AN OVERVIEW ON ADVANCED CHEMOMETRIC APPROACHES FOR (N)IR SPECTROSCOPY

Federico Marini

Dept. Chemistry, University of Rome "La Sapienza", Rome, Italy

SAPIENZA
UNIVERSITÀ DI ROMA

24èmes Rencontres

Hélio SPIR

Retour à Montpellier

du 13 au 15 Juin 2023

La spectroscopie proche-infrarouge : de la mesure à l'analyse des spectres

Classification

© 2003, Mike Twohy. Dist. By The Washington Post Writers Group
M2Ecomics@aol.com
VERY
SORT OF
VERY SORT OF
M. Twohy
1-22

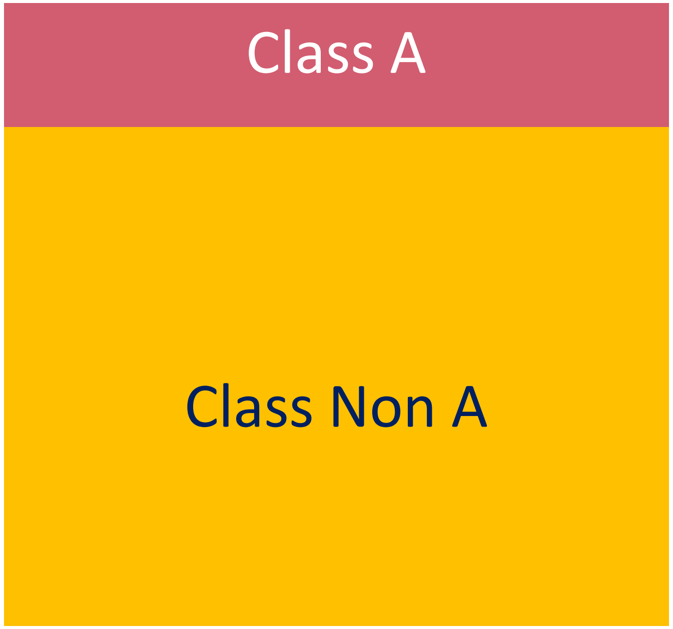
"Who's that, flyin' up there?
Is it a bird? no
Is it a plane? no
Is it the twister? Yeah"

Chubby Checker

- "To find a criterion to assign an object (sample) to one category (class) based on a set of measurements performed on the object itself"
- Category or class is a (ideal) group of objects sharing similar characteristics
- In classification categories are defined a priori

What if….

- Classes are not well defined, or
- There is only a single class of interest to be discriminated from all the rest (asymmetric classification)

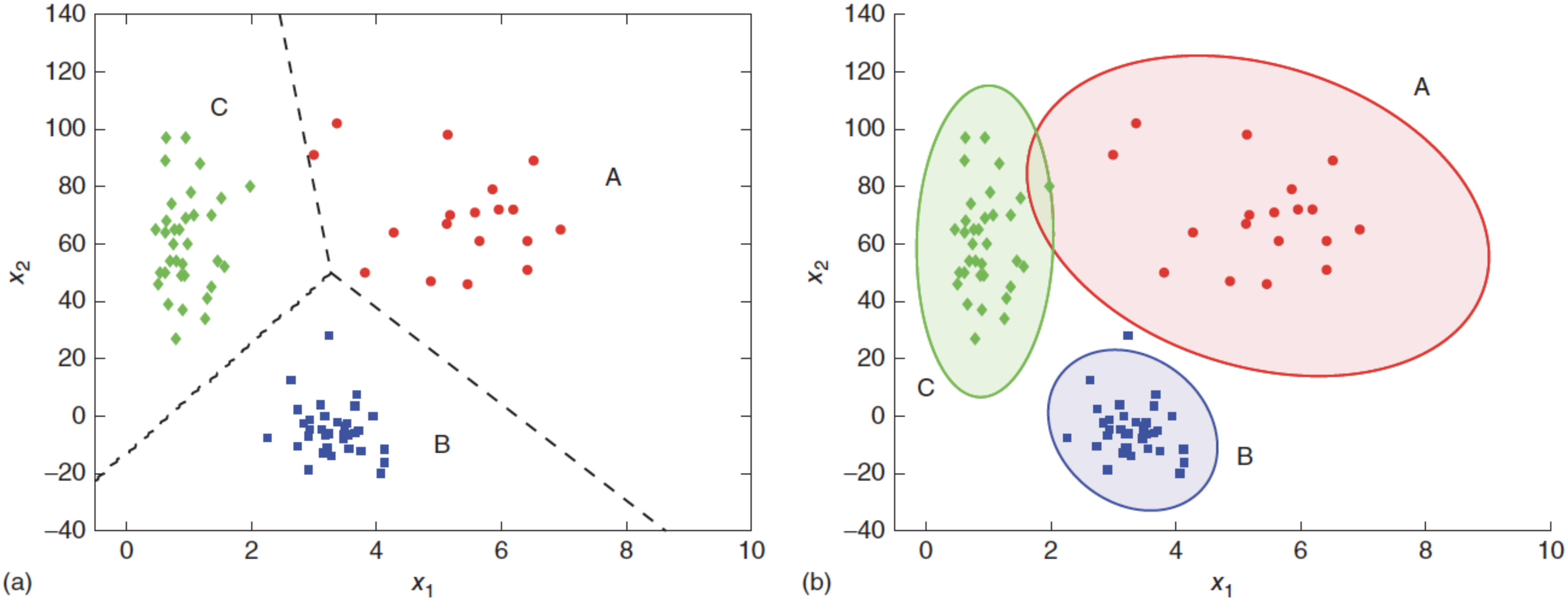


Class A

Class Non A

GOOD DOG!
BAD DOG!
THEY RARELY TALK TO US LIKE THAT BECAUSE WE'RE FAR TOO COMPLEX TO BE CLASSIFIED IN SUCH A SIMPLE, BINARY WAY.
THAVES 8-6
E-mail: ThavesOne@aol.com

A different approach: Class modeling



discriminant methods

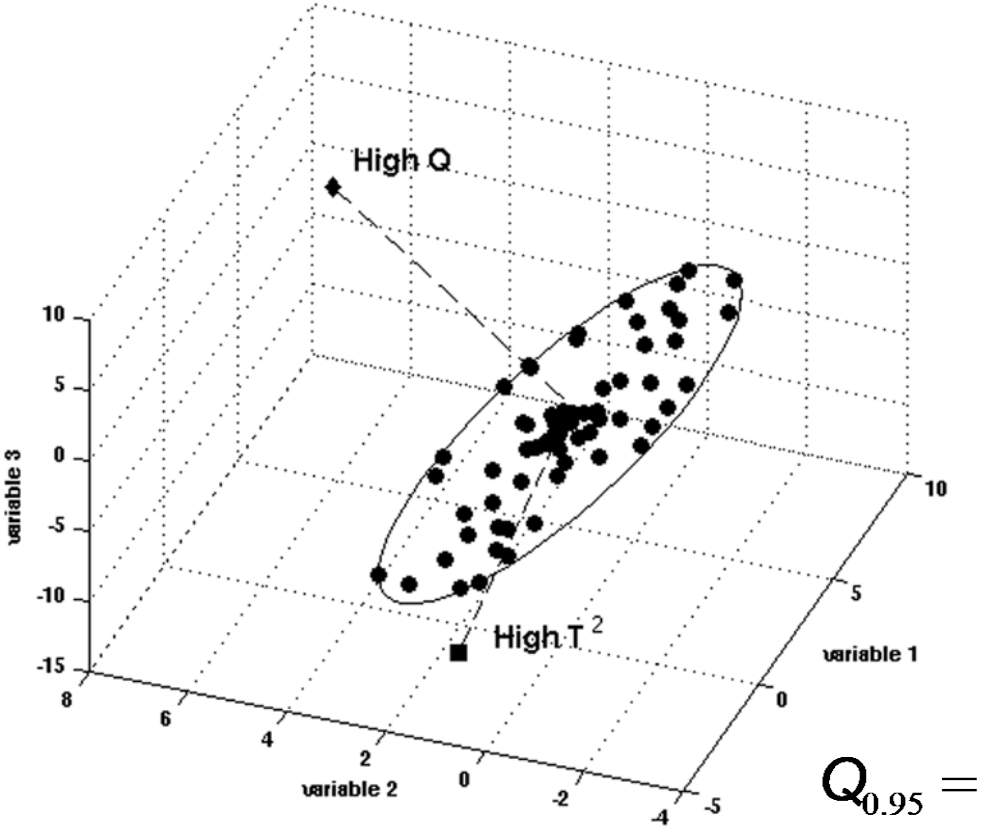
class-modeling

SIMCA

- Originally proposed by Wold in 1976
  - **SOFT**: No assumption of the distribution of variable is made (bilinear modeling)
  - **INDEPENDENT**: Each category is modeled independently
  - **MODELING of CLASS ANALOGIES**: Attention is focused on the similarity between object from the same class rather then on differentiating among classes.
- To build the individual category models, PCA is used.
  - The number of significant components A (defining the "inner space") can be different from class to class.
  - The remaining M-A components represent the residuals ("outer space")

SIMCA – Defining the model space

$$d_k^C = \sqrt{\left(T^2_{red,k}\right)_C^2 + \left(Q_{red,k}\right)_C^2} = \sqrt{\left(\frac{T_k^2}{T_{0.95}^2}\right)_C^2 + \left(\frac{Q_k}{Q_{0.95}}\right)_C^2}$$



$$T^2_{0.95} = F_{0.95,A,N-A}\frac{A(N^2-1)}{N(N-A)}$$

$$Q_{0.95} = \theta_1\left[1-\frac{\theta_2 h_0\left(1-h_0\right)}{\theta_1^2}+\frac{z_{0.95}\left(2\theta_2 h_0^2\right)^{\frac{1}{2}}}{\theta_1}\right]^{\frac{1}{h_0}}$$

$$\theta_k = \sum_i \lambda_i^k \qquad h_0 = 1-2\theta_1\theta_3/3\theta_2^2$$

KDE & SIMCA-like approaches

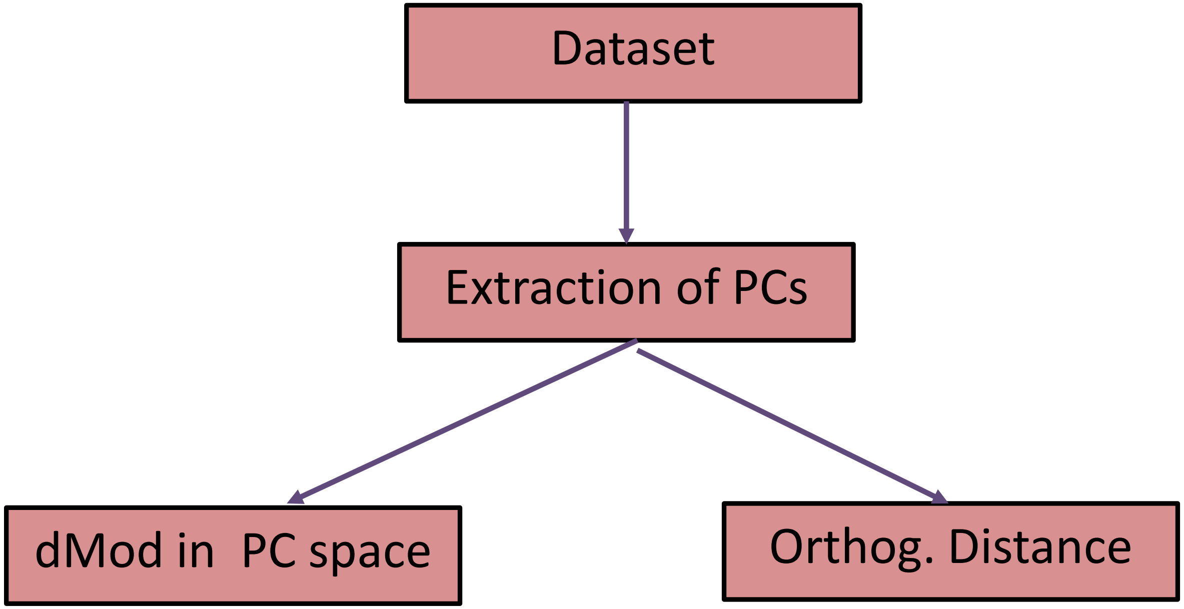
- CLASSY[1]: Kernel density estimation of the pdf in the scores space
  - Meant to achieve «probabilistic» classification
  - Discriminant approach: calculation of the posterior probabilities for each class through Bayes' theorem

- PLS-DM[2]: Class-modeling achieved by combining KDE-based scores distance and orthogonal distance
  - PLS-based bilinear decomposition
  - Model space estimation analogous to «SIM»-SIMCA ($\frac{SD}{SD_{crit}} \leq 1$ & $\frac{OD}{OD_{crit}} \leq 1$)

[1] H. Van der Voet and D.A. Doornbos, *Anal. Chim. Acta* **161** (1984) 115.

[2] P. Oliveri et al., *Anal. Chim. Acta* **851** (2014) 30.

SIMCA – A unified approach for single and multiple blocks

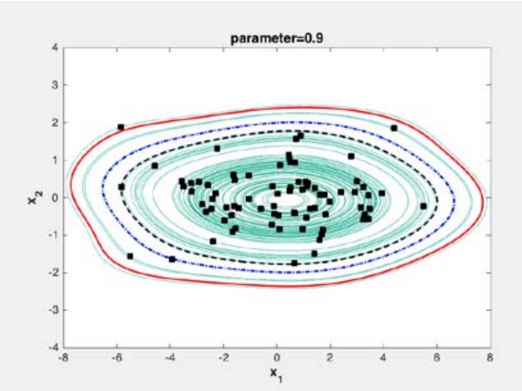


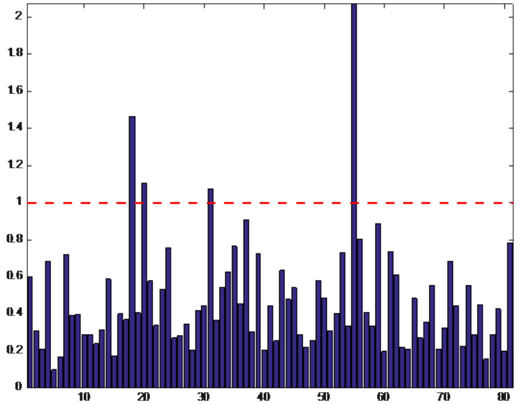
Dataset

Extraction of PCs

dMod in PC space

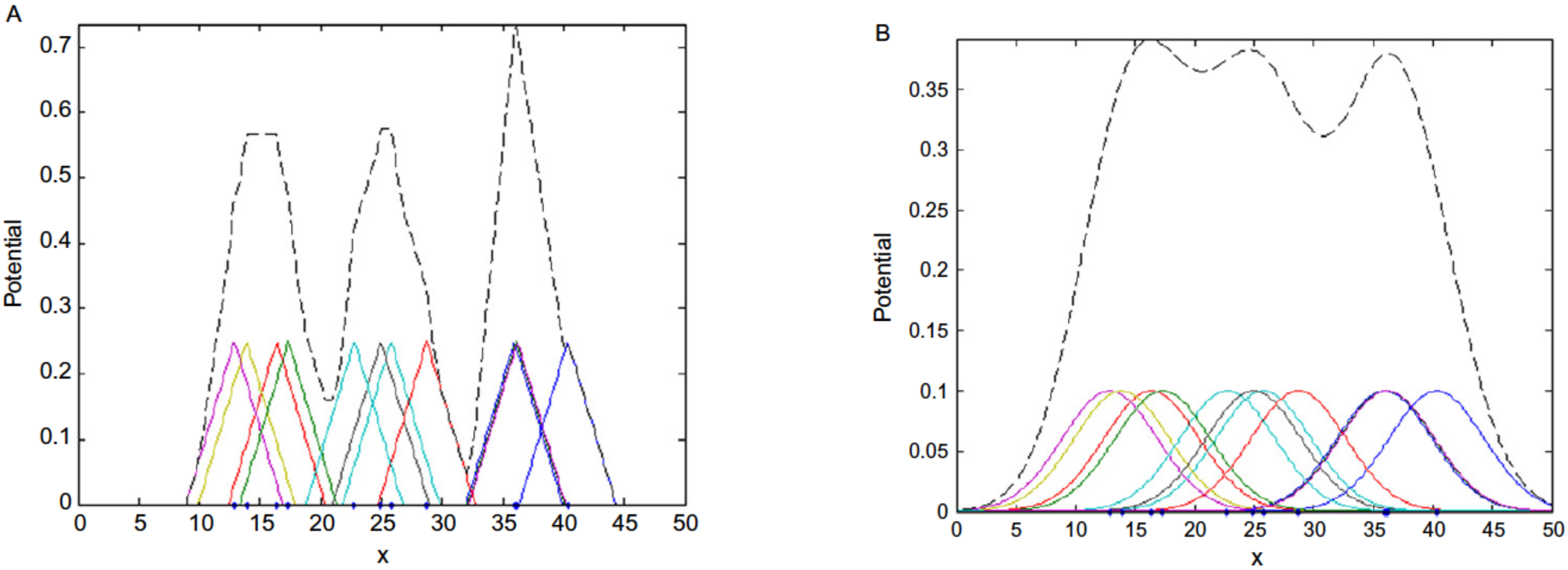
Orthog. Distance





Potential

$$p(\mathbf{x}|g) = P_g(\mathbf{x}) = \frac{\sum_{i=1}^{n_g} p_{g,i}(\mathbf{x})}{n_g}$$



$$p_{g,i}(\mathbf{x}) = \begin{cases} 0 & \text{if } \|\mathbf{x}-\mathbf{x}_{g,i}\| > d_{\max} \\ \dfrac{d_{\max} - \|\mathbf{x}-\mathbf{x}_{g,i}\|}{d_{\max}^2} & \text{if } \|\mathbf{x}-\mathbf{x}_{g,i}\| \le d_{\max} \end{cases}$$

$$p_{g,i}(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{m}{2}}|\mathbf{S}_g|} e^{-\frac{1}{2}(\mathbf{x}-\mathbf{x}_{g,i})^{\mathrm{T}} \mathbf{S}_g^{-1}(\mathbf{x}-\mathbf{x}_{g,i})}$$

