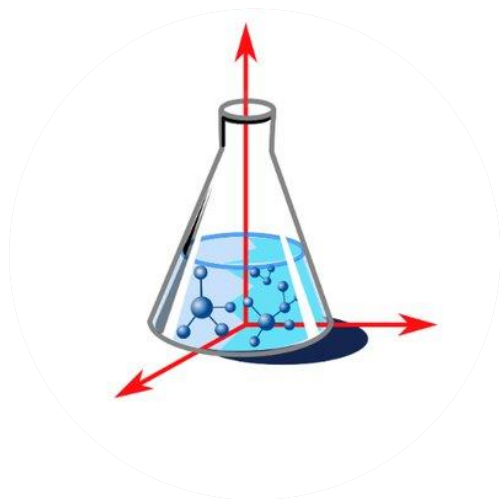


# Hands on data: Chemometrics

Antares

ANTARES 7th ANNUAL WORKSHOP

27<sup>th</sup> October, Mercedes Bertotto



# Workshop agenda: Chemometrics

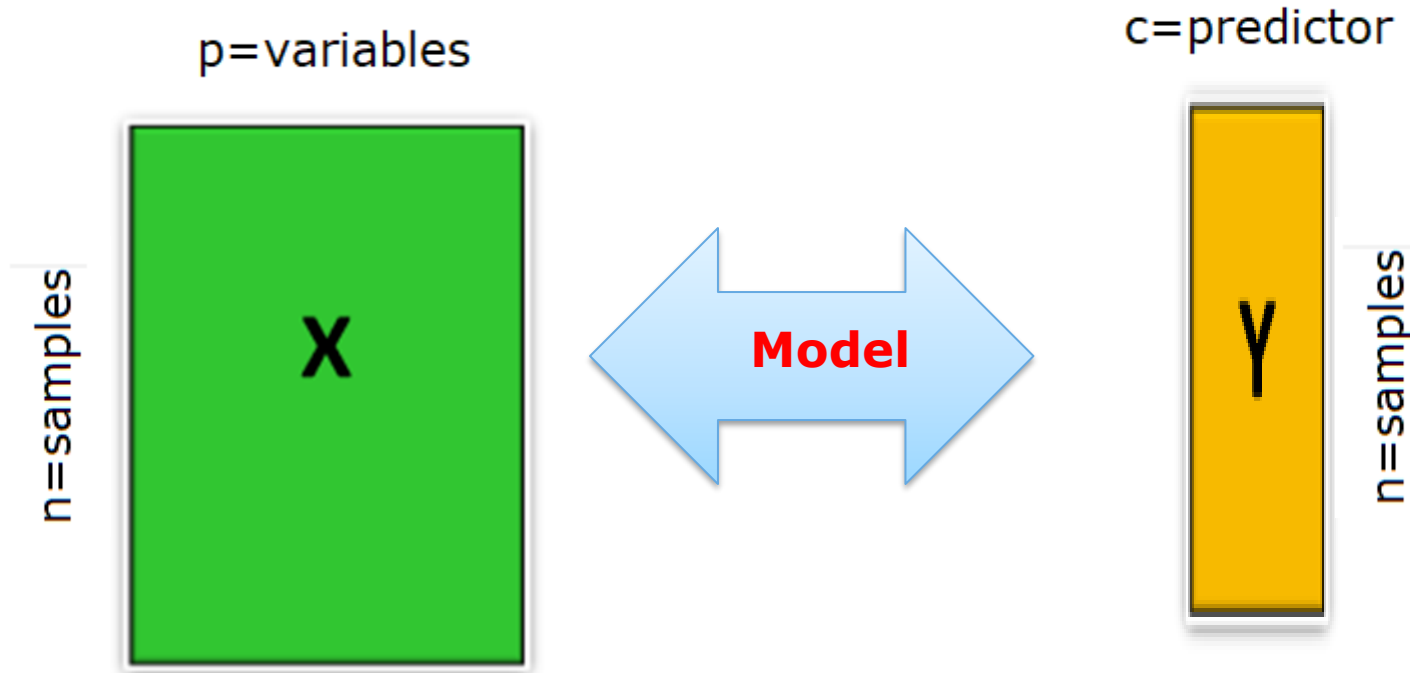
1. Introduction to Chemometrics
2. Exploratory analysis and Outlier Detection
3. Pretreatments on raw spectra
4. Feature selection
5. Cross validation or Data Split
6. Discrimination (PLSDA)
7. Hands on Data (30 min)

# Statistics versus chemometrics

- Focus on distributions
- Inference based on hypothesis tests of parameters
- Density estimation
- Mandatory courses at every biological education

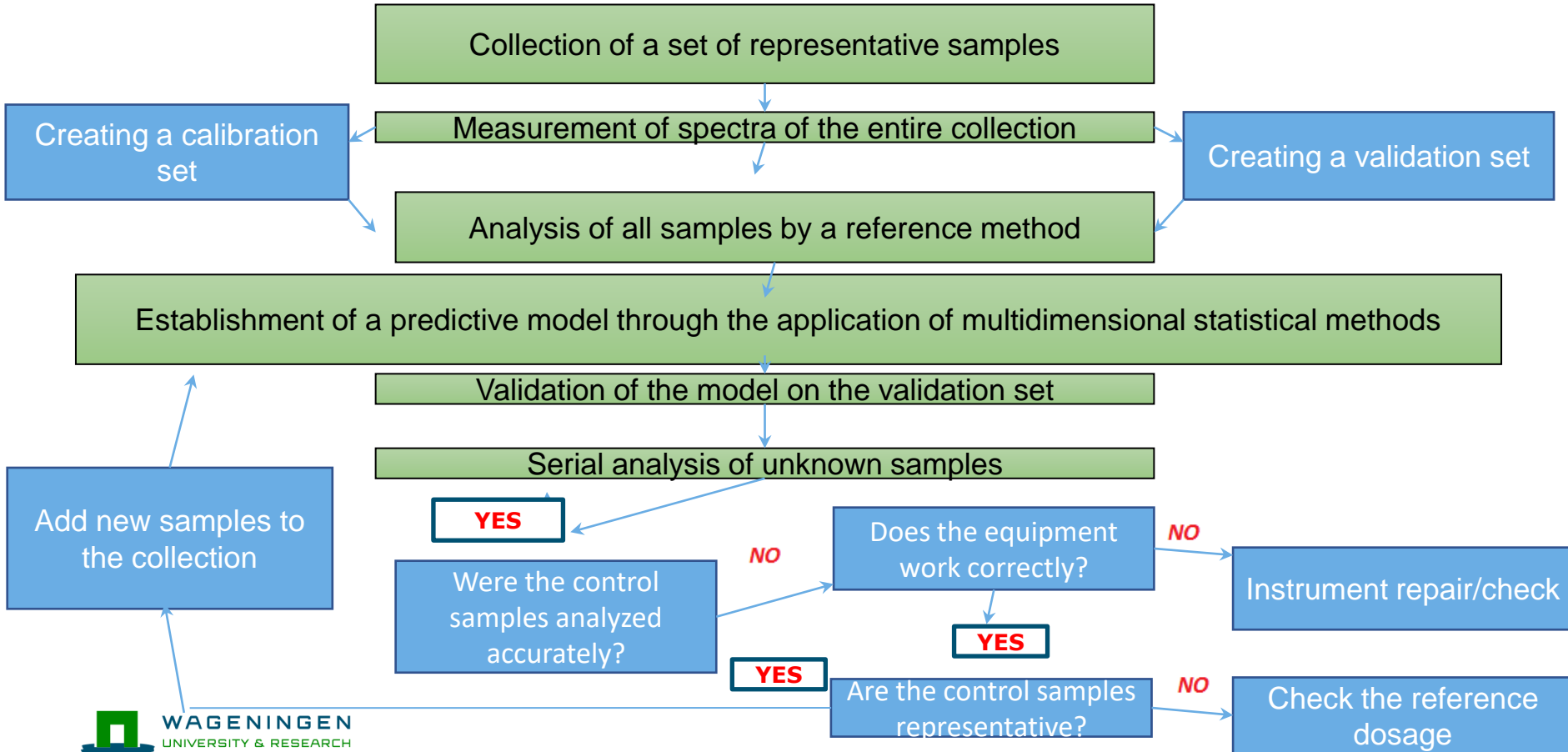
- Focus on individual samples
- Inferences based on future performance
- Clustering
- Voluntary course at only few institutions

# Multivariate analysis



**The predictor variables are partially selective of the chemical properties**

# GOOD MODELING PRACTICES (GMP)



## Part 2: Exploratory analysis and Outlier Detection

*"Delve deep enough into anything, and you will find mathematics" -*

Dean Schlicter

# Reducing dimensionality

- ❑ Identifying and Removing Irrelevant Variables to:
- ❑ Enhance Computational Performance
- ❑ Improve Model Interpretability and Result Comprehension
- ❑ Avoid overfitting and collinearity

One way to solve the problem: Transforming the initial set of variables into a lower-dimensional set while retaining most of the information

# Principal Component Analysis (PCA)

## Objectives:

- ❖ Exploration/ Description/ **Visualization**
- ❖ Dimensionality Reduction
- ❖ Preparation and Cleaning (**Outlier Identification**, Noise)
- ❖ Discrimination of Individual Groups
- ❖ Determination of Relationships Among Individuals
- ❖ **Time evolution of scores is used to detect deviations in process monitoring**
- ❖ **Preliminary Stage for Further Chemometric Treatment**



