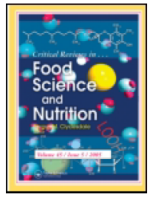
Data preprocessing



Taylor & Francis

Critical Reviews in Food Science and Nutrition

ISSN: 1040-8398 (Print) 1549-7852 (Online) Journal homepage: http://www.tandfonline.com/loi/bfsn20

Applications of Near-infrared Spectroscopy in Food Safety Evaluation and Control: A Review of Recent Research Advances

Jia-Huan Qu, Dan Liu, Jun-Hu Cheng, Da-Wen Sun, Ji Ma, Hongbin Pu & Xin-An Zeng

To cite this article: Jia-Huan Qu, Dan Liu, Jun-Hu Cheng, Da-Wen Sun, Ji Ma, Hongbin Pu & Xin-An Zeng (2015) Applications of Near-infrared Spectroscopy in Food Safety Evaluation and Control: A Review of Recent Research Advances, Critical Reviews in Food Science and Nutrition, 55:13, 1939-1954, DOI: 10.1080/10408398.2013.871693

To link to this article: http://dx.doi.org/10.1080/10408398.2013.871693

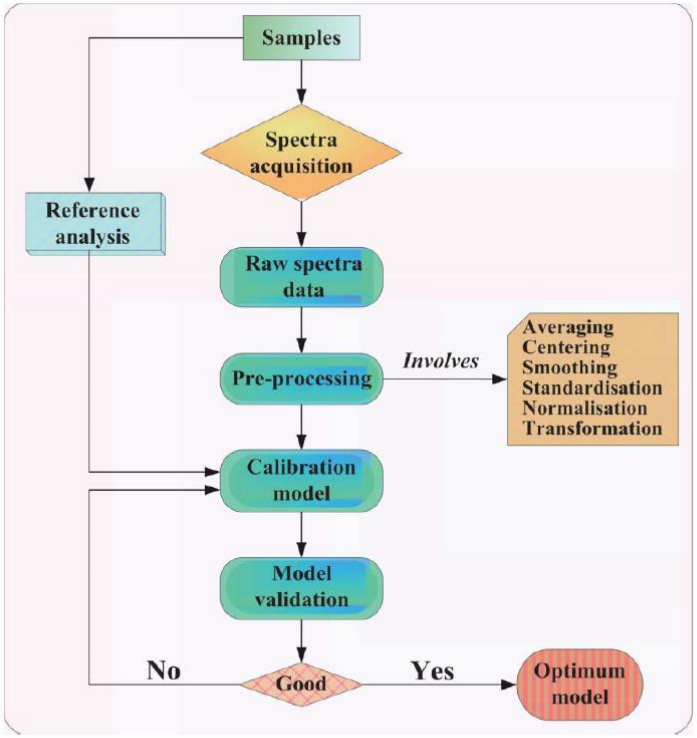


Figure 4 A series of steps in the procedure of NIRS analysis from sampling to an optimal model for prediction.

ผศ.ดร. ศิลา กิตติวัชนะ และคณะนักศึกษา
ภาควิชาเคมี คณะวิทยาศาสตร์ มหาวิทยาลัยเชียงใหม่
E-mail: silacmu@gmail.com
Tel: 087-9166692

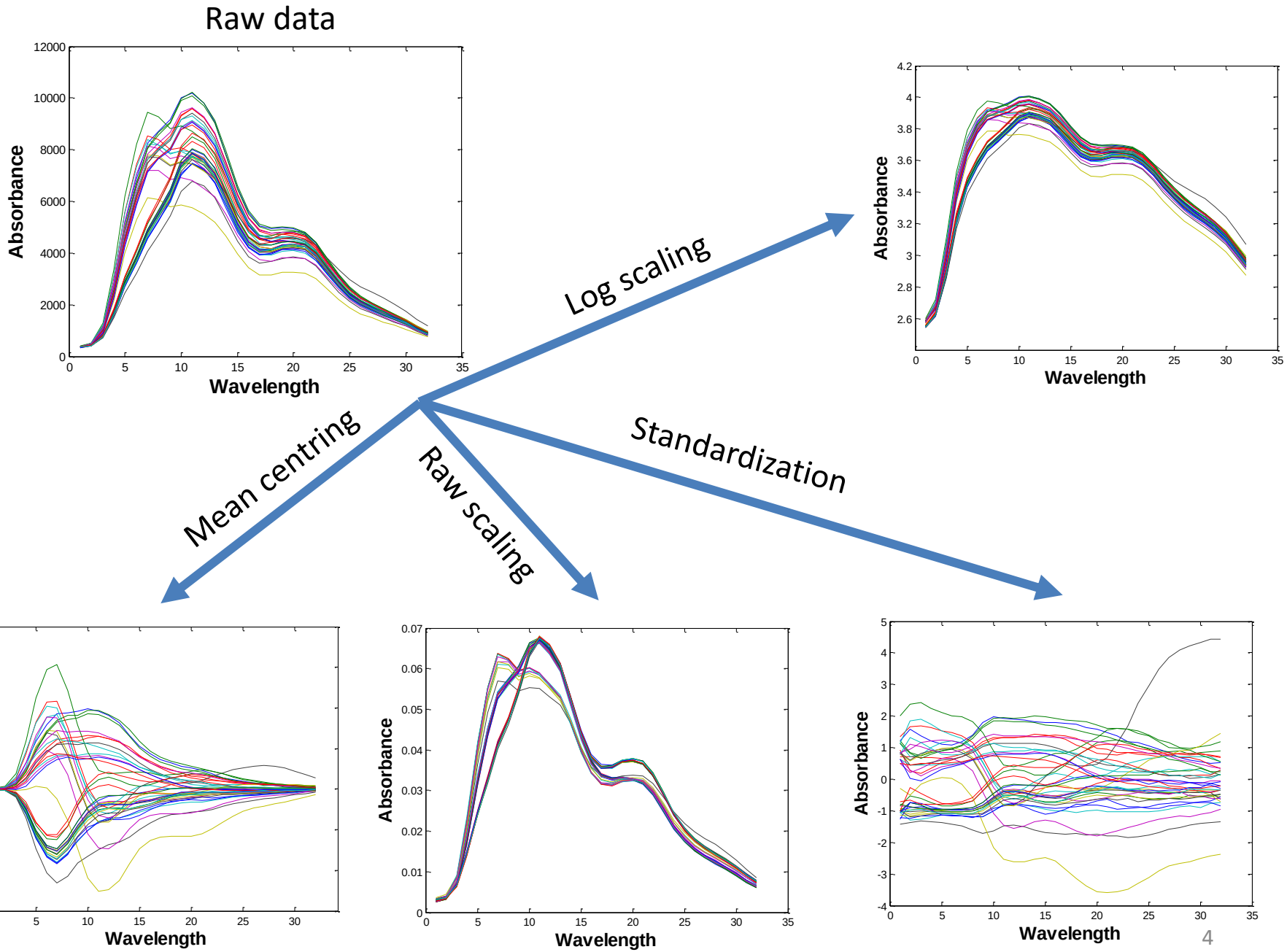
The problems are….