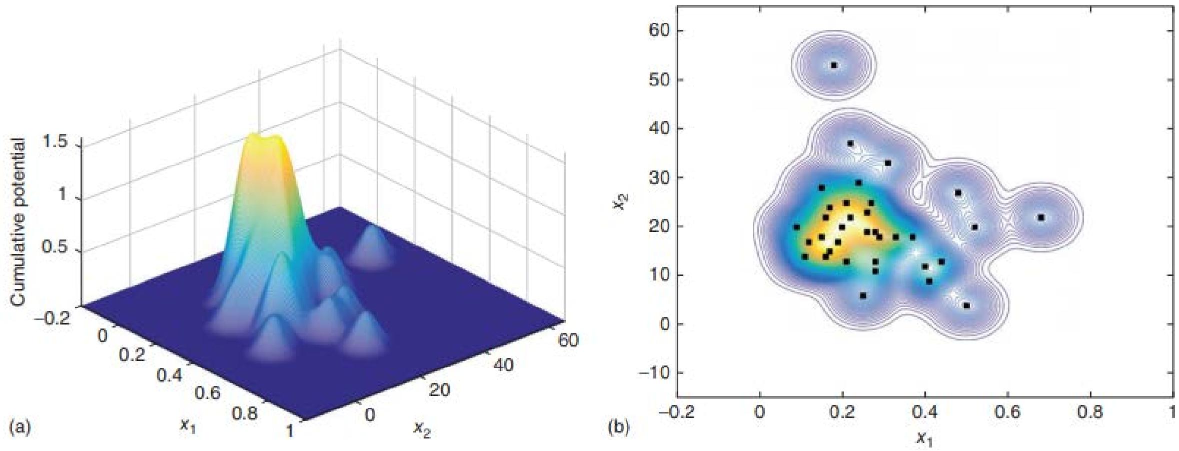
Potential functions

- Estimate the global pdf for the class as the sum of individual multivariate pdfs centered on each training sample.

$$f(\boldsymbol{x}) = \sum_{i=1}^{N_g} f_i(\boldsymbol{x}) = \frac{1}{N_g}\sum_{i=1}^{N_g} \frac{1}{(2\pi)^{\frac{v}{2}}\left|\boldsymbol{S}_g\right|^{\frac{1}{2}}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{x}_i)^T \boldsymbol{S}_g^{-1}(\boldsymbol{x}-\boldsymbol{x}_i)}$$

$$\boldsymbol{S}_g = \gamma \begin{bmatrix} s_1^2 & 0 & 0 & 0 \\ 0 & s_2^2 & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & s_v^2 \end{bmatrix}$$

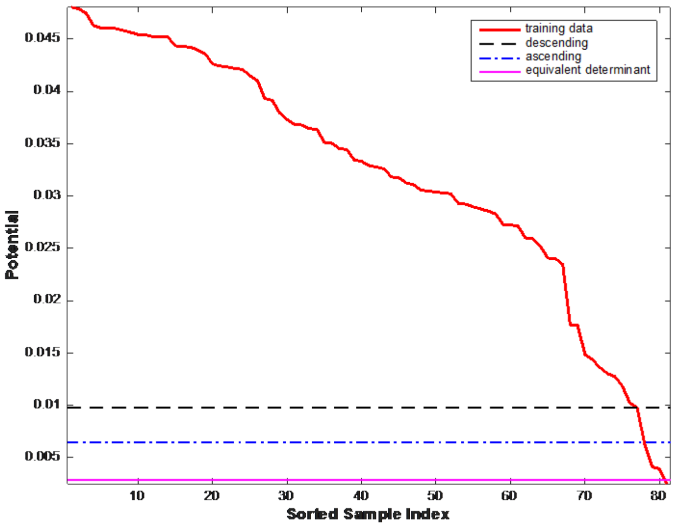
γ=smoothing parameter



- The class boundary is defined by setting the critical value (f_{crit}) of the pdf $f(\boldsymbol{x})$, at a selected confidence level (f_{crit})

Potential function – Class modeling

Percentile



$$P_{\gamma,g}(\boldsymbol{x}) = P_g(\boldsymbol{x}_k) + (q-k)[P_g(\boldsymbol{x}_{k+1}) - P_g(\boldsymbol{x}_k)]$$

with

$$q = \frac{\gamma N_g}{100} \quad k = \mathrm{int}(q)$$

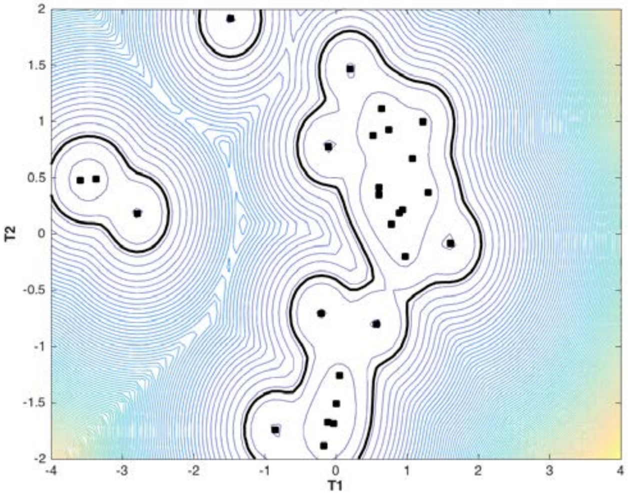
or

$$P_{100-\gamma,g}(\boldsymbol{x}) = P_g(\boldsymbol{x}_j) + (u-j)[P_g(\boldsymbol{x}_{j+1}) - P_g(\boldsymbol{x}_j)]$$

with

$$u = \frac{(100-\gamma)N_g}{100} \quad j = \mathrm{int}(q)$$

Equivalent determinant



The determinant of the covariance matrix of a Gaussian distribution having the same value of mean probability density function as the one of the current kernel density model

$$f_{crit} = \frac{1}{(2\pi)^{\frac{v}{2}}\left|\widehat{\boldsymbol{C}}\right|^{\frac{1}{2}}} e^{-\frac{\chi_\alpha^2}{2}}$$

$$\left|\widehat{\boldsymbol{C}}\right|^{\frac{1}{2}} = \frac{N_g}{2^v \pi^{\frac{v}{2}} \sum_{i=1}^{N_g} f_i(\boldsymbol{x})}$$

M. Forina et al., *J. Chemometr.* **5** (1991) 435-453.

Data sets analyzed

PGI Sicilian oranges

Peel: NIR Spectroscopy
Juice: NIR, MIR,UV and Vis Spectroscopy

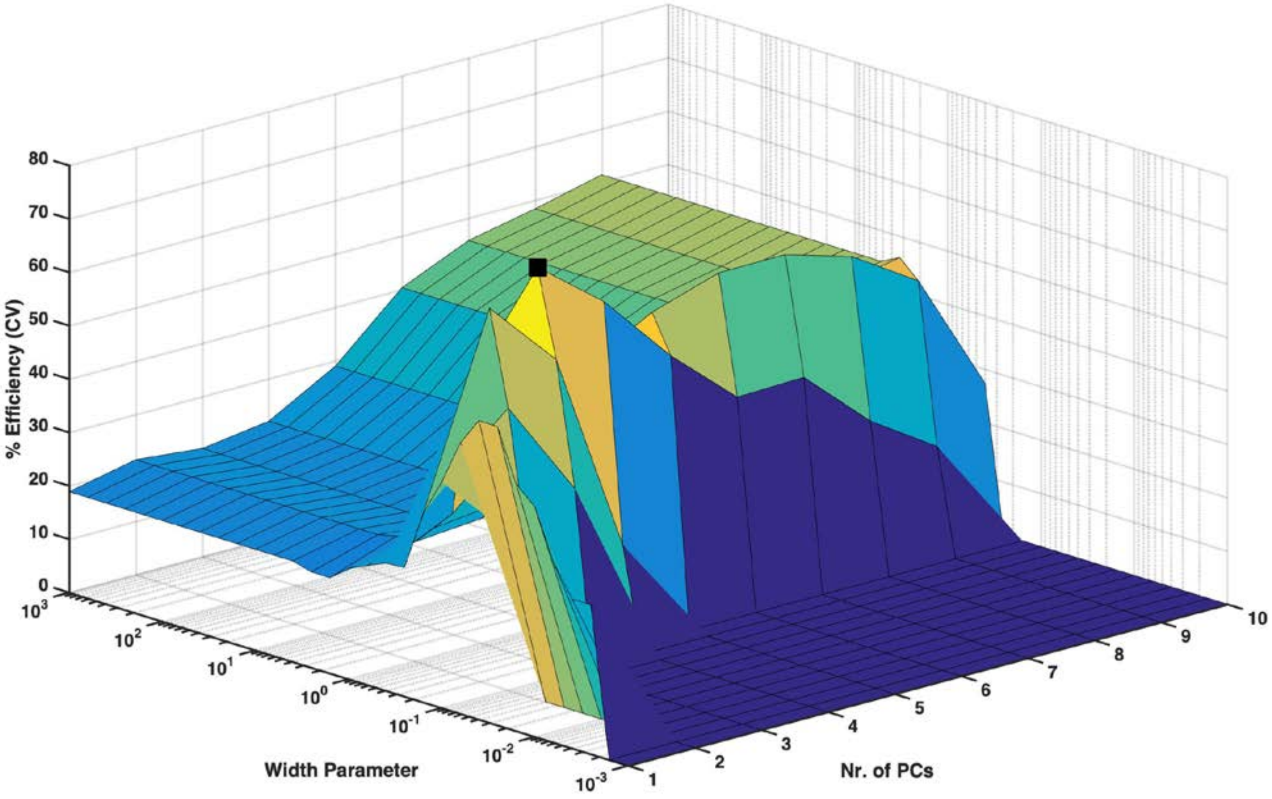
Senise Bell Pepper

NIR e MIR Spectroscopy

Borgo Reale beer

NIR, MIR, UV and Vis Spectroscopy

Optmizing model parameters



Best SIMCA and SIMCA pf results of the pure ground Senise bell pepper class

Best SIMCA Model: MIR Spectroscopy

	PC	SensCal	SpecCal	EffCal	SensCV	SpecCV	EffCV
1st Derivative	10	100.00	64.00	80.00	70.00	69.00	69.50

Best SIMCA Prediction: MIR Spectroscopy

	SensPred	SpecPred	EffPred
1st Derivative	80.00	86.67	83.27

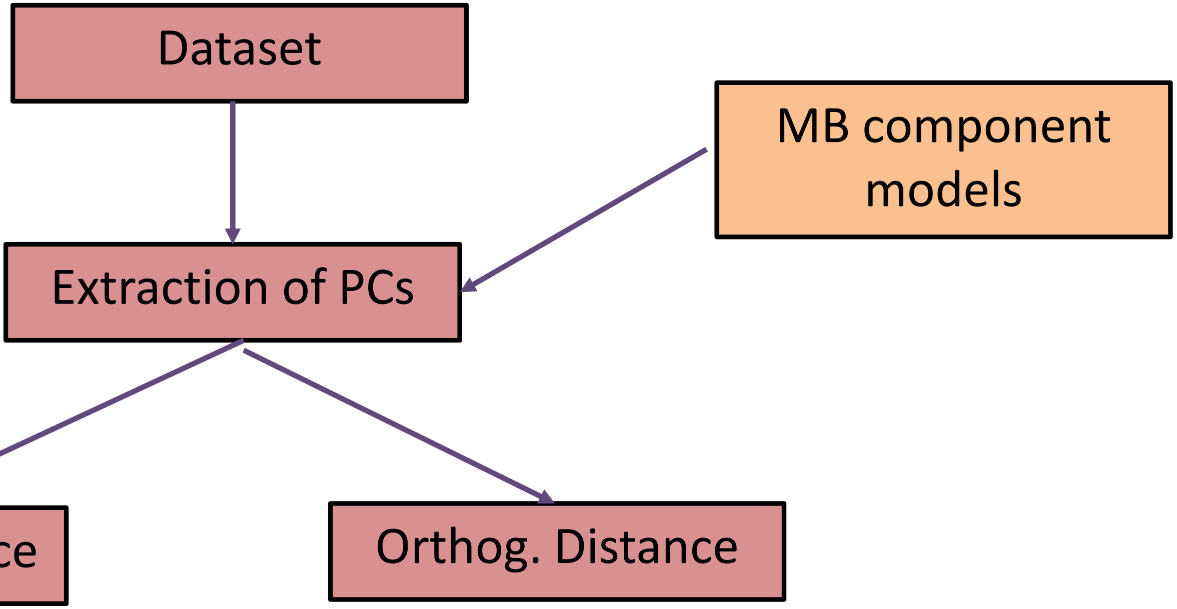
Best SIMCApf Model: MIR Spectroscopy

	PC	Optwidth	SensCal	SpecCal	EffCal	SensCV	SpecCV	EffCV
1st DerivativeED	8	0.6310	100.00	64.00	80.00	75.00	67.00	70.89

Best SIMCApf Prediction: MIR Spectroscopy

	SensPred	SpecPred	EffPred
1st DerivativeED	80.00	93.33	86.41

Moving to multiple blocks



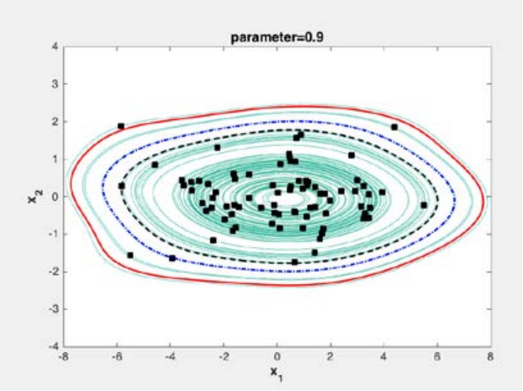
Dataset

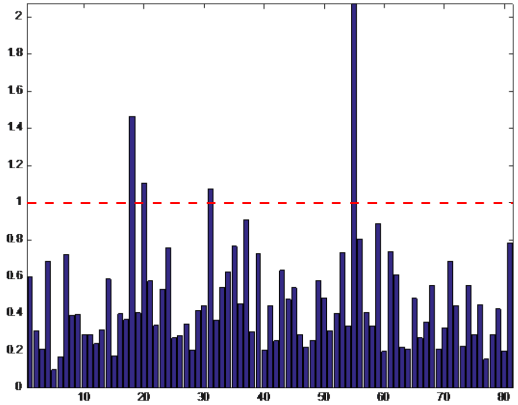
MB component models

Extraction of PCs

dMod in PC space

Orthog. Distance





Multi-block analysis results on pure ground Senise bell pepper samples with Low-Level approach

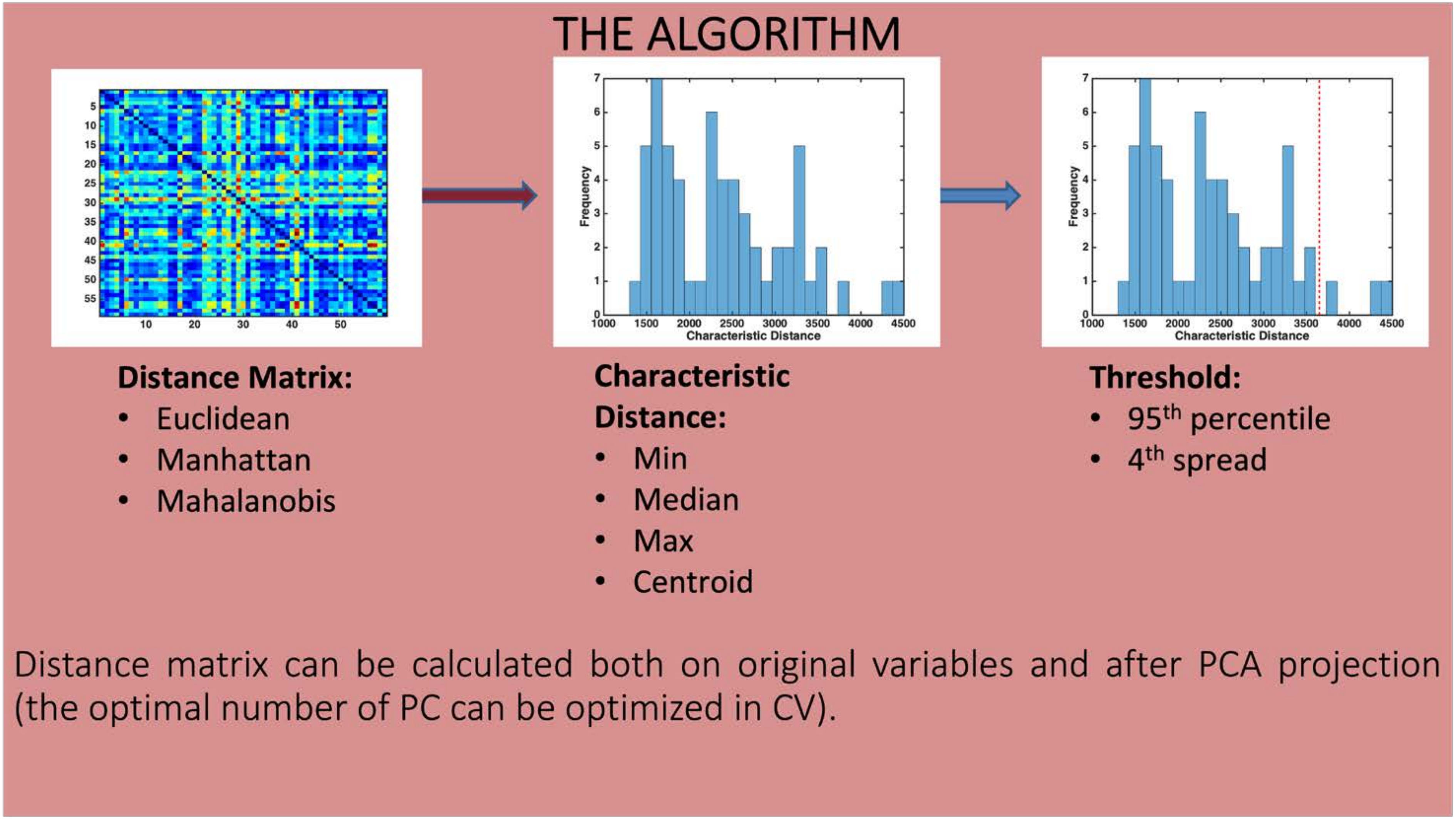
Multi-block analysis results

	PC	Optwidth	SensCal	SpecCal	EffCal	SensCV	SpecCV	EffCV
ED	1	0.1585	100.00	100.00	100.00	100.00	100.00	100.00
A	1	2.5119	100.00	100.00	100.00	100.00	100.00	100.00
D	1	2.5119	100.00	100.00	100.00	100.00	100.00	100.00