# Principal Component Analysis (PCA)



Original



R



G



B



PC1



PC2

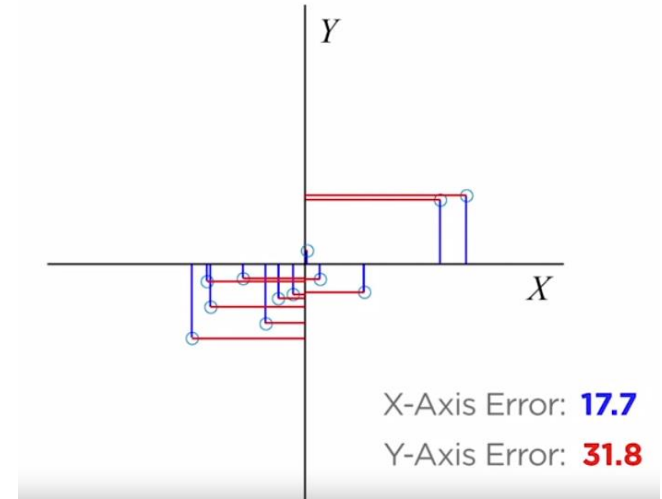
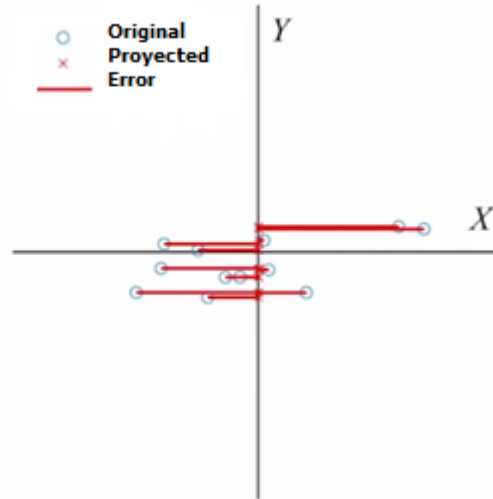
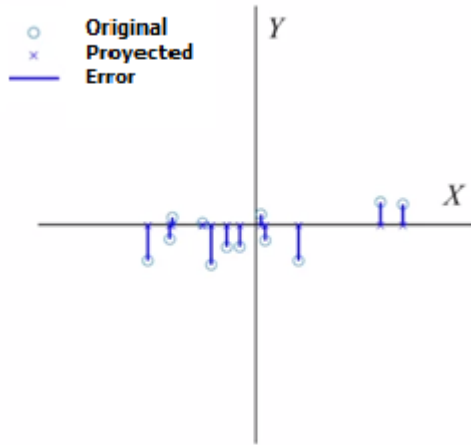


PC3

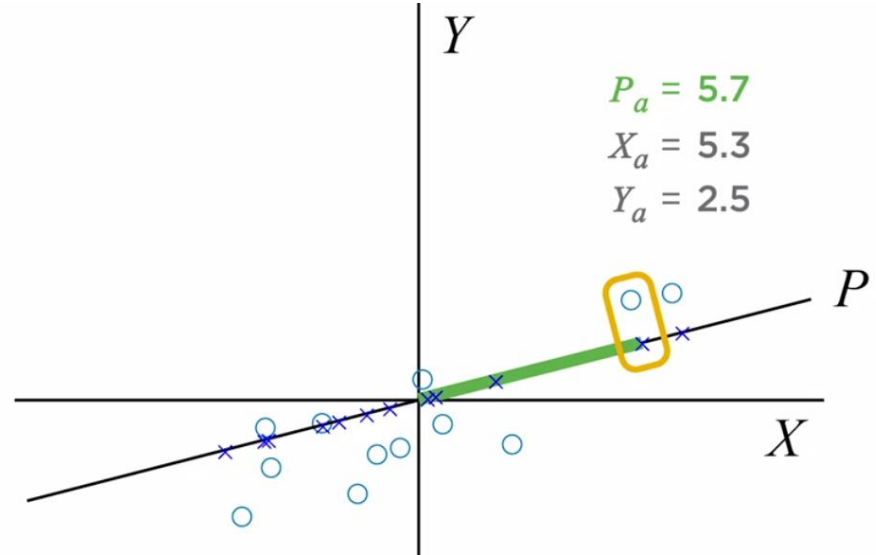
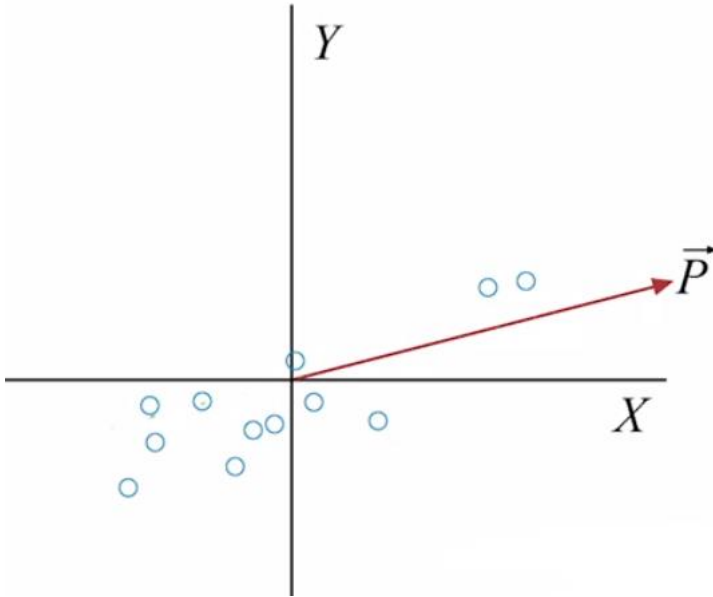


Residual

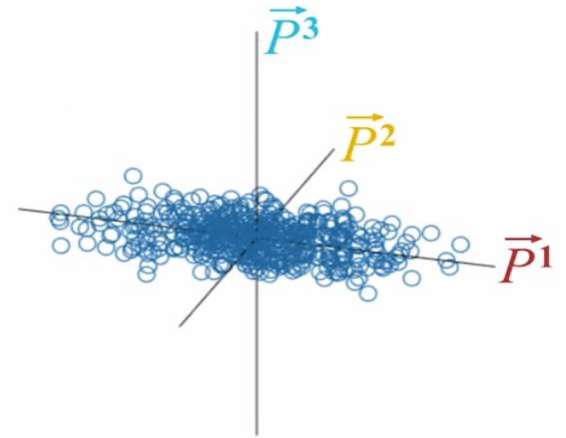
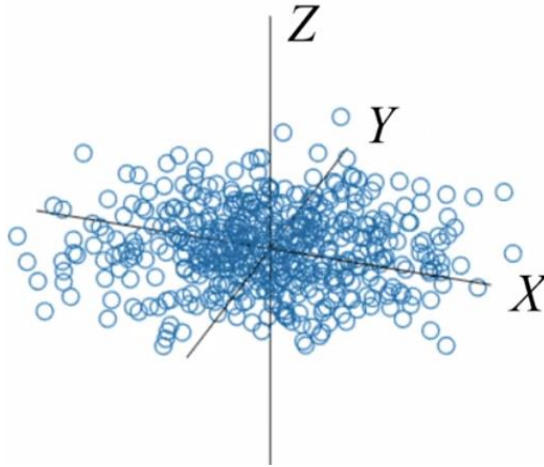
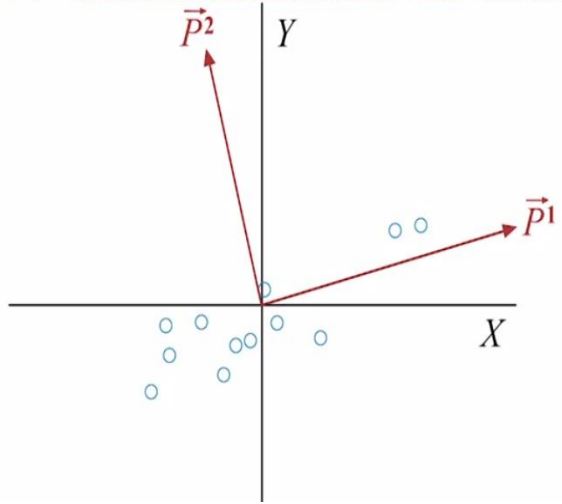
# Principal Component Analysis (PCA)



# Principal Component Analysis (PCA)



# Principal Component Analysis (PCA)



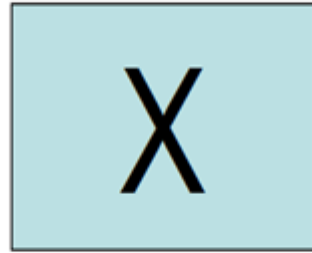
# PCA: Matrix decomposition

$(N \times P)$

$$X = T \cdot P' + E$$

$(N \times A)$

T: scores



$(P \times A)$

loadings P'

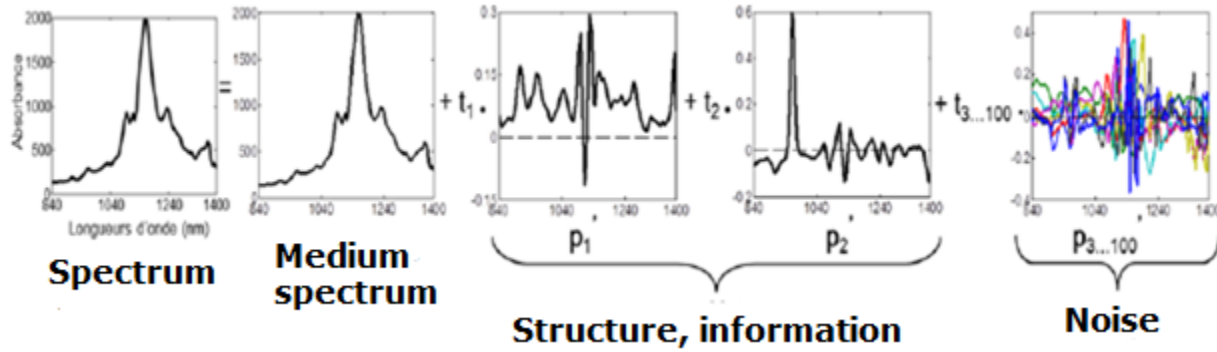


A: Number of PCs

$$X = TP^T + E = t_1 p_1^T + t_2 p_2^T + \dots + E$$

# PCA: Matrix decomposition

$$\begin{array}{c}
 \begin{array}{c} P \\ \boxed{\mathbf{X}} \\ N \end{array} = \begin{array}{c} \begin{array}{c} N \\ \boxed{\mathbf{t}_1} \end{array} \cdot \begin{array}{c} P \\ \boxed{\mathbf{p}_1'} \end{array} + \begin{array}{c} \begin{array}{c} N \\ \boxed{\mathbf{t}_2} \end{array} \cdot \begin{array}{c} P \\ \boxed{\mathbf{p}_2'} \end{array} + \begin{array}{c} \begin{array}{c} N \\ \boxed{\mathbf{t}_3} \end{array} \cdot \begin{array}{c} P \\ \boxed{\mathbf{p}_3'} \end{array} + \begin{array}{c} \begin{array}{c} N \\ \boxed{\mathbf{E}} \end{array} \end{array}
 \end{array}$$



