IMPROVING SPECTROSCOPIC-BASED AUTHENTICATION BY DATA FUSION

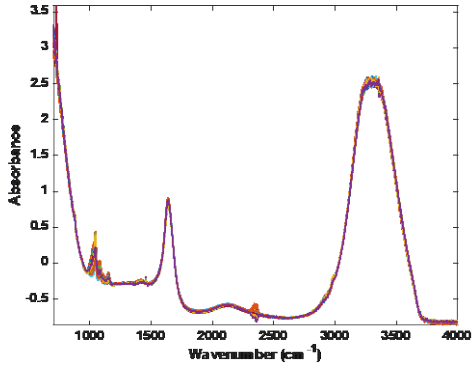
Federico Marini

Dept. Chemistry, University of Rome "La Sapienza"

SAPIENZA
UNIVERSITÀ DI ROMA

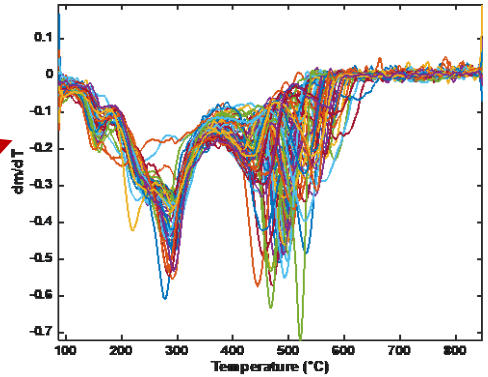
Multi-block data

- More and more situation/cases where different sets of (usually multivariate) data are available to characterize a system/process.
- E.g.: Same set of samples characterized by different analytical platforms

