Data Preprocessing

D.N. Rutledge, AgroParisTech

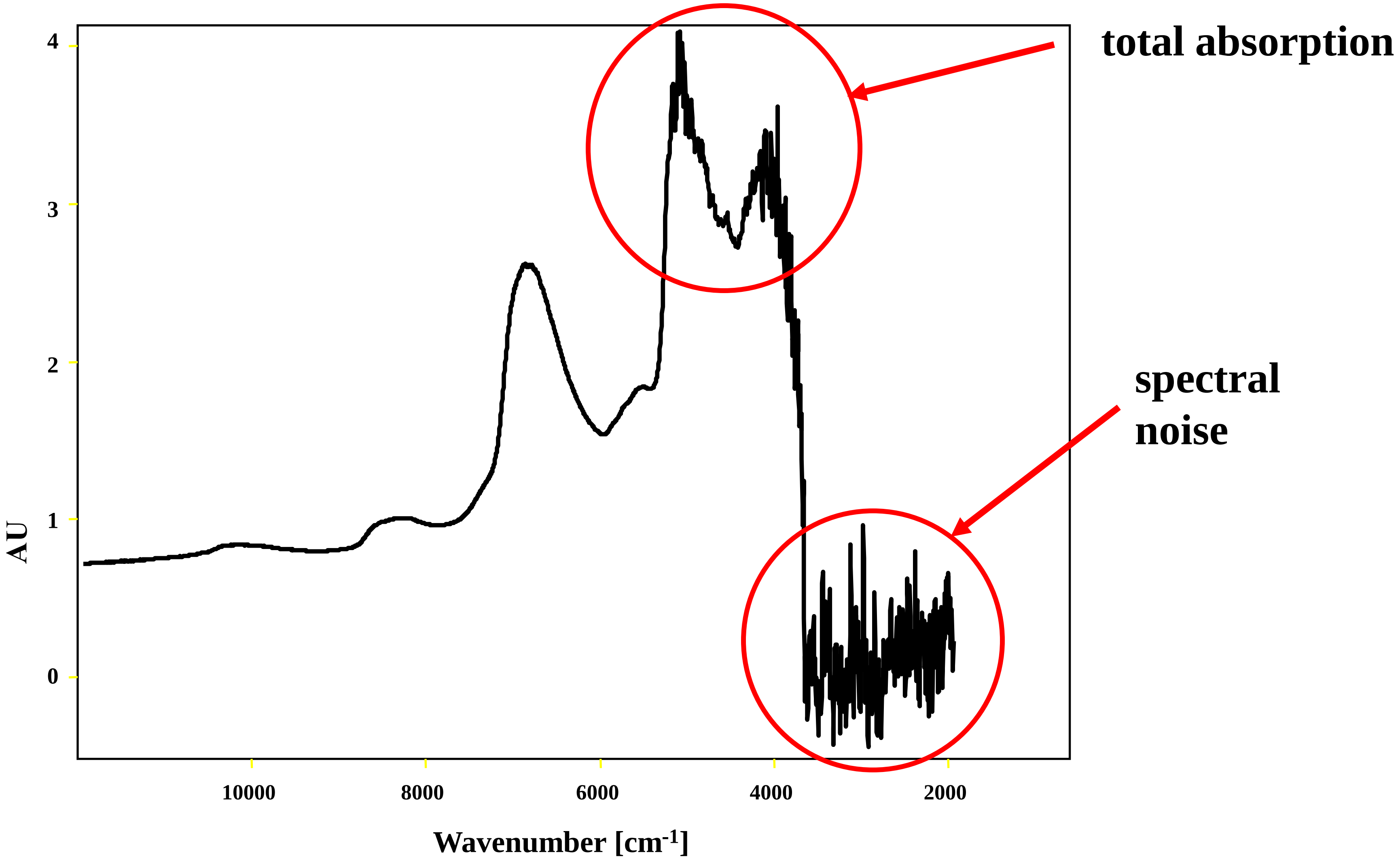
12 Rencontres Héliospir / 30 septembre 2011

Outline

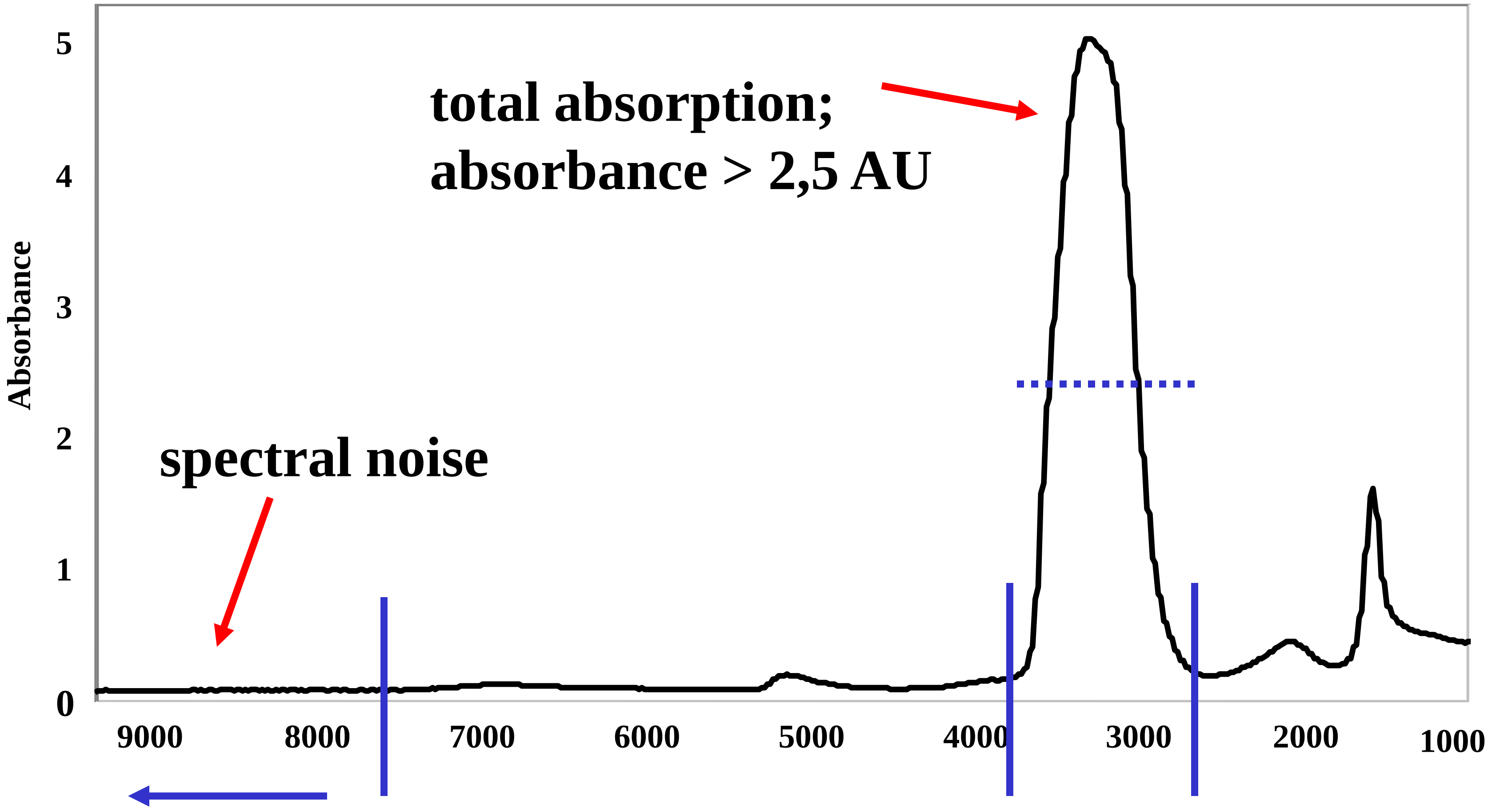
- Zone selection
- Examining raw data
- The importance of pre-treatment of data

- Common pre-treatment methods

Selection of proper frequency ranges

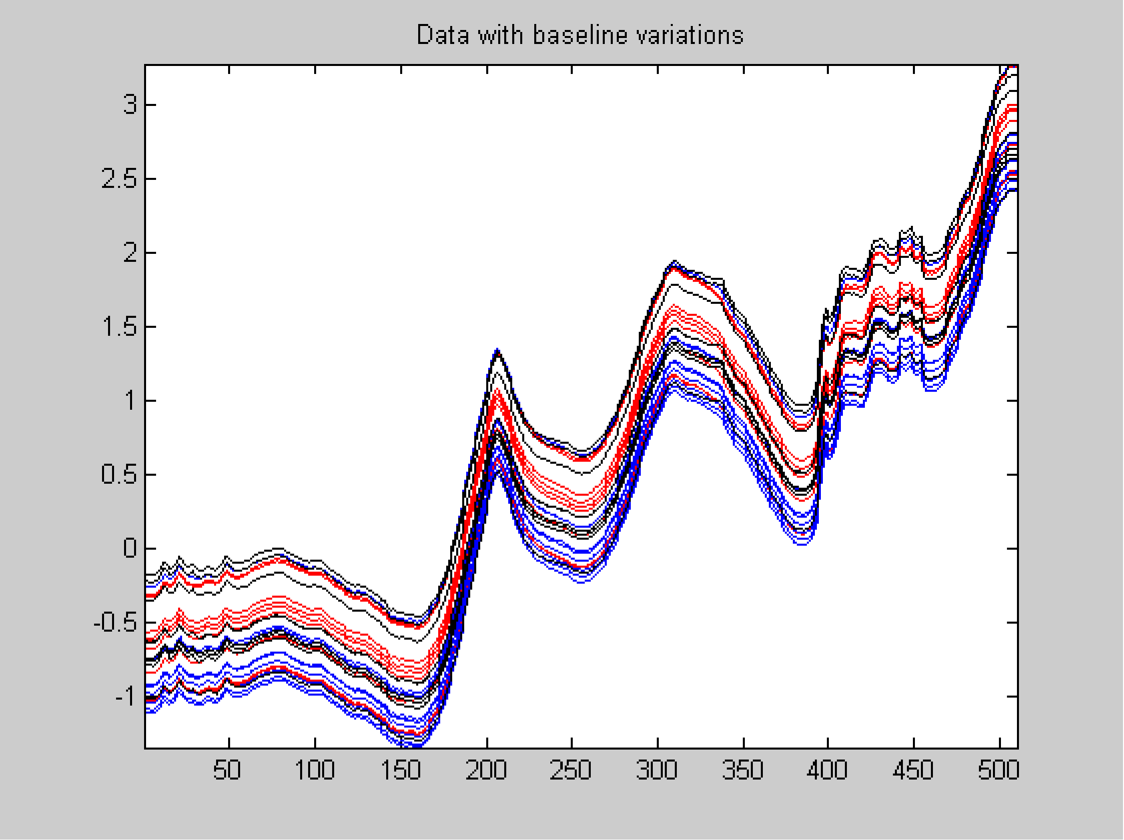


Selection of proper frequency ranges



Raw data check

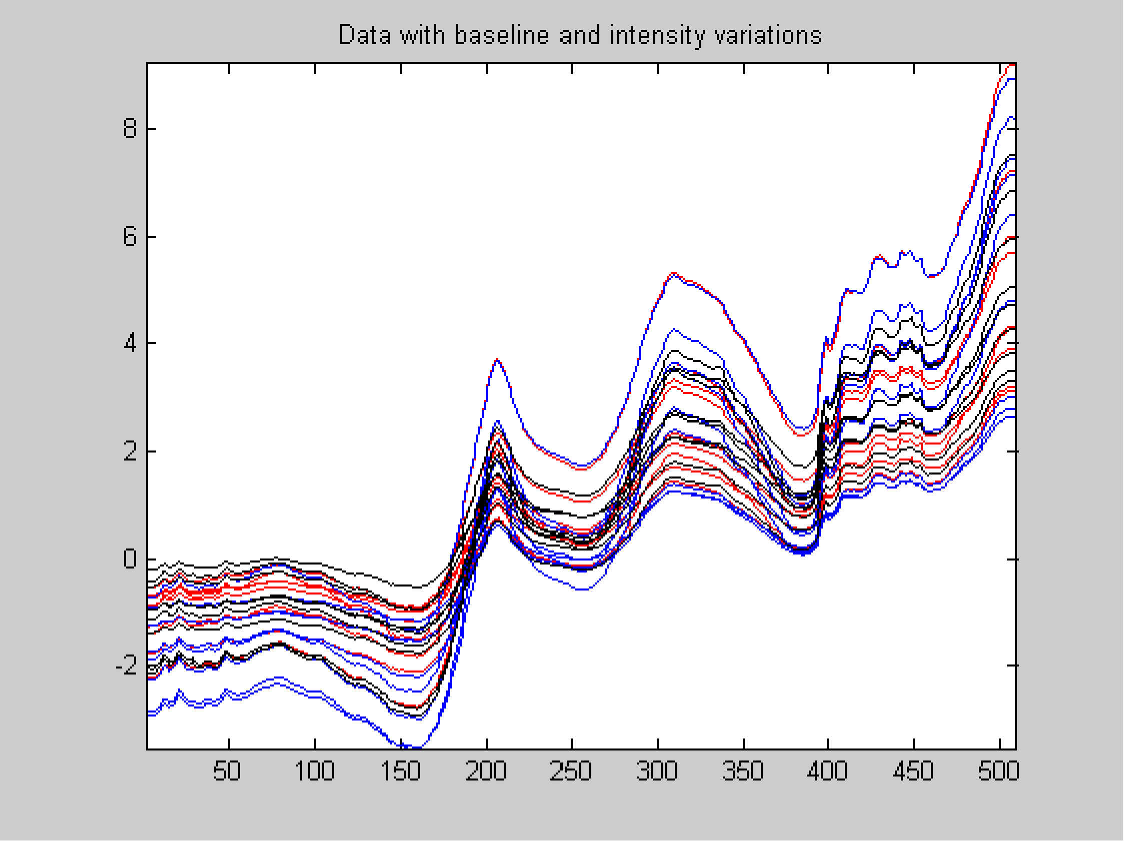
The black, red and blue curves indicate different concentration levels