Predictions on the test set

	SensPred	SpecPred	EffPred
ED	95.00	100.00	97.47
A	95.00	100.00	97.47
D	95.00	100.00	97.47

NPCM


Frequency
Characteristic Distance
1000
1500
2000
2500
3000
3500
4000
4500

THE ALGORITHM

Distance Matrix:

- Euclidean
- Manhattan
- Mahalanobis

Characteristic Distance:

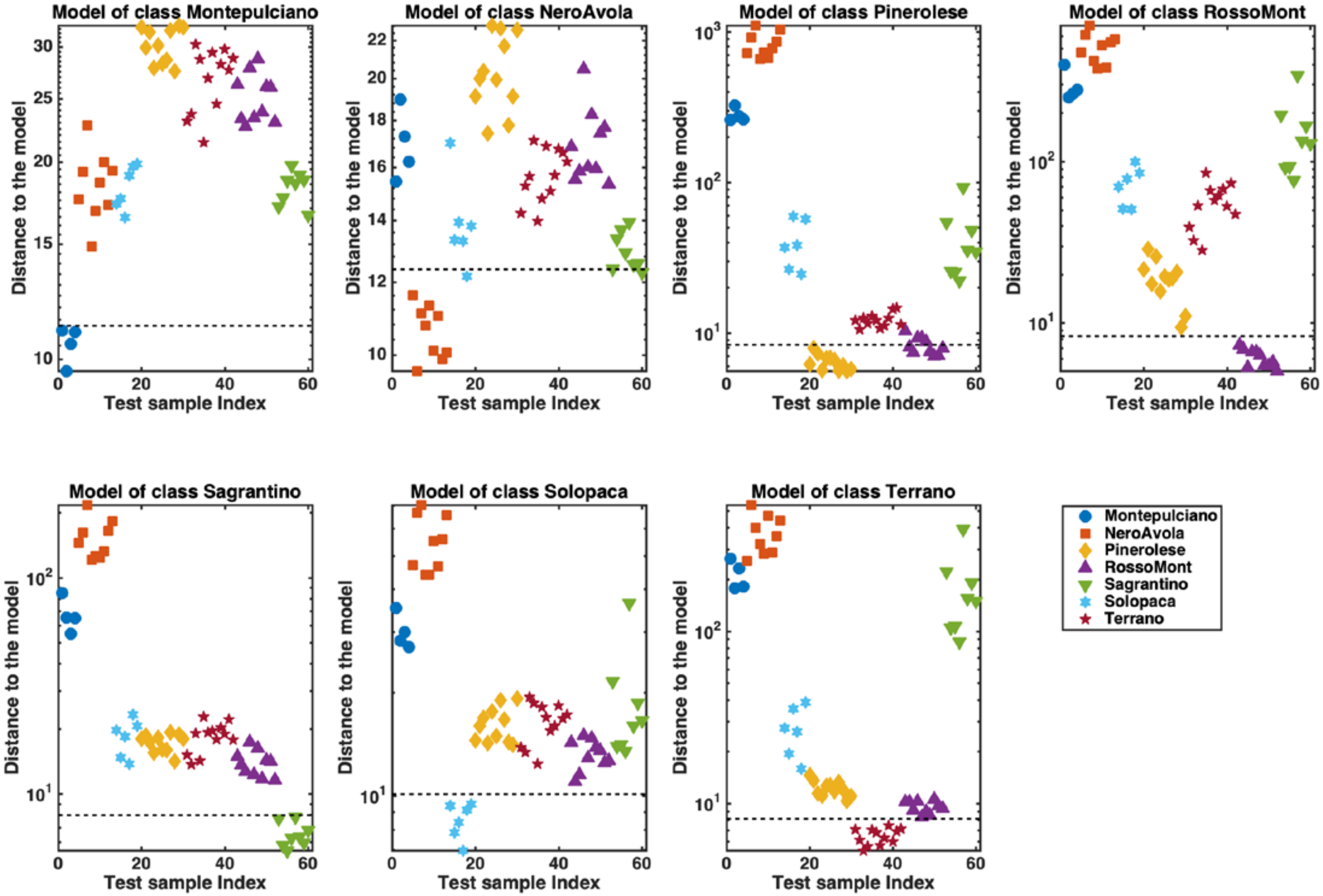
- Min
- Median
- Max
- Centroid

Threshold:

- 95th percentile
- 4th spread

Distance matrix can be calculated both on original variables and after PCA projection (the optimal number of PC can be optimized in CV).

NPCM – Italian Wines



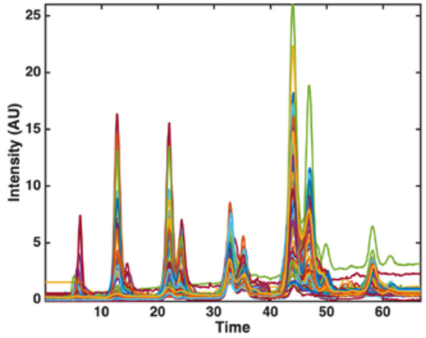
NPCM – Results

Italian wines

	Montepulciano			Nero D'Avola			Pinerolese			Rosso di Montalcino			Sagrantino			Solopaca			Terrano		
	Sens.	Spec.	Eff.	Sens.	Spec.	Eff.	Sens.	Spec.	Eff.	Sens.	Spec.	Eff.	Sens.	Spec.	Eff.	Sens.	Spec.	Eff.	Sens.	Spec.	Eff.
NPCM	100.0	100.0	100.0	100.0	96.1	95.4	100.0	89.8	94.8	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
SIMCA	100.0	100.0	100.0	100.0	98.4	99.0	100.0	85.1	92.3	100.0	100.0	100.0	75.0	100.0	86.6	71.4	98.1	83.7	80.0	100.0	89.4

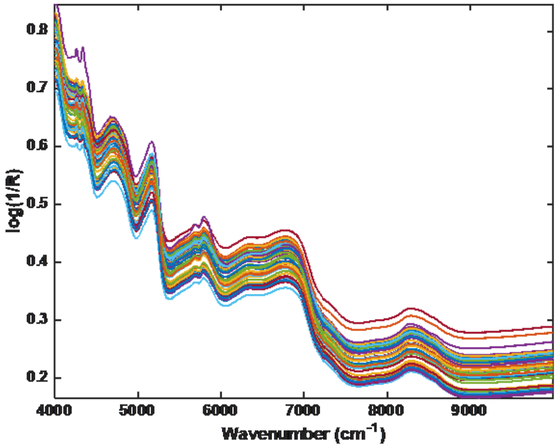
Vegetable oils

	Olive			Other vegetable		
	Sens.	Spec.	Eff.	Sens.	Spec.	Eff.
NPCM	93.3	100.0	96.6	90.0	93.3	91.6
SIMCA	93.3	100.0	96.6	80.0	100.0	89.4



Pre-processing

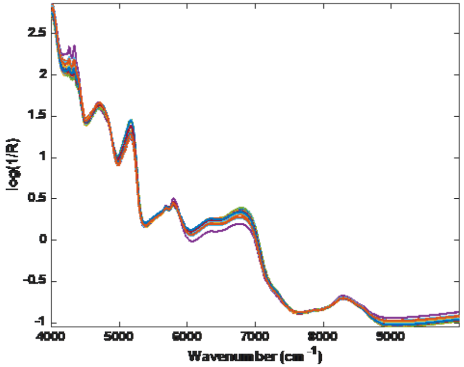
- Data copy may be deformed by artefacts due to factors (physical, chemical and environmental) not of interest for the characterization of the samples under study

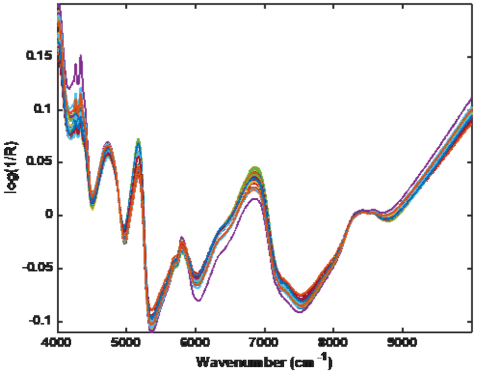


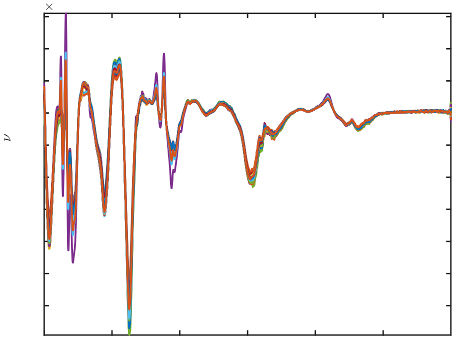
SNV

Detrending

1st derivative





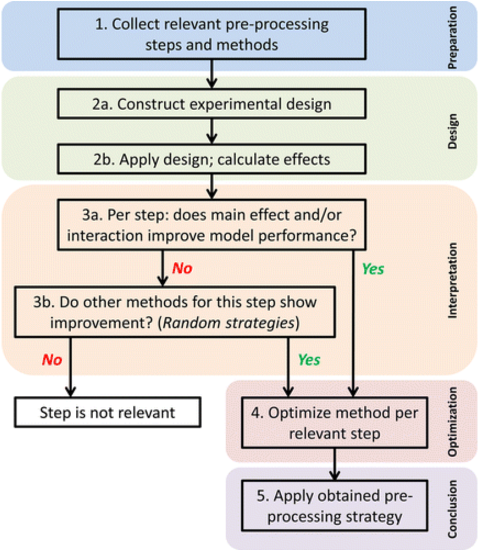


Choice of best preprocessing

- Trial and error: explore multiple preprocessing options and select the one leading to the lowest error/best performance (usually in CV)

Pre-processing	LV	RMSECV
SNV	7	0.98
1st derivative	8	0.86
SNV+1st derivative	6	0.83

- Experimental design (Gerretzen et al., *Anal. Chem.* **87** (2015) 12096-12103)



Preparation

1. Collect relevant pre-processing steps and methods

Design

2a. Construct experimental design

2b. Apply design; calculate effects

Interpretation

3a. Per step: does main effect and/or interaction improve model performance?

No

Yes

3b. Do other methods for this step show improvement? (*Random strategies*)

No

Yes

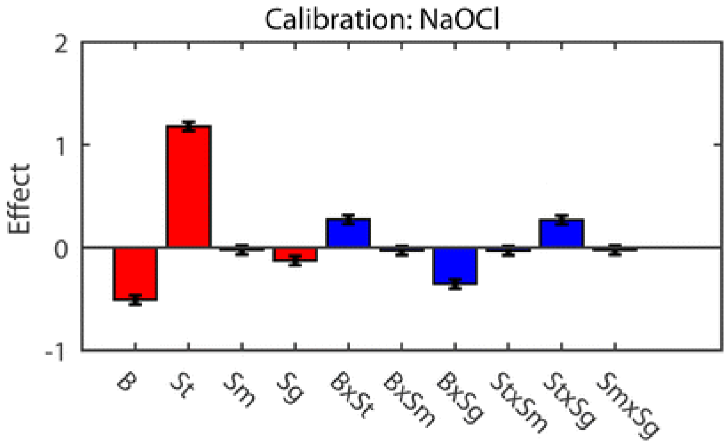
Step is not relevant

Optimization

4. Optimize method per relevant step

Conclusion

5. Apply obtained pre-processing strategy



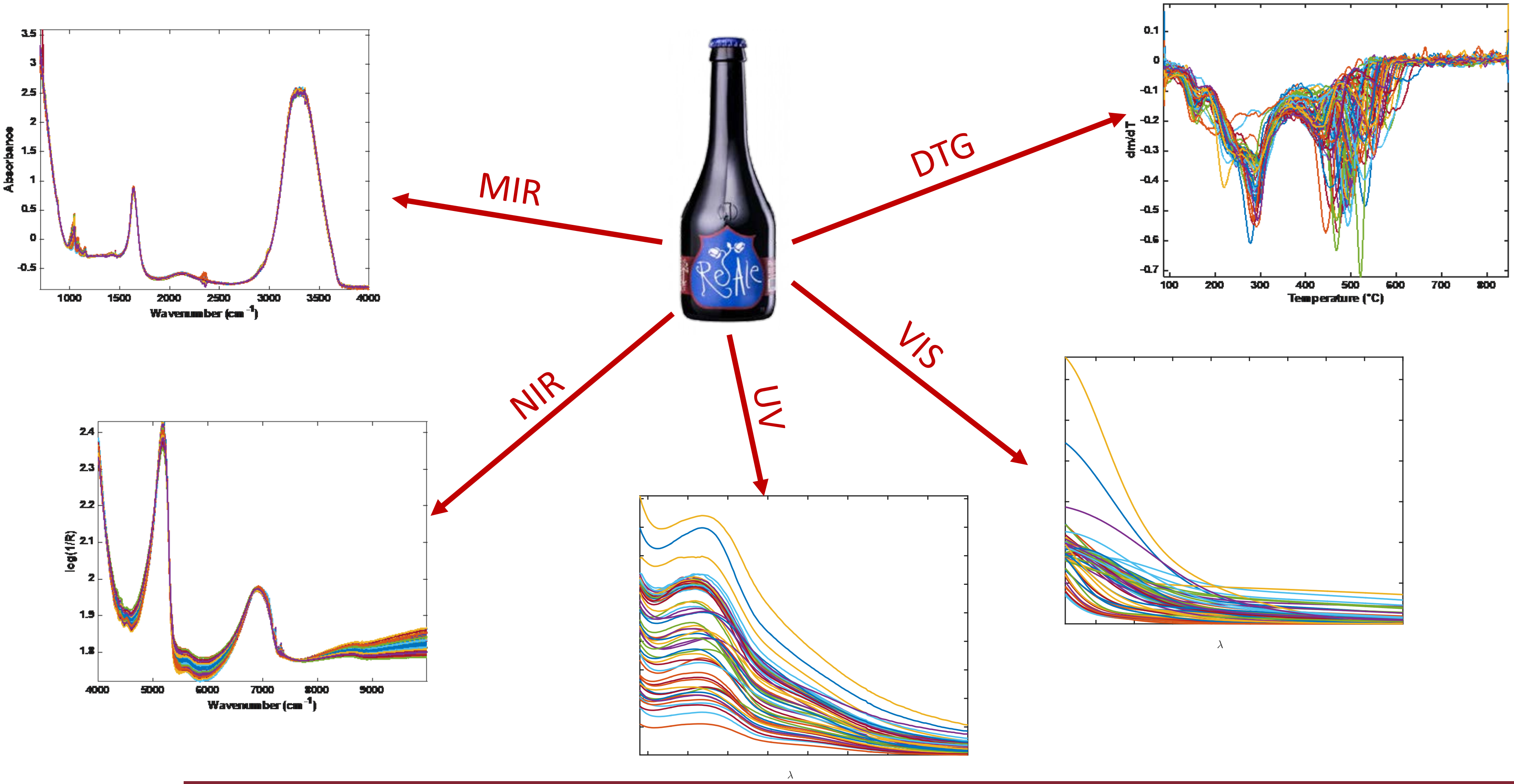
Choice of best preprocessing: Boosting approaches

- Ensemble learning: stack different PLS models using different pretreatments on the same data
  - The output of this approach is computed by averaging the predicted values computed by its constituent learners.
  - Examples are, e.g.:
    - L. Xu, et al. Ensemble preprocessing of near-infrared (NIR) spectra for multivariate calibration, Anal. Chim. Acta **616** (2008) 138-143: Twenty different pre-processing operations, based on first and second derivatives, smoothing, SNV, MSC and their combinations
    - R. Reda et al. A comparative study between a new method and other machine learning algorithms for soil organic carbon and total nitrogen prediction using near infrared spectroscopy, Chemometr. Intell. Lab. Syst. **195** (2019) 103873: Six PLS models calculated on data preprocessed by diverse preprocessing approaches, raw, log(1/R), 1st and 2nd derivative, MSC, SNV

