Data treatment and analysis

Vincent Larat, HORIBA Scientific

Ecole WURM de spectroscopie Raman,

Lyon, du 3 au 7 novembre 2014
Outline

- Data treatment
  - Baseline correction
  - Smoothing
  - Normalisation
  - Others

- Data evaluation
  - Univariate analysis
  - Multivariate analysis
Why data treatment?

- High Background
Why data treatment?

- Noisy data

![Graph showing Raman shift (cm\(^{-1}\)) vs. Intensity (counts)](image-url)

© 2014 HORIBA All rights reserved.
Data preprocessing

- Large number of data preprocessing exist
Outline

- Data treatment
  - Baseline correction
  - Smoothing
  - Normalisation
  - Others

- Data evaluation
  - Univariate analysis
  - Multivariate analysis
Baseline correction

■ Origin of the background

• Fluorescence
  – From the sample
  – From the impurities, solvent, other compounds

• External stray light

• Poor rejection of the laser filtering
Baseline correction

- **Linear Fit**
  - uses straight line segments to fit the data.
  - interactive positioning of the baseline points.
Baseline correction

- Polynomial fit
  - suited for curvy baselines
  - uses a polynom of degree n to fit the curve
  - the degree n depends on the shape / complexity of the curve
Baseline correction

- **Polynomial fit**
  - the result depends on the number of degrees used

![Graph showing baseline correction with polynomial fit examples]

- Too small degree
- Too high degree
Baseline correction

- Taking into account the noise
  - Noisy data influence the background
Baseline correction

- Taking into account the noise
  - Noisy data influence the background

![Graph showing baseline correction in Raman spectroscopy](image.png)
Baseline correction

Other methods

- Based on spectral derivation
  - Derivation removes offsets
  - Spectra less usable
- Using least squares
  - Often needs an explicit baseline model
- Wavelet transform / Fourier transform
- Fluorescence removal with double excitation (SERDS)
  - 2 close excitation wavelengths (separated < 1 nm) are used to get 2 Raman spectra
  - Subtracting the 2 spectra will remove the fluorescence but leave the Raman information
- Use the photobleaching to minimize / cancel the fluorescent background
Outline

- Data treatment
  - Baseline correction
  - Smoothing
  - Normalisation
  - Others

- Data evaluation
  - Univariate analysis
  - Multivariate analysis
Smoothing

**Median**

- Using a moving window of a given size, each data point is replaced by the median of all data within that window.

Original data

Original data with median smoothing (size window = 3 points).

Use of the median of the 3 points to smooth the data.
Smoothing

Polynomial

- Based on the Savitzky-Golay smoothing algorithm
- uses a polynom of degree n to fit the curve over few adjacent pixels (window) and replaces these pixels with the polynomial curve

\[ p(n) = \sum_{k=0}^{N} a_k n^k \]

- the fitting of the curve is based on the least square method

\[ E_M = \sum_{n=-M}^{N} (p(n) - x[n])^2 \]
\[ = \sum_{n=-M}^{M} \left( \sum_{k=0}^{N} a_k n^k - x[n] \right)^2. \]

- The smaller the degree and the higher the window size, the more significant is the smoothing
Smoothing

- Polynomial

Small window; 1 pt

Large window; 10 pts
Smoothing

- Fourier transform
  - Conversion from wavenumber space to frequency space by applying FT
  - Removing the high frequency components (corresponding to noise)
  - Applying reverse FFT to get back to wavenumber space

- Specific algorithms: ex: Denoiser
  - Detection of noise on every pixel using a moving window (size adjustable by the user)
  - Smooth of the data using a quadratic function within a self-adapting moving window:
    - If a peak is detected the window reduces to minimize the effect of smoothing on peaks (change in shapes, position, width)
    - Over areas with no detected peaks, the window is larger
Smoothing

Comparison

Raman shift (cm⁻¹)

Intensity (counts)

500 1 000 1 500 2 000 2 500 3 000

0 500 1 000 1 500 2 000 2 500

Original

Denoise, 9 pts

FFT

Median, 9 pts

Polynomial, 9 pts
Smoothing

- Comparison
  - Alteration of peak widths and intensities

![Graph showing different smoothing techniques: Original, Denoise, 9 pts, FFT, Median, 9 pts, Polynomial, 9 pts.](image)
Smoothing