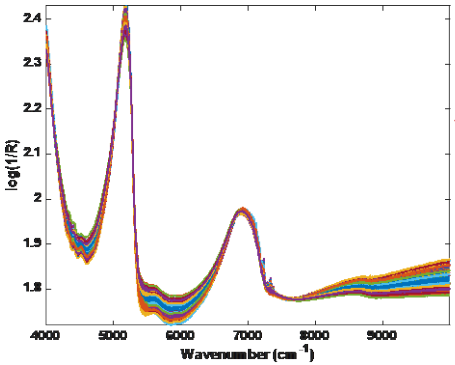


MIR

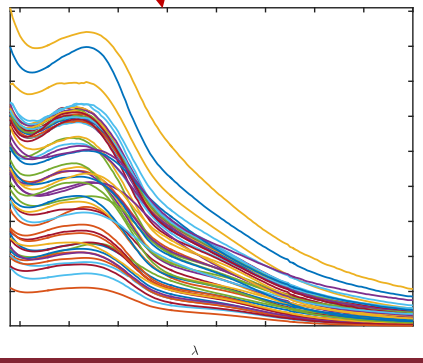
DTG



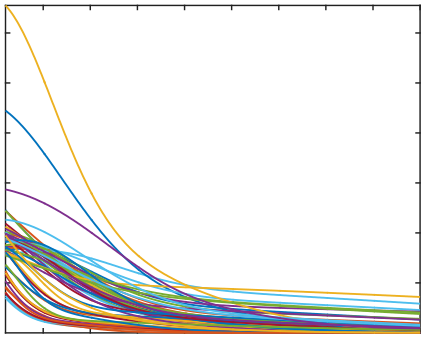
NIR



UV



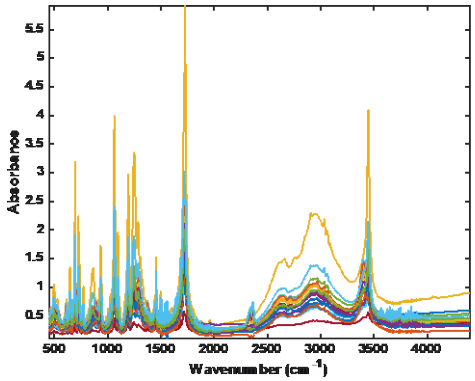
VIS



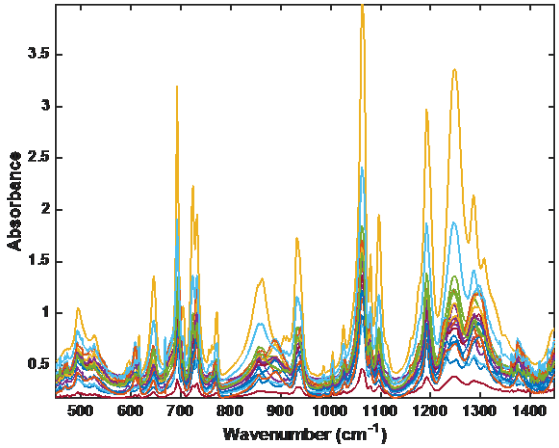
Multi-block data - 2

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

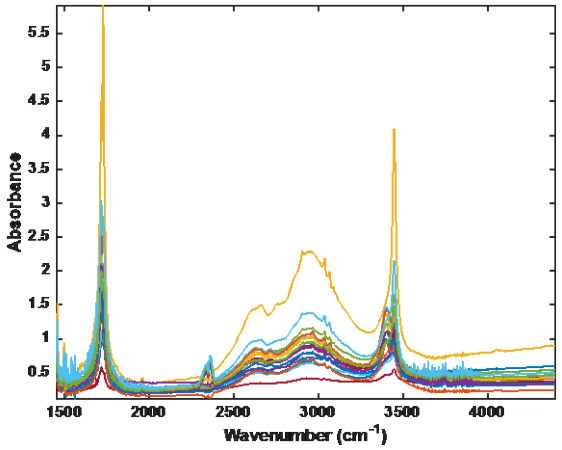
MIR



Fingerprint region

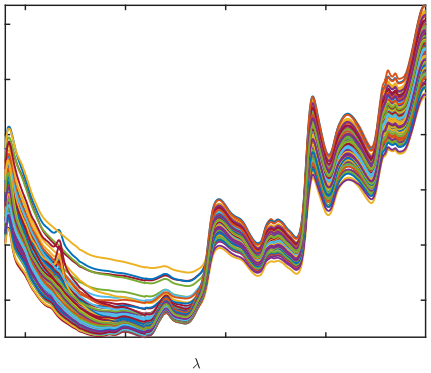


Group frequencies

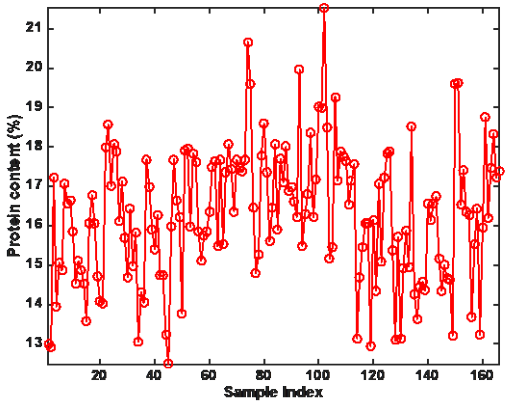


Multi-block data - 3

- Blocking can also be induced by asymmetric relations between the data:

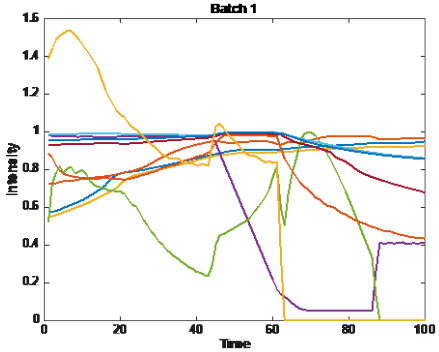


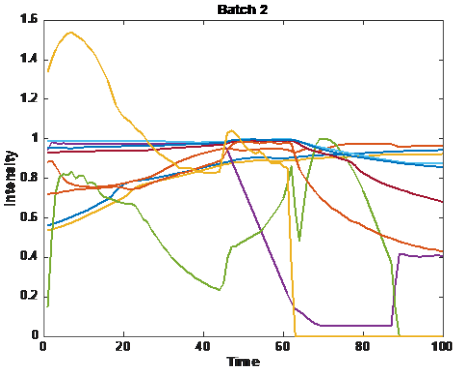
Independent variables



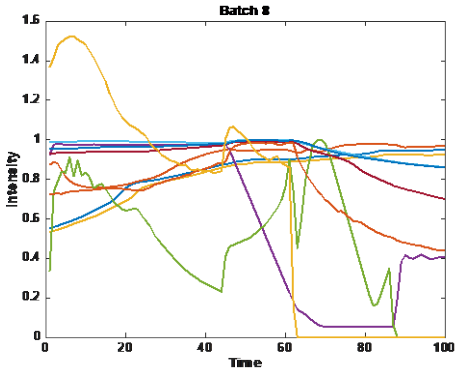
Dependent variables

- Or the same set of variables measured on different samples (e.g., groups or batches).



