Well suited when the dataset comprises similar spectra having small differences
  - Loadings not easily interpretable
  - Not usable with MCR due to the non-negativity constraint

- Autoscale
  - For each wavelength, the variables (intensities) are subtracted from their mean and divided by the standard derviation.
  - Loadings not interpretable
  - Not usable with MCR due to the non-negativity constraint

- All these preprocessing functions can be combined to add their effects.

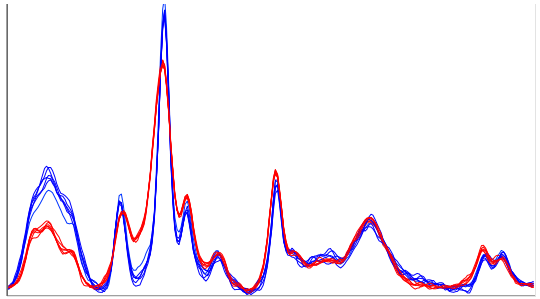
3- Data preprocessing

Prepocessings included in the MVA module

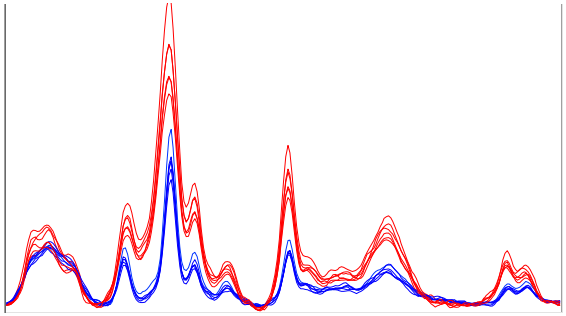


- Example: 6 capsules filled with form 1 and 6 capsules filled with form 2

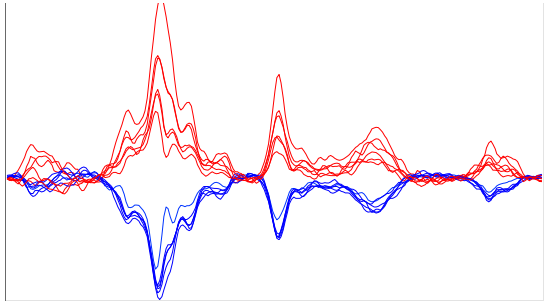
Normalized data



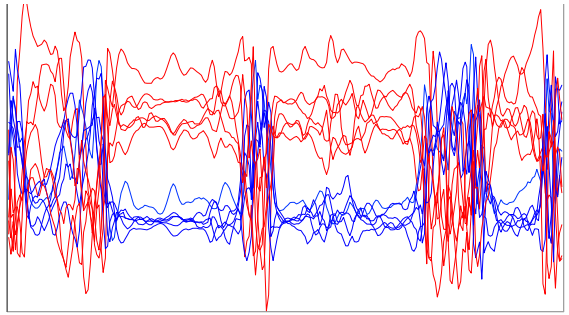
Raw data



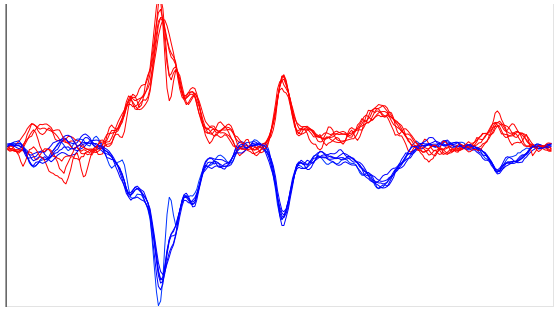
Mean centered data



Autoscaled data



Mean centered + normalized data



3- Data preprocessing

Other prepocessings

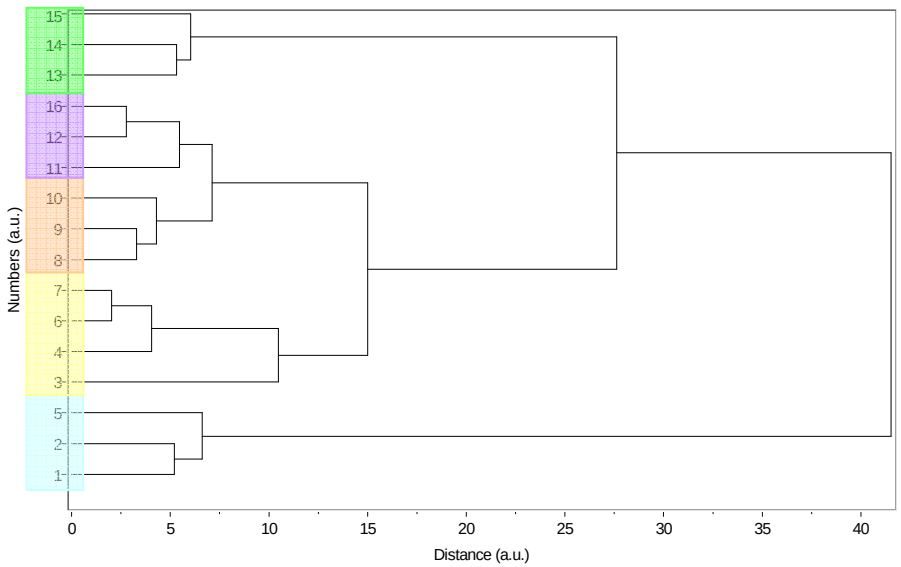
- Baseline correction
  - Should be applied when fluoresecent backgrounds are observed
- Derivative
  - Should be applied when spectral background are observed