- Data are from different types/scales
- Variation during experiments
- Errors: Systematic and non-systementic errors
- Noise: Homoscedastic and heteroscedastic noise
- Characteristics of the acquisition methods such as light scattering in NIR
- ……etc

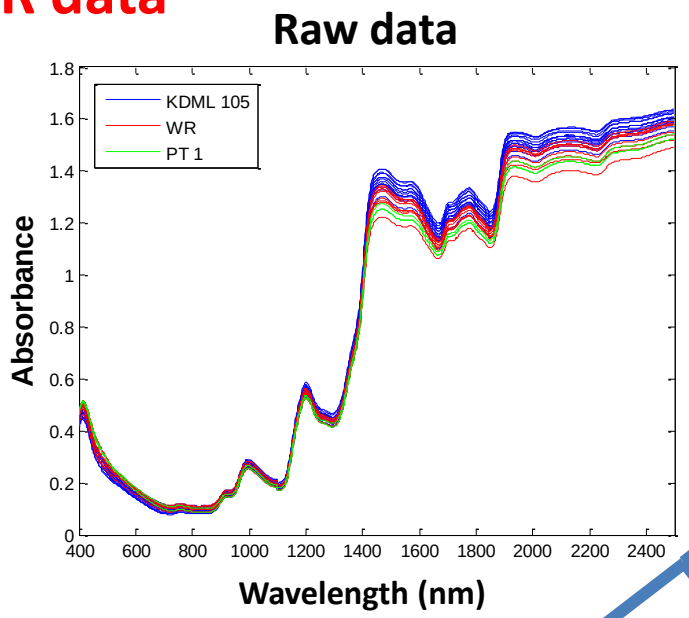
Data preprocessing

- Data pre-processing is to apply a mathematical modification to the values of a matrix prior to formal data analysis methods.
- A raw data may be shifted/rescaled along the variables/samples to place emphasis on the aspects wanted from the data analysis or to improve the interpretation of the data.
- This is to ensure that the right trends are being studied, and the analysis methods do not get confused by non-essential information.

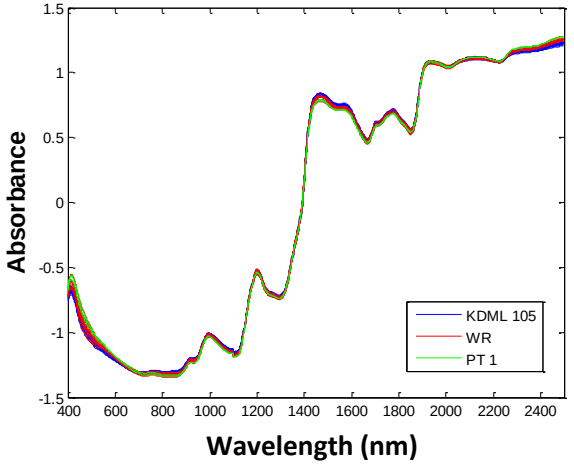
UV-Vis data


Raw data
Absorbance
Wavelength
Log scaling
Mean centring
Raw scaling
Standardization