Baseline offset

Raw data check

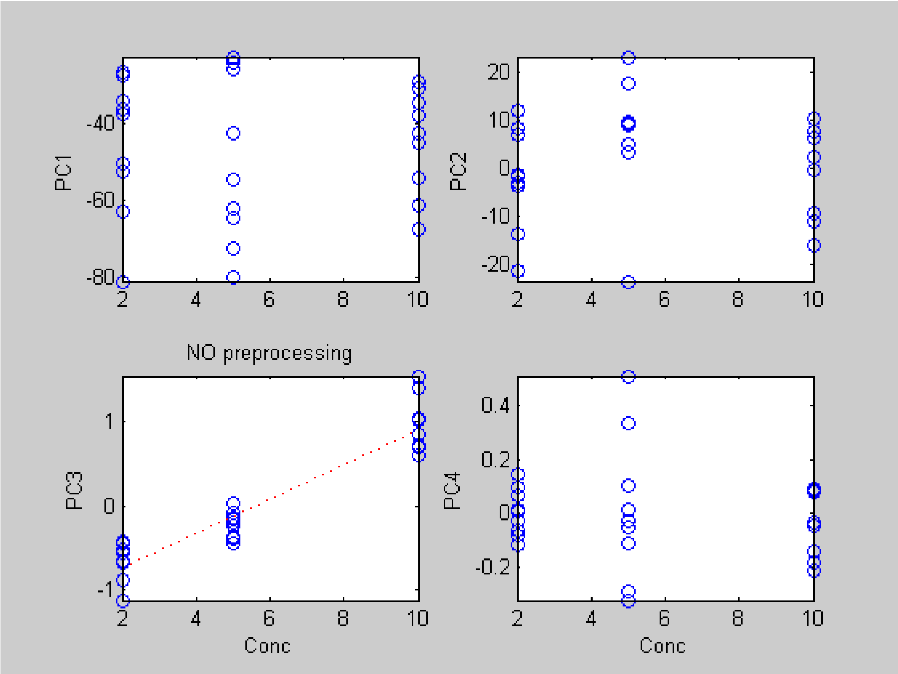
The black, red and blue curves indicate different concentration levels

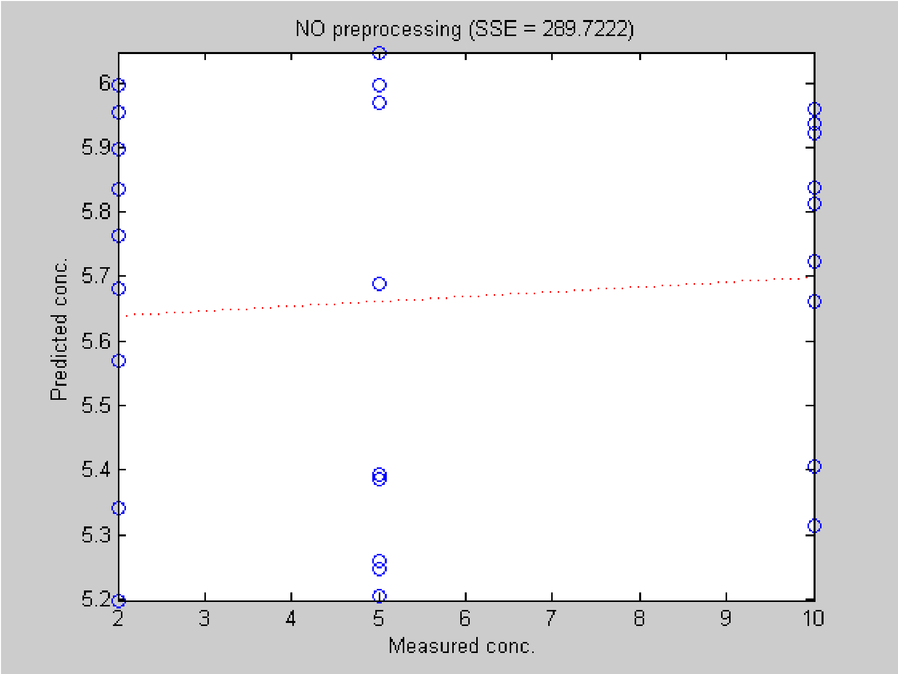


Baseline offset and global intensity variations

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
  - 1st & 2nd order Derivatives
  - Smoothing

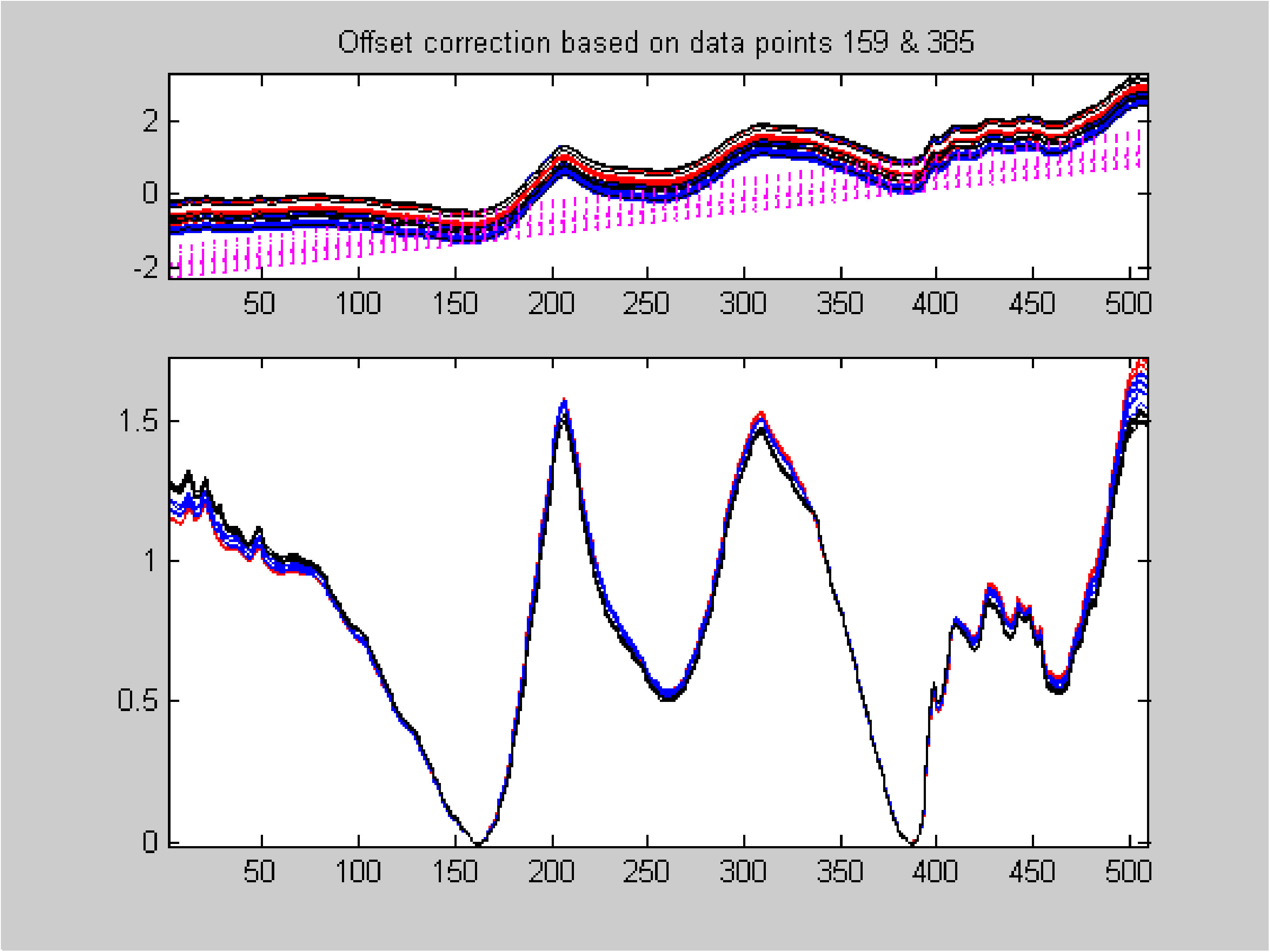
- Orthogonalisation
  - Direct Othogonalisation
  - O-PLS
  - OSC
  - DOSC
  - …

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

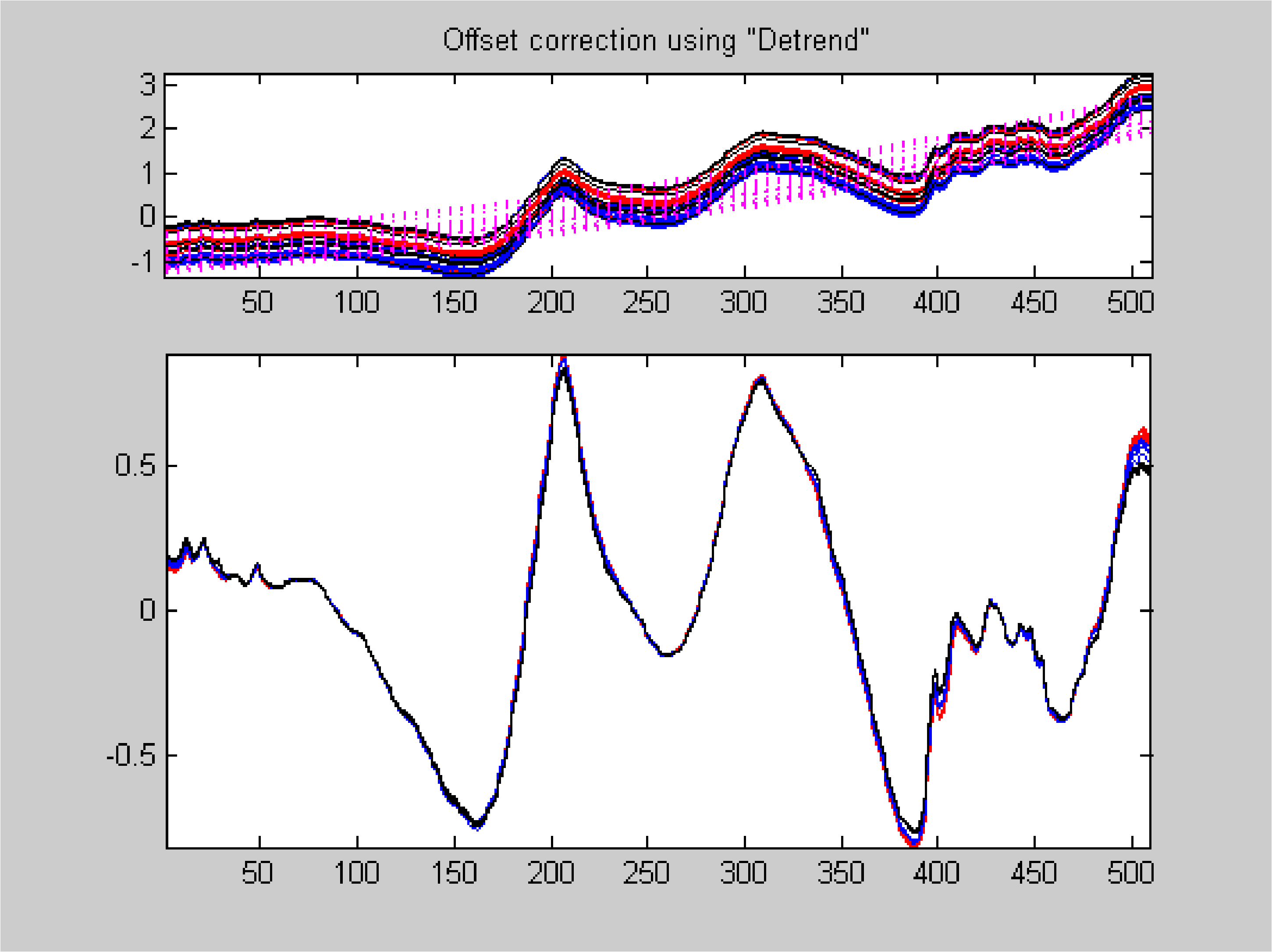


Detrend correction

- Subtract 2nd degree polynomial baseline from signals
  - Automatically calculated from data points