  - Derivation enhances the spectral differences (shoulder becomes a peak) and brings constants to zero
  - Loadings are not much usable after derivation
- Normalization
  - Normalization to normal variate:
    - weighted normalization
    - for each spectrum the variables (intensities) are divided by the standard deviation after subtraction of the mean value.
    - Well suited when the dataset comprises similar spectra having small differences
- Spectral range
  - Reducing the spectral range may improve the results
    - Remove the saturated parts
    - Focus on the most interesting spectral regions

3- Data preprocessing

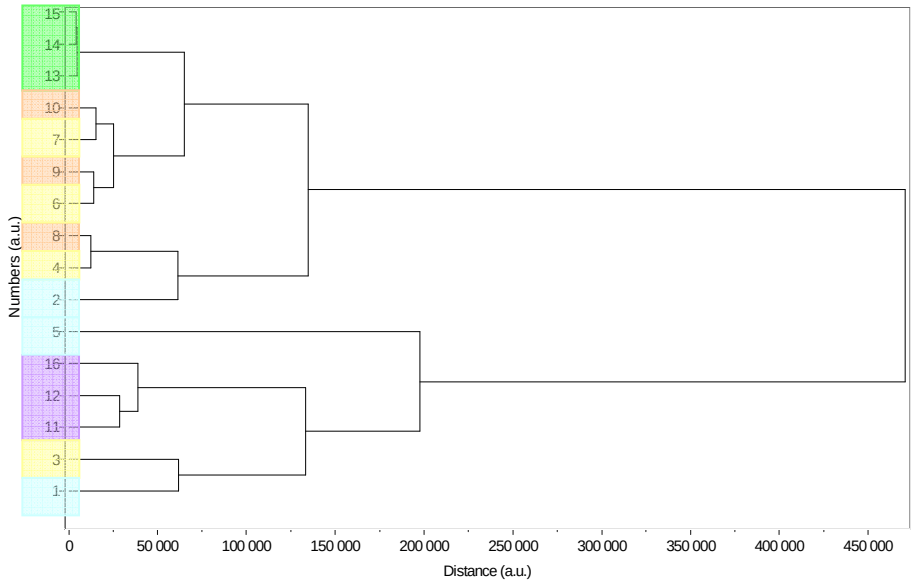
Effect on data preprocessing over the results

- Spectra of skin: importance of preprocessing

First derivative and normalization



No preprocessing



Take away message

- MVA is a very powerful tool for data treatment
  - Decomposition
  - Classification
  - Helps a lot for analyzing large datasets, improving the image rendering, finding out groups and clusters.

- But it's not a 'magic box'. Don't forget that the spectra always contain all the information and in case of doubt, you should go back to the fundamentals: spectra!

http://www.horiba.com/raman

Thank you !!