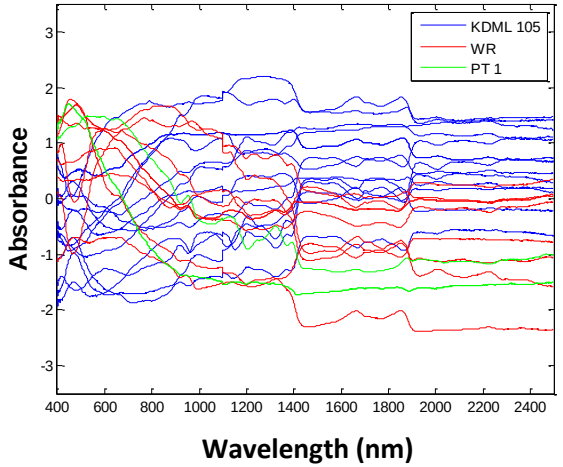
NIR data



SNV

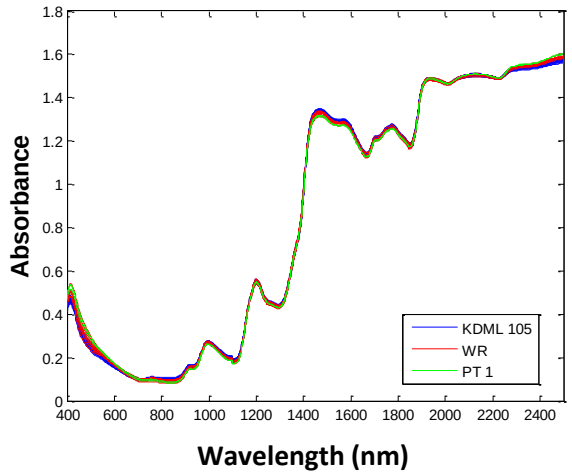


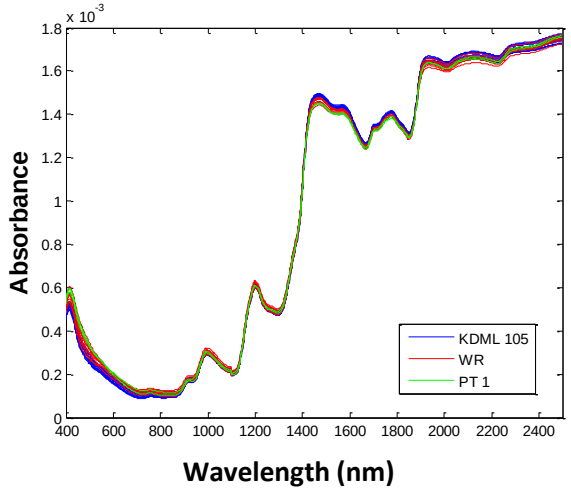
Standardization



Raw scaling

MSC



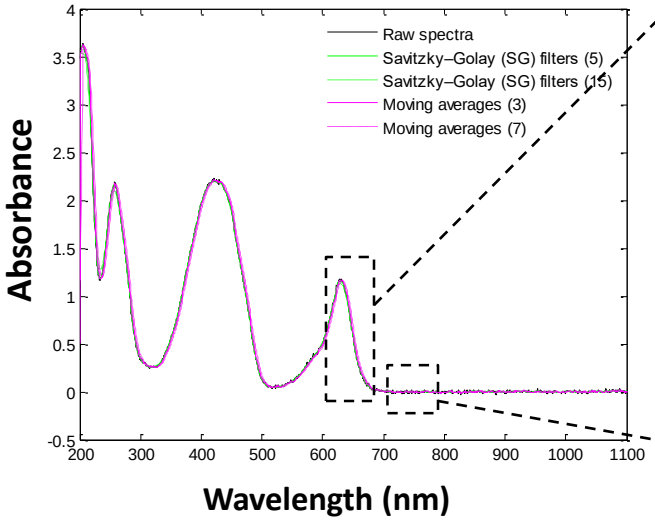


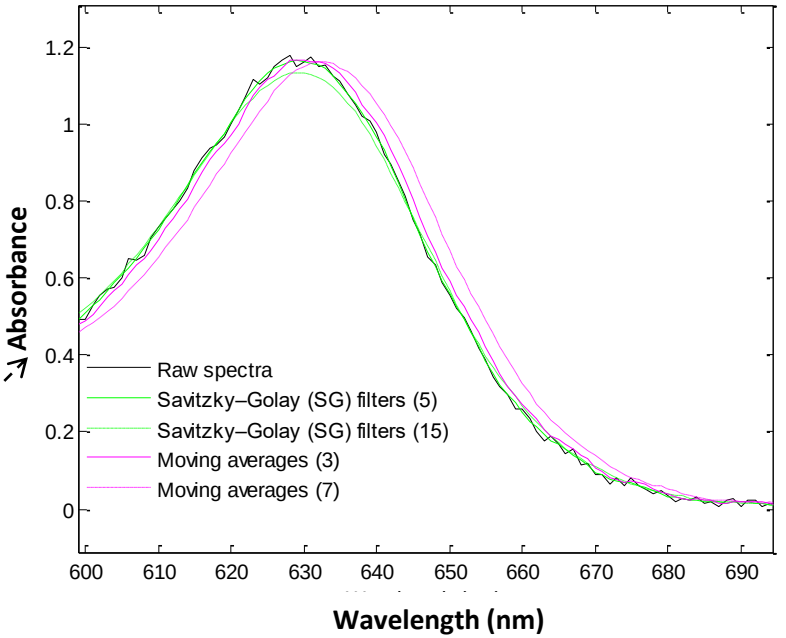
Data preprocessing

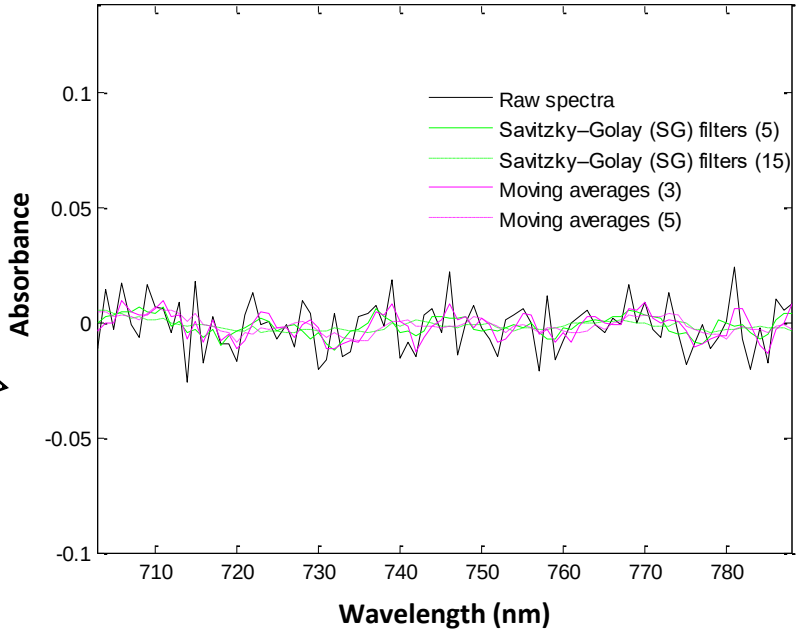
1. **Data smoothing**
   - Savitzky–Golay (SG) filters
   - Moving averages
2. **Data scaling**
   - Square root scaling
   - Log scaling
3. **Normalization**
   - Row scaling
4. **Auto scaling**
   - Mean centring
   - Standardization
5. **Others**
   - SNV
   - MSC
   - Derivatives

1. Data Smoothing

- Savitzky–Golay (SG) filters
- Moving averages



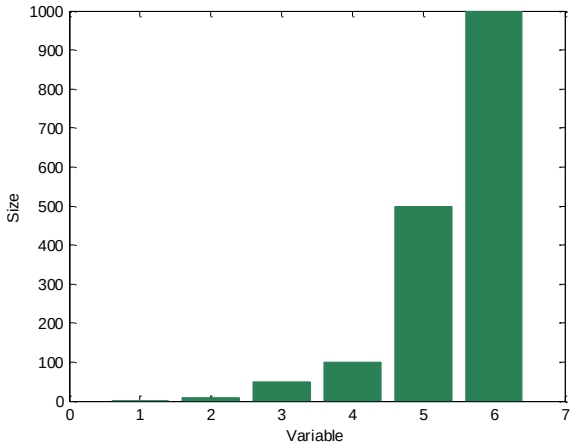




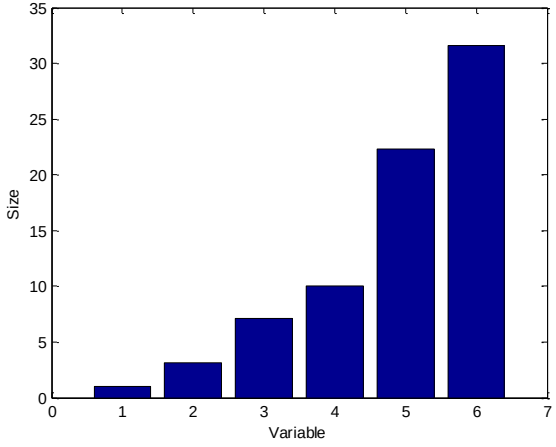
2. Data scaling

- Square root scaling $\boldsymbol{X}_{sqrt} = \sqrt{\boldsymbol{X}}$
- Log scaling $\boldsymbol{X}_{log} = \log \boldsymbol{X}$

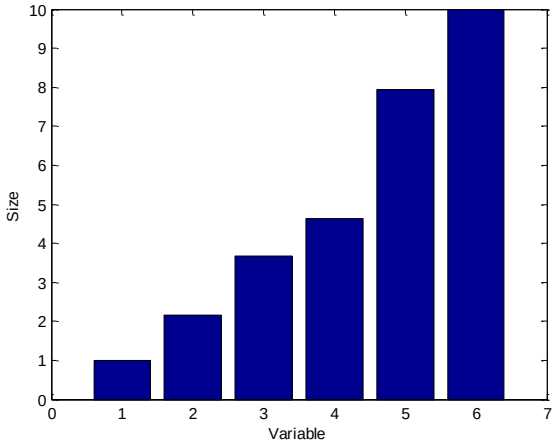
Raw data ($\boldsymbol{X}$)