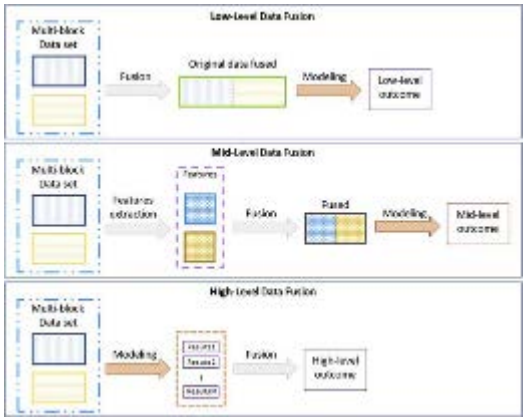


...

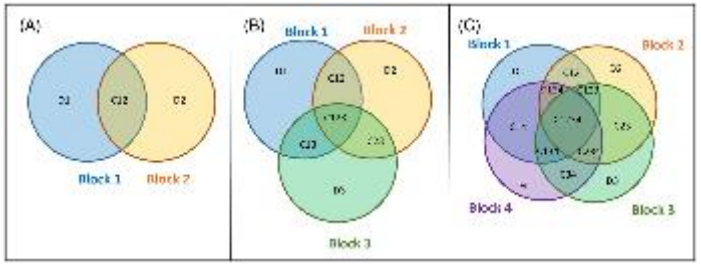


Multi-block data analysis

- Ignoring the block structure may blur the final results
- Multi-block models:
  - Keep the natural ordering of the data
  - Explain relation between blocks
  - Describe variation within blocks
  - Assess block contribution to the overall variability
- Hierarchy of MB models:
  - Based on the level of fusion



  - Based on the kind of information extracted (common/partly common/distinct)



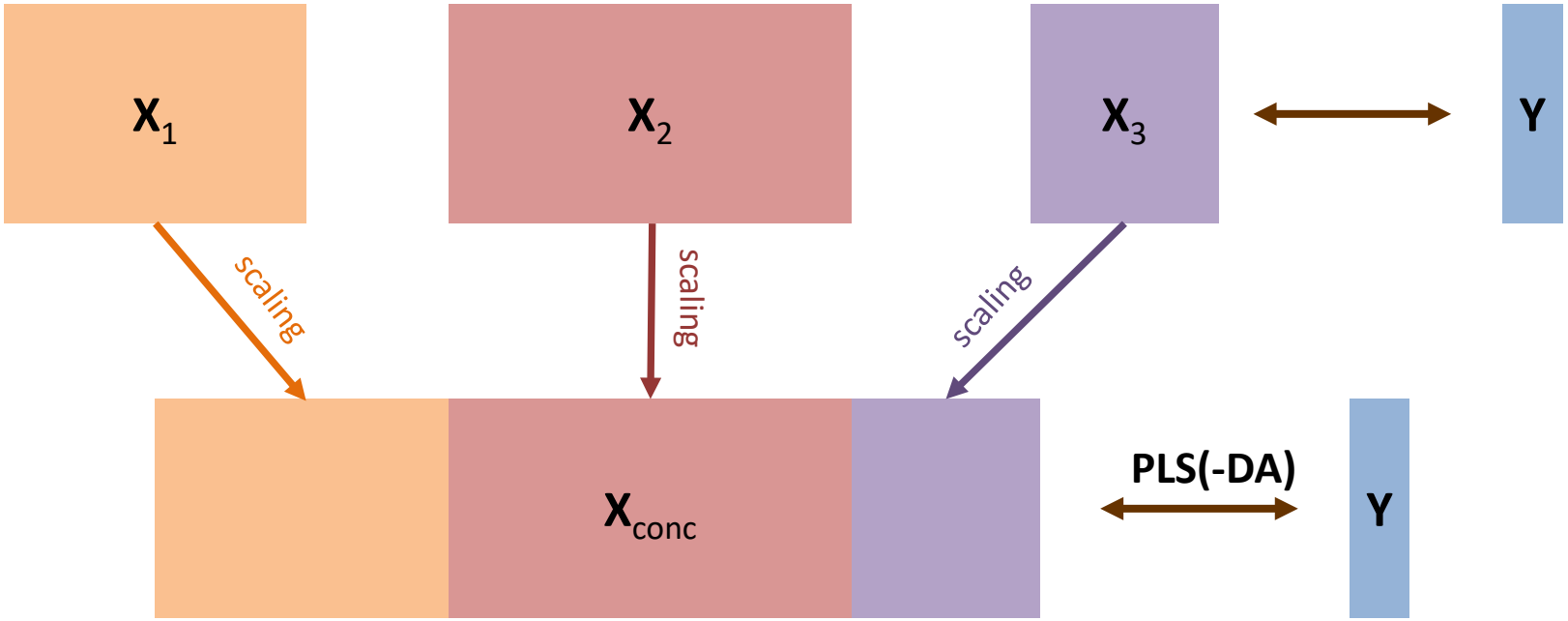
Reproduced from: I. Måge et al., *J. Chemometr.* **33** (2019) e3085.