- Multi-block approaches
  - Data are preprocessed by different techniques
  - The differently preprocessed matrices are used as input to a multi-block (data fusion) algorithm

Multi-block data

- Different sets of (usually multivariate) data collected on the same samples
- E.g.: Same set of samples characterized by different analytical platforms



MIR

DTG

NIR

UV

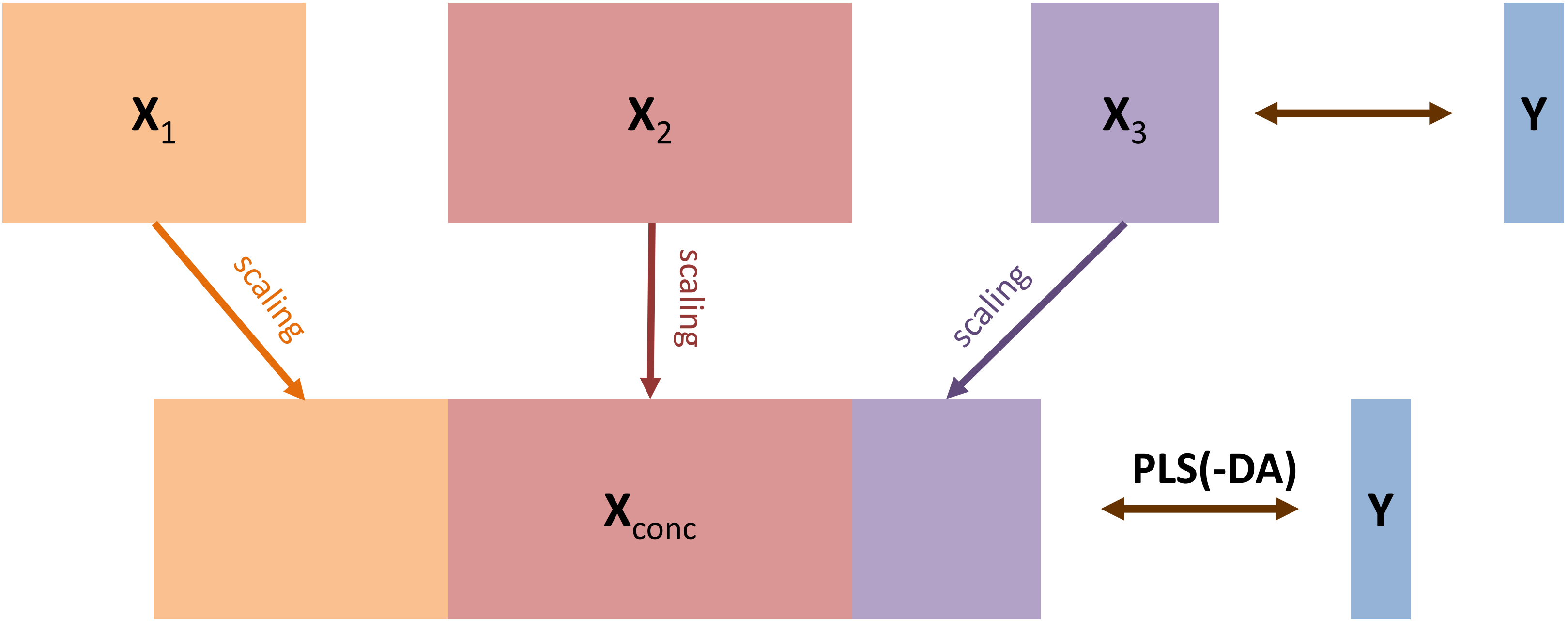
VIS

Multi-block PLS(-DA)

- Straightforward generalization of standard PLS(-DA)
- Low-level approach:
  - Assumes that global (super-scores) are weighted combination of block scores:

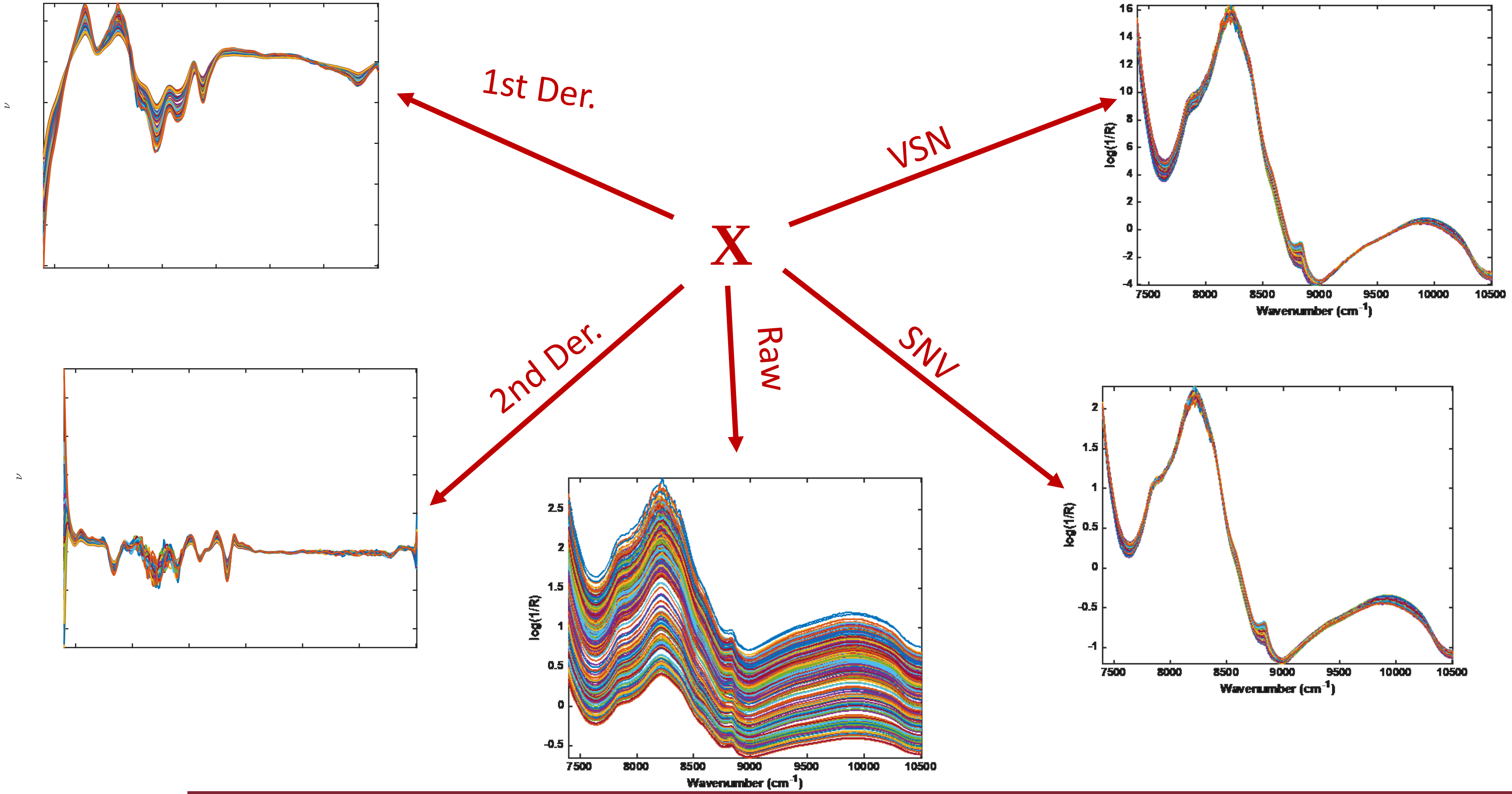
$$\boldsymbol{t}_i = \boldsymbol{X}_i\boldsymbol{w}_i \qquad \boldsymbol{t}_{super} = [\boldsymbol{t}_1 \quad \boldsymbol{t}_2 \quad \cdots \quad \boldsymbol{t}_B]\boldsymbol{w}_{super}$$

  - PLS on the concatenated data matrices after suitable scaling.
  - Block scores, weights and loadings and super-weights can be obtained a posteriori



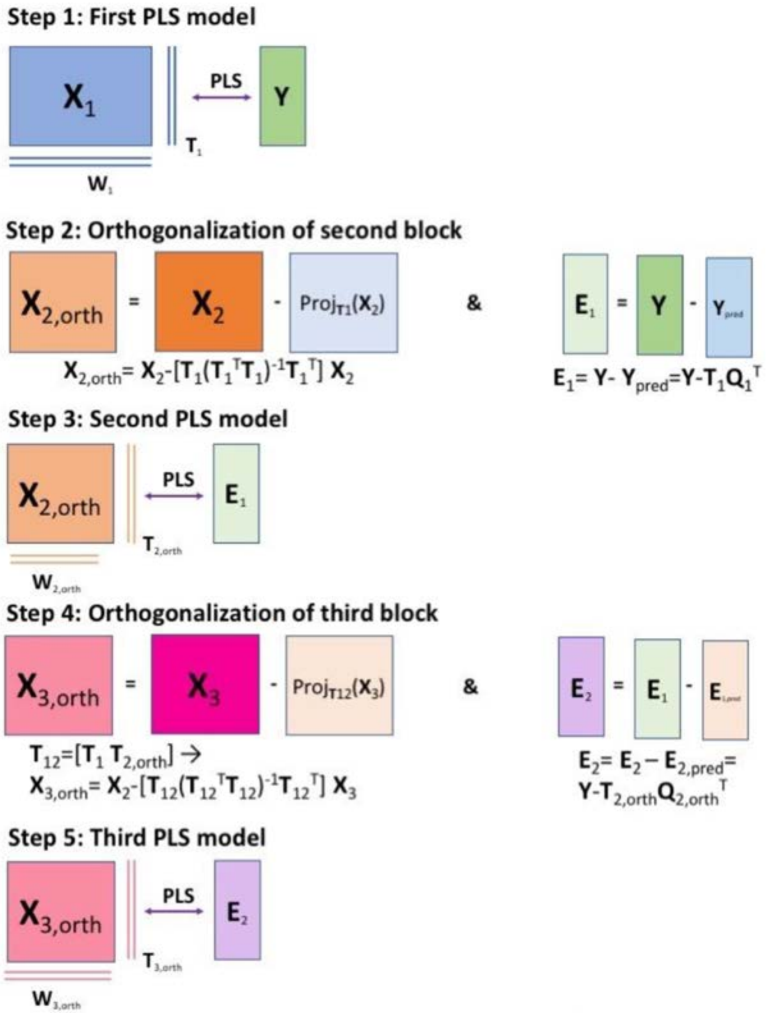
Multi-block data

- The same data matrix pre-processed with different approaches



The SPORT approach

- SO-PLS is used as the modeling method



Step 1: First PLS model

Step 2: Orthogonalization of second block

$X_{2,orth}= X_2-[T_1(T_1^TT_1)^{-1}T_1^T] X_2$

$E_1= Y- Y_{pred}=Y-T_1Q_1^T$

Step 3: Second PLS model

Step 4: Orthogonalization of third block

$X_{3,orth}$ = X_3 - $Proj_{T12}(X_3)$
&
E_2 = E_1 - $E_{1,pred}$

$T_{12}=[T_1\ T_{2,orth}] \rightarrow$
$X_{3,orth}= X_2-[T_{12}(T_{12}^TT_{12})^{-1}T_{12}^T] X_3$

$E_2= E_2 - E_{2,pred}=$
$Y-T_{2,orth}Q_{2,orth}^T$

Step 5: Third PLS model

Global model: $Y_{pred}=T_1Q_1^T+ T_{2,orth}Q_{2,orth}^T + T_{3,orth}Q_{3,orth}^T$

Chemometrics and Intelligent Laboratory Systems 199 (2020) 103975

Contents lists available at ScienceDirect

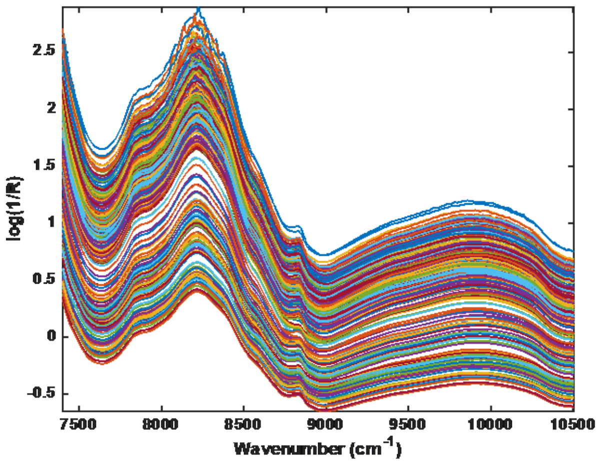
Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemometrics

Sequential preprocessing through ORThogonalization (SPORT) and its application to near infrared spectroscopy

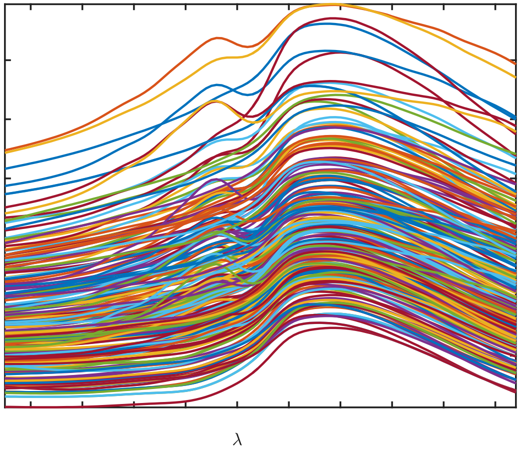
Jean-Michel Roger[a,b,*], Alessandra Biancolillo[c], Federico Marini[d]

Data sets



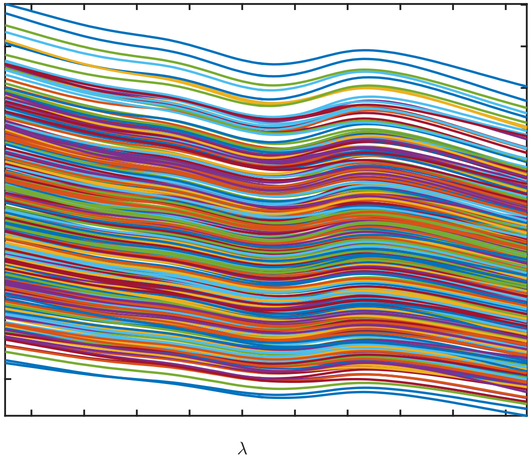
Tablets

M. Dyrby et al., *Appl. Spectrosc.* **56** (2002) 579-585.



Meat

C. Borggaard and H.H.Thodberg, *Anal. Chem.* **64** (1992) 545-551.

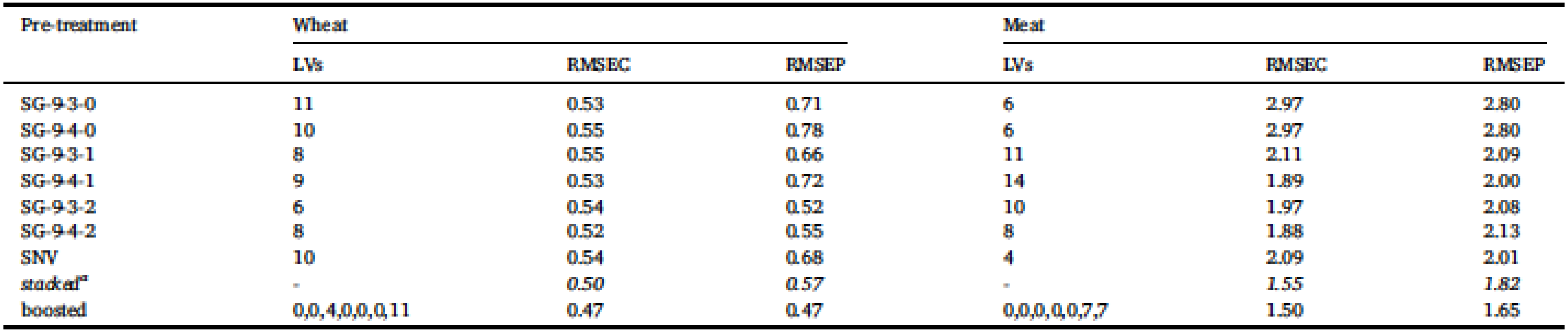


Wheat

D.K. Pedersen et al., *Appl. Spectrosc.* **56** (2002) 1206-1214.