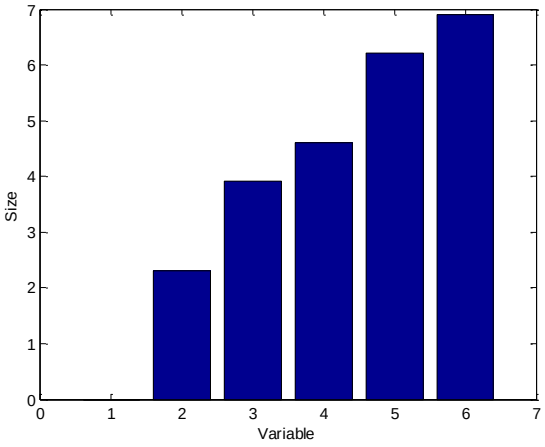
$X_{sqrt.5} = \sqrt{X}$



$X_{sqrt.33} = \sqrt[1/3]{X}$



$X_{log} = \log X$



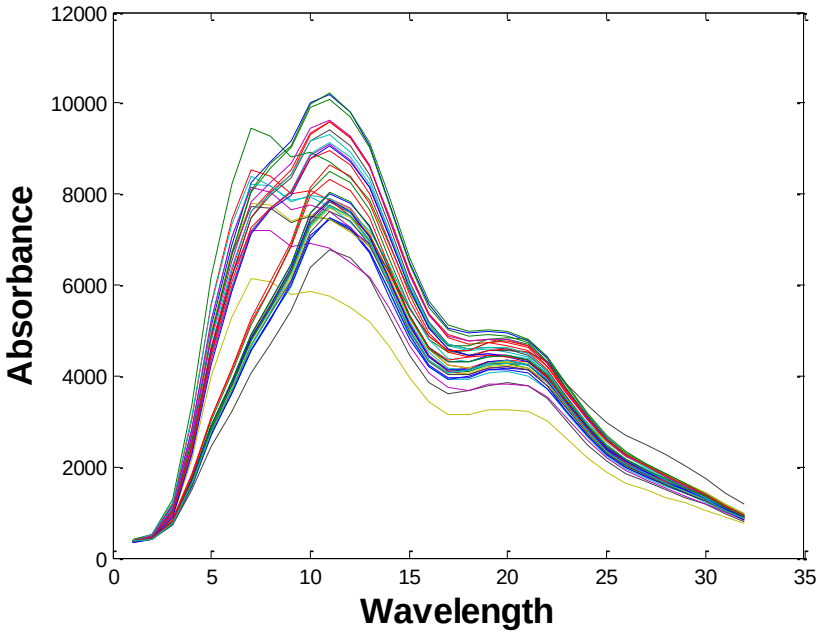
3. Normalization

– Row scaling

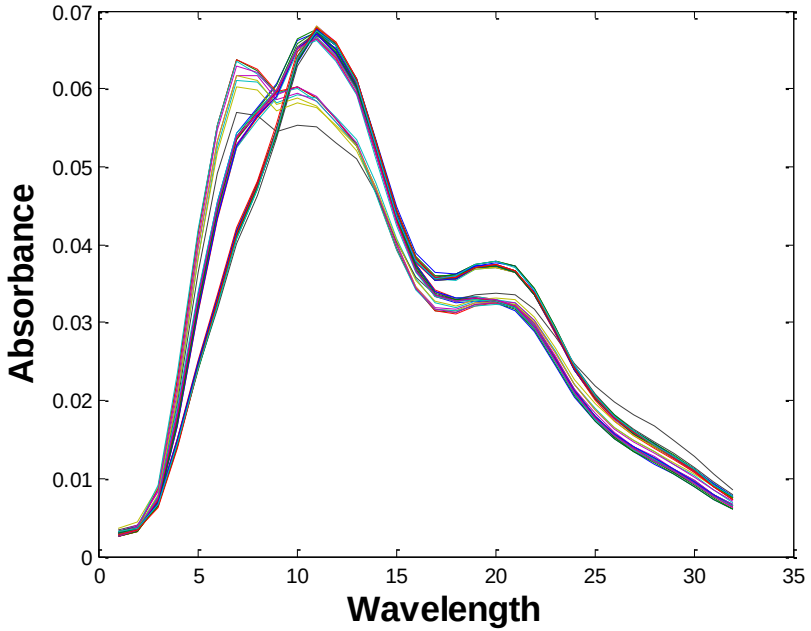
$$ {}^{rs}x_{ij} = \frac{x_{ij}}{\sum_{J=1}^{J} x_{ij}} $$

${}^{rs}x_{ij}$ = normalized value of x_{ij}

J = number of parameter



Raw data



Row scaling

4. Auto scaling

– Mean centring

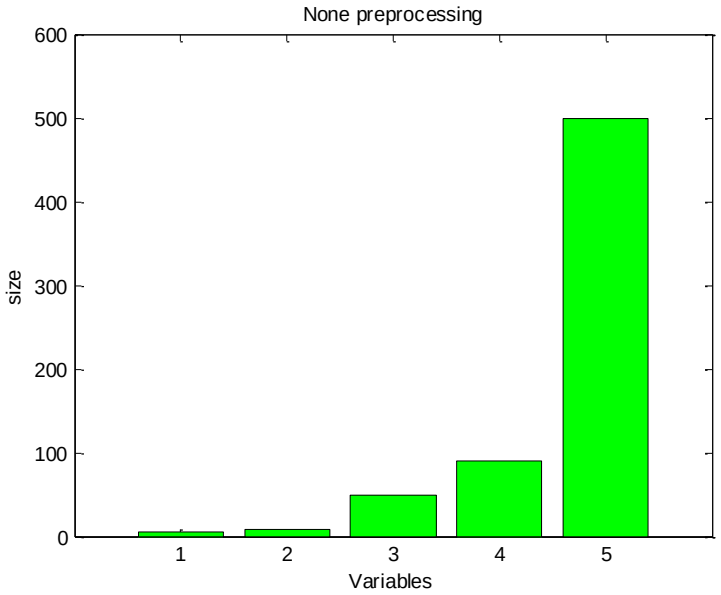
$$ {}^{cent}x_{ij} = x_{ij} - \bar{x}_j $$

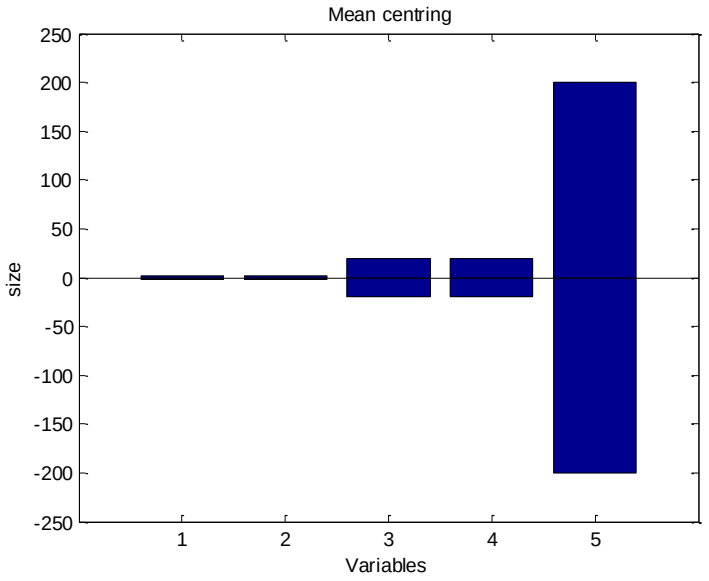
$$ \bar{x}_j = \frac{\sum_{i=1}^{I} x_{ij}}{I} $$

Raw data

1	5	10	50	100
2	6	20	60	200
3	7	30	70	300
4	8	40	80	400
5	9	50	90	500

${}^{cent}x_{ij}$= mean centring value of x_{ij} with average value of $\bar{x}_j$





4. Auto scaling

– Standardization

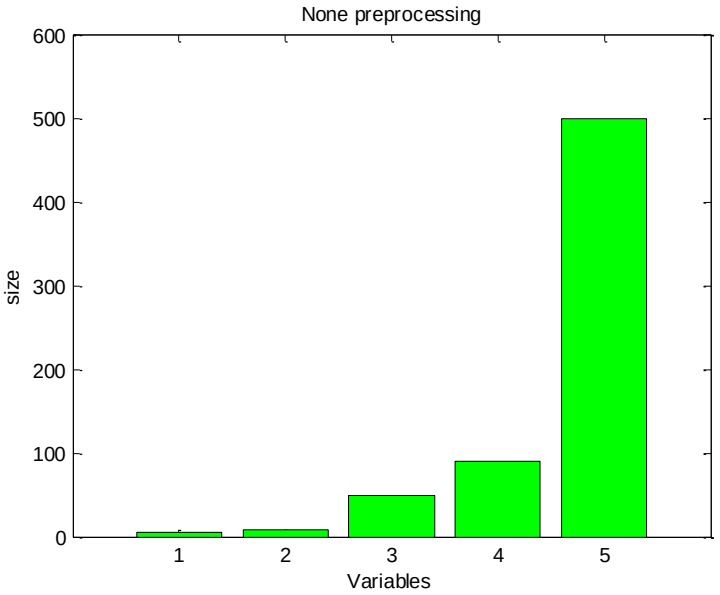
$$ {}^{std}x_{ij} = \frac{x_{ij} - \bar{x_j}}{S_j} $$