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

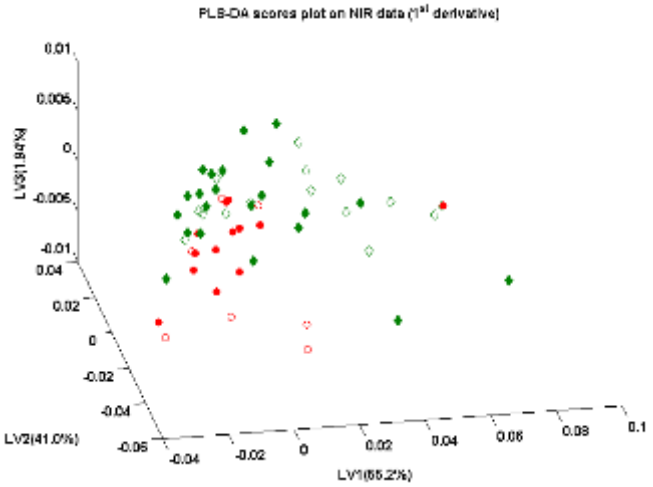


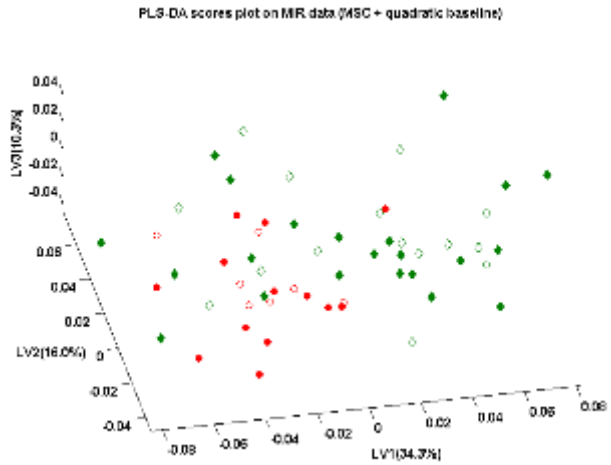
A first example: Authentication of extra virgin olive oils from PDO Sabina by NIR&MIR

Results of individual techniques (external validation)

- Best results with MSC + quadratic bl.:
  - 100.0% (Sabina)
  - 100.0% (other origins)

NIR



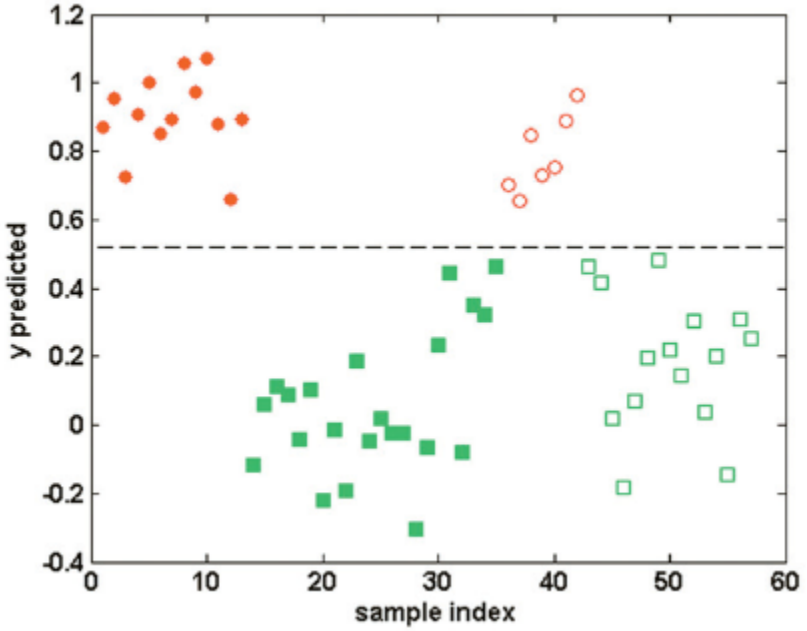


- Best results with MSC + quadratic bl.:
  - 85.7% (Sabina)
  - 86.7% (other origins)