# Change of basis: Linear transformation

$$\begin{bmatrix} 2 & -1 \\ 1 & 1 \end{bmatrix}$$

Vector in our  
"language"

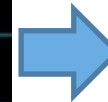


$$\begin{bmatrix} -1 \\ 2 \end{bmatrix}$$

Transformation



$$\begin{bmatrix} ? \\ ? \end{bmatrix} = -1 \begin{bmatrix} 2 \\ 1 \end{bmatrix} + 2 \begin{bmatrix} -1 \\ 1 \end{bmatrix} = \begin{bmatrix} -4 \\ 1 \end{bmatrix}$$



Vector in  
another  
"language"

# Eigenvalues and eigenvectors

- An eigenvalue of a matrix is a scalar that represents how a particular transformation, described by that matrix, stretches or compresses a vector in space
- Eigenvalues quantify how much the direction of a vector changes when it is multiplied by a matrix
- if  $A$  is a square matrix, and  $\lambda$  is a scalar, then  $\lambda$  is an eigenvalue of  $A$  if there exists a non-zero vector  $v$  (called an eigenvector) such that the following equation holds:

$$A * v = \lambda * v$$

**The value  $\lambda$  is the eigenvalue associated with the eigenvector  $v$**



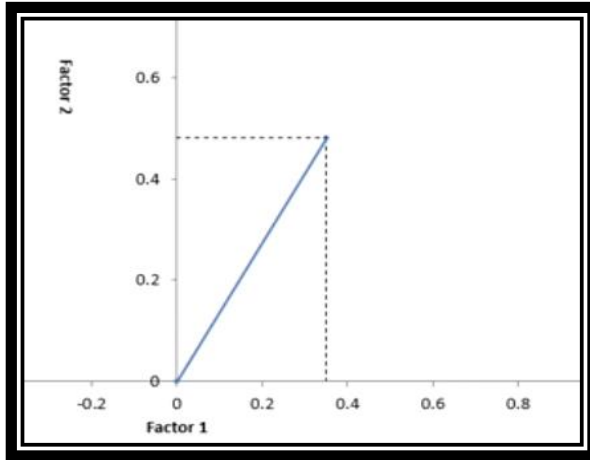
PCA explains the original variables through the strength of their relationship with the factors

Eigenvalues	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6	Factor 7	Factor 8
	4.349726834	1.92281186	1.290080831	1.149885024	1.129518656	1.014206025	0.977356029	0.9318098

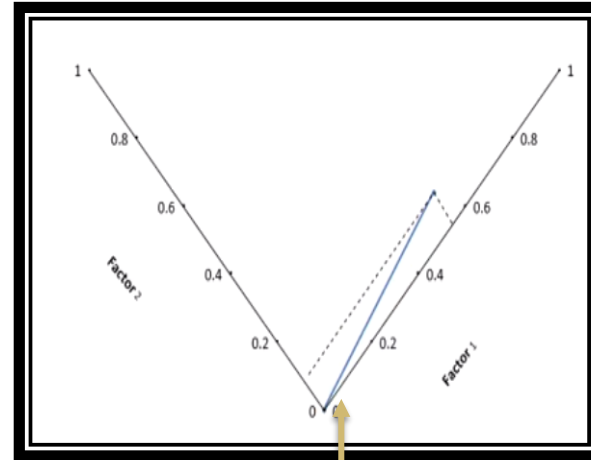
- ✓ The factors are extracted from the data **sequentially**
- ✓ The factor that explains the highest % of variance is extracted first
- ✓ Then, the factor that explains the second highest % of variance is extracted, and so on
- ✓ The factors are **orthogonal** to each other
- ✓ They have zero correlation between them, each representing something **unique**

# Factor rotation

Ideal



Real



The variables that are weakly related to the component are located near the center of the graph

## Loadings

### Correlation coefficients

-1.0 : Perfect negative correlation

0.0: No correlation

1.0: Perfect positive correlation

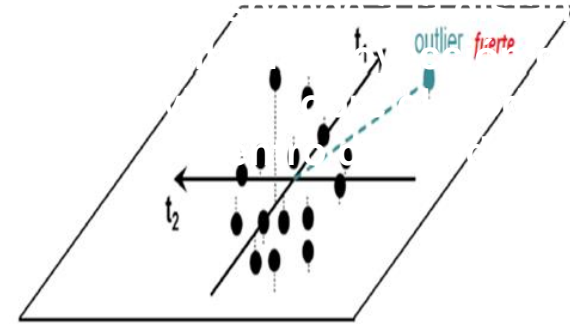
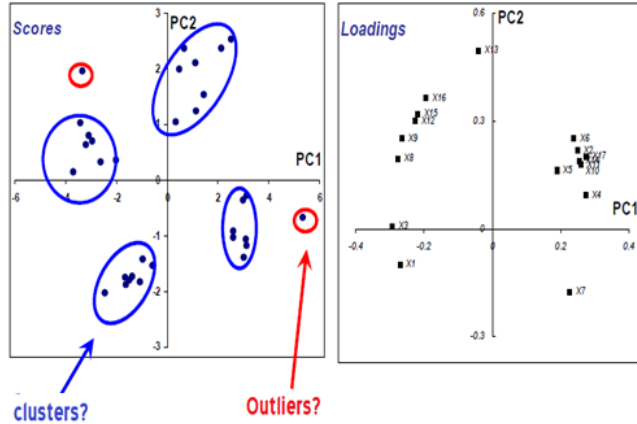
In yellow: Loadings > 0.5

The variables that load strongly on the same factor may share something in common. For example, products C, D, J, and U load strongly on the first factor. In this example, they are antibiotics

	Factor 1	Factor 2	Factor 3
Product A	0.338392	0.359125	-0.35551
Product B	-0.03337	0.359674	-0.44017
Product C	0.709268	0.313117	-0.10372
Product D	0.866839	0.026961	-0.09743
Product E	0.037134	-0.10703	-0.40144
Product F	0.140315	0.202643	-0.10759
Product G	0.118608	0.235802	-0.35763
Product H	0.029118	0.683191	0.2143
Product I	0.027812	0.522297	-0.23995
Product J	0.915644	-0.03241	-0.04724
Product K	-0.01446	0.280687	0.000529
Product L	0.138604	0.709171	-0.05589
Product M	0.017498	0.246222	-0.16969
Product N	0.329753	0.132799	-0.40502
Product O	0.136257	0.141867	-0.5055
Product P	0.189386	0.041223	-0.52597
Product Q	0.298836	0.418047	-0.32678
Product R	0.112088	0.08413	-0.55576
Product S	0.001687	0.16686	-0.55969
Product T	0.024188	0.358303	-0.17977
Product U	0.715734	-0.03241	-0.23856

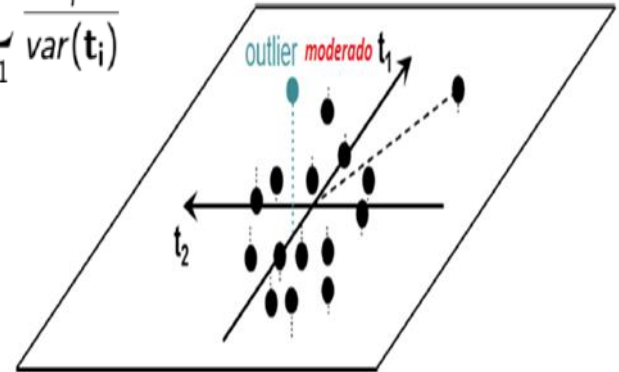
# Outlier detection

Hotelling or T2



$$T^2 = \sum_{i=1}^A \frac{t_i^2}{\text{var}(t_i)}$$

Q or residual



"The **test hypotheses** are:

- **Null hypothesis** ( $H_0$ ): the two samples are from populations with the same multivariate mean
- **Alternate hypothesis** ( $H_1$ ): the two samples are from populations with different multivariate means"

<https://www.statisticshowto.com/hotellings-t-squared/>

$$Q_i = e_i e_i^T = x_i (I - P_k P_k^T) x_i^T$$

# Isolation Forest

- We pick a column at random, and we select an arbitrary threshold within its range
- We divide all the samples into two groups according to that threshold
- There is a higher chance that the outlier point would end up in the smaller group or alone
- This randomized splitting process is repeated recursively
- The more common the point is, the more splits it will take to be isolated

Isolation depth: Number of partitions that it takes to isolate a point

One tree has a lot of  
variability



A final score is obtained by  
averaging results of many trees

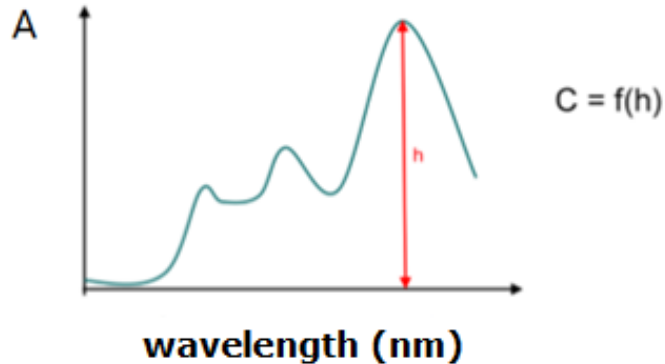
## Part 3: Pretreatments on Raw Spectra

# Near Infrared Spectroscopy

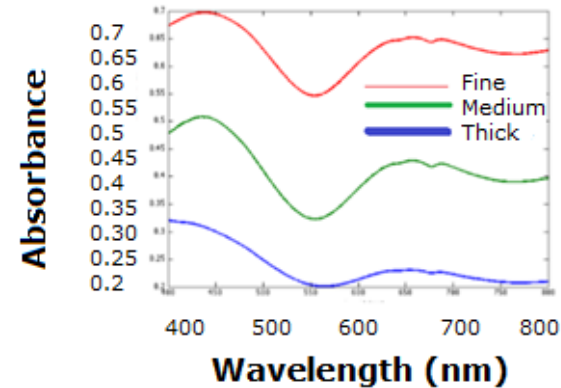
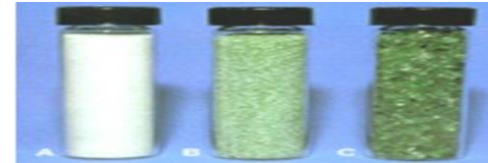
Ideal conditions

$$A(\lambda) = \varepsilon(\lambda)CL$$

Lambert-Beer Law



Real conditions



Pretreatments eliminate spectral deformations to approach the pure contribution of the studied parameter