Baseline correction

Detrend correction

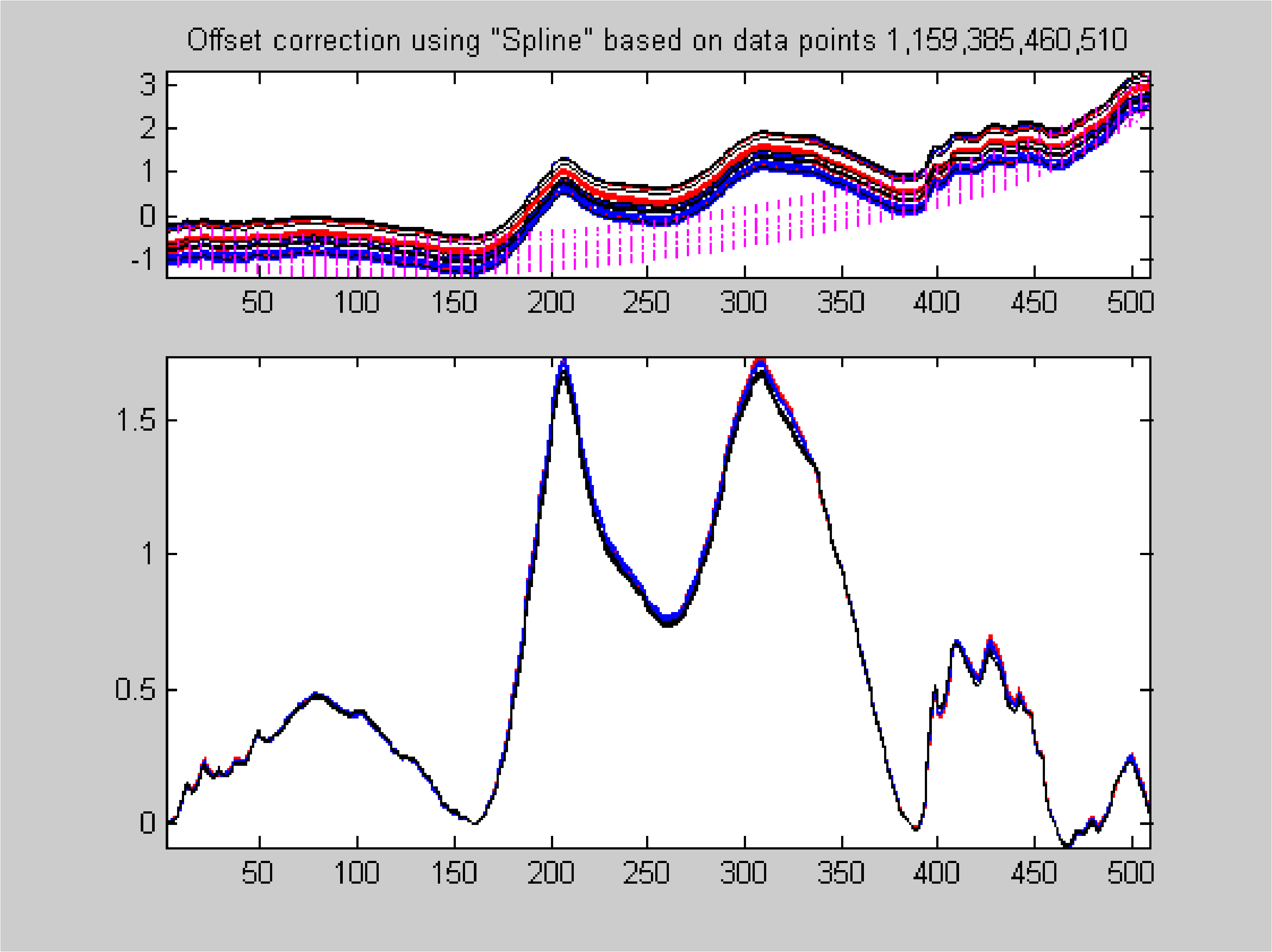


Spline correction

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

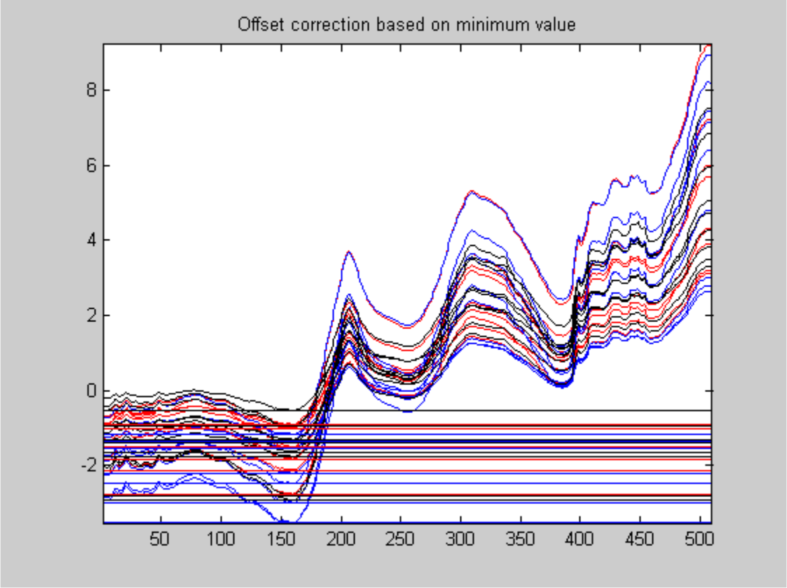
Baseline correction

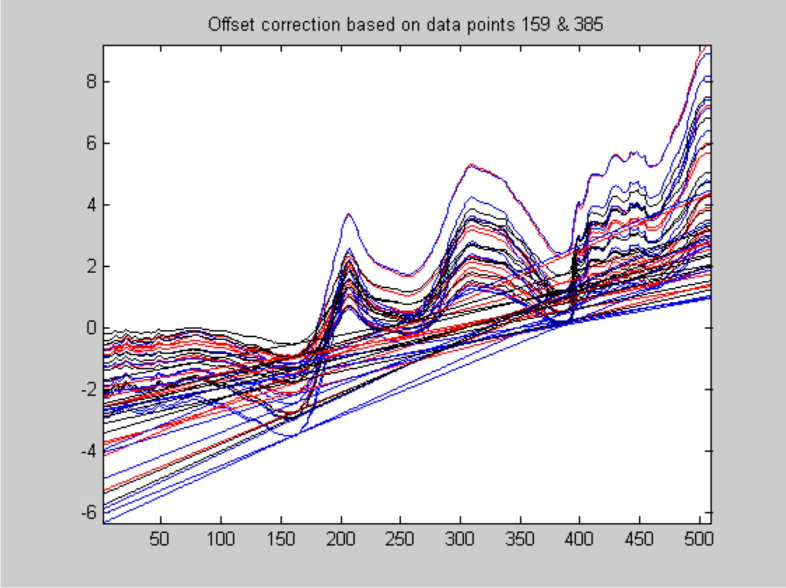
Spline correction

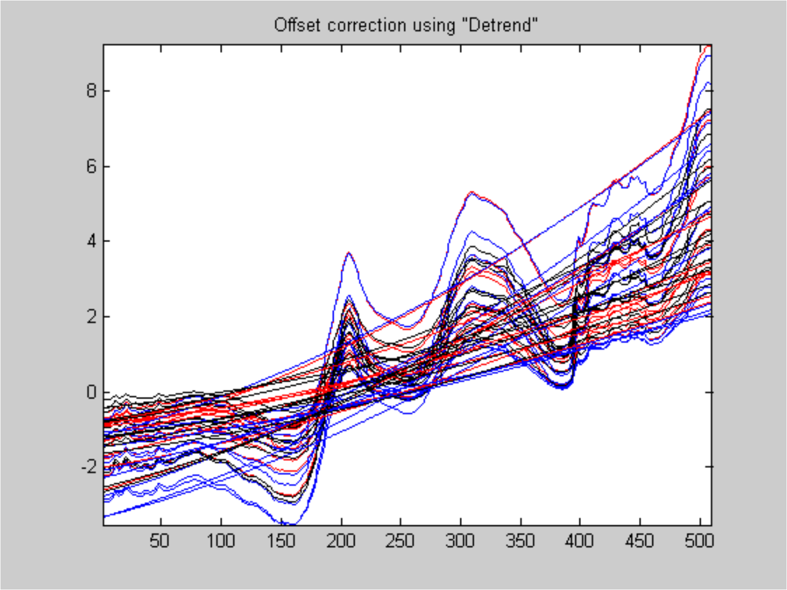


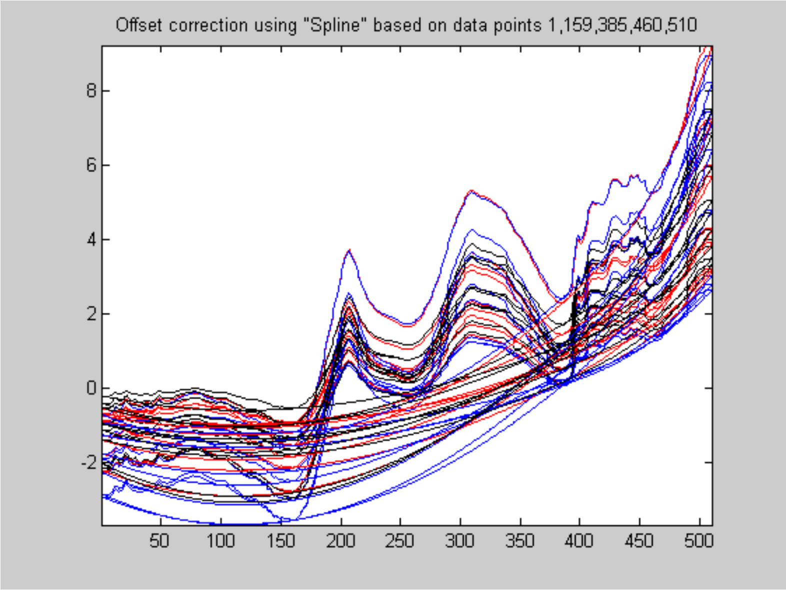
Baseline correction

Baseline correction methods



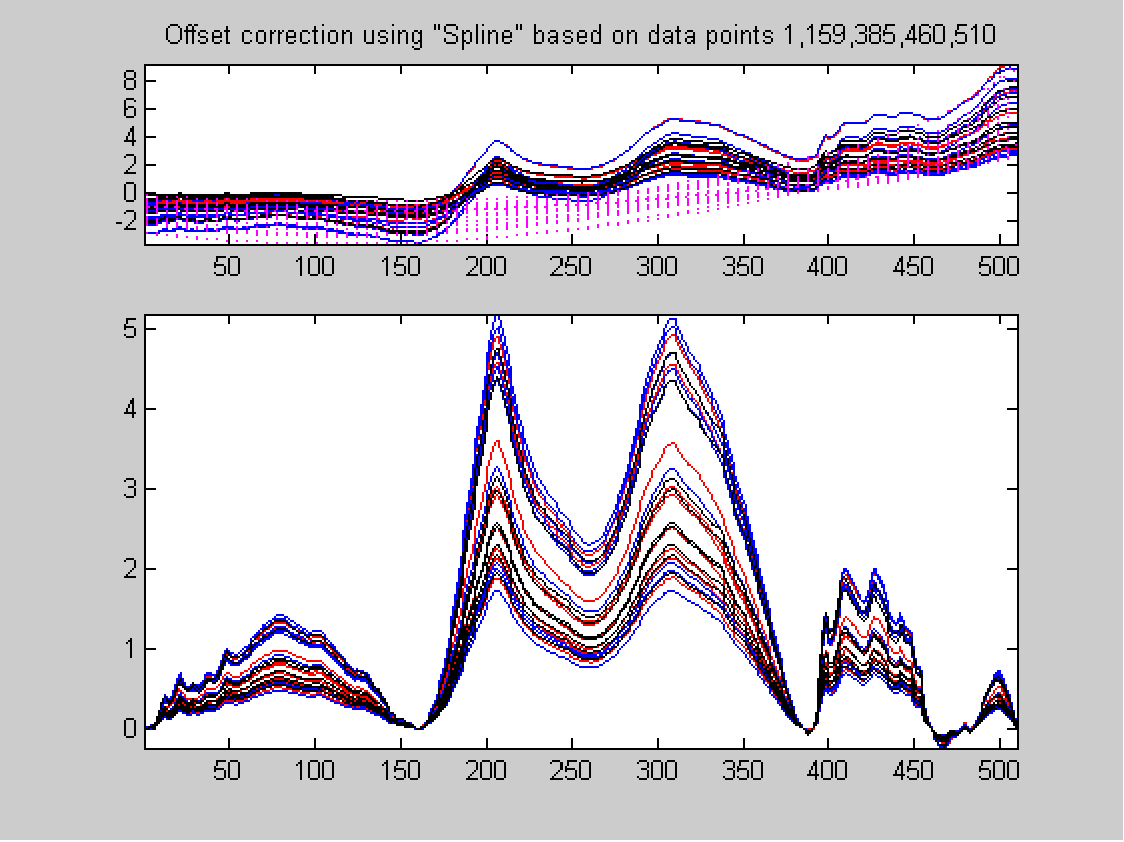






Baseline correction

