Data Preprocessing

D.N. Rutledge, AgroParisTech

12  Rencontres Héliospir / 30 septembre 2011
Outline

• Zone selection
• Examining raw data
• The importance of pre-treatment of data
• Common pre-treatment methods
Selection of proper frequency ranges

![Graph showing wavenumber vs. AU with highlighted areas for total absorption and spectral noise.](image-url)
Selection of proper frequency ranges

- total absorption; absorbance > 2.5 AU
- spectral noise
Raw data check

The black, red and blue curves indicate different concentration levels

Baseline offset
Raw data check

The black, red and blue curves indicate different concentration levels.

Baseline offset and global intensity variations
PCA & PLS on Raw data

- It is not easy to separate three concentration levels
- Need to correct the spectra
Preliminary conclusion

• Pre-treatment of data is crucial

• But it is not always simple ...
Common pre-treatment methods

- Baseline correction
  - Offset
  - Detrend
  - Spline
  - MSC and EMSC

- Scale correction
  - Standard Normal Variates (snv)
  - MinMax
  - Log
Common pre-treatment methods

• Data enhancement
  – Centering
  – Standardising
  – 1\textsuperscript{st} & 2\textsuperscript{nd} order Derivatives
  – Smoothing

• Orthogonalisation
  – Direct Orthogonalisation
  – O-PLS
  – OSC
  – DOSC
  – ...

Offset correction

• Subtract linear baseline from each signal
  ▪ Intensity of lowest point
  ▪ Intensity of a user-chosen point
  ▪ Intensities calculated between 2 points
Baseline correction

Offset correction

Offset correction based on data points 159 & 385
Baseline correction

**Detrend correction**

- Subtract $2^{nd}$ degree polynomial baseline from signals
  - Automatically calculated from data points
Baseline correction

Detrend correction

Offset correction using "Detrend"
Baseline correction

**Spline correction**

- Subtract a cubic piece-wise polynomial baseline from each signal
  - Requires input of a series of spline nodes
  - Delicate choice with important consequences!
Baseline correction

Spline correction

Offset correction using "Spline" based on data points 1, 159, 385, 460, 510.
Baseline correction methods

Offset correction based on minimum value

Offset correction based on data points 159 & 365

Offset correction using "Detrend"

Offset correction using "Spline" based on data points 1139386, 460, 510
Baseline correction

- Only corrects for linear & non-linear baseline shifts
  - Does not correct for global intensity variations
Standard Normal Variates (SNV)

- Subtract the mean for each spectrum $i$
- Then divide by its standard deviation:

$$x_{ik}^{SNV} = (x_{ik} - m_i) / s_i$$

- SNV is a *baseline* and a *quantity* correction method
Scale correction

SNV correction
Scale correction

**PCA & PLS on SNV-corrected data**

- It is easier to separate three concentration levels
Problems with SNV

- In some cases, global intensity variations are interesting!
- SNV enhances noisy signals
Problems with SNV

- Can change relations between peaks
- If prior to SNV one peak varies, after SNV all peaks vary
Subtract first signal

- Highlights *evolution* of signals
- Not often used, but can be very interesting
- Increases apparent noise level
Column Centering

- Often used
- Enhances differences among samples
- Increases apparent noise level
Column Centering and Scaling

- Gives equal importance to all parts of signals
  - Both peaks and baseline
  - Makes results difficult to interpret spectrally
PCA & PLS on centred and on scaled data

- Scaled data noisier
- More difficult to interpret
- But multivariate data analysis results are better

Signal Enhancement
Derivatives

- Computing derivatives of various orders is a classical technique widely used for spectroscopic applications.

- Information in a spectrum may be more easily revealed when working on a 1st or 2nd order derivative.
2\textsuperscript{nd} Derivative is preferred

- 2\textsuperscript{nd} derivatives is the most common preprocessing
- Removes background drift due to scattering
- Can help resolve nearby peaks
- Peak positions are at the same place as in the original spectra.
- Can improve spectral resolution:

Invisible Peaks at 0.0 and 1.3 2\textsuperscript{nd} derivative
Signal Enhancement

Example: NIR Spectrum of Coal

![Graph showing absorption vs. wavenumber for original and second derivative spectra of a coal sample. The graph illustrates the absorption peaks and troughs at different wavenumbers, with enhanced signal visibility after applying a second derivative technique.]
Savitsky-Golay Derivatives

Windows size: Noise decrease vs. loss of resolution
Signal Smoothing

- Reduces effects of random noise

- Several algorithms:
  - Boxcar smoothing
  - Savitsky-Golay polynomial smoothing
  - PCA smoothing
Signal Enhancement

**Savitsky-Golay smoothing**

Windows size: Noise decrease *vs.* loss of resolution

![Graph showing 2° derivative Moving Average smoothed (Window=21)]
Signal Enhancement

PCA smoothing

2° derivative smoothed using PCA (1 to 2)


Othogonalisation

- Eliminate variability in signals not related to studied factor

- Eliminate that part of $X$ which is orthogonal to $y$
  - Direct Othogonalisation
  - O-PLS
  - OSC
  - DOSC
  - ...

Othogonalisation
Direct Orthogonalisation

- Calculate space orthogonal to $y = y_o$
- Project $X$ onto $y_o = X_o$
- Do PCA on $X_o = T_o$ and $P_o$
- Use $T_o$ and $P_o$ to calculate interesting part of $X_o = X_o'$
- $X_{DO} = X - X_o'$
Direct Orthogonalisation

DO on SNV data
Orthogonalisation

PCA & PLS on DO-corrected data

- It is easier to separate the three concentration levels
- Need to determine optimal number of PCs for DO
Orthogonalisation

Orthogonal-PLS

- Do PLS between $X$ and $y$
- Calculate $w$, $t$, $p$
- Project $w$ orthogonal to $p = w_o$
- Use $w_o$ to calculate orthogonal part of $t$ and $p = t_o, p_o$
- Use $t_o$ and $p_o$ to calculate orthogonal part of $X = X_o$
- $X_{O-PLS} = X - X_o'$
Orthogonalisation

O-PLS

O-PLS on SNV data
Orthogonalisation

**PCA & PLS on O-PLS data**

- It is easier to separate the three concentration levels
- Need to determine optimal number of LVs for O-PLS and for PLS!
- No real improvement in the model, just in its interpretability
Can pretreatment of spectra improve regression models?

<table>
<thead>
<tr>
<th>Preprocessing</th>
<th>SSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>289.7</td>
</tr>
<tr>
<td>SNV</td>
<td>0.103</td>
</tr>
<tr>
<td>SNV-2 Deriv.</td>
<td>0.005</td>
</tr>
<tr>
<td>SNV-Centering</td>
<td>0.103</td>
</tr>
<tr>
<td>SNV-Center/Std</td>
<td>0.053</td>
</tr>
<tr>
<td>SNV-DO</td>
<td>0.013</td>
</tr>
<tr>
<td>SNV-OPLS</td>
<td>0.103</td>
</tr>
</tbody>
</table>
Conclusions

- Pretreatments can eliminate interferences
- Pretreatments can facilitate extraction of information
- The optimal pretreatment depends on the data
Reference

- M. Zeaiter, D. N. Rutledge
  Chapter 2: “Preprocessing”
  Section: “Linear Regression Modeling”
  in “Comprehensive Chemometrics”, Elsevier 2009