## Multiplicative and additive effects

Real spectrum may be influenced by:

Photon diffusion



- 1) Enlarges the medium length of the optical path by factor **k**
- 2) Will cause a certain number of photons to escape from the captor, creating a leak term

Measurement noise



It is due to a set of random phenomena present throughout the measurement chain

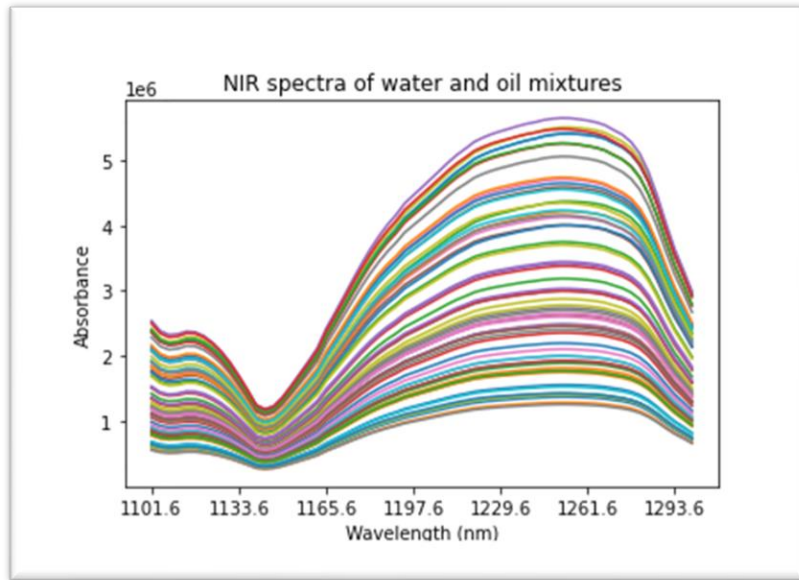
Effects  
Effects  
Additive effect

$$A(\lambda) = k\varepsilon(\lambda)LC + A_f(\lambda) + A_b(\lambda)$$

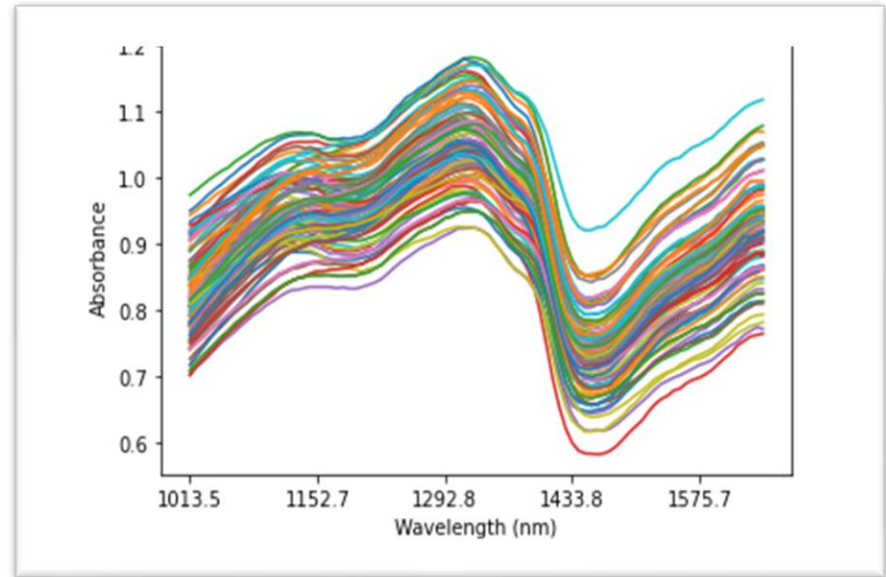


# How to distinguish additive and multiplicative effects

NIR spectra of water and oil mixtures showing a multiplicative effect. Source: WUR.

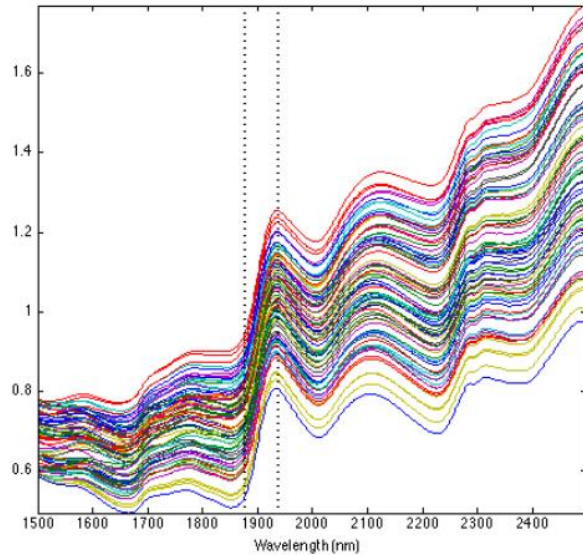


NIR spectra showing an additive effect. Source: WUR.

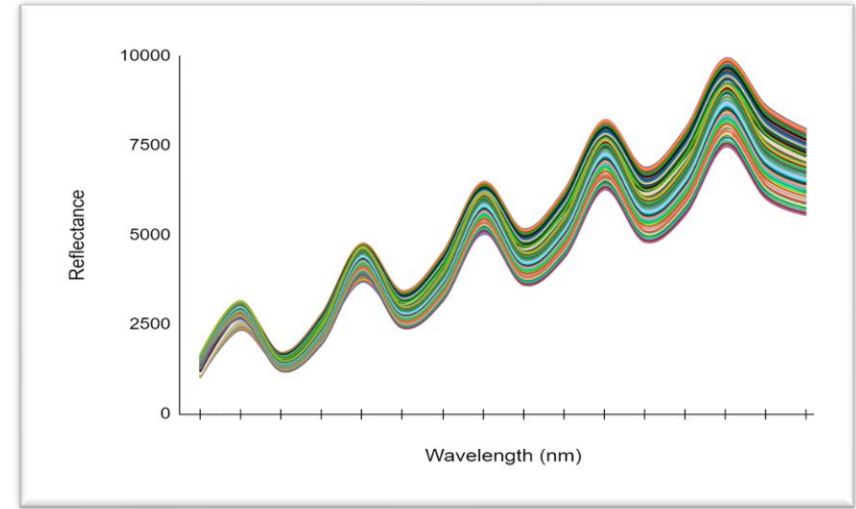


# How to distinguish additive and multiplicative effects

(Fictional) example of an additive effect of a rising baseline. Source: IRSTEA Montpellier

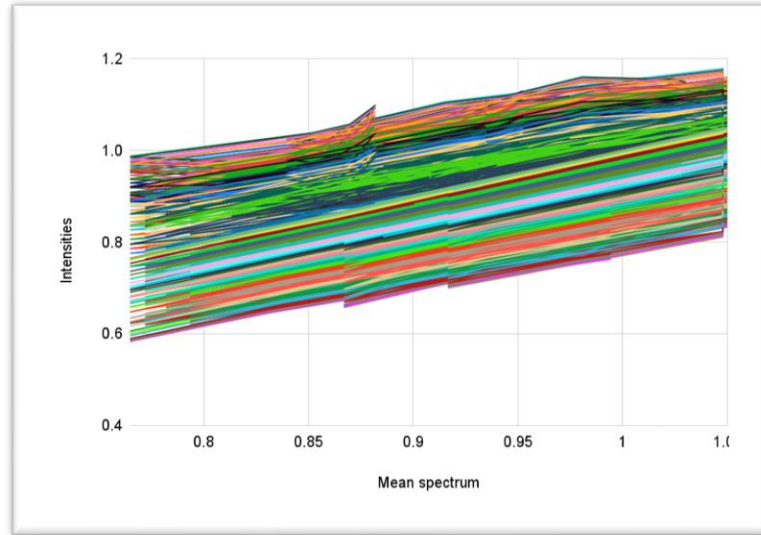


Handmade example of NIR spectra with combined effect (additive and multiplicative)  
Source: WUR

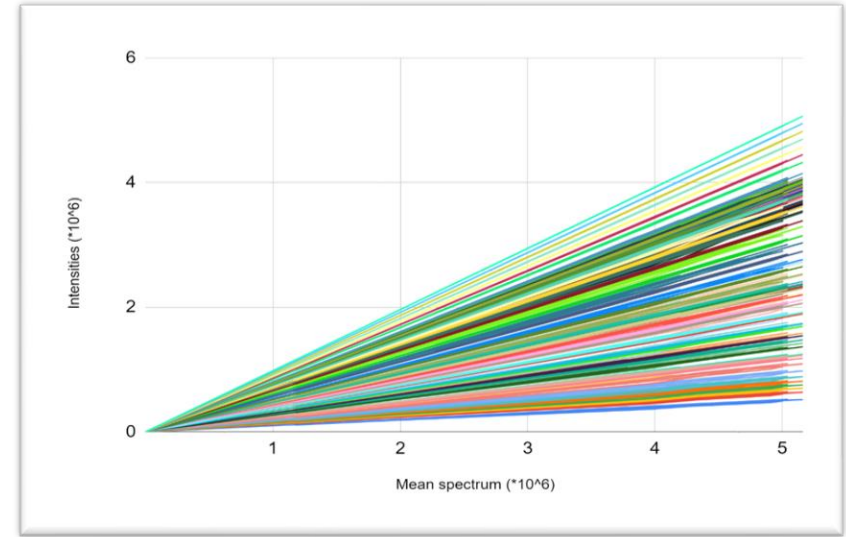


# How to distinguish additive and multiplicative effects

A millefeuille shape observed when plotting NIR spectra with additive effect versus the mean spectrum. Source: WUR.