- 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*

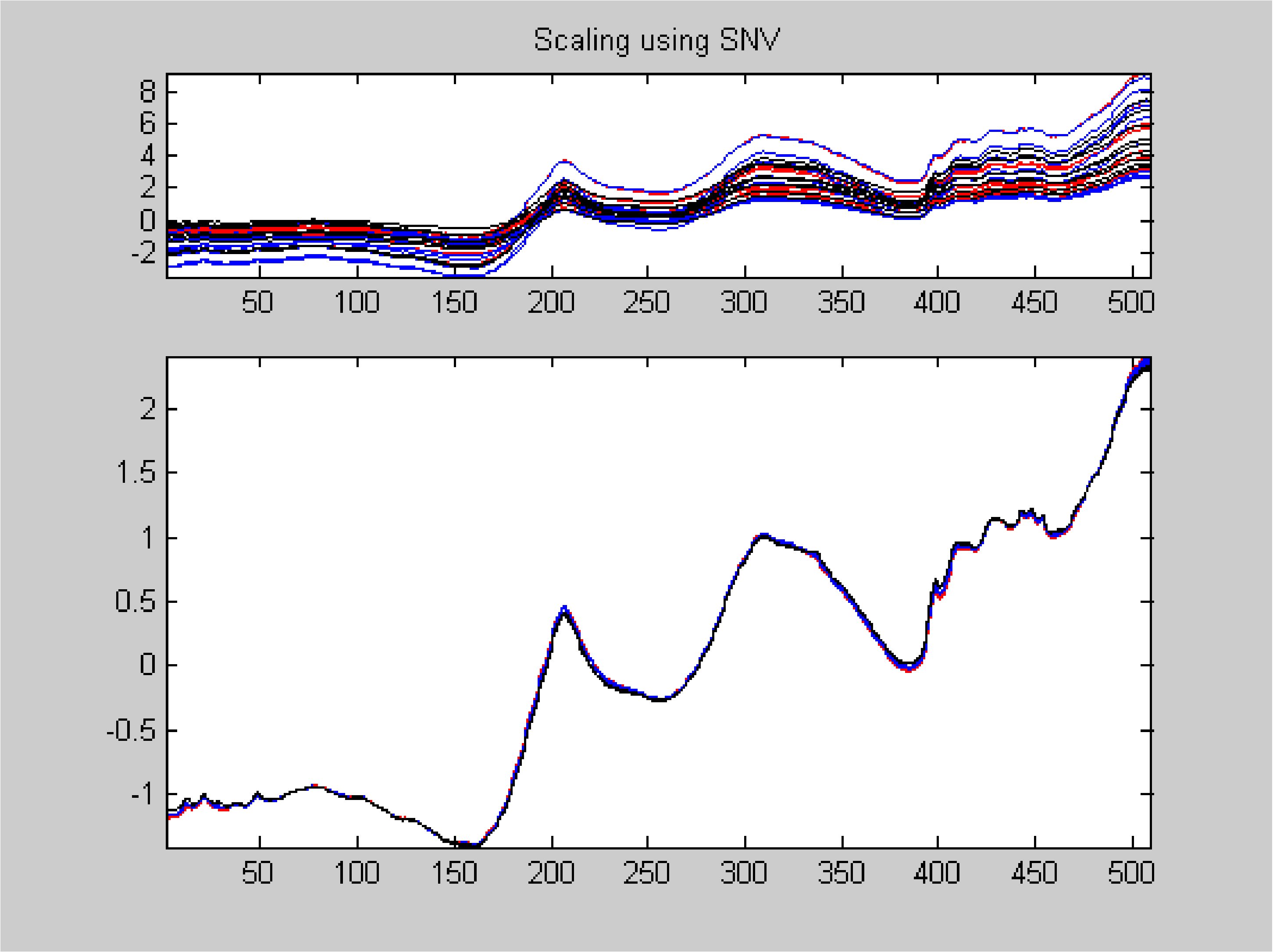
- Then divide by its standard deviation :

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

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

Scale correction

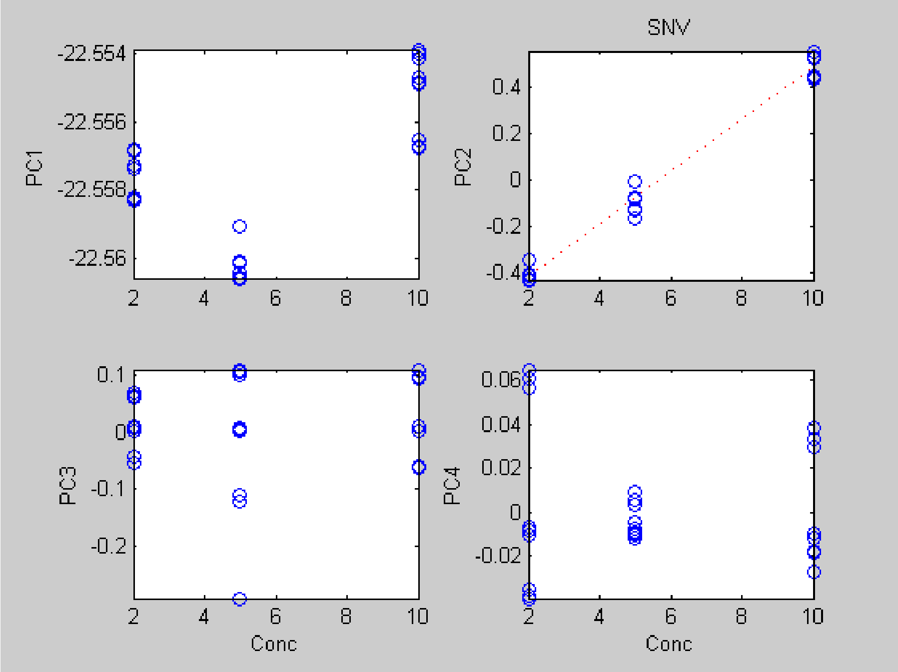
SNV correction

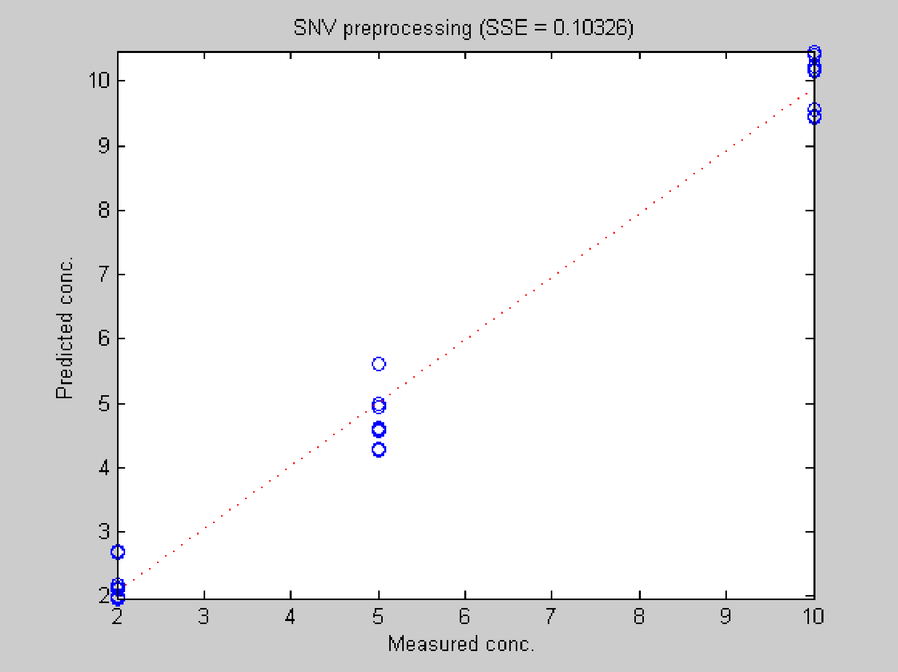


Scale correction

PCA & PLS on SNV-corrected data

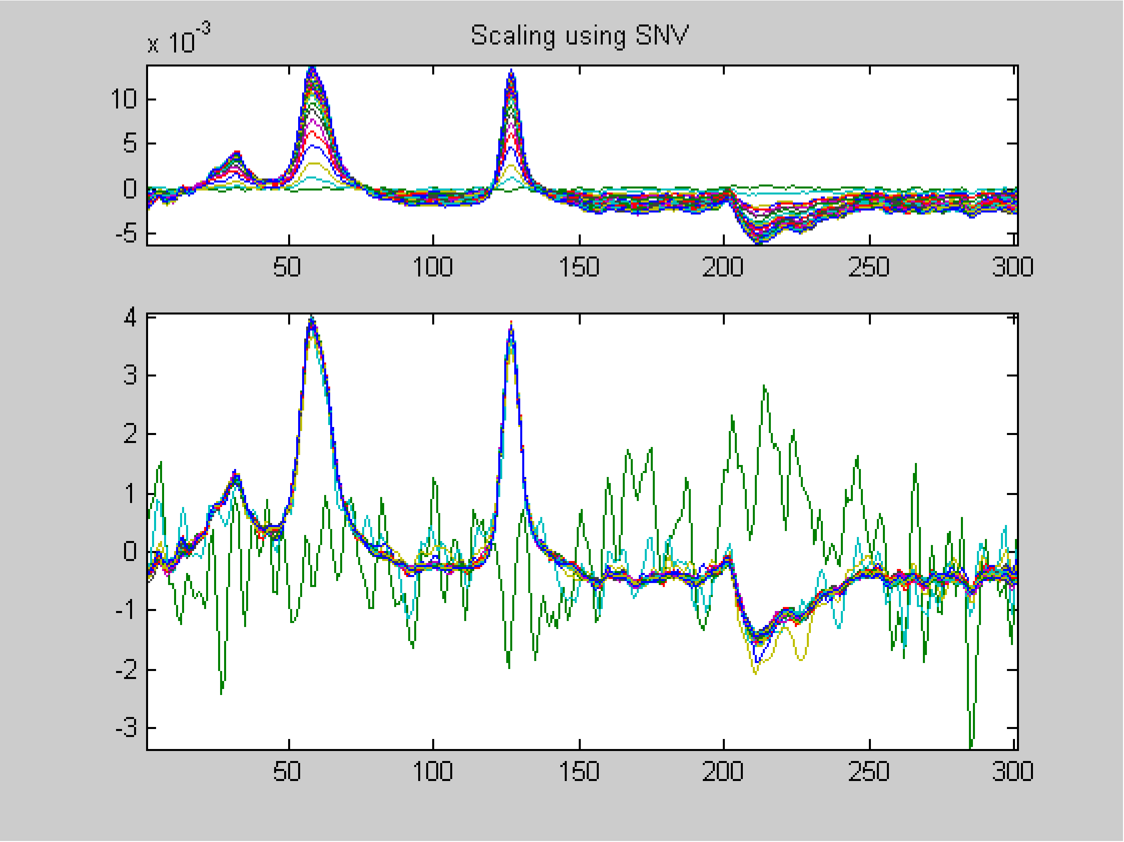
- It is easier to separate three concentration levels