Wheat & Meat

- Results are compared to those of the stacking approach in L. Xu, et al., *Anal. Chim. Acta* **616** (2008) 138-143:

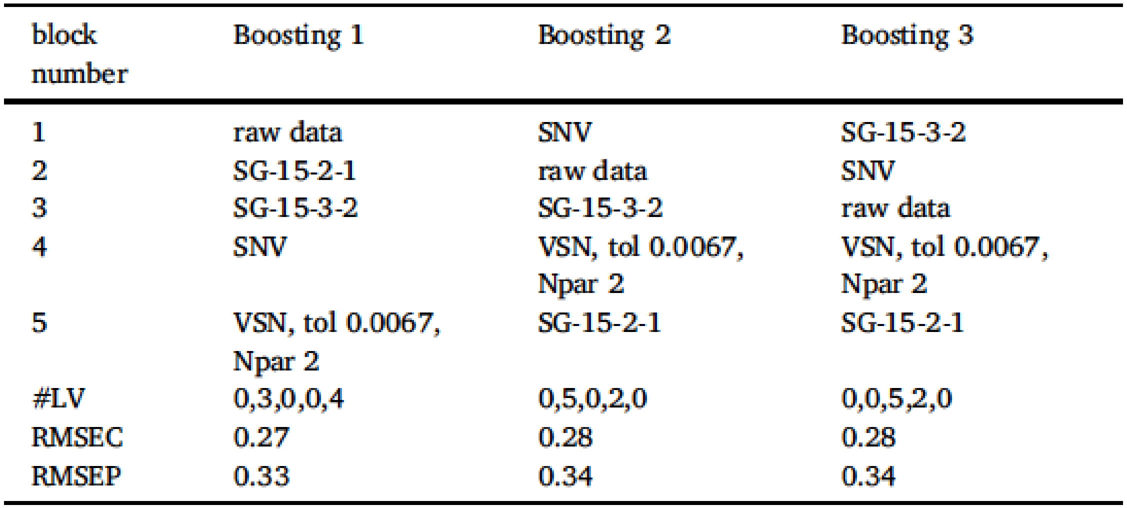


Pre-treatment	Wheat			Meat		
	LVs	RMSEC	RMSEP	LVs	RMSEC	RMSEP
SG-9-3-0	11	0.53	0.71	6	2.97	2.80
SG-9-4-0	10	0.55	0.78	6	2.97	2.80
SG-9-3-1	8	0.55	0.66	11	2.11	2.09
SG-9-4-1	9	0.53	0.72	14	1.89	2.00
SG-9-3-2	6	0.54	0.52	10	1.97	2.08
SG-9-4-2	8	0.52	0.55	8	1.88	2.13
SNV	10	0.54	0.68	4	2.09	2.01
stacked[a]	-	*0.50*	*0.57*	-	*1.55*	*1.82*
boosted	0,0,4,0,0,0,11	0.47	0.47	0,0,0,0,0,7,7	1.50	1.65

- SPORT approach performs better than any single pretreatment model and of the stacked approach
- Very parsimonious selection → only two blocks are included in each model

Tablets

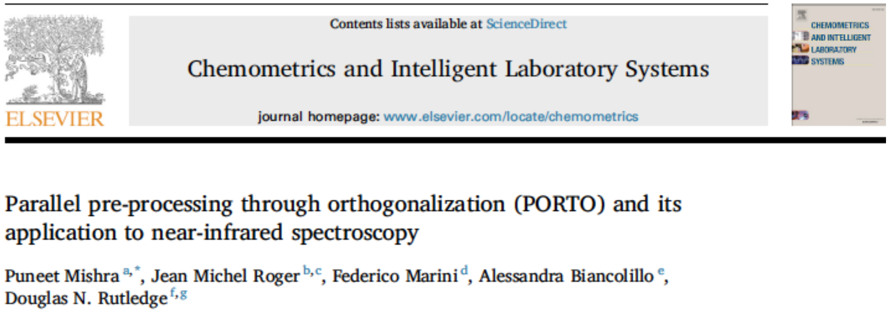
- By exchanging the order of the blocks, it is possible to explore common and distinctive information



block number	Boosting 1	Boosting 2	Boosting 3
1	raw data	SNV	SG-15-3-2
2	SG-15-2-1	raw data	SNV
3	SG-15-3-2	SG-15-3-2	raw data
4	SNV	VSN, tol 0.0067, Npar 2	VSN, tol 0.0067, Npar 2
5	VSN, tol 0.0067, Npar 2	SG-15-2-1	SG-15-2-1
#LV	0,3,0,0,4	0,5,0,2,0	0,0,5,2,0
RMSEC	0.27	0.28	0.28
RMSEP	0.33	0.34	0.34

- Exchanging the order of the blocks has little effect on the predictivity, but impacts the selected pre-processings

Recent developments



Contents lists available at ScienceDirect

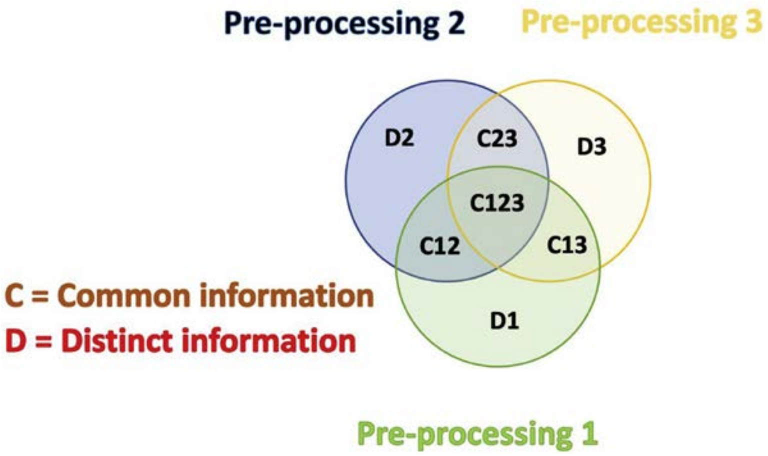
Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemometrics

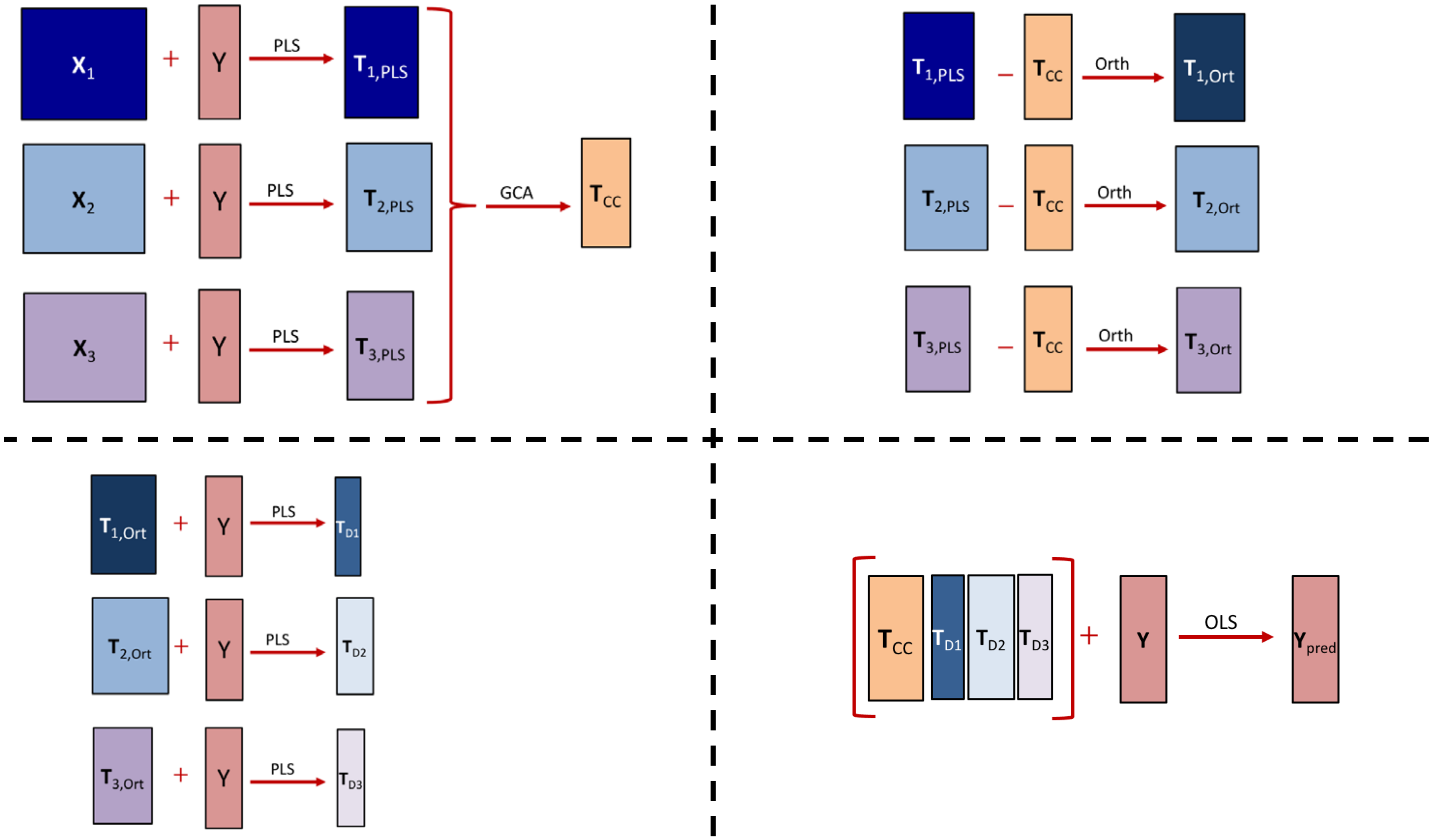
Parallel pre-processing through orthogonalization (PORTO) and its application to near-infrared spectroscopy

Puneet Mishra[a,*], Jean Michel Roger[b,c], Federico Marini[d], Alessandra Biancolillo[e], Douglas N. Rutledge[f,g]

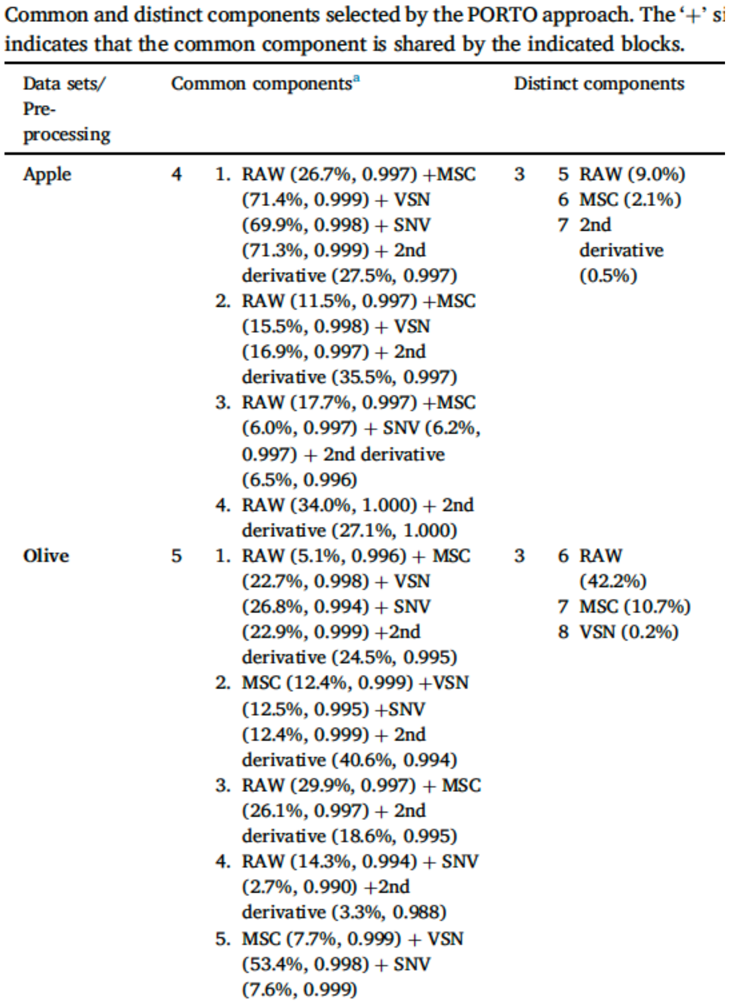
- SO-PLS is not the only multi-block method which can be used to fused different pre-treatments of the same data matrix
- The same concept has been exploited in PORTO, where the MB *engine* is represented by PO-PLS
- Straightforward exploitation of the concept of common and distinct components and lower impact of the order of the blocks



PO-PLS scheme

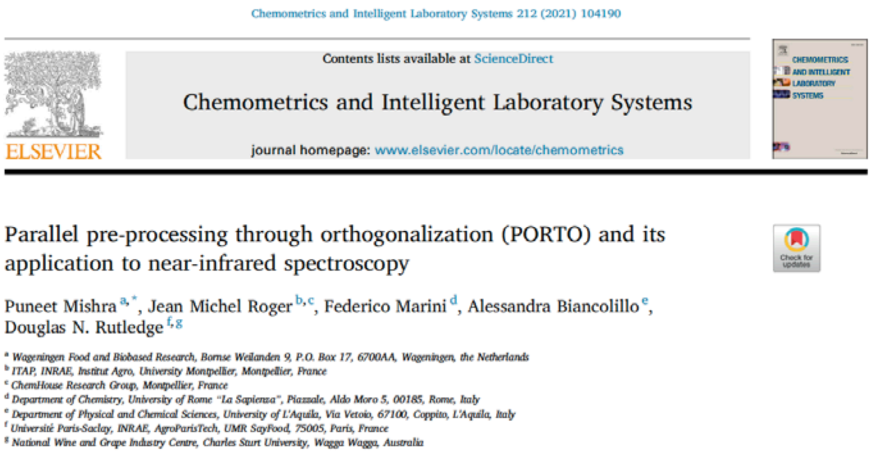

X_1 + Y $\xrightarrow{PLS}$ $T_{1,PLS}$
X_2 + Y $\xrightarrow{PLS}$ $T_{2,PLS}$
X_3 + Y $\xrightarrow{PLS}$ $T_{3,PLS}$
GCA
T_{CC}
$T_{1,PLS}$ − T_{CC} $\xrightarrow{Orth}$ $T_{1,Ort}$
$T_{2,PLS}$ − T_{CC} $\xrightarrow{Orth}$ $T_{2,Ort}$
$T_{3,PLS}$ − T_{CC} $\xrightarrow{Orth}$ $T_{3,Ort}$
$T_{1,Ort}$ + Y $\xrightarrow{PLS}$ T_{D1}
$T_{2,Ort}$ + Y $\xrightarrow{PLS}$ T_{D2}
$T_{3,Ort}$ + Y $\xrightarrow{PLS}$ T_{D3}
[T_{CC} T_{D1} T_{D2} T_{D3}] + Y $\xrightarrow{OLS}$ Y_{pred}

PORTO



Common and distinct components selected by the PORTO approach. The '+' si
indicates that the common component is shared by the indicated blocks.

Data sets/ Pre-processing	Common components[a]		Distinct components	
Apple	4	1. RAW (26.7%, 0.997) +MSC (71.4%, 0.999) + VSN (69.9%, 0.998) + SNV (71.3%, 0.999) + 2nd derivative (27.5%, 0.997) 2. RAW (11.5%, 0.997) +MSC (15.5%, 0.998) + VSN (16.9%, 0.997) + 2nd derivative (35.5%, 0.997) 3. RAW (17.7%, 0.997) +MSC (6.0%, 0.997) + SNV (6.2%, 0.997) + 2nd derivative (6.5%, 0.996) 4. RAW (34.0%, 1.000) + 2nd derivative (27.1%, 1.000)	3	5 RAW (9.0%) 6 MSC (2.1%) 7 2nd derivative (0.5%)
Olive	5	1. RAW (5.1%, 0.996) + MSC (22.7%, 0.998) + VSN (26.8%, 0.994) + SNV (22.9%, 0.999) +2nd derivative (24.5%, 0.995) 2. MSC (12.4%, 0.999) +VSN (12.5%, 0.995) +SNV (12.4%, 0.999) + 2nd derivative (40.6%, 0.994) 3. RAW (29.9%, 0.997) + MSC (26.1%, 0.997) + 2nd derivative (18.6%, 0.995) 4. RAW (14.3%, 0.994) + SNV (2.7%, 0.990) +2nd derivative (3.3%, 0.988) 5. MSC (7.7%, 0.999) + VSN (53.4%, 0.998) + SNV (7.6%, 0.999)	3	6 RAW (42.2%) 7 MSC (10.7%) 8 VSN (0.2%)



Chemometrics and Intelligent Laboratory Systems 212 (2021) 104190

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

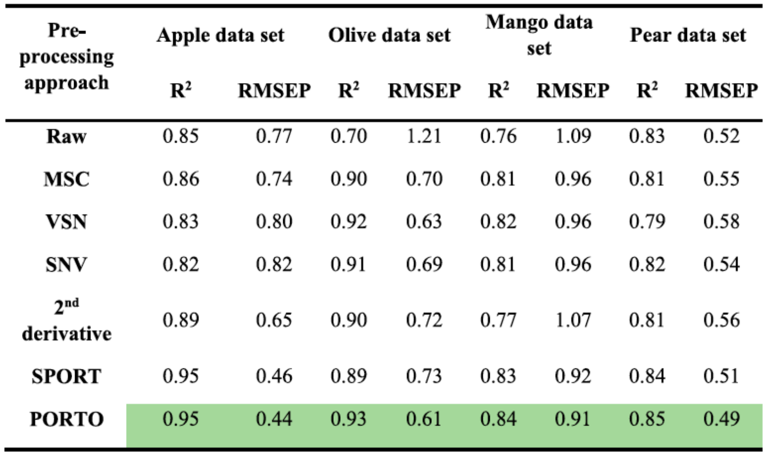
journal homepage: www.elsevier.com/locate/chemometrics

ELSEVIER
Check for updates

Parallel pre-processing through orthogonalization (PORTO) and its application to near-infrared spectroscopy