- Smoothing of images
Normalisation

- Spectra corresponding to the measurements of 2 sets of capsules having different API forms (form I, form II)

Raw data

Normalized data
Normalisation

Goal of the normalisation: to compensate from variations induced by the measurement itself:
- Variation of focus during maps (surface not flat, defocusing)
- Variation of laser power
- Global loss of intensity when going in depth

Or to monitor the relative variations of spectral feature
- Normalization over a given peak

Types of normalisation
- Over an area (full spectrum or a peak)
- Min-Max normalization
- SNV – standard normal variate
- Scattering correction
Normalisation

- Map of a cross section of a multilayer film
  - Monitoring of 2 different spectral features before and after normalization of the whole spectrum.
Others

- **Despike**
  - Multiple acquisitions
  - Smoothing function (median) can be applied
  - Despike algorithm (based on peak width recognition)

![Despike example graphs](image-url)
Others

- **Deconvolution**
  - Choice in the peak shape type (gaussian, lorentzian, asymertric, etc…)
  - Possibility to fix / limit some parameters
  - A priori knowledge of the composition of the sample is recommended
Outline

- Data treatment
  - Baseline correction
  - Smoothing
  - Normalisation
  - Others

- Data evaluation
  - Univariate analysis
  - Multivariate analysis
Univariate Analysis

- Univariate analysis
- Use a single parameter to plot an image
  - Peak height / amplitude
  - Peak area
  - Peak position
  - Peak width (FWHM)
Monitoring the Height/Area of peaks that can be assigned to a single compound.
Univariate Analysis

- Monitoring the peak shift (position) to map the strains

Figure 1. Color-coded Raman map showing the pattern of strain-induced up-shift of the main LO=TO diamond Raman band around a larnite (β-Ca₂SiO₄) inclusion in a diamond from the Kankan district, Guinea. Observed Raman shifts as high as 1338 cm⁻¹ indicate remnant pressures up to 2.7 GPa [3].

Univariate Analysis

- Monitoring the peak width to map the crystallinity

- Metamictization of Zircon (gradual destruction of the crystalline structure, turning to amorphous)
Univariate Analysis

- Exemple of the Silicon with combined maps
Outline

Data treatment
- Baseline correction
- Smoothing
- Normalisation
- Others

Data evaluation
- Univariate analysis
- Multivariate analysis
  - Decomposition
  - Clustering
  - Fitting procedures (supervised)
Multivariate analysis

- Large number of algorithms available
Multivariate analysis

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):

  \[ \text{Spectrum 1} = \text{Score}_{1,1} \times \text{Loading 1} + \text{Score}_{1,2} \times \text{Loading 2} + \ldots + \text{Score}_{1,n} \times \text{Loading n} \]

  \[ \text{Spectrum 2} = \text{Score}_{2,1} \times \text{Loading 1} + \text{Score}_{2,2} \times \text{Loading 2} + \ldots + \text{Score}_{2,n} \times \text{Loading n} \]

  \[ \ldots \]

  \[ \text{Spectrum n} = \text{Score}_{n,1} \times \text{Loading 1} + \text{Score}_{n,2} \times \text{Loading 2} + \ldots + \text{Score}_{n,n} \times \text{Loading n} \]

- 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).

  \[ S_i = \sum_j s_{ij} \times L_j + E \]

  residuals
Multivariate analysis

Decomposition principle

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

<table>
<thead>
<tr>
<th>Components</th>
<th>% variations</th>
<th>accum. % variation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>PCA_1</td>
<td>96.03</td>
</tr>
<tr>
<td>2</td>
<td>PCA_2</td>
<td>3.79</td>
</tr>
<tr>
<td>3</td>
<td>PCA_3</td>
<td>0.04</td>
</tr>
<tr>
<td>4</td>
<td>PCA_4</td>
<td>0.04</td>
</tr>
<tr>
<td>5</td>
<td>PCA_5</td>
<td>0.02</td>
</tr>
<tr>
<td>6</td>
<td>PCA_6</td>
<td>0.02</td>
</tr>
</tbody>
</table>

Loading 1: 96.03 % of the variance

Loading 2: 3.79 % of the variance

Loading 6: 0.02 % of the variance
Decomposition principle

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

Spectrum 1 from the dataset

\[
\text{Score 1,1} = 2.99^* + \text{Loading 1} + (-0.50)^* + \text{Loading 2} + \ldots
\]

Spectrum 7 from the dataset

\[
\text{Score 7,1} = 3.81^* + \text{Loading 1} + 0.61^* + \text{Loading 2} + \ldots
\]
Multivariate analysis

PCA and MCR

- **PCA**
  - **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 1st 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 mappings

- **MCR**
  - **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 mappings  
    - When interpretable components are required  
    - MCR component can be searched in libraries.
PCA and MCR

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

- Loadings
Multivariate analysis

PCA and MCR

**PCA: Score plots**

1. 