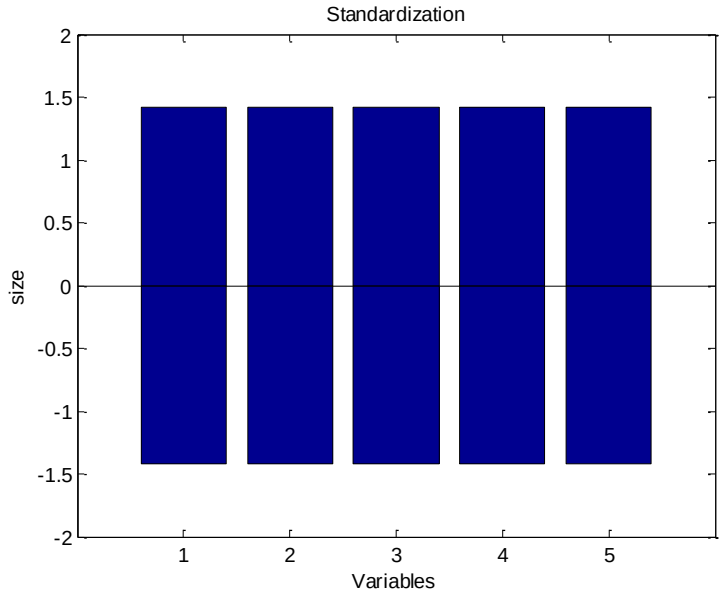
$$ S_j = \frac{\sum_{i=1}^{I} \left(x_{ij} - \bar{x_j} \right)^2}{I} $$

Raw data

1	5	10	50	100
2	6	20	60	200
3	7	30	70	300
4	8	40	80	400
5	9	50	90	500

${}^{std}x_{ij}$ = standardized element of x_{ij} using mean of $\bar{x_j}$ standard deviation of S_j





5. Others

- Standard normal variate (SNV)

$$^{SNV}A_{ij} = \frac{A_{ij} - \bar{x}_i}{SD_{ev}}$$

i = spectrum counter
j = absorbance value counter of i^{th} spectrum
$^{SNV}A_{ij}$ = corrected absorbance value
A_{ij} = measured absorbance value
$\bar{x}_i$ = mean absorbance value
SD_{ev} = standard deviation

- Multiplicative scatter-correction (MSC)

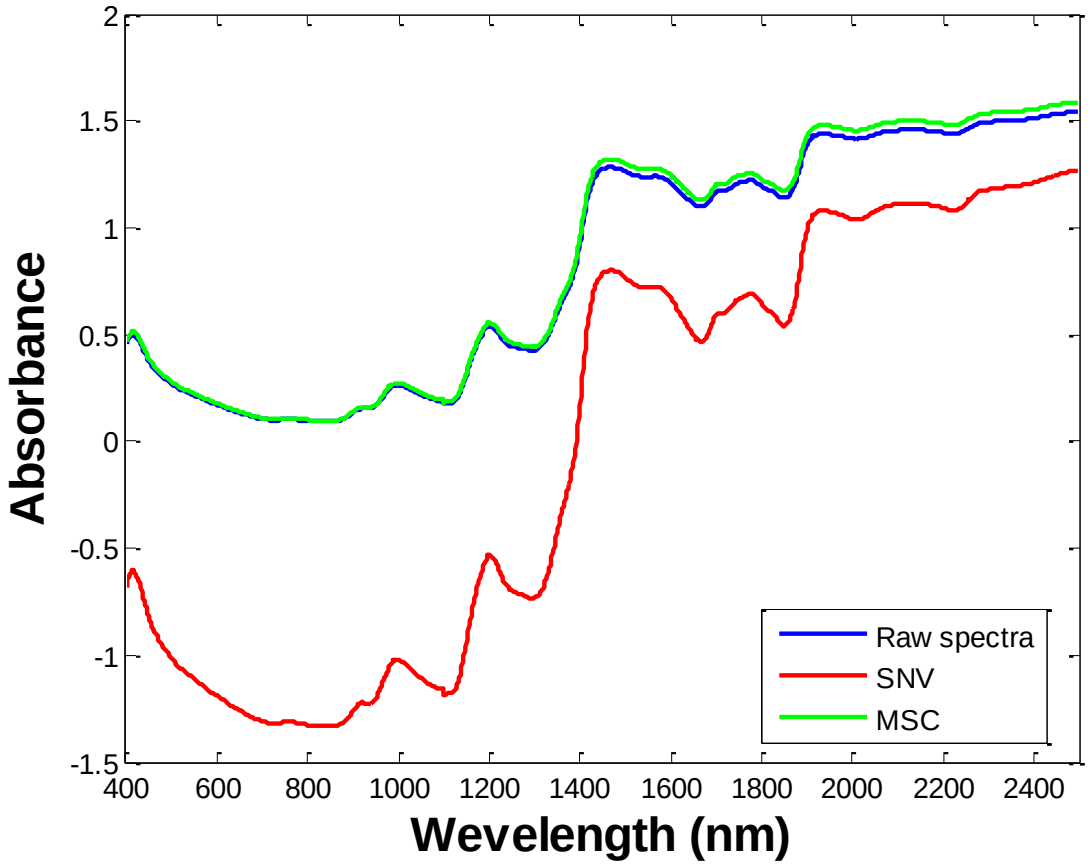
$$x_i = a + b\, y_{average} + \varepsilon$$

$$^{MSC}x_{ij} = \frac{x_{ij} - a}{b}$$

x_i = the i^{th} spectrum collection
x_{ij} = the absorbance of the i^{th} spectrum and j^{th} wavelength of the collection

For each sample, *a* and *b* are estimated by ordinary least-squares regression of spectrum x_i versus $y_{Average}$ over the available wavelengths *j*

https://www.labcognition.com/onlinehelp/en/standard_normal_variate_correction.htm
https://www.labcognition.com/onlinehelp/en/multiplicative_scatter_correction.htm