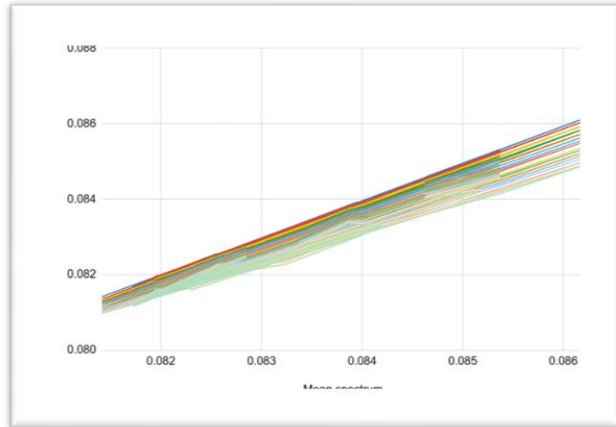


A cone shape observed when plotting spectra with multiplicative versus their mean spectrum. Source: from author

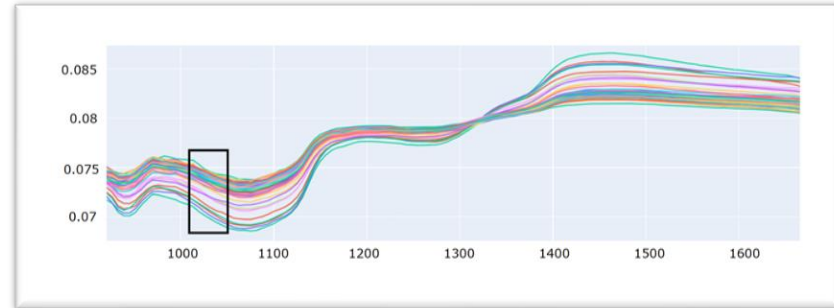


# How to distinguish additive and multiplicative effects

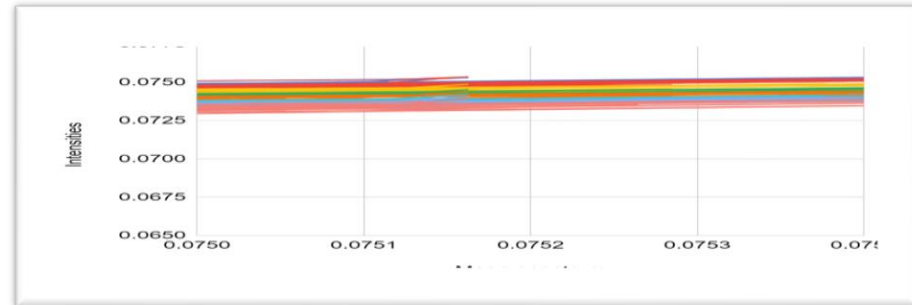
No clear cone neither millefeuille shape shown; difficult to tell if there is additive or multiplicative effect present in spectra. Source: WUR



Intensities from the black rectangle can be plotted versus their mean spectrum, to understand the type of effect present. Source: WUR



**This shows an additive effect in that spectral range. Source: WUR.**

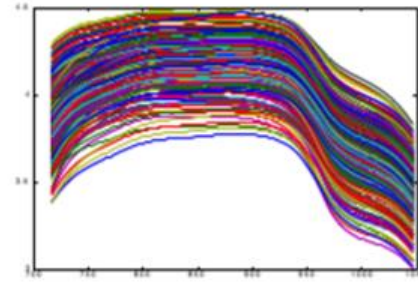


## How to reduce a multiplicative effect ( $k$ )?

$$\log(ab) = \log(a) + \log(b).$$

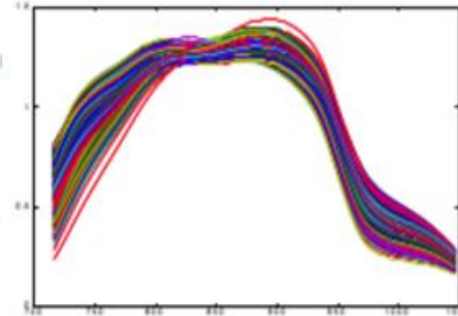
$$(k\mathbf{x}) \quad (\log(k)\mathbf{1} + \log(\mathbf{x})).$$

Logarithm



+ Detrend

Normalization

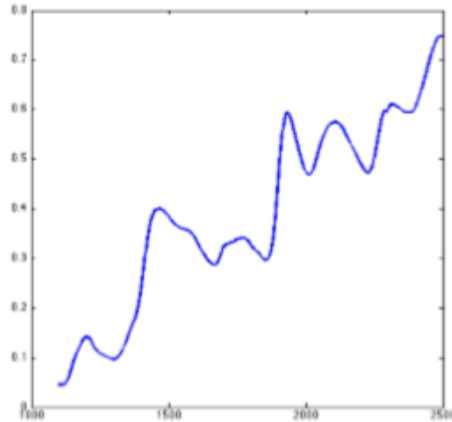


To remove noise

$$A_f(\lambda)$$

Detrend

Spectrum of Wheat. Source: IRSTEA Montpellier



Delete the global tendency from the spectrum, modeled by a polynomium

Order 0  
Average

Orden 1  
Line

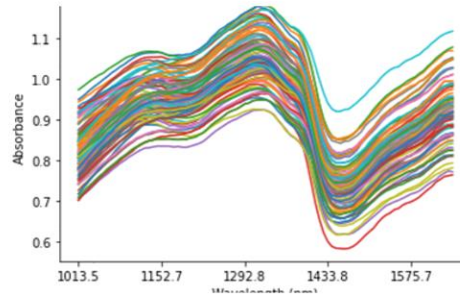
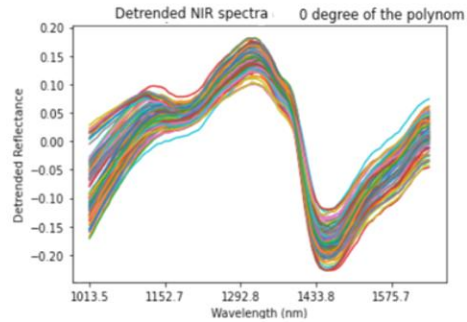
Orden 2  
Parabola

The residuals that remain after removing the line correspond to the absorbance peaks related to the chemical components of the sample

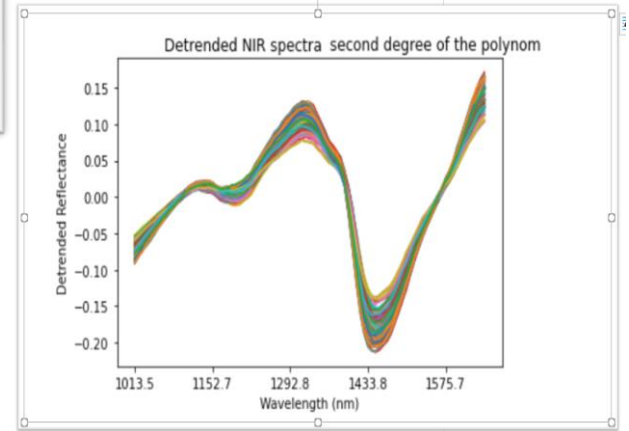
# Detrend

Raw spectra

Detrend  
order 0



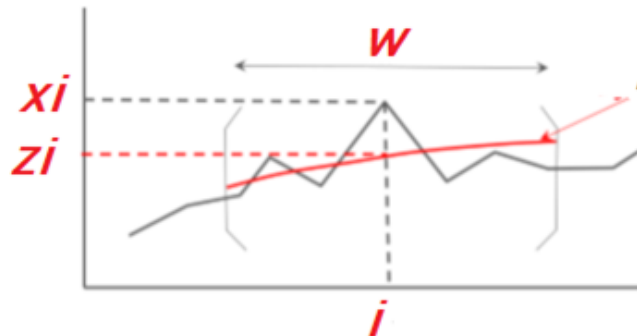
Detrend  
order 2



To remove noise

$$\mathcal{A}_b(\lambda)$$

polynomial of degree **d**



Savitsky Golay

$$w(\text{odd}) > d$$

At each point  $i$  in the spectrum, the raw value  $x_i$  is replaced by  $z_i$  from a polynomial fitted over a window around point  $i$ .