   ![PCA Score Plot 1](image1)

   ![PCA Score Plot 2](image2)

   ![PCA Score Plot 3](image3)

**MCR: Score plots**

1. 

   ![MCR Score Plot 1](image4)

   ![MCR Score Plot 2](image5)

   ![MCR Score Plot 3](image6)

<table>
<thead>
<tr>
<th>Components</th>
<th>% variations</th>
<th>accum. % variation</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCA_1</td>
<td>85.26</td>
<td>85.26</td>
</tr>
<tr>
<td>PCA_2</td>
<td>8.31</td>
<td>93.56</td>
</tr>
<tr>
<td>PCA_3</td>
<td>4.08</td>
<td>97.64</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Loadings</th>
<th>% variations</th>
<th>accum. % variation</th>
</tr>
</thead>
<tbody>
<tr>
<td>MCR_1</td>
<td>9.80</td>
<td>9.80</td>
</tr>
<tr>
<td>MCR_2</td>
<td>19.71</td>
<td>29.51</td>
</tr>
<tr>
<td>MCR_3</td>
<td>68.13</td>
<td>97.64</td>
</tr>
</tbody>
</table>
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 (HCA)
  - Divisive clustering (DCA)
  - ....
Hierarchical Cluster Analysis (HCA)

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)

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

<table>
<thead>
<tr>
<th>Components</th>
<th>Class</th>
<th>Counts</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>HCA_1</td>
<td>3</td>
<td>3</td>
<td>18.75</td>
</tr>
<tr>
<td>HCA_2</td>
<td>4</td>
<td>3</td>
<td>25.00</td>
</tr>
<tr>
<td>HCA_3</td>
<td>6</td>
<td>3</td>
<td>37.50</td>
</tr>
<tr>
<td>HCA_4</td>
<td>3</td>
<td>3</td>
<td>18.75</td>
</tr>
</tbody>
</table>

Class 1
- Clean skin
- Class 2
- Treated with cream and washed
- Treated with shampoo and washed
- Treated with silicone and washed

Class 4
- Clean skin other person
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)

Partitioning processes are used to divide the space and group each sample into groups. 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)

- Pharmaceutical tablet

3 Classes

<table>
<thead>
<tr>
<th>Components</th>
<th>class</th>
<th>counts</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>DCA_1</td>
<td>1</td>
<td>1760</td>
<td>18.34</td>
</tr>
<tr>
<td>DCA_2</td>
<td>2</td>
<td>1360</td>
<td>14.18</td>
</tr>
<tr>
<td>DCA_3</td>
<td>3</td>
<td>5811</td>
<td>60.56</td>
</tr>
<tr>
<td>DCA_4</td>
<td>4</td>
<td>664</td>
<td>6.92</td>
</tr>
</tbody>
</table>

4 Classes

<table>
<thead>
<tr>
<th>Components</th>
<th>class</th>
<th>counts</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>DCA_1</td>
<td>1</td>
<td>2097</td>
<td>21.86</td>
</tr>
<tr>
<td>DCA_2</td>
<td>2</td>
<td>6832</td>
<td>71.20</td>
</tr>
<tr>
<td>DCA_3</td>
<td>3</td>
<td>666</td>
<td>6.94</td>
</tr>
</tbody>
</table>

© 2014 HORIBA All rights reserved.
Fitting procedure (least squares)

- Supervised method. The pure compounds are known and the least squares procedure finds the linear combination (coefficients) to fit the raw data as best as possible.

This method gives quantitative indications
Fitting procedure (least squares)

- Mixture of inks. Each pure material is known and CLS fitting can be applied.
- Important to have **all** the pure compounds otherwise the fitting is not correct and the information are biased.
Example

Example of data treatment and evaluation of a map