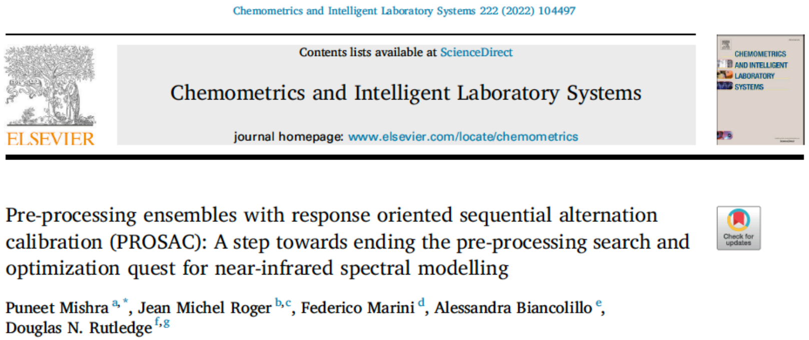
Puneet Mishra[a,*], Jean Michel Roger[b,c], Federico Marini[d], Alessandra Biancolillo[e], Douglas N. Rutledge[f,g]

[a] Wageningen Food and Biobased Research, Bornse Weilanden 9, P.O. Box 17, 6700AA, Wageningen, the Netherlands
[b] ITAP, INRAE, Institut Agro, University Montpellier, Montpellier, France
[c] ChemHouse Research Group, Montpellier, France
[d] Department of Chemistry, University of Rome "La Sapienza", Piazzale, Aldo Moro 5, 00185, Rome, Italy
[e] Department of Physical and Chemical Sciences, University of L'Aquila, Via Vetoio, 67100, Coppito, L'Aquila, Italy
[f] Université Paris-Saclay, INRAE, AgroParisTech, UMR SayFood, 75005, Paris, France
[g] National Wine and Grape Industry Centre, Charles Sturt University, Wagga Wagga, Australia



Pre-processing approach	Apple data set		Olive data set		Mango data set		Pear data set	
	R^2	RMSEP	R^2	RMSEP	R^2	RMSEP	R^2	RMSEP
Raw	0.85	0.77	0.70	1.21	0.76	1.09	0.83	0.52
MSC	0.86	0.74	0.90	0.70	0.81	0.96	0.81	0.55
VSN	0.83	0.80	0.92	0.63	0.82	0.96	0.79	0.58
SNV	0.82	0.82	0.91	0.69	0.81	0.96	0.82	0.54
2nd derivative	0.89	0.65	0.90	0.72	0.77	1.07	0.81	0.56
SPORT	0.95	0.46	0.89	0.73	0.83	0.92	0.84	0.51
PORTO	0.95	0.44	0.93	0.61	0.84	0.91	0.85	0.49

ROSA →PROSAC



Chemometrics and Intelligent Laboratory Systems 222 (2022) 104497

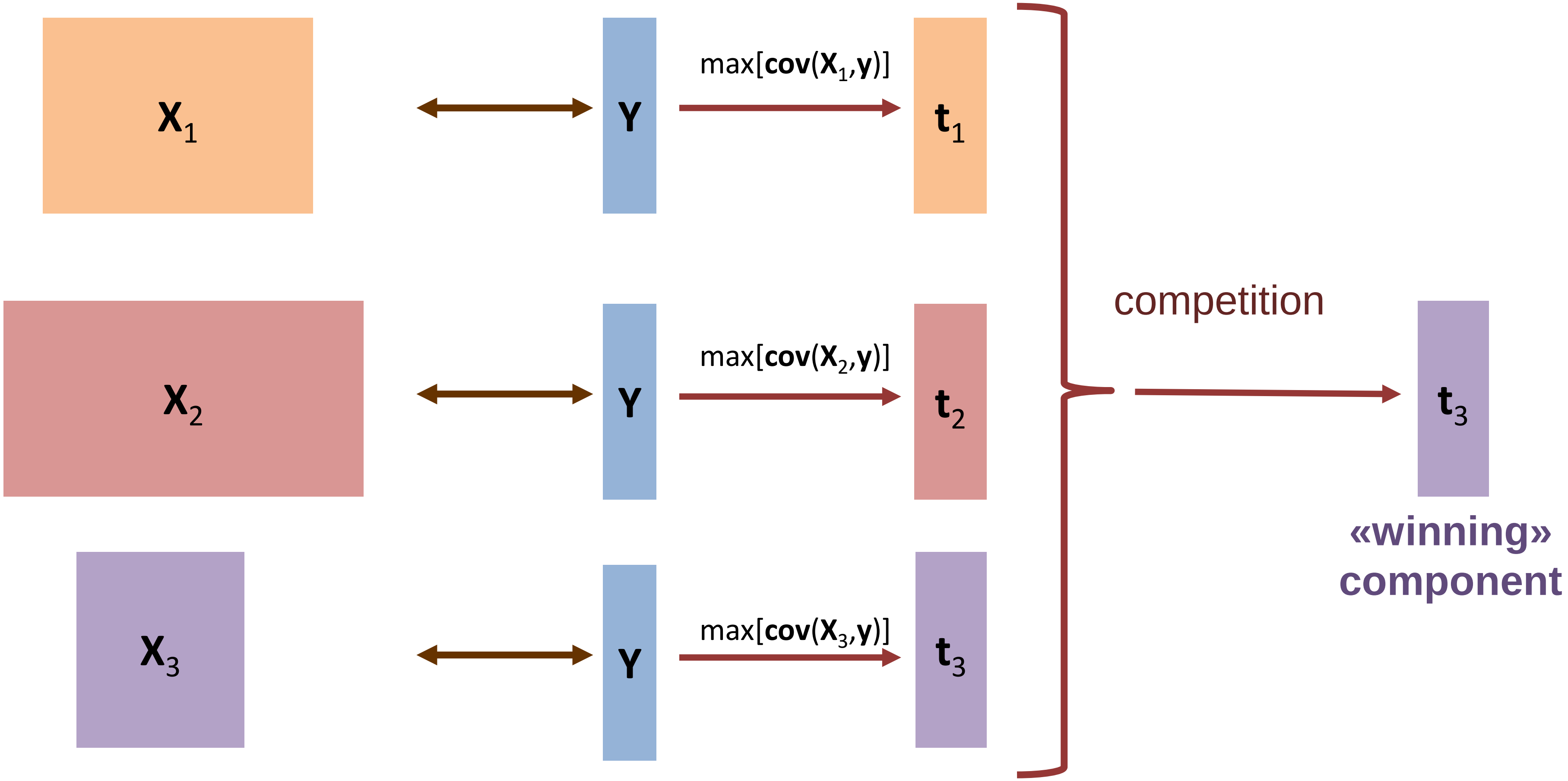
Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemometrics

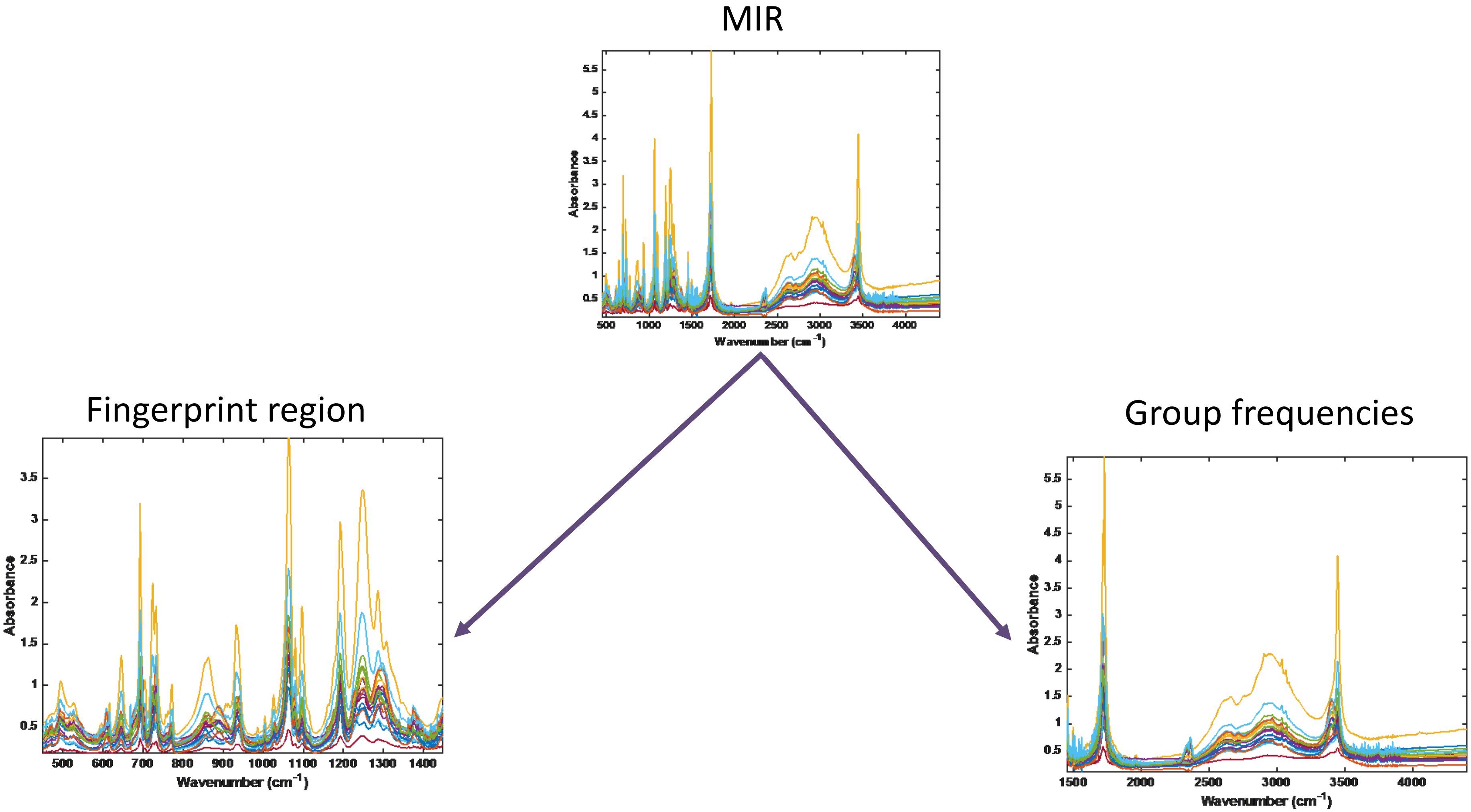
Pre-processing ensembles with response oriented sequential alternation calibration (PROSAC): A step towards ending the pre-processing search and optimization quest for near-infrared spectral modelling

Puneet Mishra [a,*], Jean Michel Roger [b,c], Federico Marini [d], Alessandra Biancolillo [e], Douglas N. Rutledge [f,g]



Multi-block data - 2

- Sometimes blocking can be induced within the same data set, due to physical or chemical reasons:


MIR

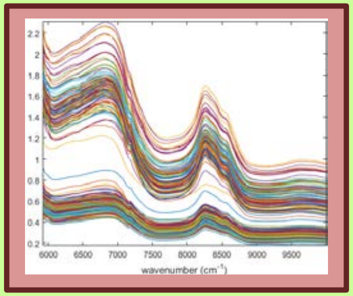
Fingerprint region

Group frequencies

Hazelnuts data set

2112 Variables

286 Samples

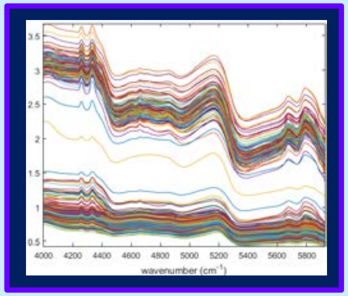


X

Overtone region

1000 Variables

286 Samples



Z

Combination bands region

221 PDO Romana Hazelnut

155 Other Hazelnut

2 Classes

Training Set of 286 samples

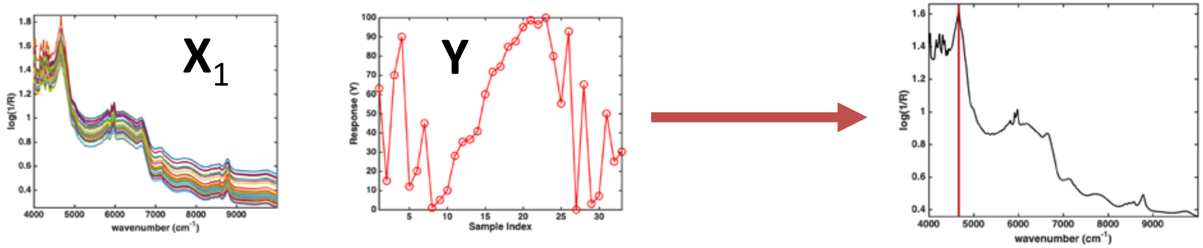
Test Set of 90 samples

49 Romana PDO 41 Others

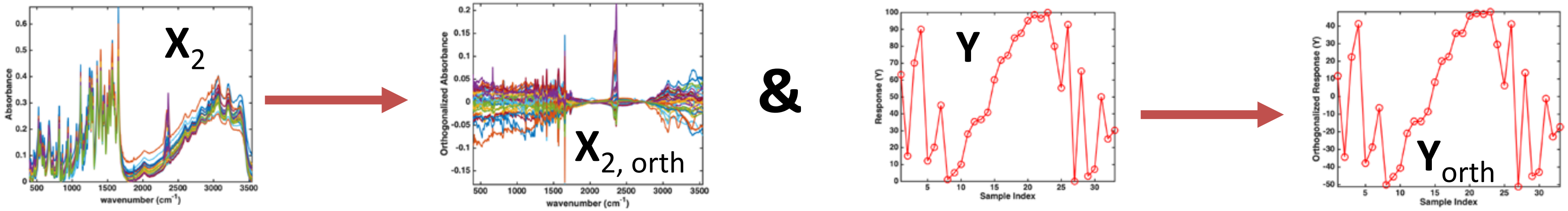
Adding variable selection: SO-CovSel

- Variables from the 1st block are selected by CovSel

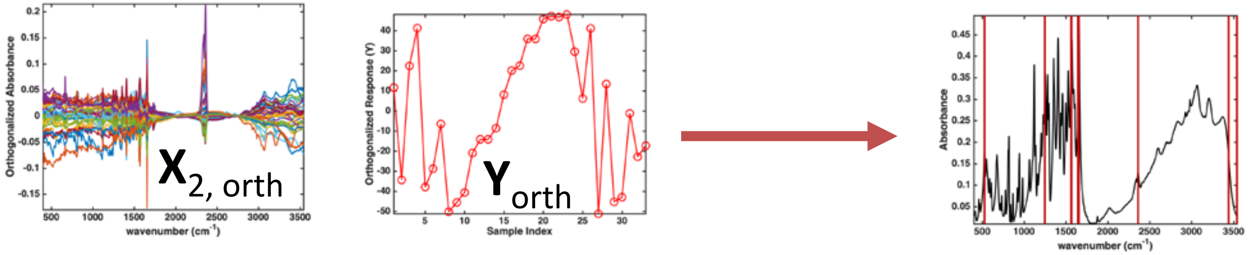
A. Biancolillo et al., *J. Chemometr.* **34** (2020) e3120



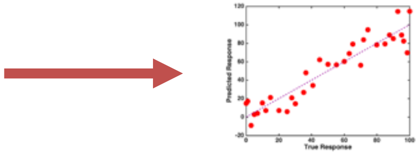
- The second X block and the Y are orthogonalized wrt the selected variables



- Variables from the orthogonalized 2nd block are selected by CovSel



- An overall regression model is calculated between the Y and the selected variables