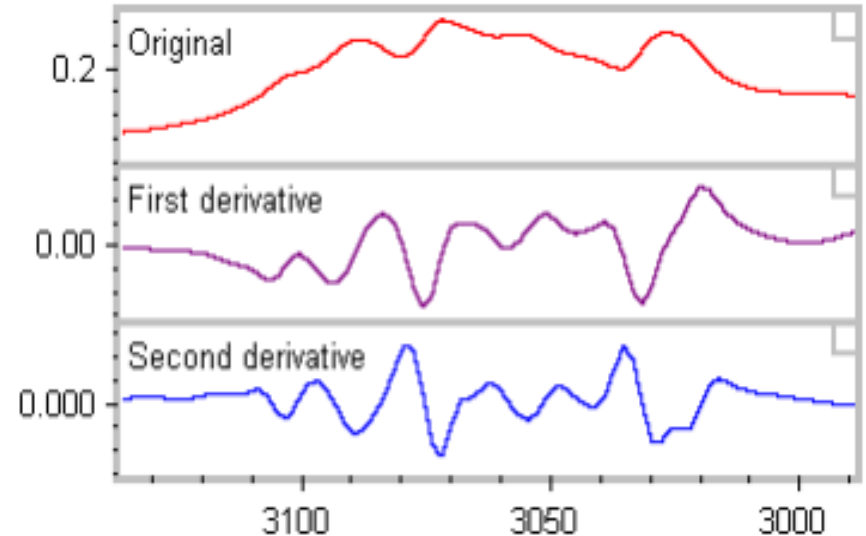
# Derivatives

- They are also used to decrease the baseline. If the baselines are polynomials of degree **K**, then the derivatives of order **k+1** will make them disappear

$$\mathcal{A}_f(\lambda) = a\lambda + b$$

$$\mathcal{A}(\lambda) = k\varepsilon(\lambda)LC + a\lambda + b$$

$$\frac{\partial^2 \mathcal{A}(\lambda)}{\partial \lambda^2} = k \frac{\partial^2 \varepsilon(\lambda)}{\partial \lambda^2} LC$$



# Standard Normal Variate (SNV)

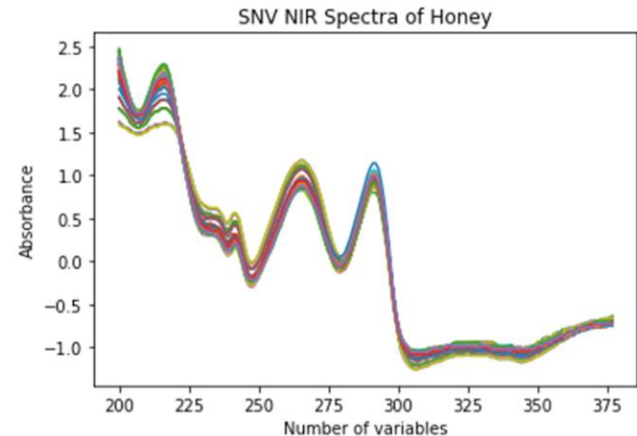
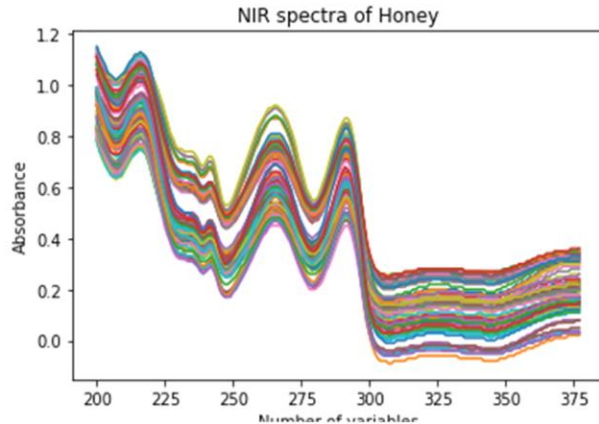
*The SNV pre-processing of a spectrum  $X$  consists of removing the mean of  $X$  at each of its points, then dividing them by the standard deviation of  $X$ .*

NIR spectra of honey, with both additive and multiplicative effects. Source: SENASA (Argentina)

$$X_c = X - \text{mean}$$
$$X_{SNV} = X_c / SD(X_c)$$

SNV NIR spectra of honey. Additive and multiplicative effects are gone.

Source: SENASA (Argentina)



# Multiplicative Scatter Correction (MSC)

- *“Mathematically, if we call  $X_m$  the mean spectrum, the multiplicative scatter correction is done in two steps.*

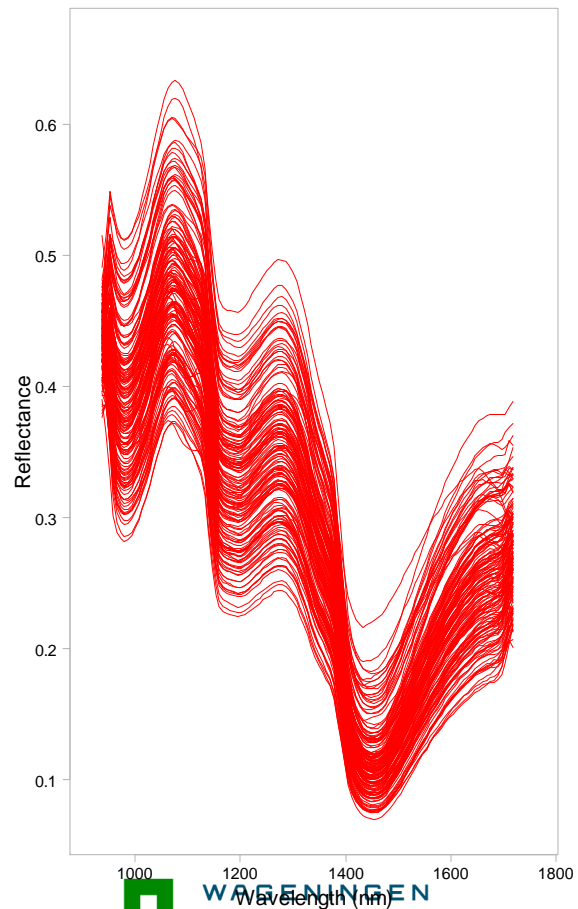
*1. We first regress each spectrum  $X_i$  against the mean spectrum*

*This is done by ordinary least squares:  $X_i \approx a_i + b_i X_m$  .*

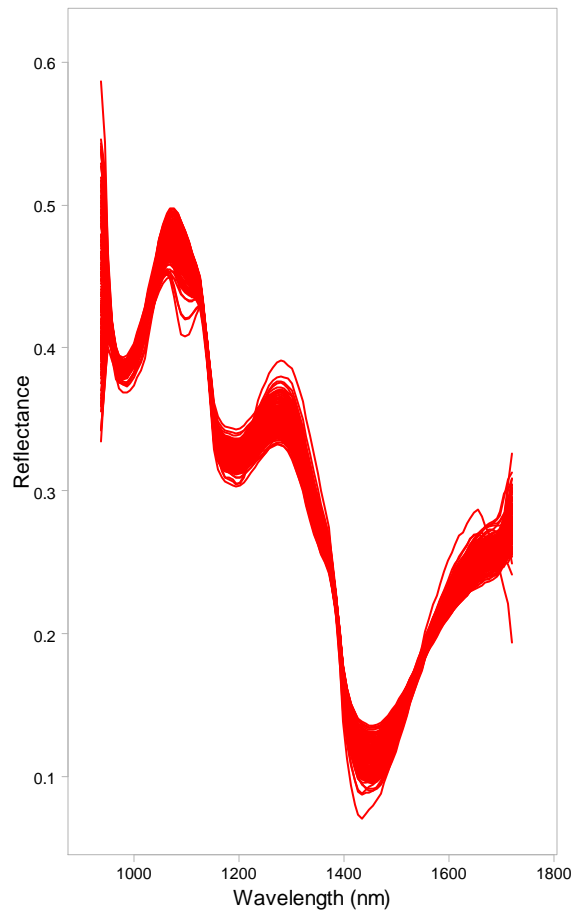
*1. We calculate the corrected spectrum  $msc = (X_i - a_i) / b_i$ ”*

Source: <https://nirpyresearch.com/two-scatter-correction-techniques-nir-spectroscopy-python/>