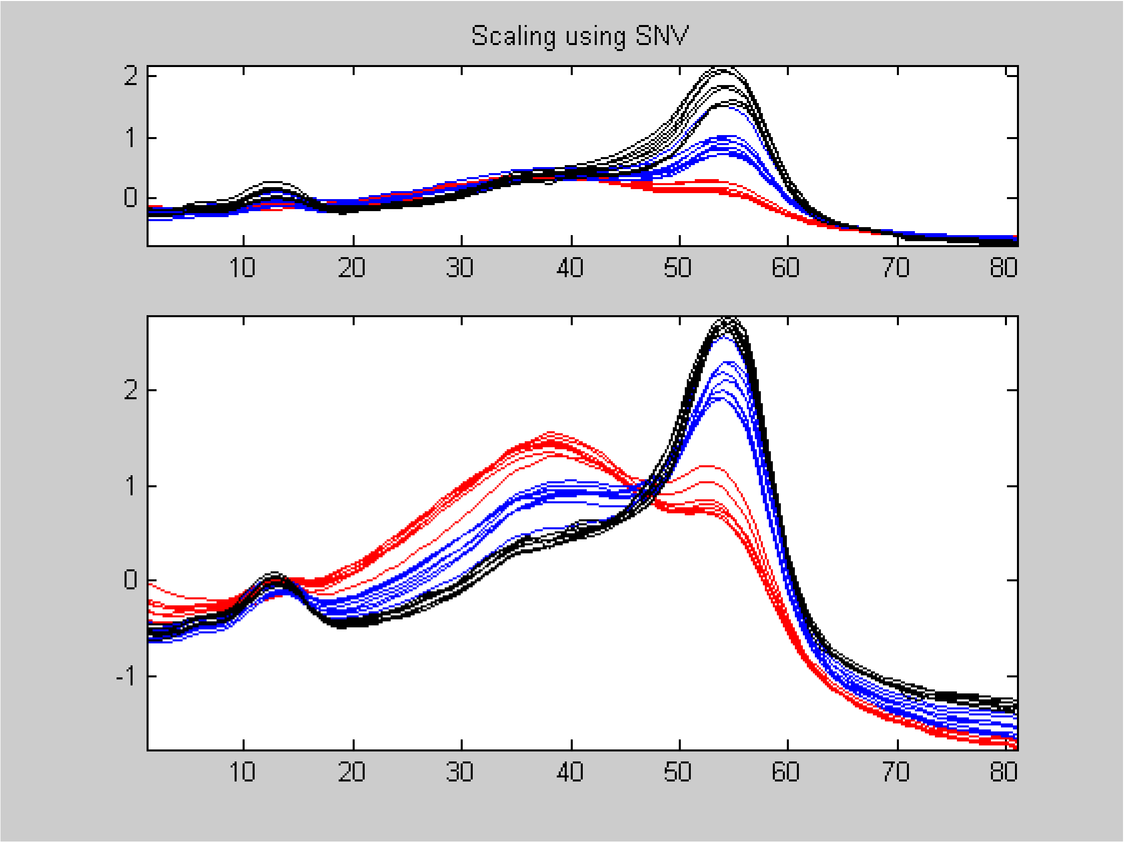
Problems with SNV

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



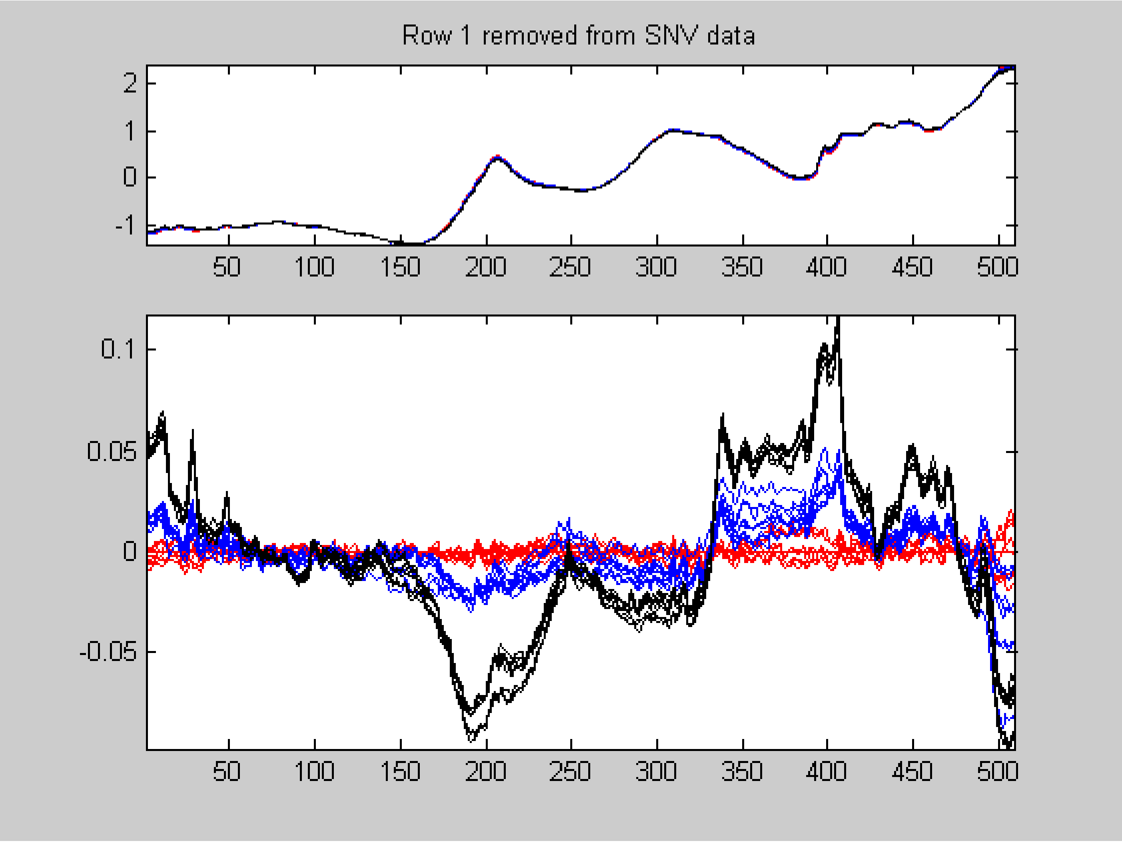
Problems with SNV

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



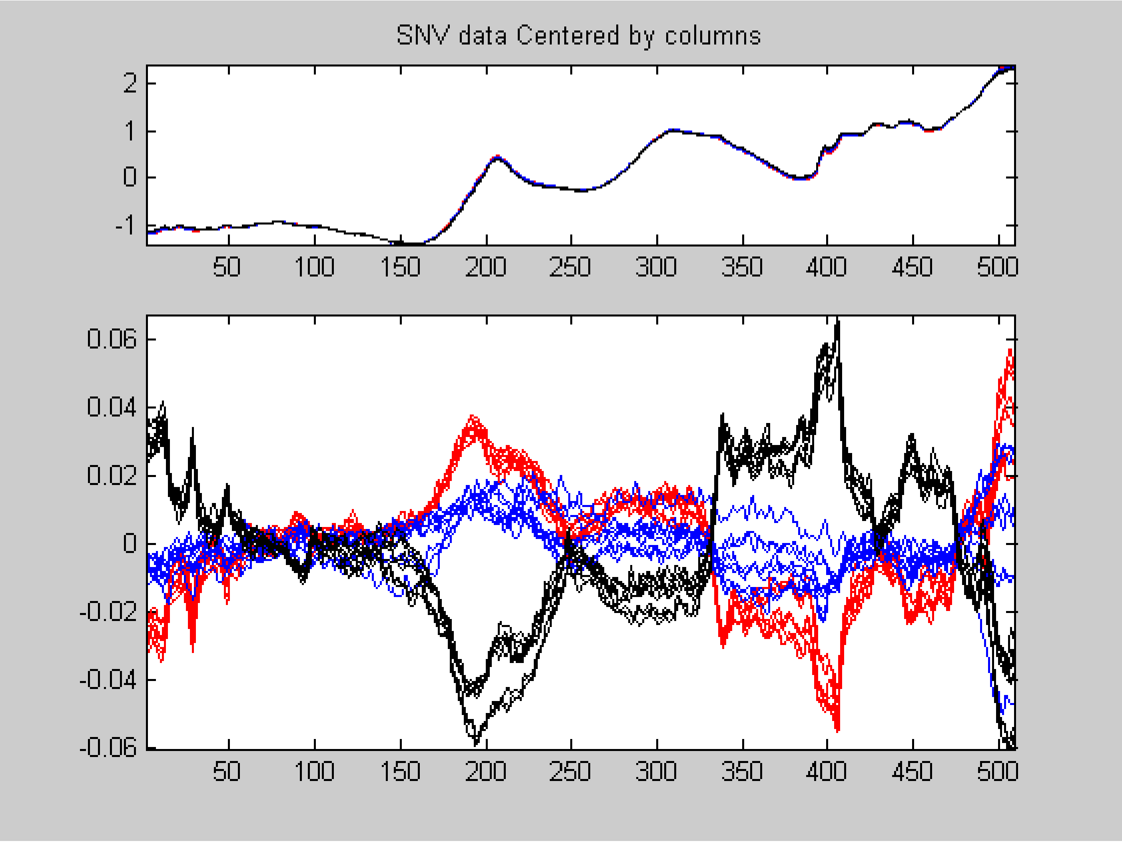
Subtract first signal

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



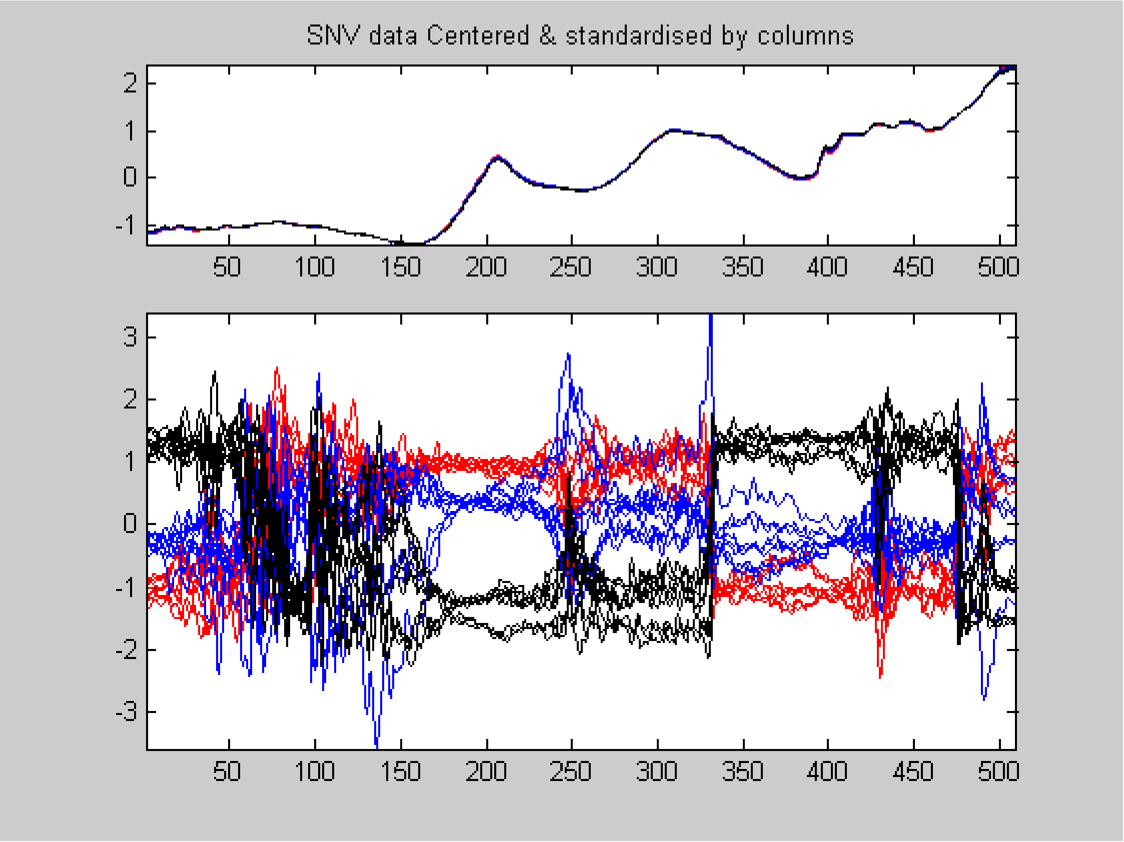
Column Centering

- Often used
- Enhances differences among samples
- Increases apparent noise level



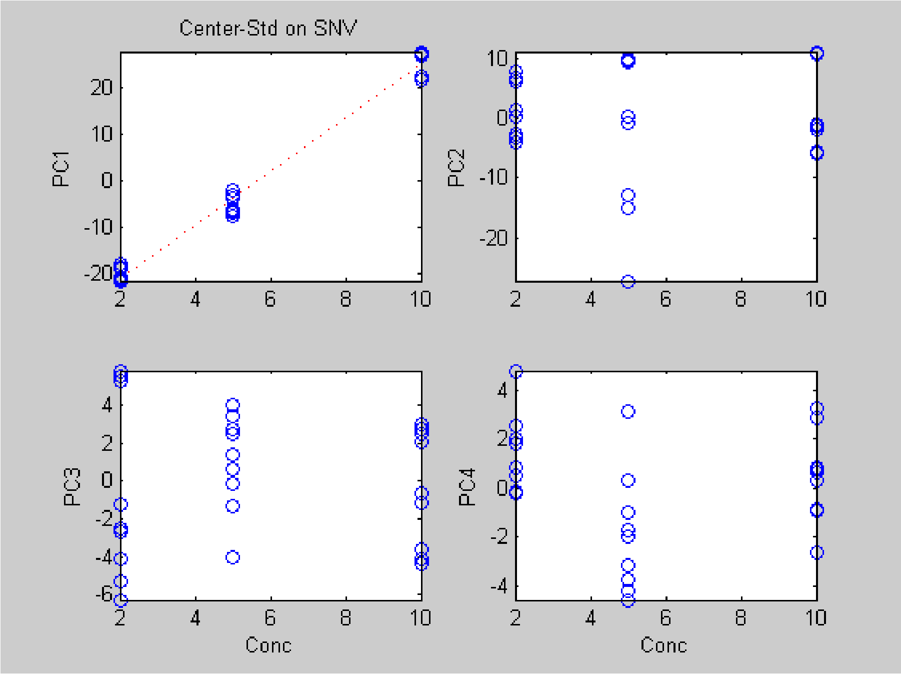
Column Centering and Scaling

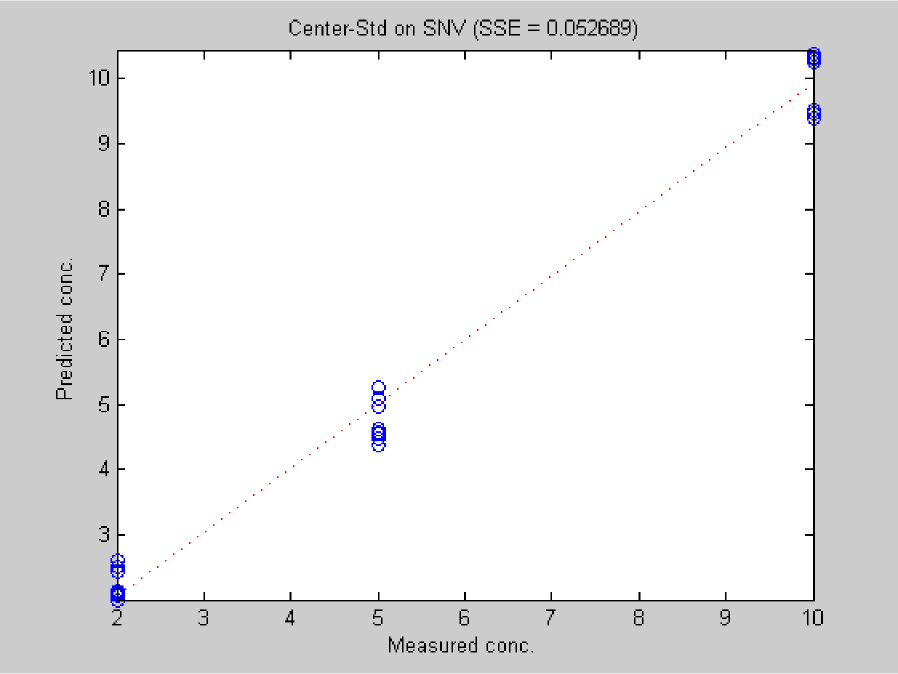
- Gives equal importance to all parts of signals
  - Both peaks and baseline
  - Makes results difficult to interpret spectrally