Hazelnuts data set: Predictions

SO-PLS-LDA		
Class	Predicted PDO	Predicted Common
PDO	38	3
Common	3	46

6 Misclassified

SO-CovSel-LDA		
Class	Predicted PDO	Predicted Common
PDO	39	2
Common	3	46

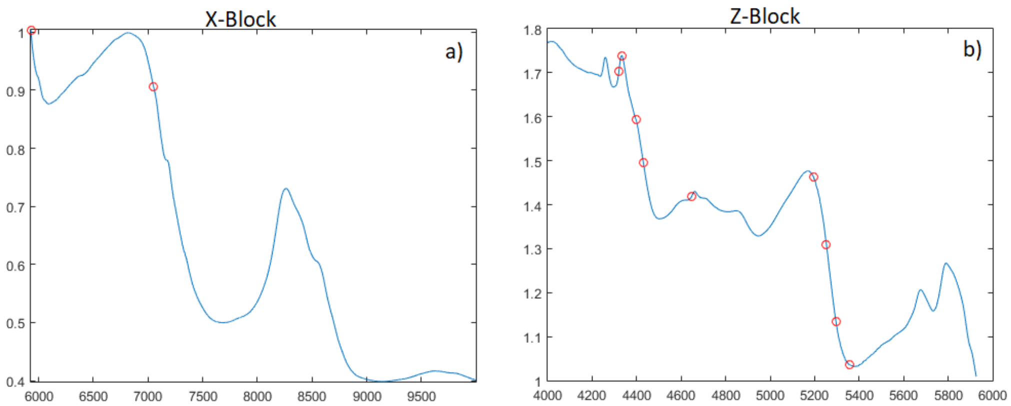
5 Misclassified

Hazelnuts data set: Interpretation

CovSel

X: 2 variables selected

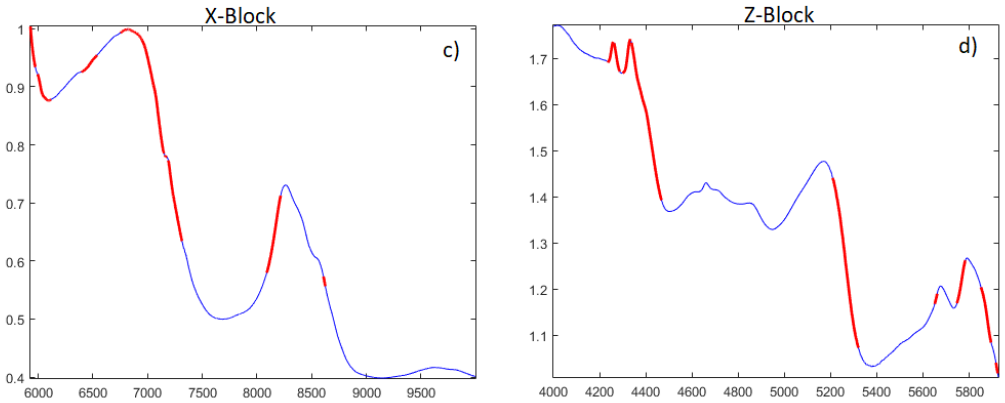
Z: 9 variables selected



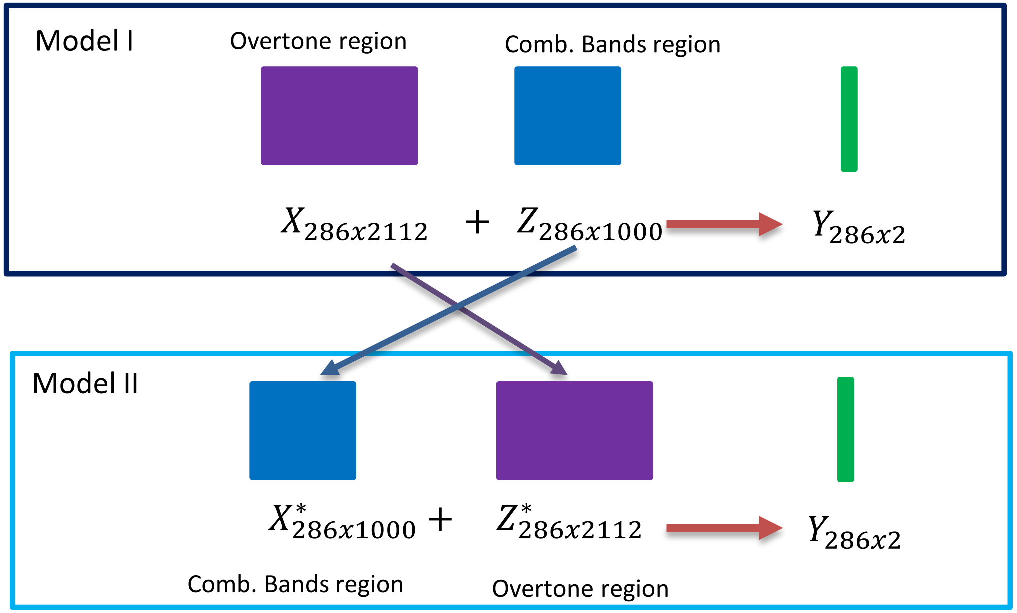
VIP

X: 526 variables selected

Z: 210 variables selected



Hazelnuts data set: Interpretation - 2



Model I

Overtone region

Comb. Bands region

$X_{286x2112} + Z_{286x1000} \rightarrow Y_{286x2}$

Model II

$X^{*}_{286x1000} + Z^{*}_{286x2112} \rightarrow Y_{286x2}$

Comb. Bands region

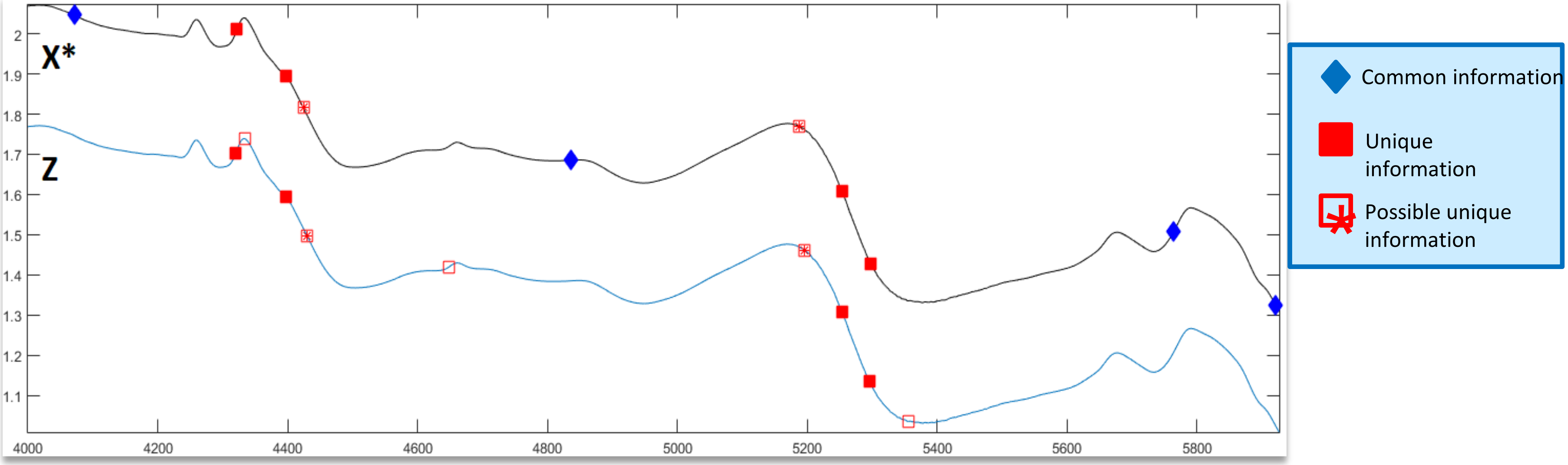
Overtone region

- If a variable is selected only when the block is the 1st input block (i.e. it is removable by orthogonalization) it is **common** between the blocks

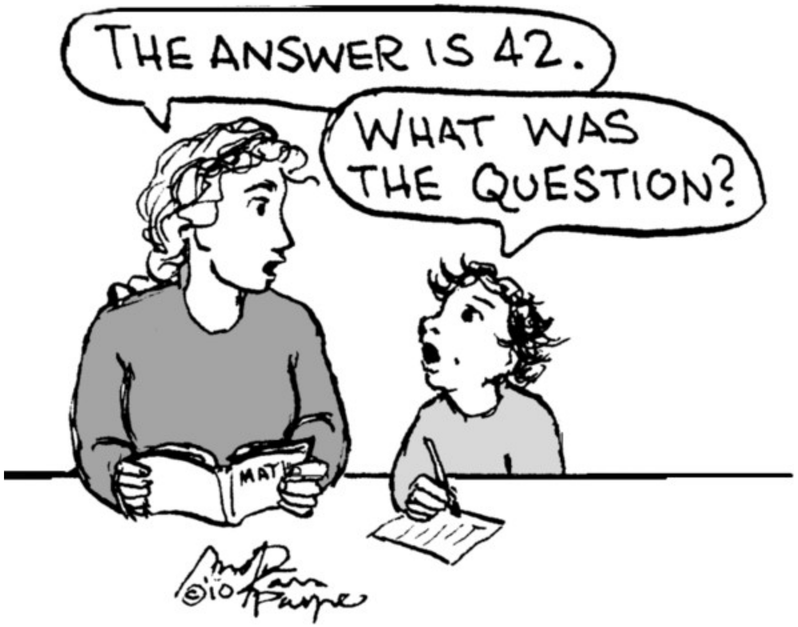
- If a variable is selected independently of the order of the blocks it represent **unique information** brought by a block and not present in the other

Hazelnuts data set: Interpretation - 4

	Data Block		Selected variables (cm^{-1})	n. of Sel.Variables
Model I	Z	Combination Bands region	4320; 4334; 4397; 4430; 4648; 5195; 5253; 5296; 5355	9
Model II	X^*	Combination Bands region	4071; 4322; 4397; 4424; 4837; 5188; 5253; 5298; 5764; 5920	10



Conclusions & Questions



Digression: Introducing VSN

Spectral pretreatment: Standard Normal Variate

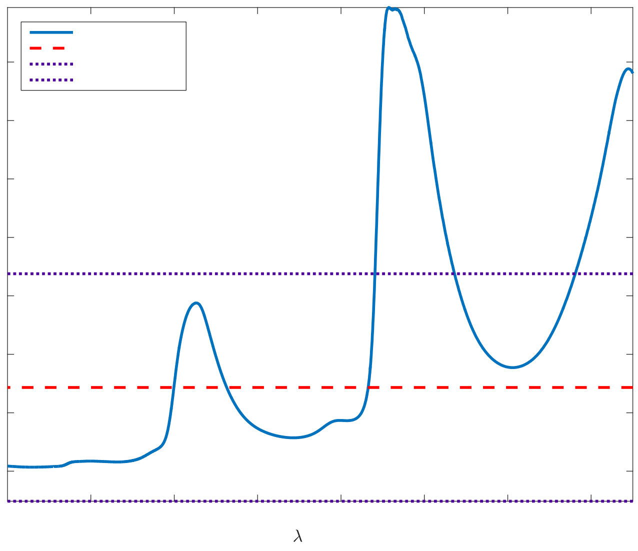
- Correction for linear and additive effects:

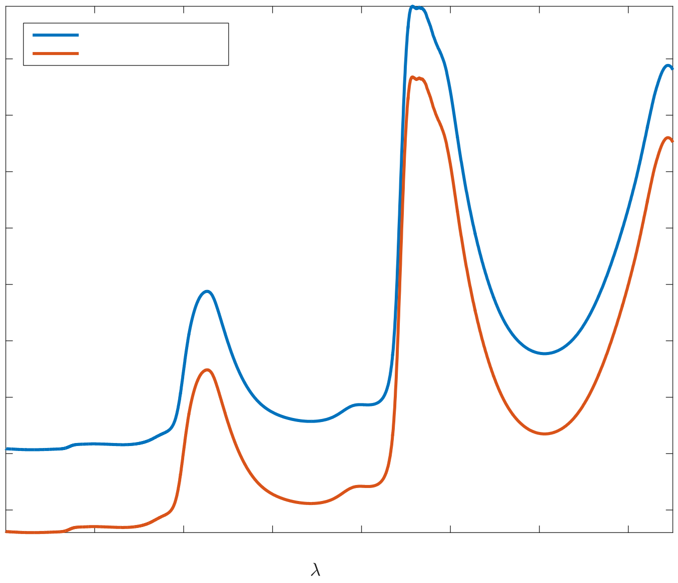
$$\mathbf{z}_i \approx a_i\mathbf{1} + b_i\mathbf{z}_{i,chem}$$

- Mean and standard deviation of measured spectra are used for the correction:

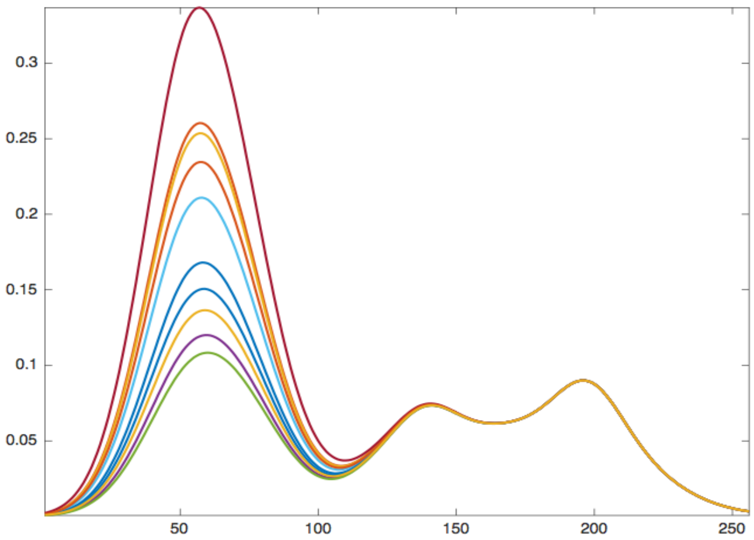
$$a_i = \mathrm{mean}(\mathbf{z}_i) \qquad b_i = \mathrm{std}(\mathbf{z}_i)$$

$$\mathbf{z}_{i,corr} = \frac{\mathbf{z}_i - a_i}{b_i}$$

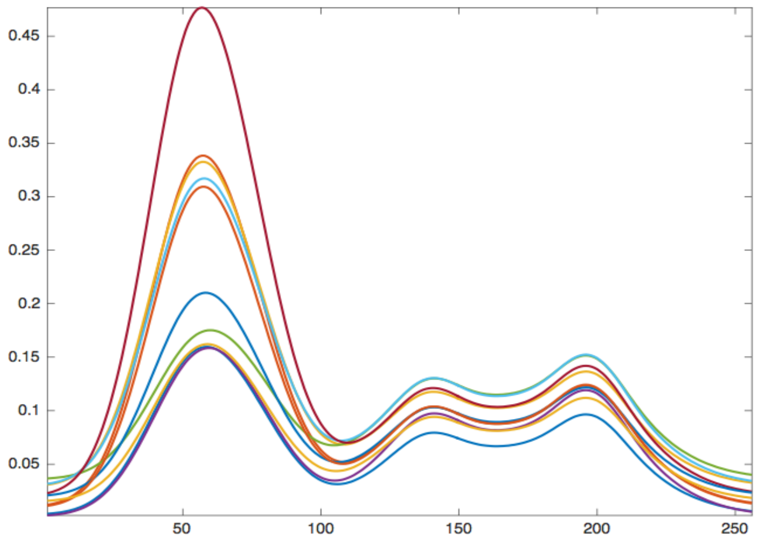




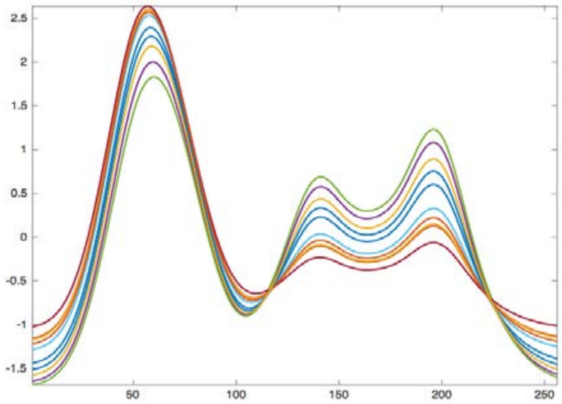
Introduction: a simulated example



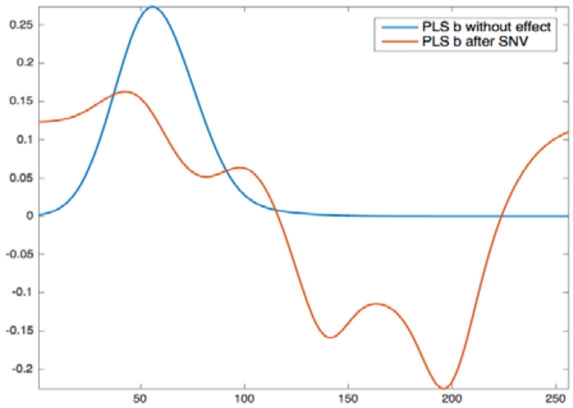
Spectra without any additive or multiplicative effect
One peak related to Y, two not



+ additive and multiplicative effect



After applying SNV



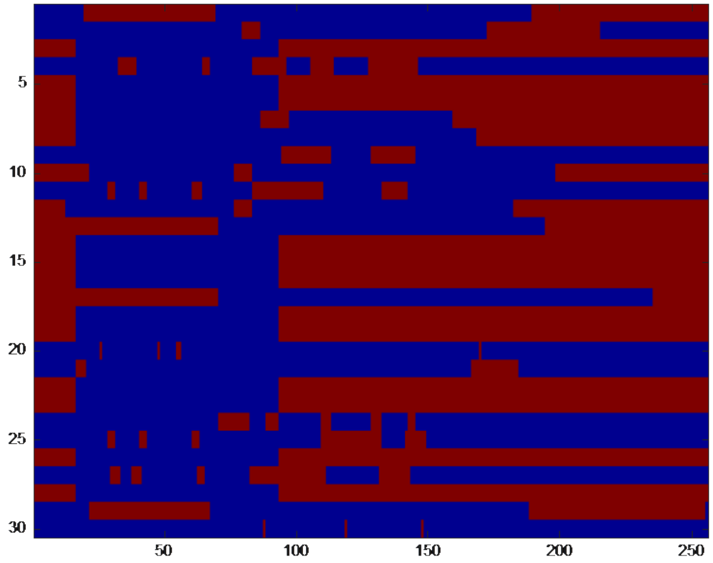
Model performances are good (on calibration set)
But the model itself is erroneous

Theory

- SNV tends to dilute the information along the whole spectrum
- A solution :
  - To calculate standard deviation and mean on wavelengths little related to **Y**
  - To normalize the spectrum with these values
- Or, more generally:
  - To calculate diagonal matrix **W** of weights between 0 (no selection) and 1 (complete selection)
  - To calculate the normalisation on **Wx** and apply it to **x**