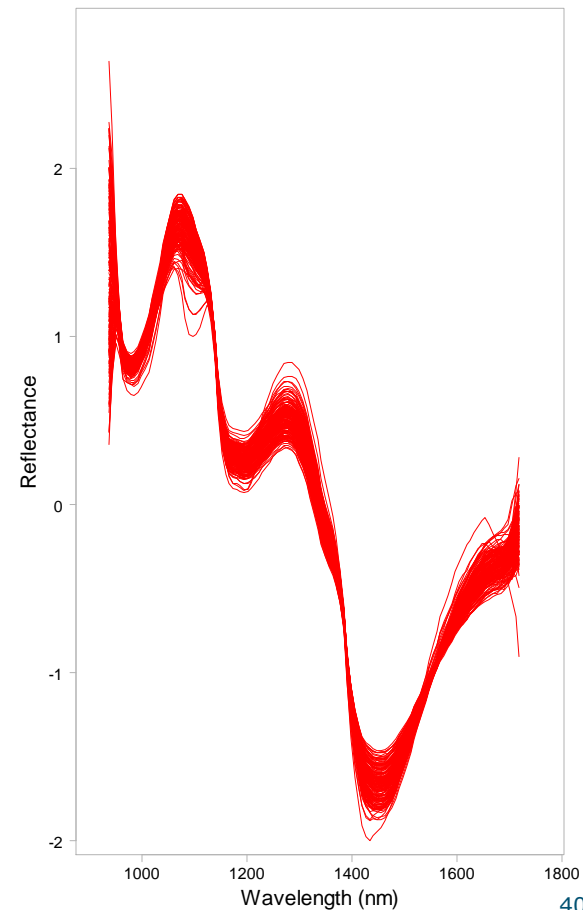
**Before MSC**



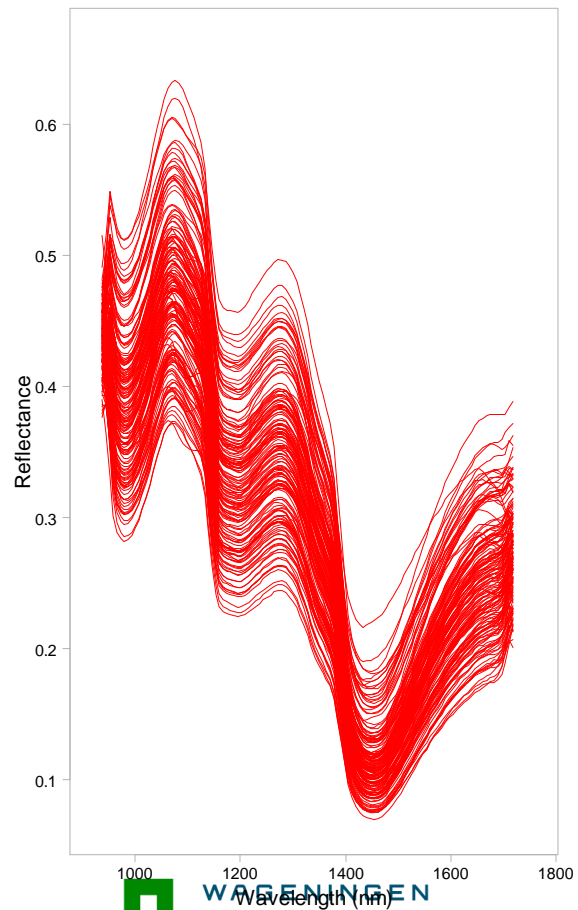
**After MSC**



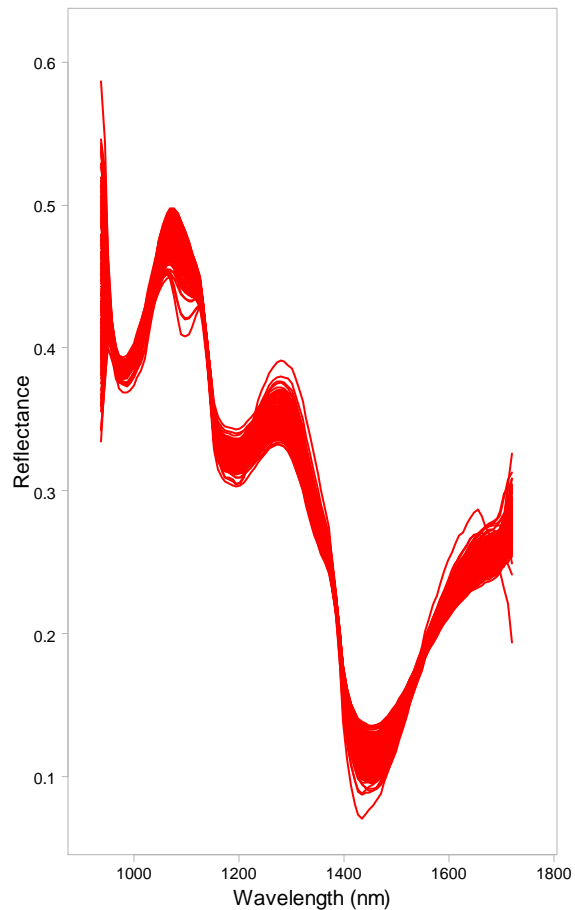
**After SNV**



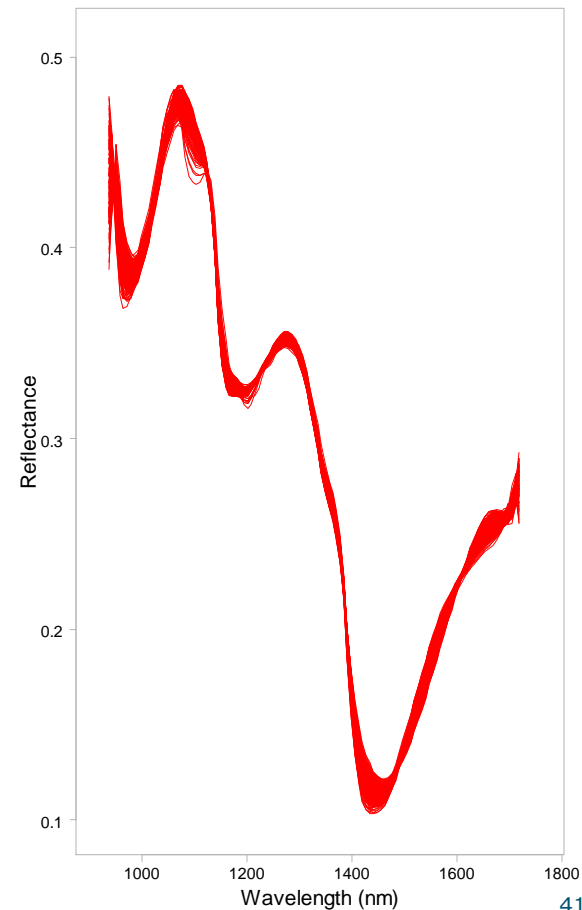
**Before MSC**



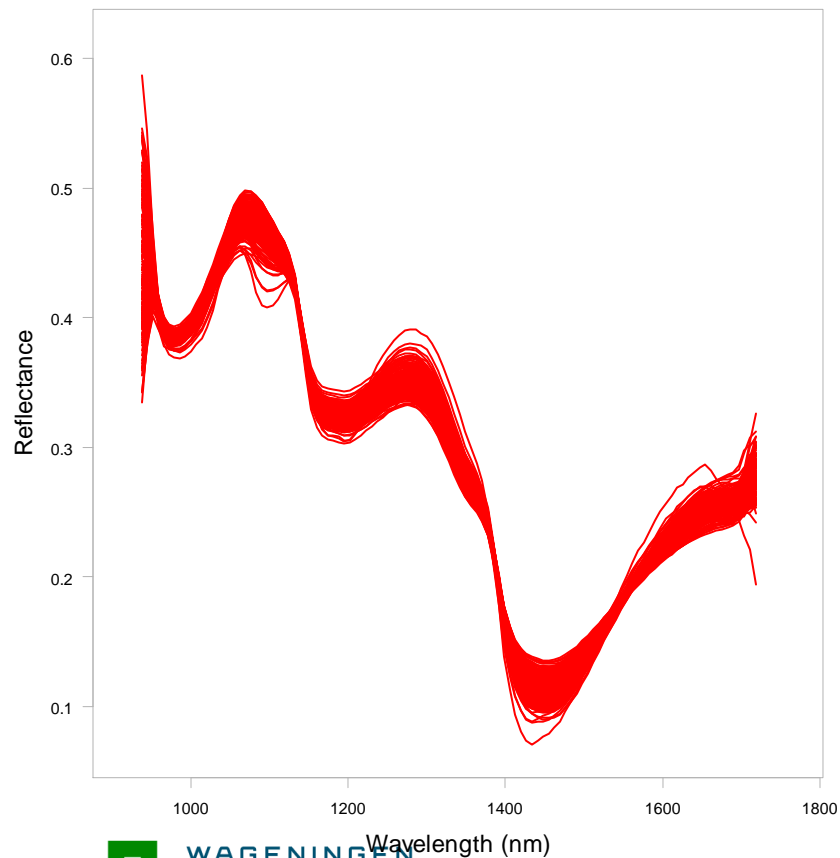
**After MSC**



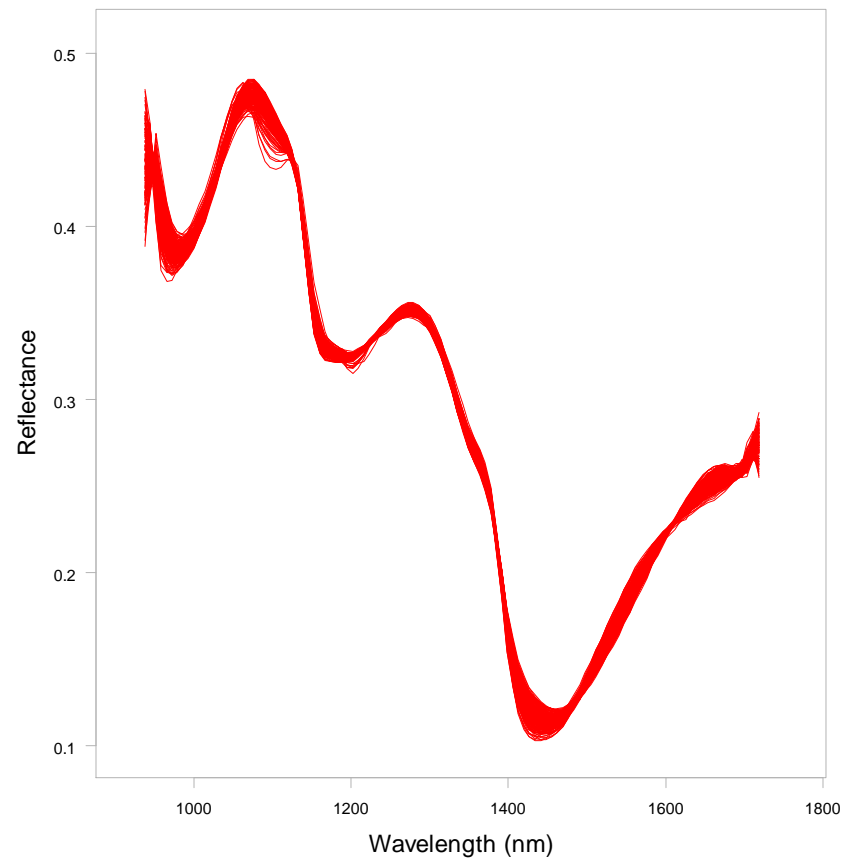
**After EMSC**



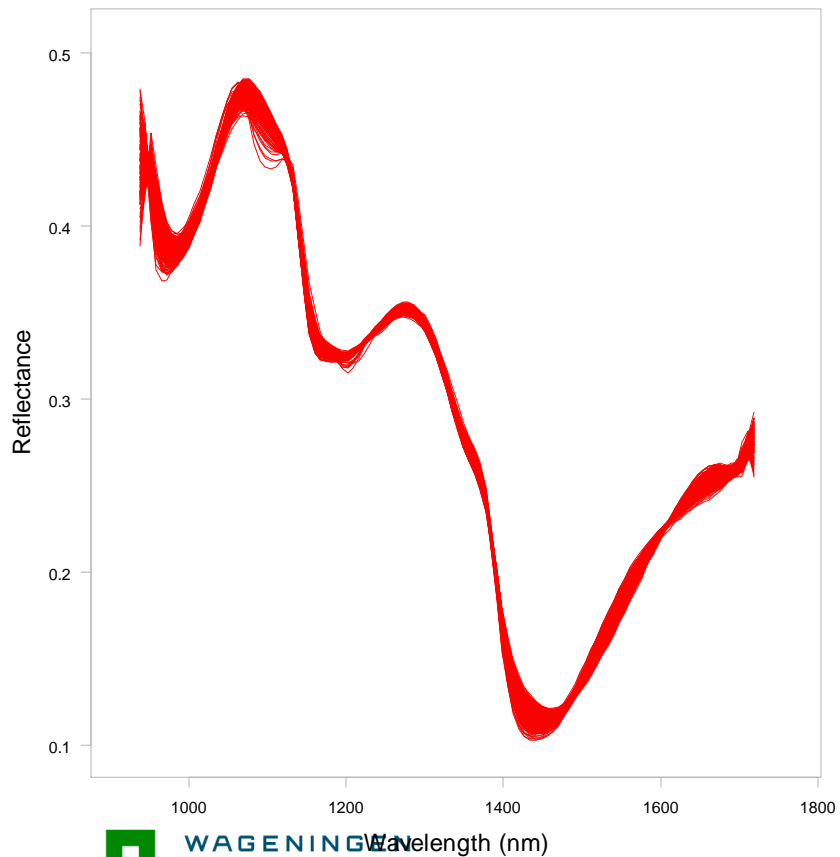
After MSC



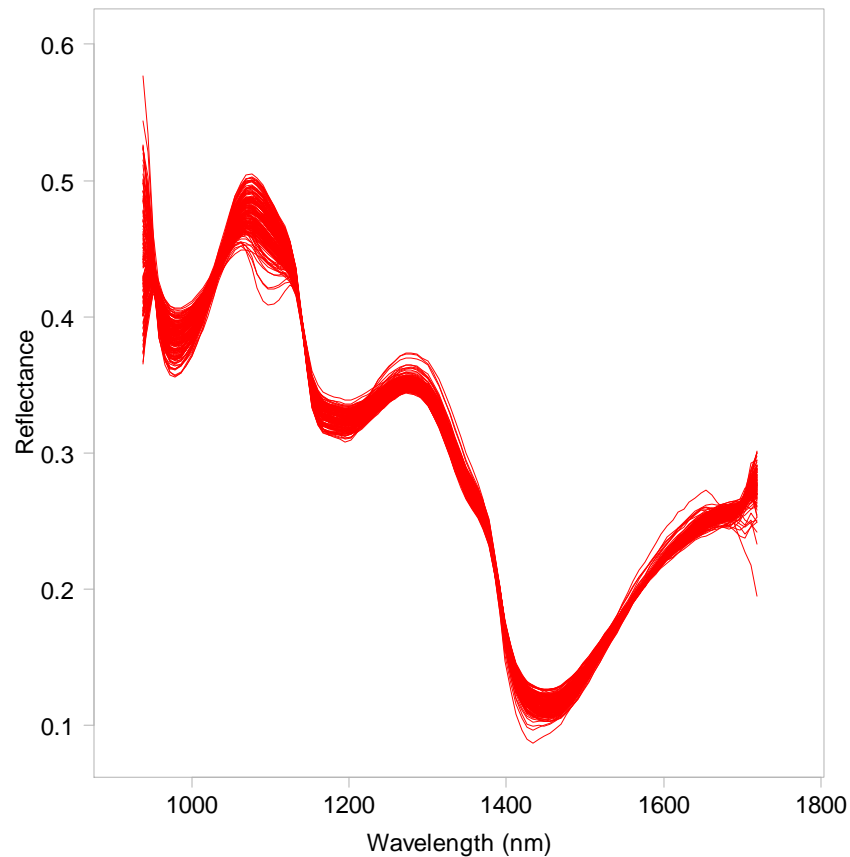
After EMSC, degree=6

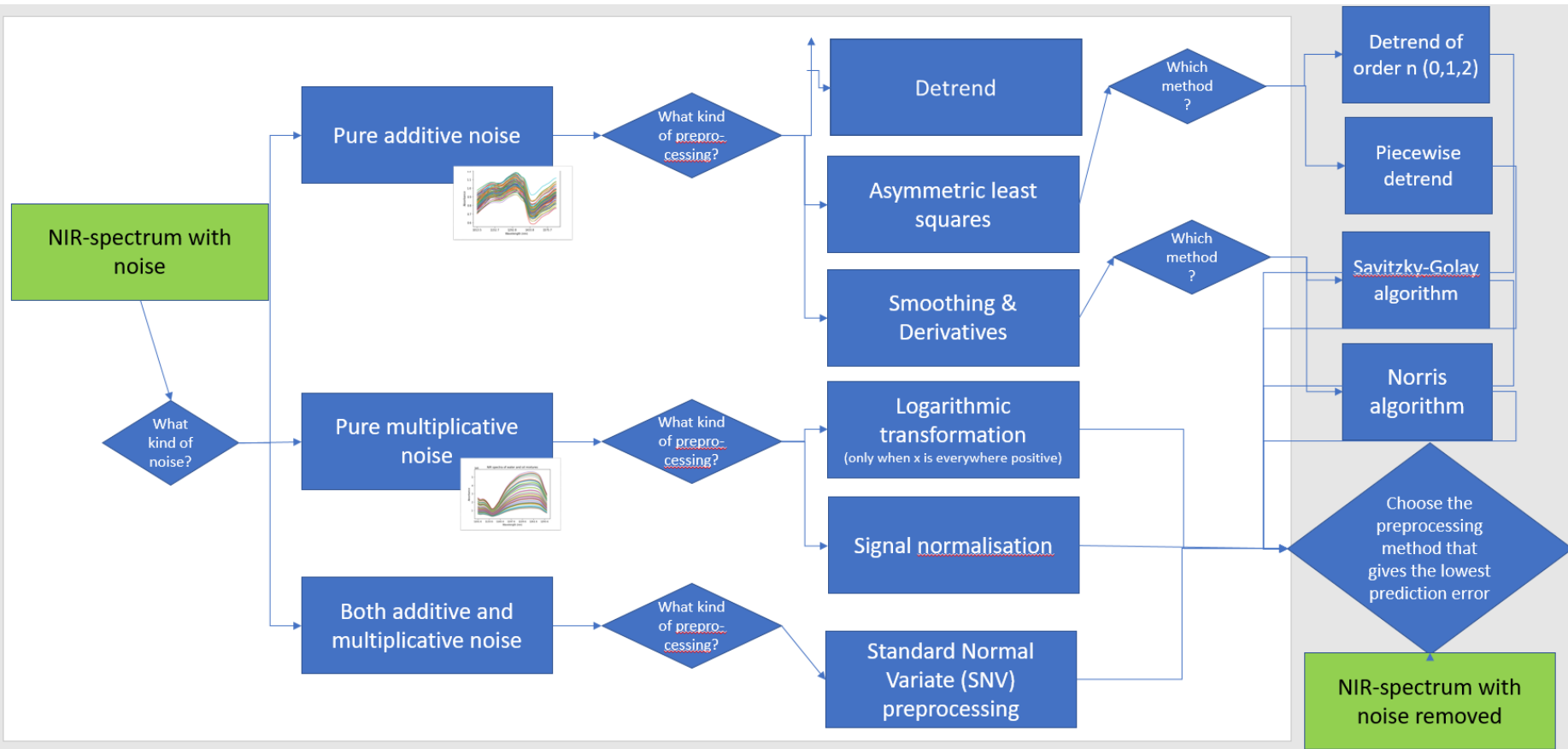


After EMSC, degree=6



After EMSC, degree=2







# Principles and Objectives of Discrimination

- We have a data matrix  $(n \times p)$  where  $n$  samples were measured for  $p$  quantitative variables, and a vector  $Y$  of size  $n$  measured on the same samples
- This vector represents the membership of each sample to each class  $K$
- Each class contains at least one sample, and each sample belongs to a single class

## Part 7: Hands on data

# Hands on data: R code

<https://github.com/MMBDD/Serbia>

Name	Functionality
OutlierDetection.R: NEEDED	Removes outliers Averages Spectra from the same Sepal into only One Spectrum per Sepal
GlobalModel.R: NEEDED	Creates PLSDA Classification models with different number of important variables as input for PLSDA
PLSDA.R: OPTIONAL	Creates PLSDA Classification models with optimized parameters using Rchemo
PLSDAmdatools.R: OPTIONAL	Creates PLSDA and SIMCA Classification models using mdatools
MSC.R: OPTIONAL	Pretreats raw data and plots the results

# Description of initial datasets

Cultivar name/ number of	Pixels per sepal	Sepals per tomato	Tomatoes per image	Images	Spectra in the initial dataset	Spectra in the averaged dataset	Variables