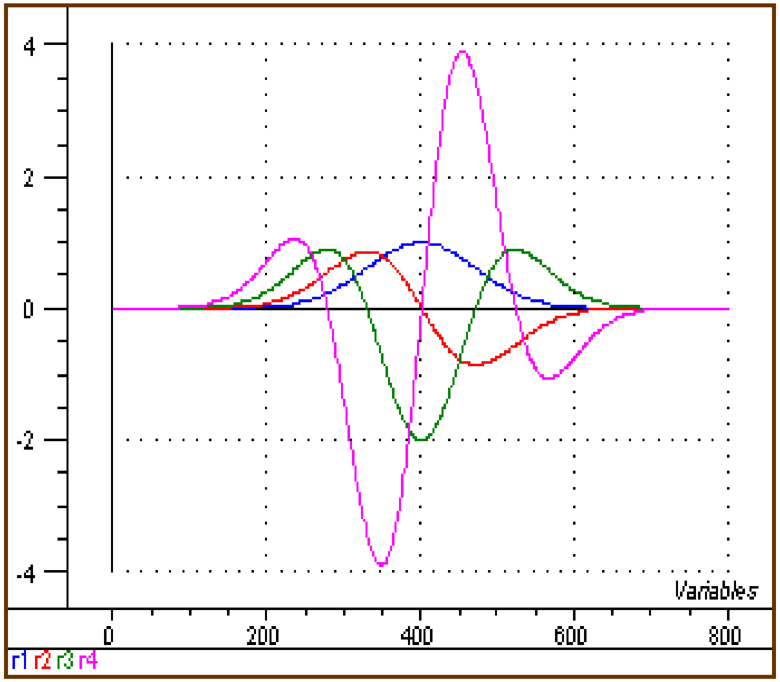
PCA & PLS on centred and on scaled data

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





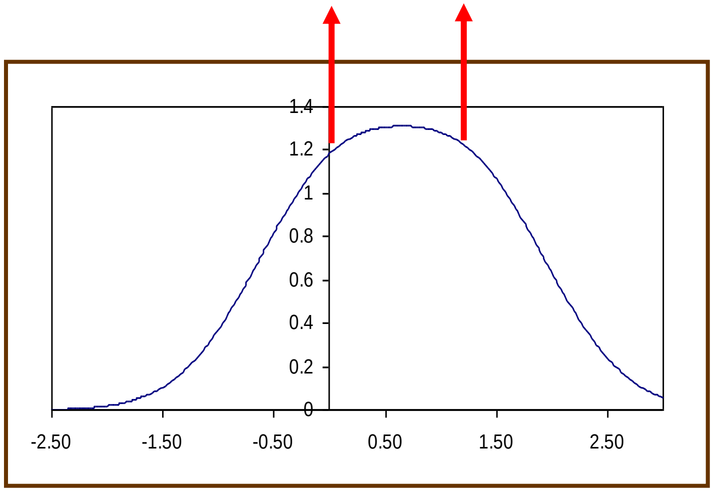
Derivatives

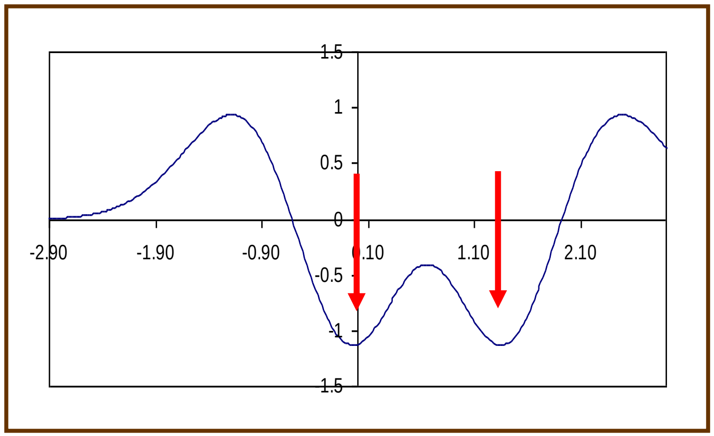


- Computing derivatives of various orders is a classical technique widely used for spectroscopic applications
- Information in a spectrum may be more easily revealed when working on a 1st or 2nd order derivative

2nd Derivative is preferred

- 2nd 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:

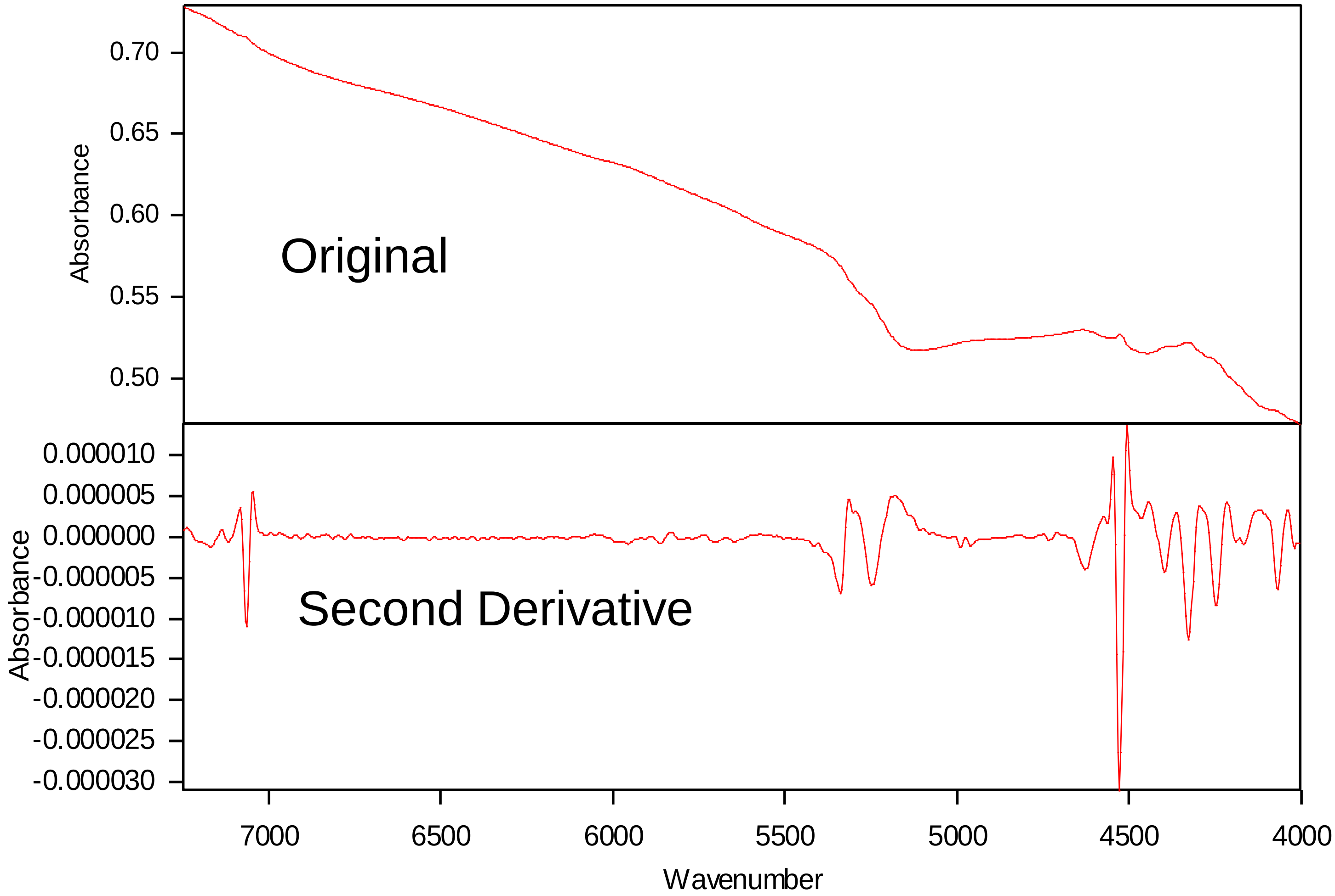




Invisible Peaks at 0.0 and 1.3

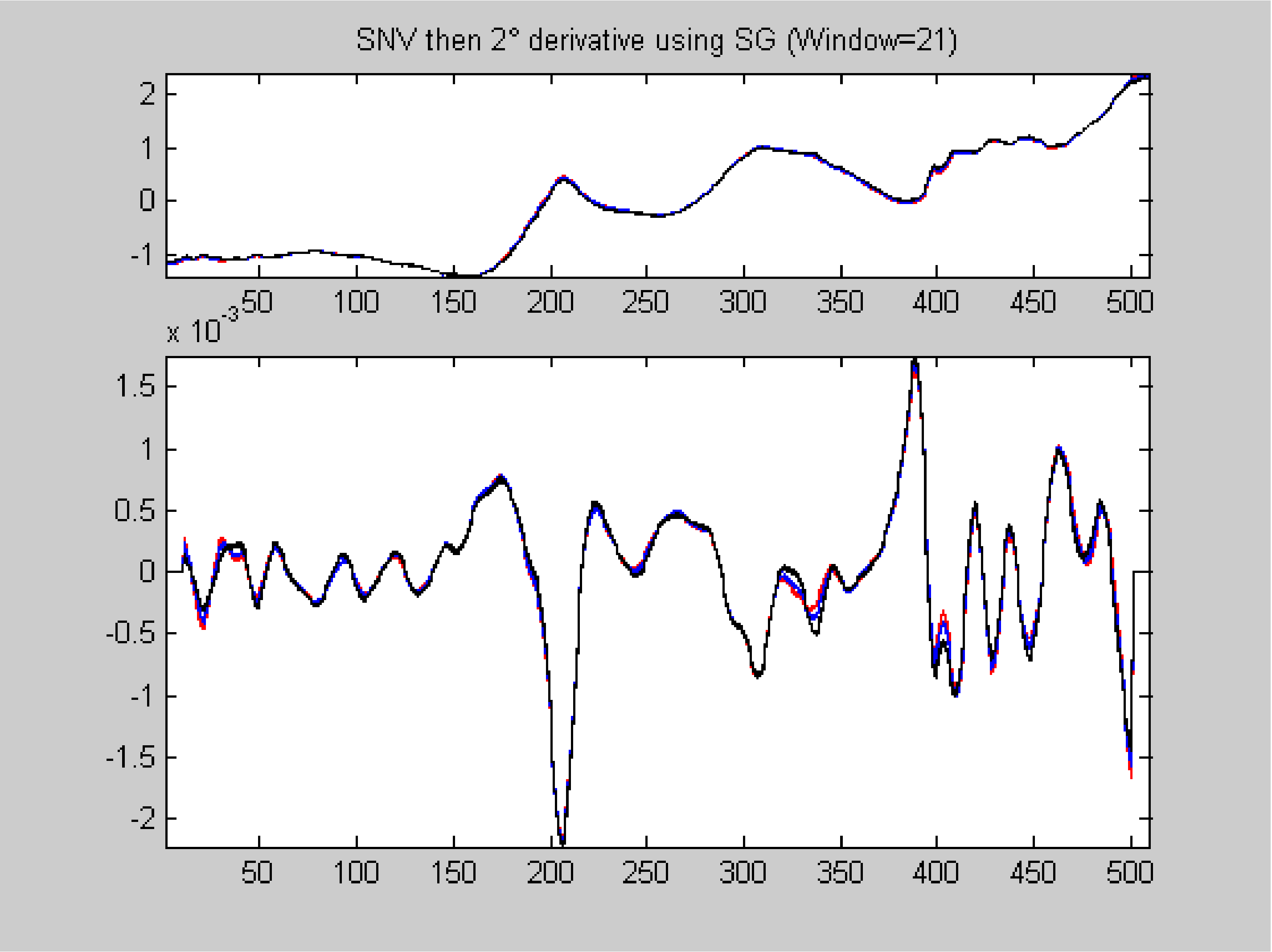
2nd derivative

Example : NIR Spectrum of Coal



Savitsky-Golay Derivatives

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

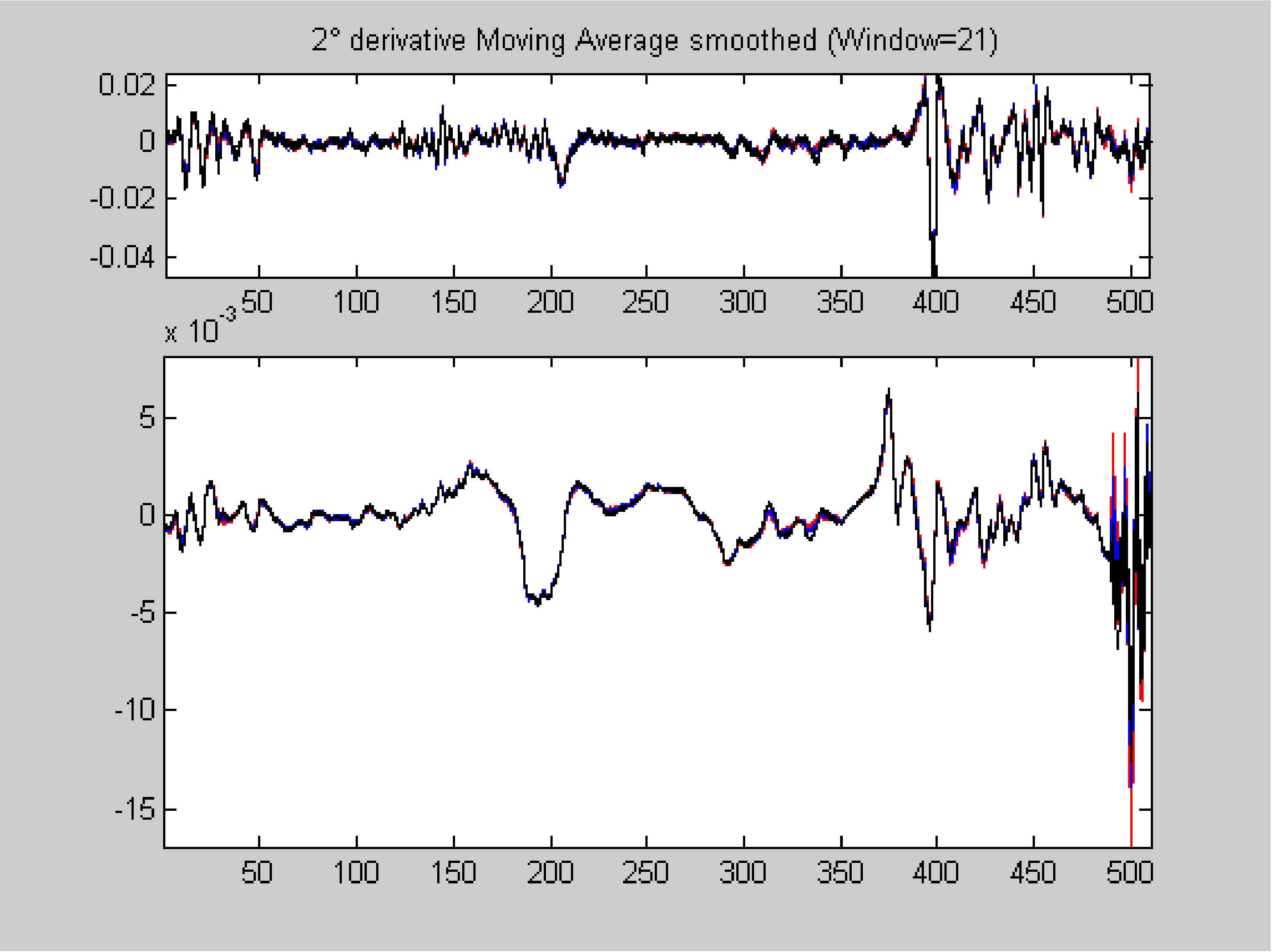


Signal Smoothing

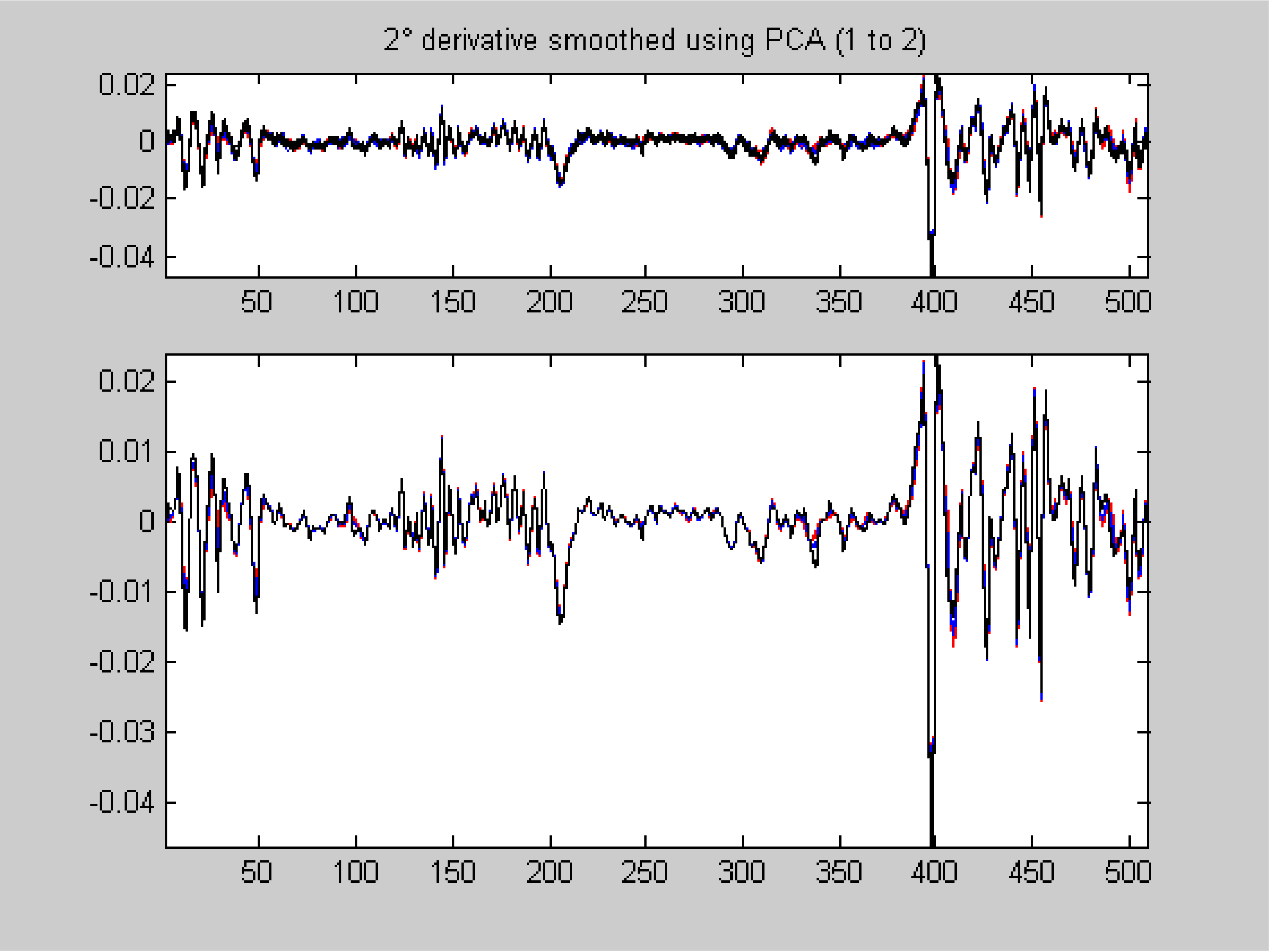
- Reduces effects of random noise
- Several algorithms :
  - Boxcar smoothing
  - Savitsky-Golay polynomial smoothing
  - PCA smoothing

Savitsky-Golay smoothing

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



PCA smoothing



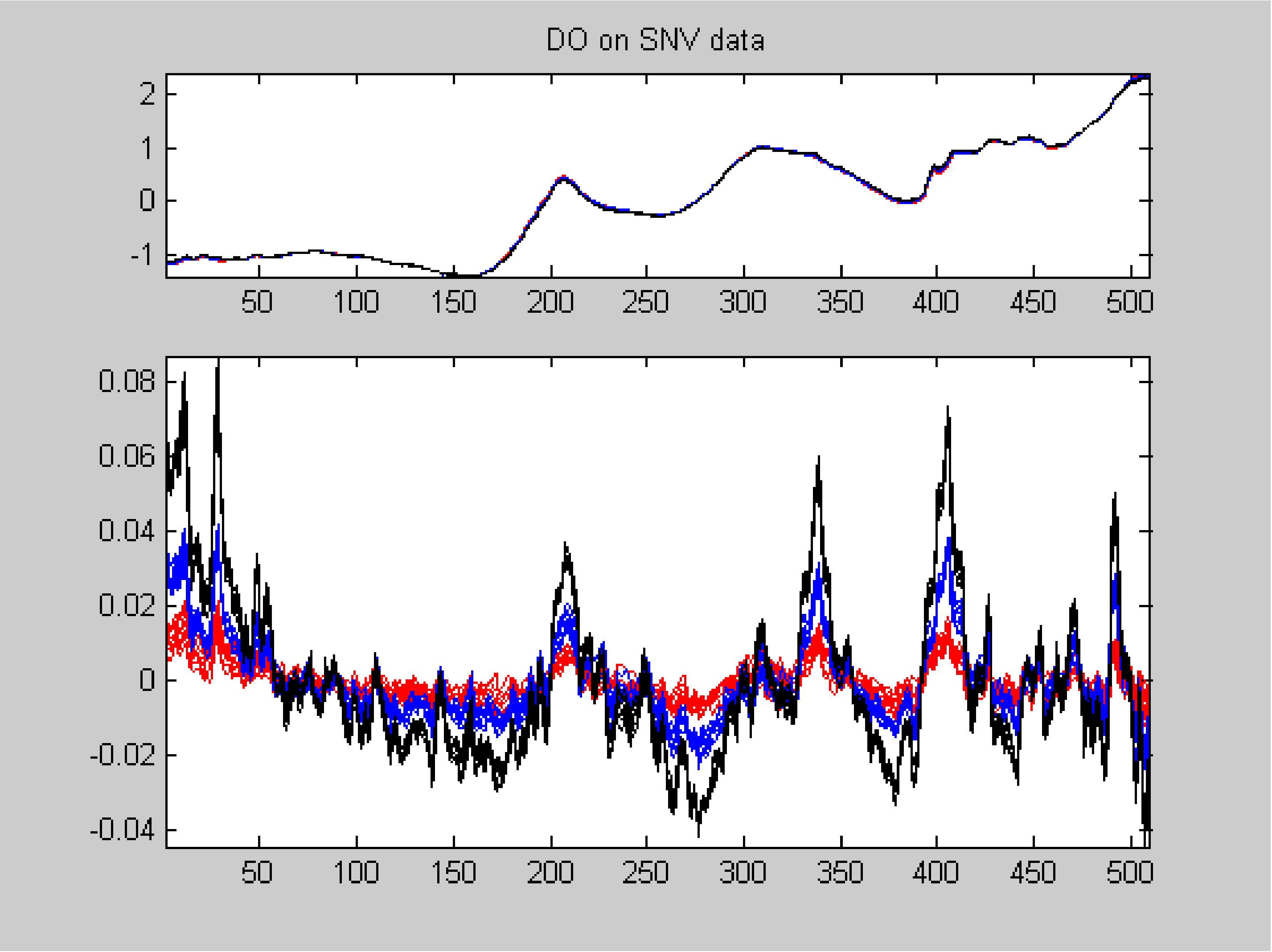
Othogonalisation

- 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
  - ...