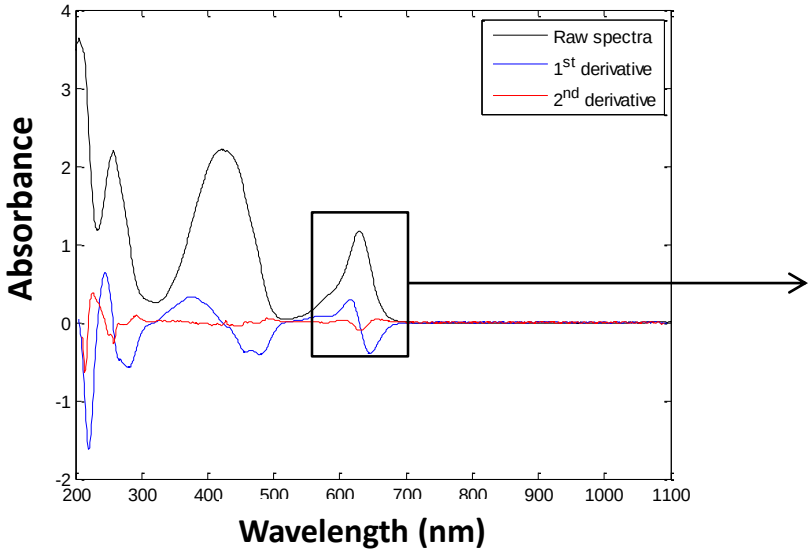

Absorbance
Wevelength (nm)
Raw spectra
SNV
MSC
2
1.5
1
0.5
0
-0.5
-1
-1.5
400
600
800
1000
1200
1400
1600
1800
2000
2200
2400

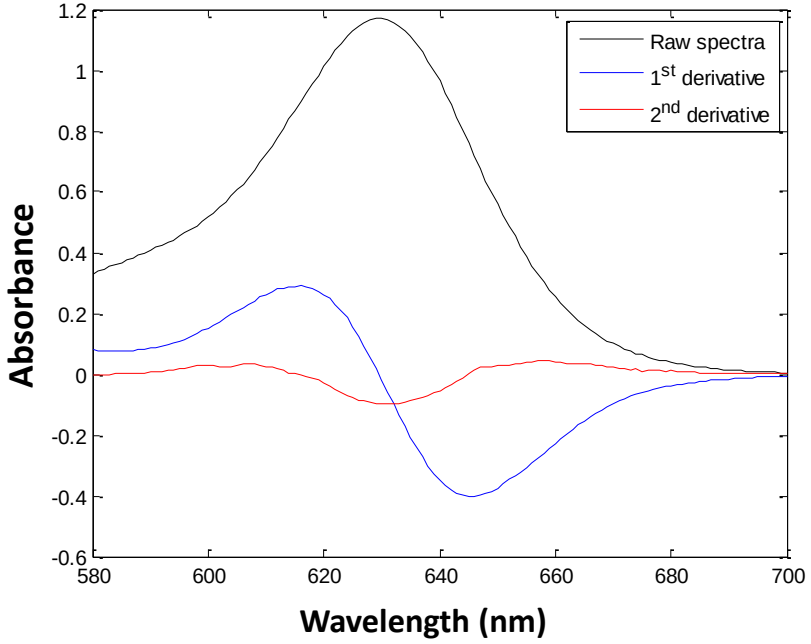
5. Others

– Derivatives

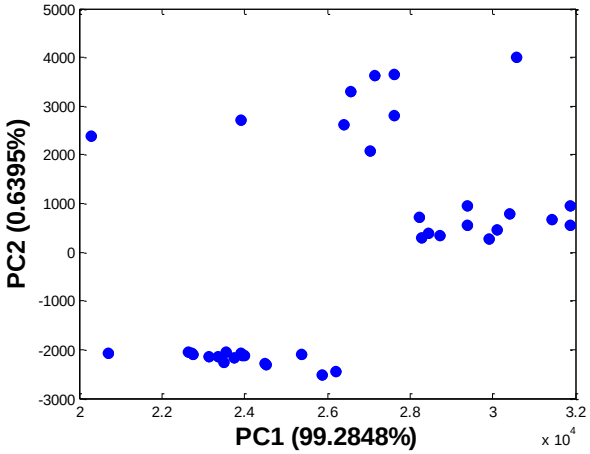
1st derivative → $dX/dY = \frac{\Delta X}{\Delta Y}$

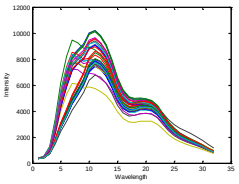
2nd derivative → $d^2X/dY^2 = \frac{\Delta^2 X}{\Delta Y^2}$