Provine	Between 119 and 90	5 or 6	16	2	16156	159	112
Brioso	Between 45 and 53	5 or 6	32	1	6497	164	112
Cappricia	Between 81 and 124	5 or 6	16	2	12816	165	112

Number of spectra in each class (Healthy: Class 1; Diseased: Class 2) when dataset was split according to different labelling scenarios (Label 1: 0/123; Label 2: 01/23 and Label 3: 0.5/123).

Cultivar	n	Label 1		Label 2		Label 3	
		Healthy	Diseased	Healthy	Diseased	Healthy	Diseased
Cappriccia	163	139	24	85	78	117	46
Brioso	153	145	8	78	75	126	27
Provine	152	137	15	72	80	129	23

# References

- <https://www.chemproject.org/ChemHouse>
- <https://www.fun-mooc.fr/en/courses/?limit=21&offset=0>
- Personal Material from the International School of Chemometrics 2023
- [https://eigenvector.com/wp-content/uploads/2022/10/Onion\\_SampleSelection.pdf](https://eigenvector.com/wp-content/uploads/2022/10/Onion_SampleSelection.pdf)
- <https://nirpyresearch.com/two-scatter-correction-techniques-nir-spectroscopy-python/>
- J.M. Roger, B. P., D. Bertrand, E. Fernandez-Ahumada (2011). "CovSel: Variable selection for highly multivariate and multi-response calibration. Application to IR spectroscopy." Chemometrics and Intelligent Laboratory Systems 106(2).
- J.M. Roger, B. P., D. Bertrand, E. Fernandez-Ahumada. (2011). CovSel: Variable selection for highly multivariate and multi-response calibration. Application to IR spectroscopy. Chemometrics and Intelligent Laboratory Systems, 106(2). doi:ff10.1016/j.chemolab.2010.10.003f

# Thank you for your attention

Please, open your R studio!

Any questions or comments?