Direct Othogonalisation

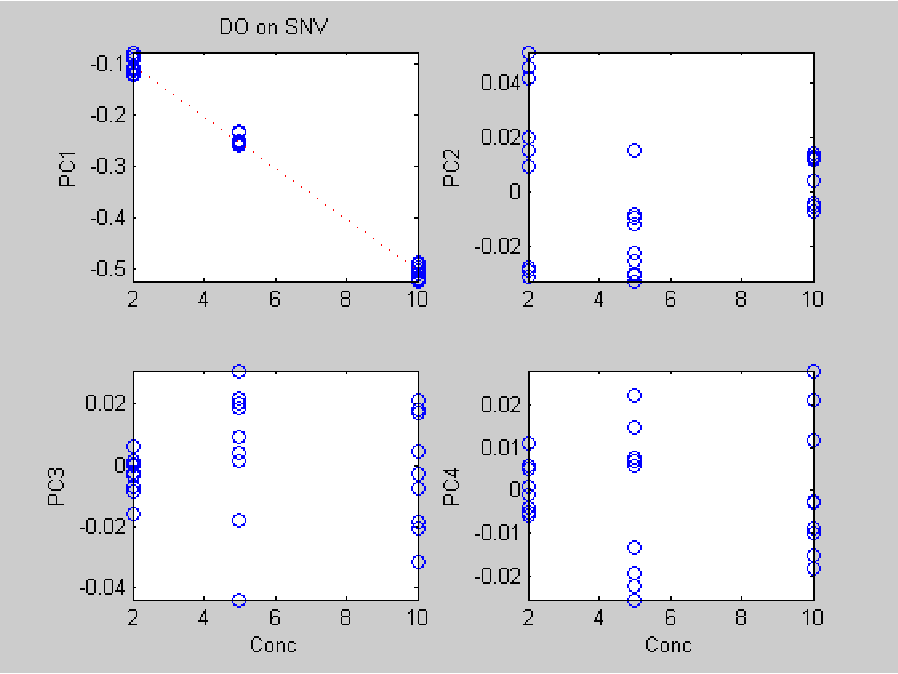
- Calculate space orthogonal to $\boldsymbol{y} = \boldsymbol{y_o}$
- Project **X** onto $\boldsymbol{y_o} = \mathbf{X_o}$
- Do PCA on $\mathbf{X_o} = \mathbf{T_o}$ and $\mathbf{P_o}$
- Use $\mathbf{T_o}$ and $\mathbf{P_o}$ to calculate interesting part of $\mathbf{X_o} = \mathbf{X_o}'$
- $\mathbf{X_{DO}}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

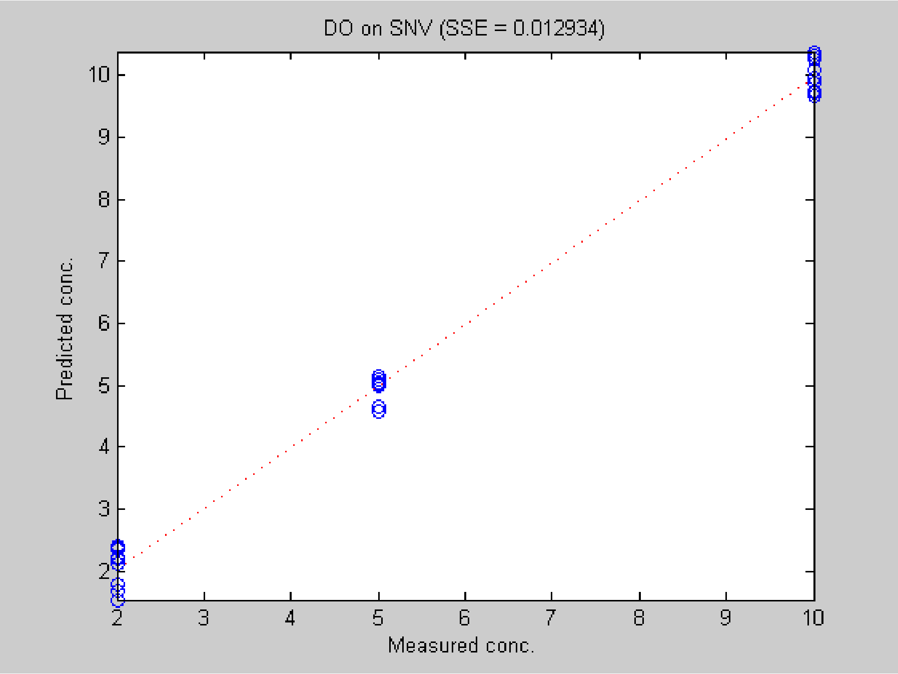
Direct Othogonalisation



PCA & PLS on DO-corrected data

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



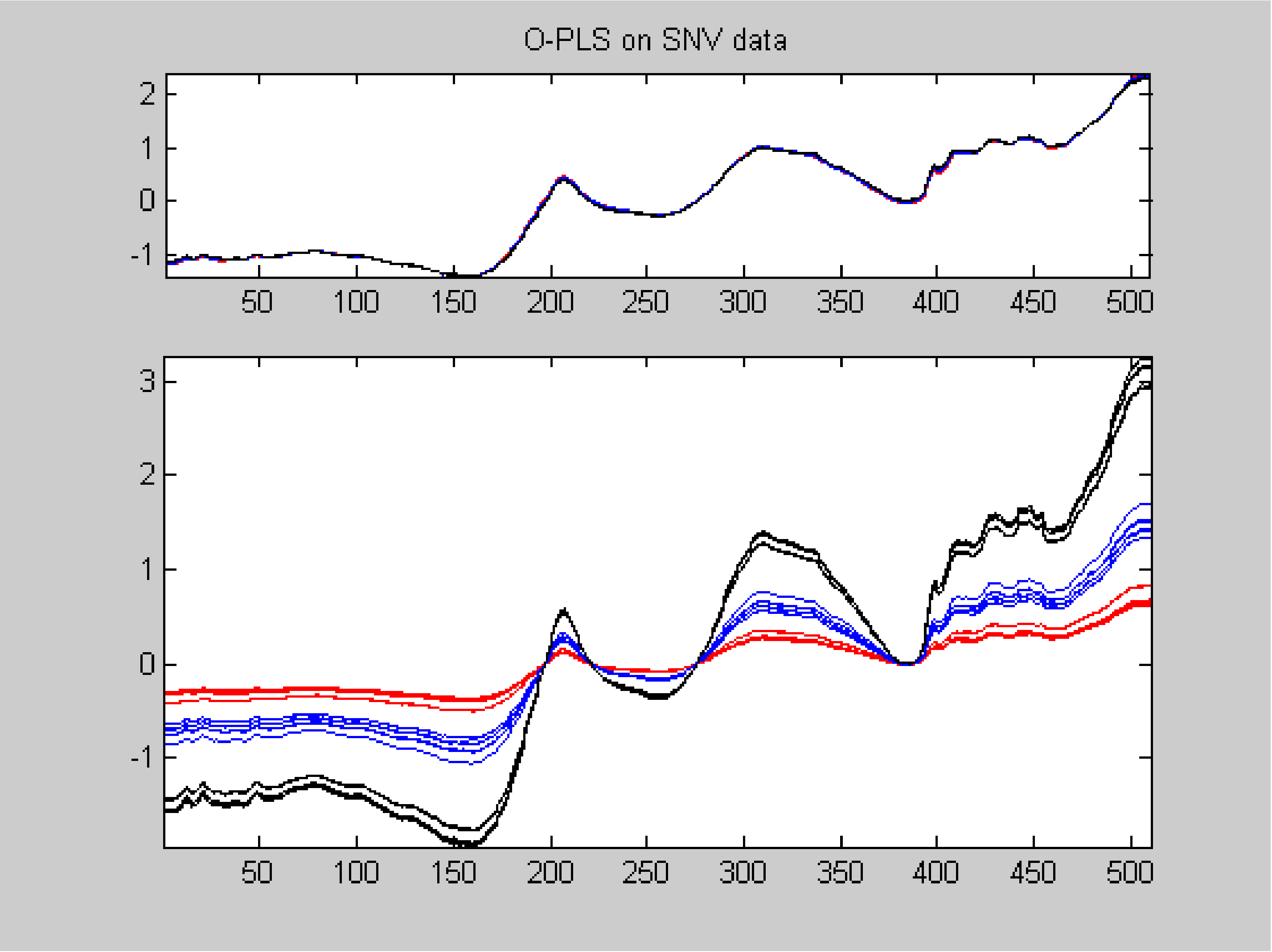


Orthogonal-PLS

- Do PLS between **X** and $\boldsymbol{y}$
- Calculate $\boldsymbol{w}$, $\boldsymbol{t}$, $\boldsymbol{p}$
- Project $\boldsymbol{w}$ orthogonal to $\boldsymbol{p}$ = $\mathbf{w}_{\boldsymbol{O}}$
- Use $\mathbf{w}_{\boldsymbol{O}}$ to calculate orthogonal part of $\boldsymbol{t}$ and $\boldsymbol{p}$ = $\mathbf{t}_{\boldsymbol{O}}$, $\boldsymbol{p}_{\boldsymbol{O}}$
- Use $\boldsymbol{t}_{\boldsymbol{O}}$ and $\boldsymbol{p}_{\boldsymbol{O}}$ to calculate orthogonal part of **X** = $\mathbf{X}_{\boldsymbol{O}}$

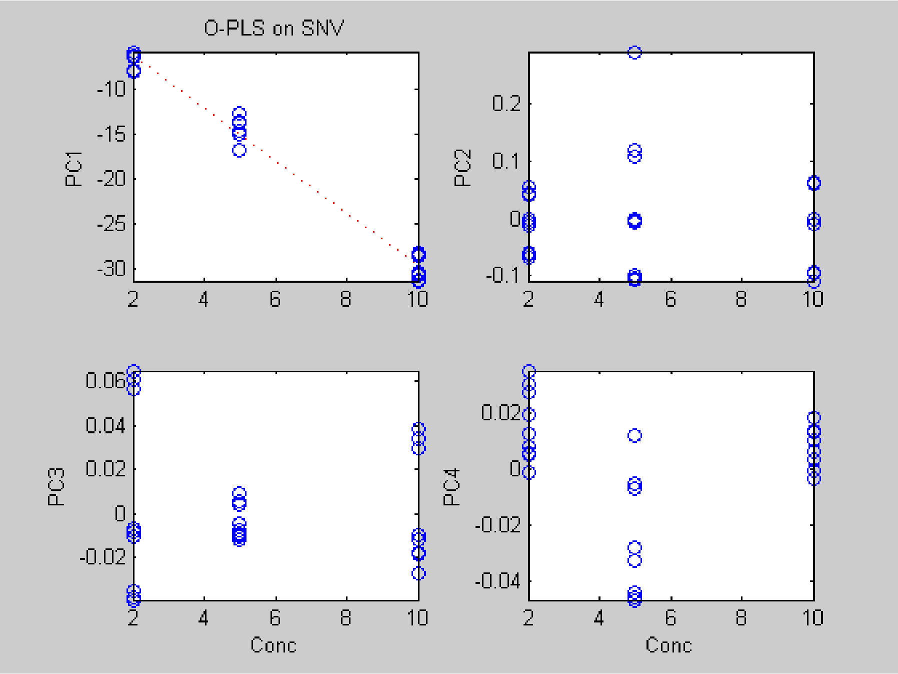
- $\mathbf{X}_{\boldsymbol{O\text{-}PLS}} = \mathbf{X} - \mathbf{X}_{\boldsymbol{O}}'$

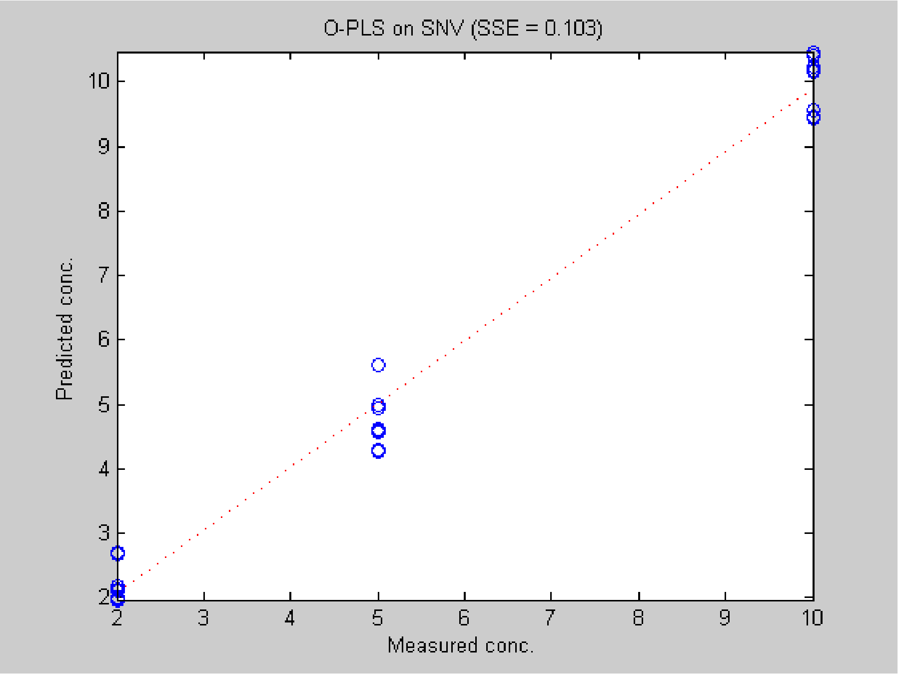
O-PLS


O-PLS on SNV data
2
1
0
-1
50
100
150
200
250
300
350
400
450
500
3
2
1
0
-1
50
100
150
200
250
300
350
400
450
500

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 ?

Preprocessing	SSE
None	289.7
SNV	0.103
SNV-2 Deriv.	0.005
SNV-Centering	0.103
SNV-Center/Std	0.053
SNV-DO	0.013
SNV-OPLS	0.103

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