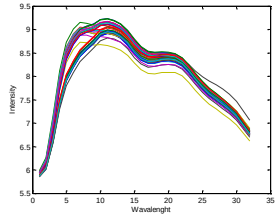


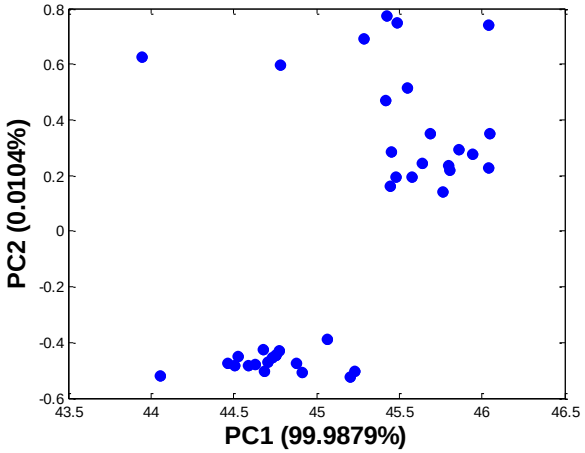


PCA score plots







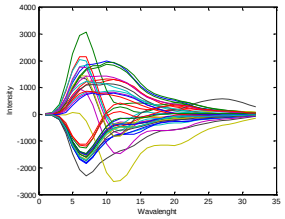


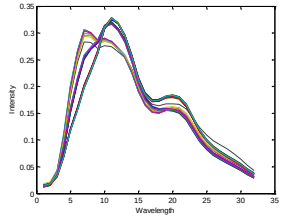
Log scaling

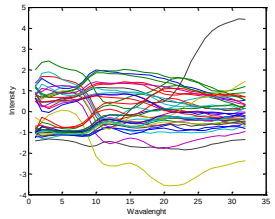
Standardization

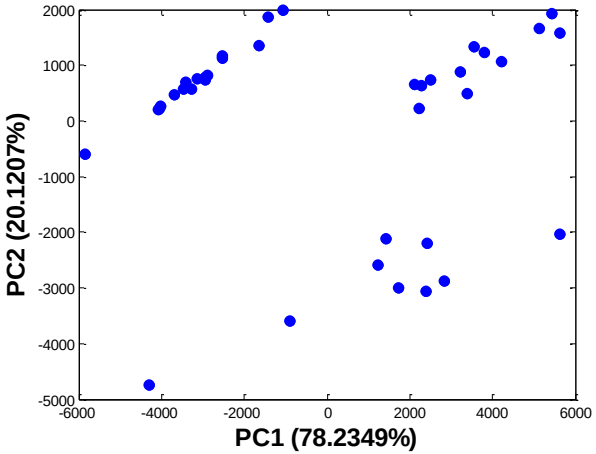
Raw scaling

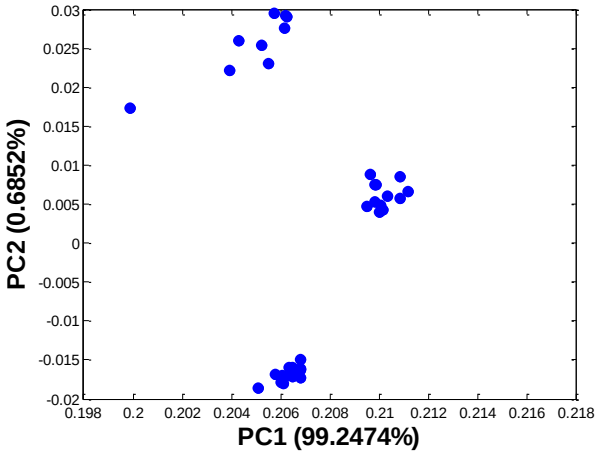
Centring

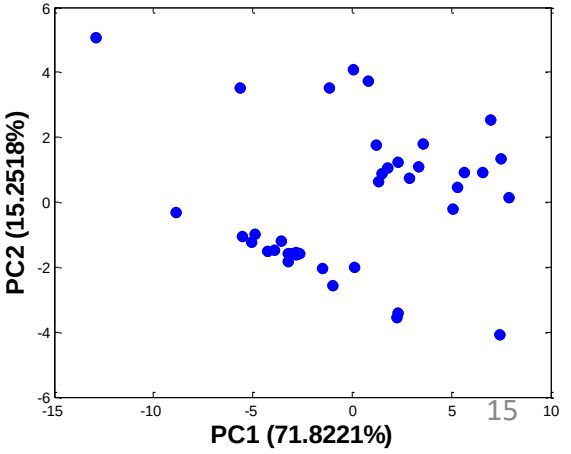




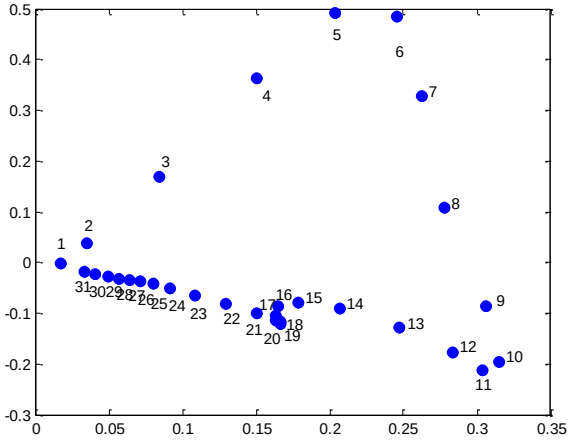


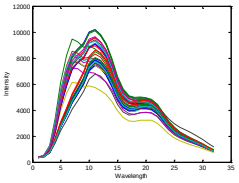


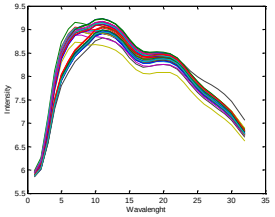


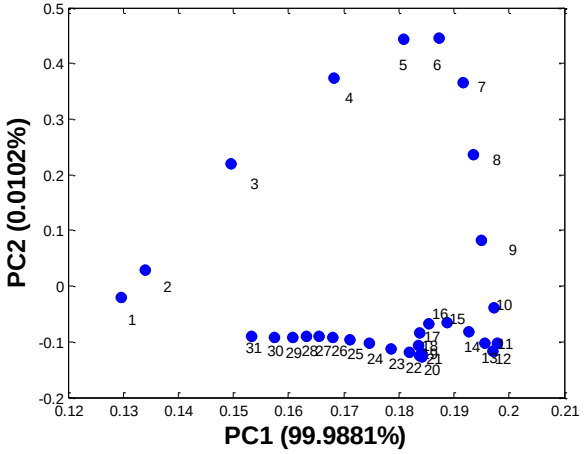


PCA loading plots







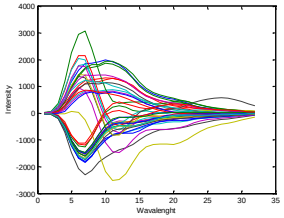


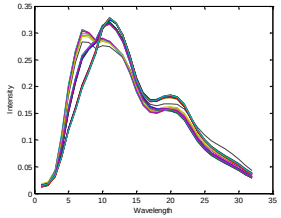
Log scaling

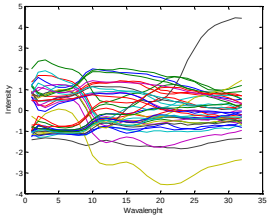
Standardization

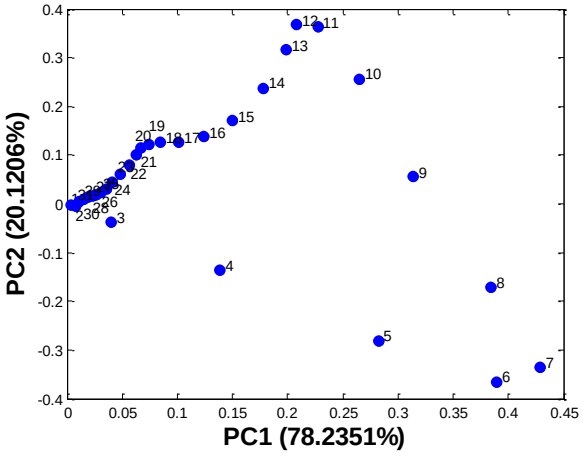
Raw scaling