An algorithm using RANSAC in practice

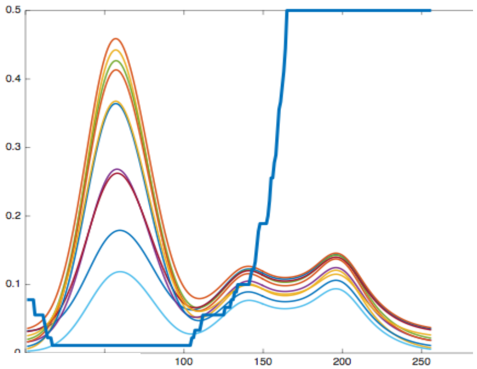
An algorithm using RANSAC in practice

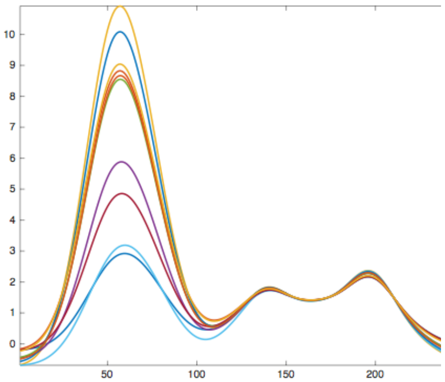


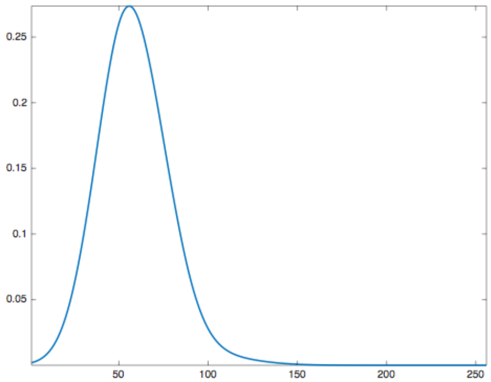
- The largest set is selected and stored
- Another pair of spectra is randomly picked
- Procedure continues iteratively until maximum number of runs
- Weights are calculated as frequency of selection in the consensus sets

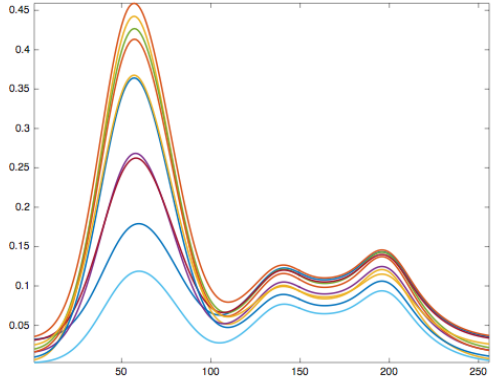
Results on simulated data

weighted SNV
tol = 0.001

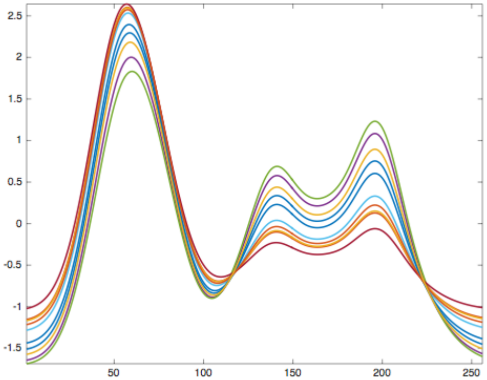


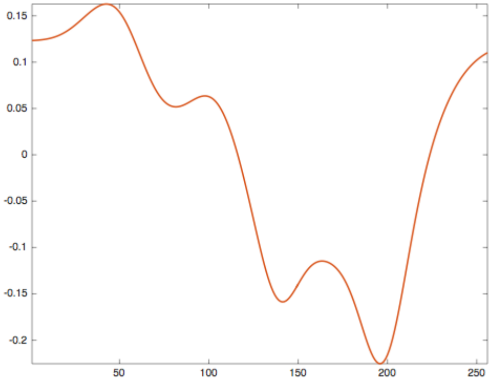






classical SNV





Spectral pretreatment: Extended MSC

- May also remove non-linear baseline or contribution from interferents:

$$\boldsymbol{z}_i \approx a_i\mathbf{1} + b_i\boldsymbol{z}_{i,chem} + c_i\lambda + d_i\lambda^2 + f_i\boldsymbol{x}_{int}$$

- Also in this case, it is easier to describe chemical variation as difference with respect to a reference spectrum, $\boldsymbol{m}$.
- Then:

$$\boldsymbol{z}_i = a_i\mathbf{1} + b_i\boldsymbol{m} + c_i\boldsymbol{\lambda} + d_i\boldsymbol{\lambda}^2 + f_i\boldsymbol{x}_{int} + e_i$$

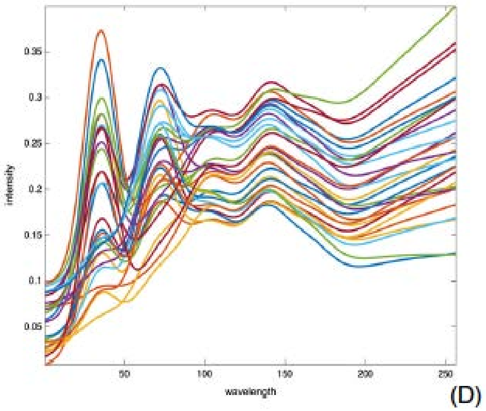
- Which becomes the regression problem:

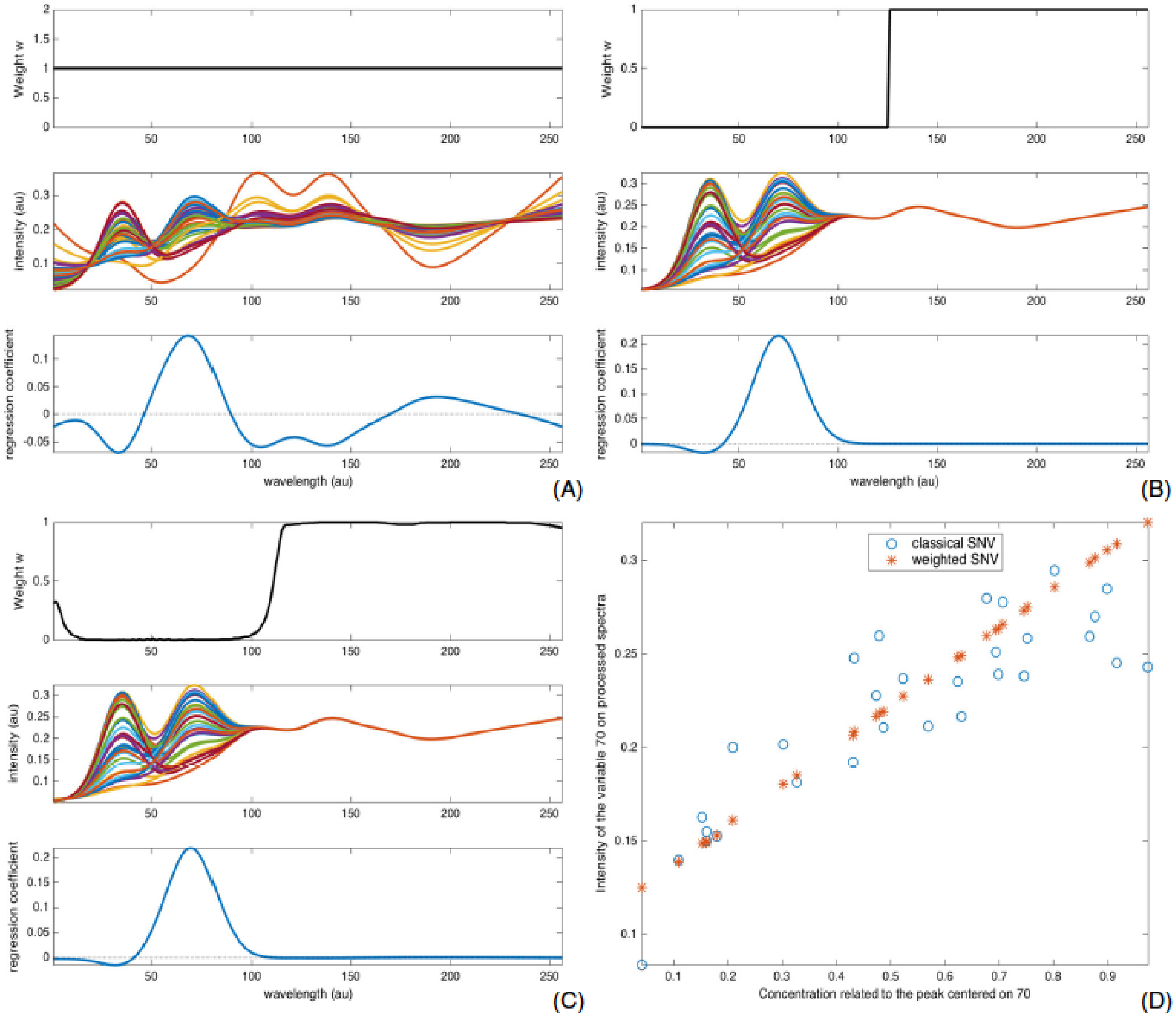
$$\boldsymbol{z}_i = \boldsymbol{M}\boldsymbol{p}_i + e_i \Longrightarrow \widehat{\boldsymbol{p}}_i = (\boldsymbol{M}^T\boldsymbol{M})^{-1}\boldsymbol{M}^T\boldsymbol{z}_i$$

with

$$\begin{aligned} p_i &= [a_i \quad b_i \quad c_i \quad d_i \quad f_i]^T \\ \boldsymbol{M} &= [\mathbf{1} \quad \boldsymbol{m} \quad \boldsymbol{\lambda} \quad \boldsymbol{\lambda}^2 \quad \boldsymbol{x}_{int}] \end{aligned}$$

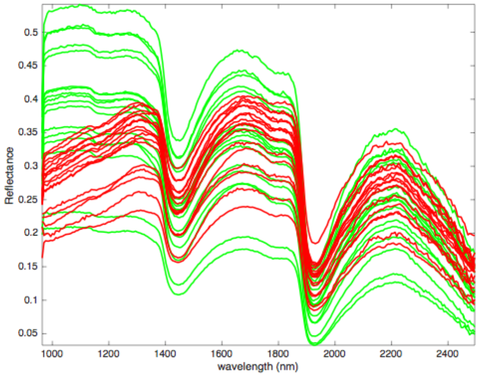
Results on simulated data – 2 EMSC



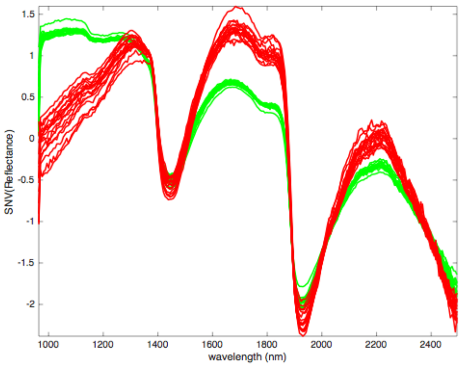


Real example – 1: Apple Leaves

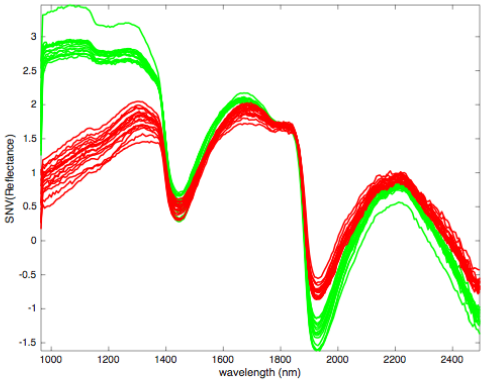
- Data: apple tree leaf spectra
- Images acquired with an NEO SWIR hyperspectral camera; 1000 - 2500 nm
- Each spectrum is the mean of pixels from an area
- Two classes :
  - healthy
  - scab disease spot



Raw data



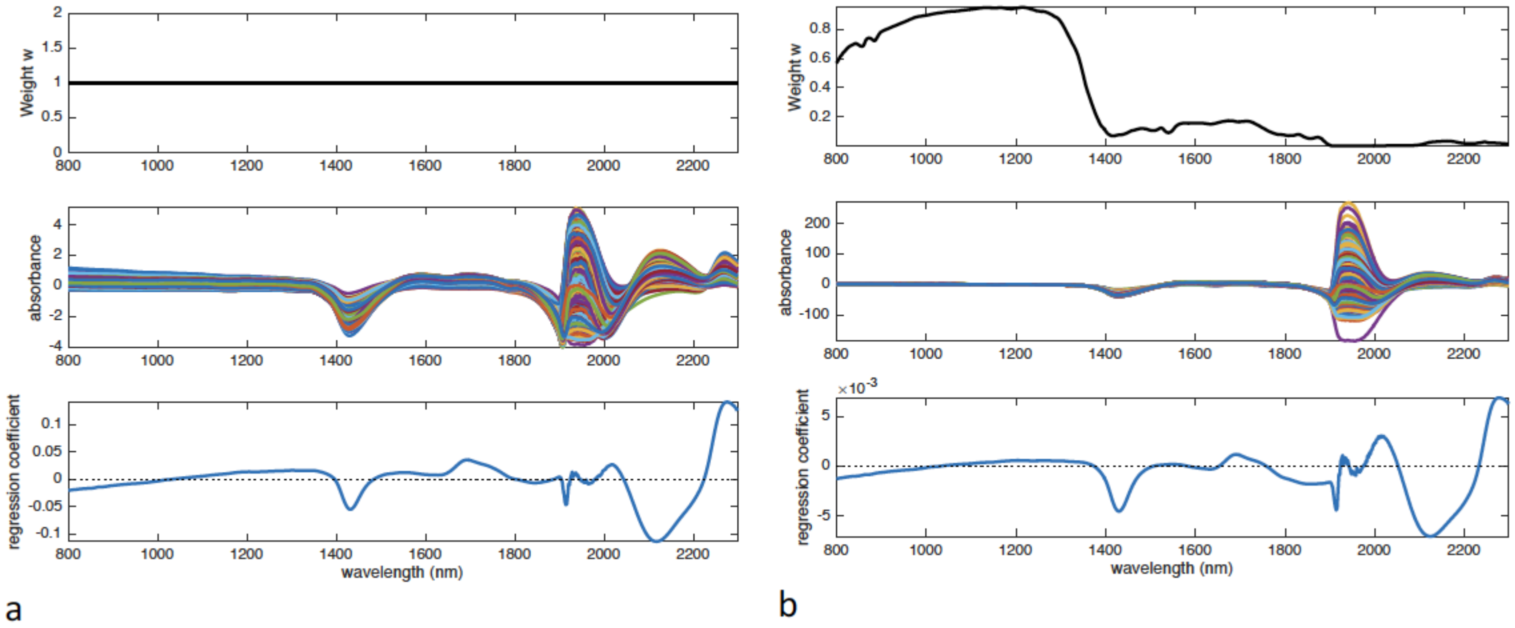
After SNV



After VSN

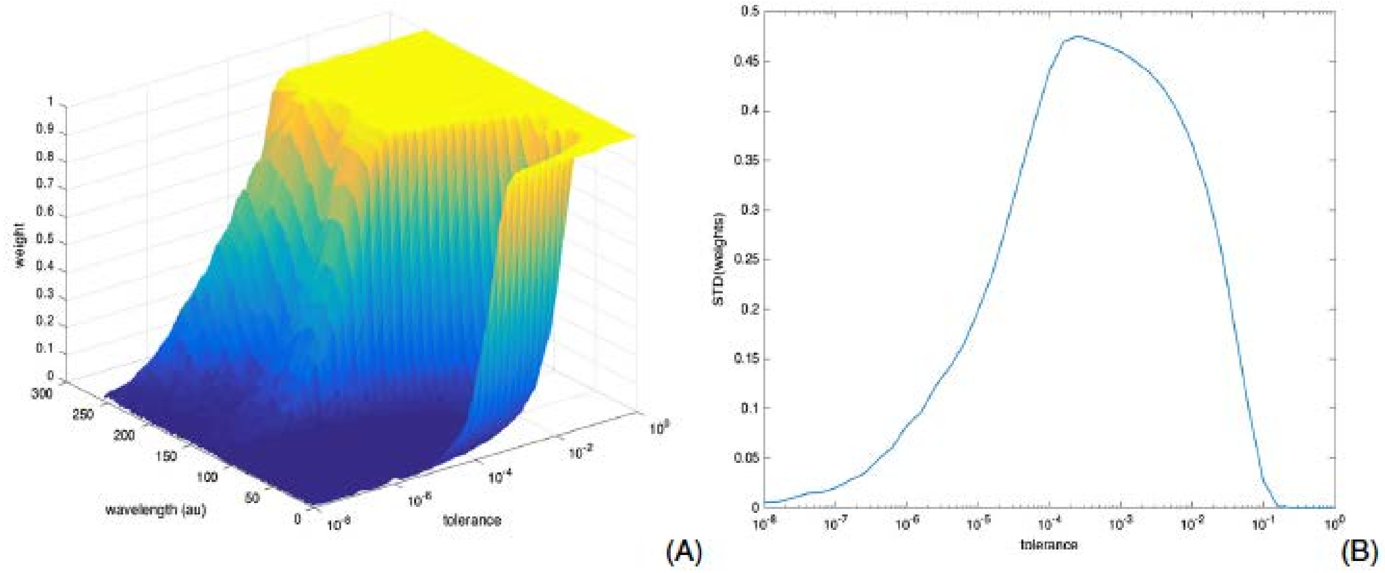
Real example – 2: Musts

- Data: NIR spectra of musts, associated to values of alcohol by volume.
- The dataset, containins 621 spectra: A calibration set of 414 samples (about 2/3) and a test set of 207 samples (about 1/3) were drawn using Duplex.



	#LV	RMSECV	R^2_{CV}	RMSEP	R^2_{TEST}
SNV	5	0.890	0.94	0.963	0.93
VSN + Weighted SNV	5	0.653	0.97	0.701	0.96

Digression: selecting the optimal tolerance



- Empirical rule: Max(std(weights))
- Seems to work on many data analyzed so far

Time for doing more SPORT