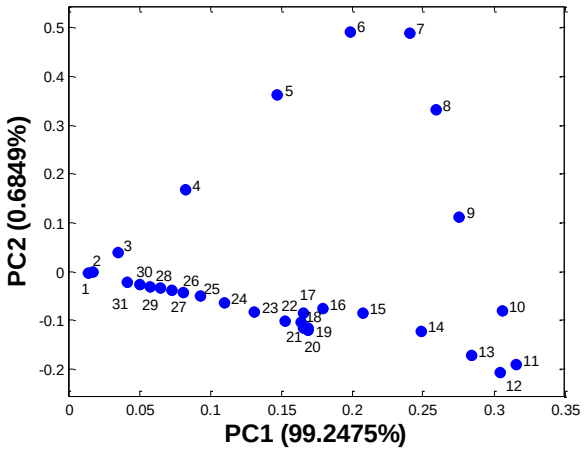
Centring

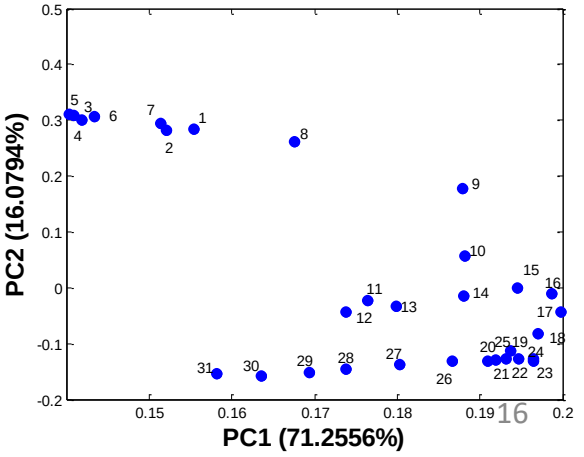




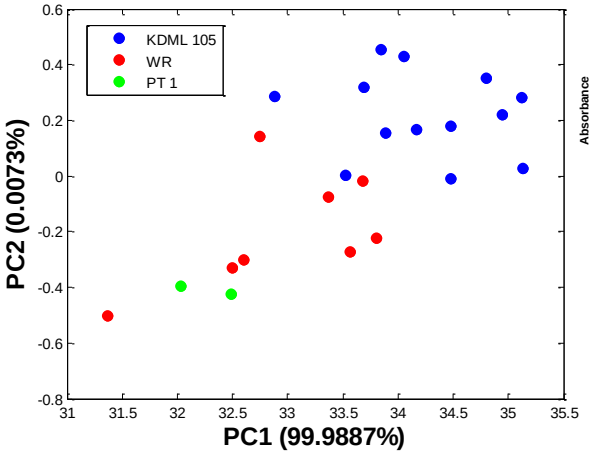


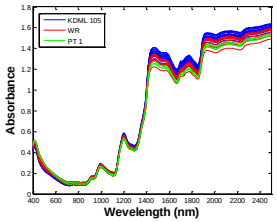




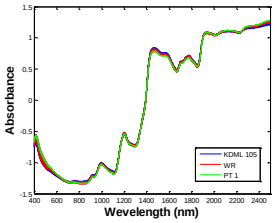


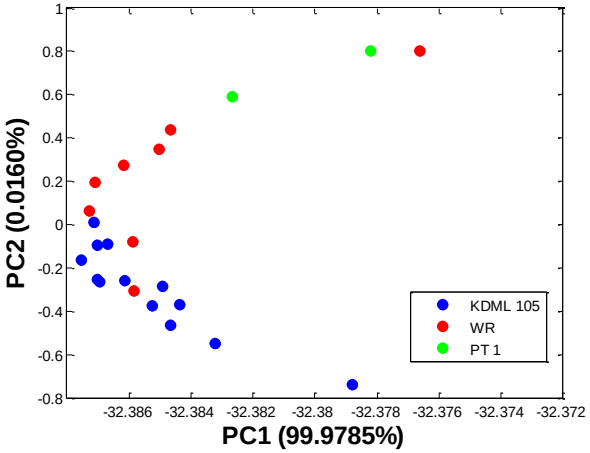
Raw data





SNV

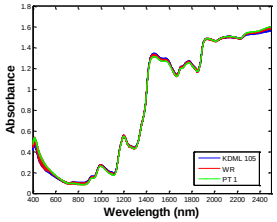


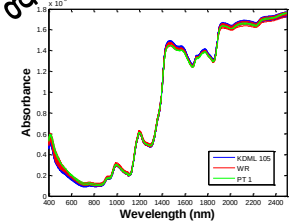


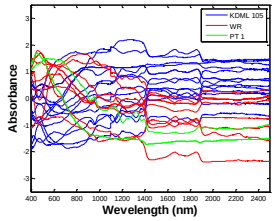
MSC

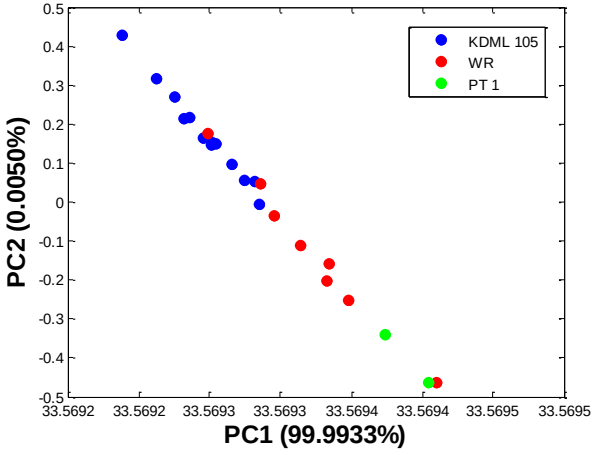
Raw scaling

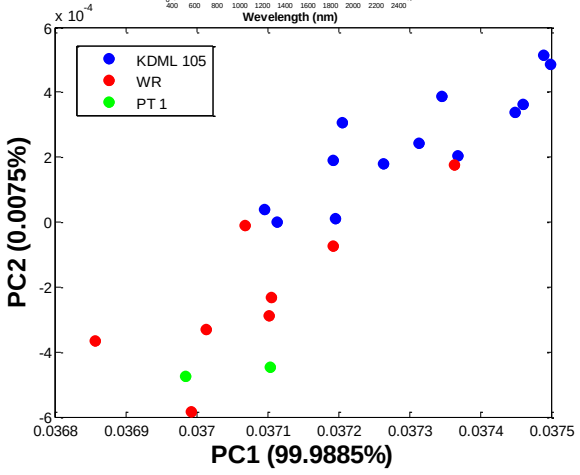
Standardization

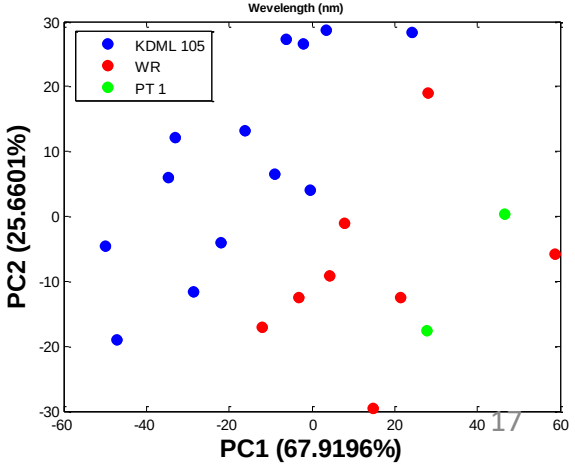




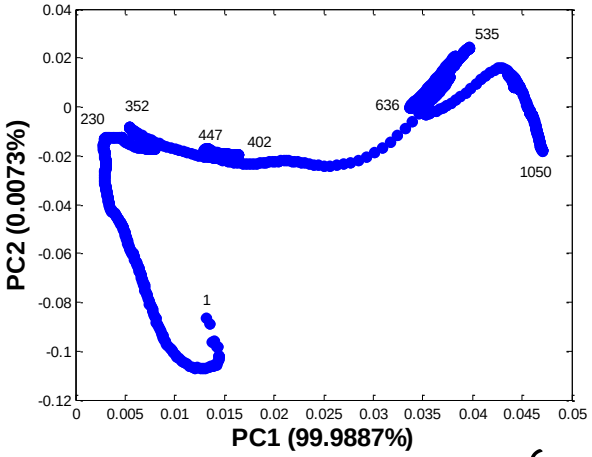


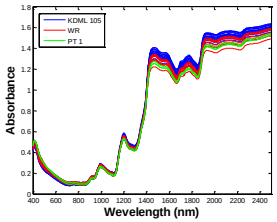


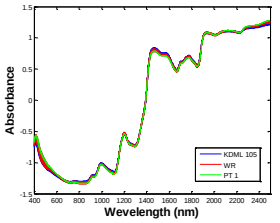




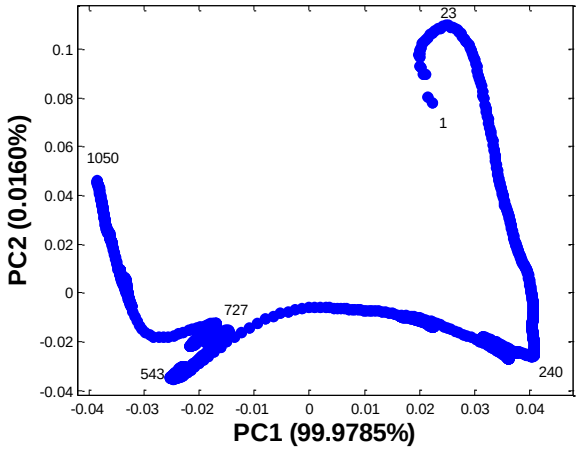
PCA loading plots







SNV



MSC

Raw scaling

Standardization