MIR

Results of data fusion (external validation)

- LOW LEVEL
  - Without block-scaling: Block with the highest variance (here MIR) governs the model
  - With block-scaling: Improved contribution of NIR but still poorer results than with NIR alone
- MID LEVEL (PLS-DA scores after autoscaling)



AUTHENTICATION OF BEER

Characterization of artisanal beer ‘‘Reale’’ and its authentication

BIRRA DEL BORGO

‘‘ *ReAle*’’ is an artisanal beer brewed by ‘‘*Birrificio del Borgo”*, an Italian microbrewery well recognized also abroad for its high quality products

Analytica Chimica Acta 820 (2014) 23–31

Contents lists available at ScienceDirect

Analytica Chimica Acta

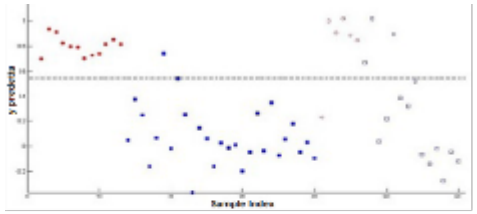
journal homepage: www.elsevier.com/locate/aca

Data-fusion for multiplatform characterization of an italian craft beer aimed at its authentication

Alessandra Biancolillo, Remo Bucci, Antonio L. Magrì, Andrea D. Magrì, Federico Marini *

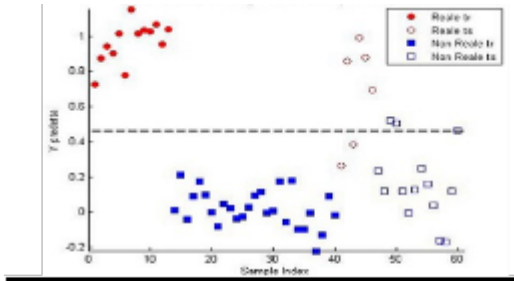
Department of Chemistry, University of Rome "La Sapienza", Rome, Italy

Results of individual techniques



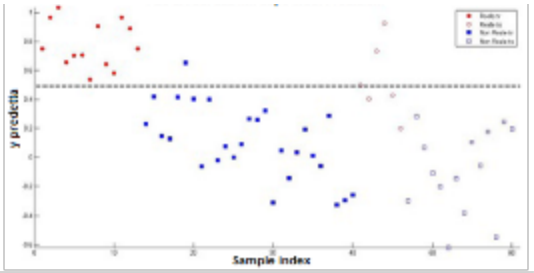
Predictions		
Pretreatment	%Correct Class. (Pred)	
	"Reale"	"Not Reale"
Deriv. I (+MC)	**83.3**	**71.4**

TG



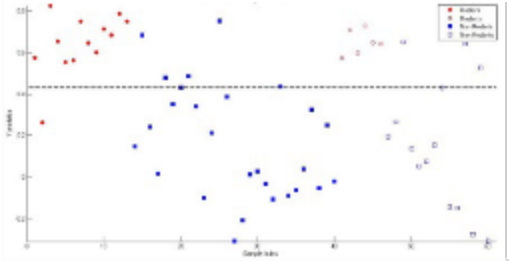
Predictions		
Pretreatment	%Correct Class. (Pred)	
	"Reale"	"Not Reale"
SNV+Detrending (+MC)	**86.7**	**78.6**

MIR



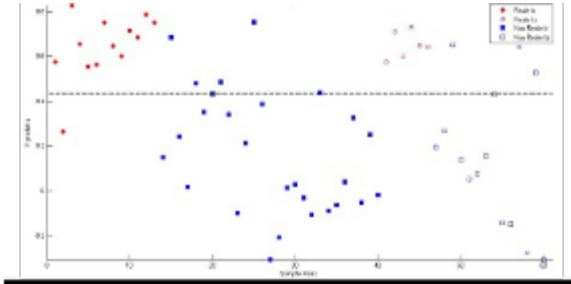
Predictions		
Pretreatment	%Correct Class. (Pred)	
	"Reale"	"Not Reale"
MSC (+MC)	**66.7**	**100.0**

NIR



Predictions		
Pretreatment	%Correct Class. (Pred)	
	"Reale"	"Not Reale"
Mean Centering	**82.3**	**77.8**

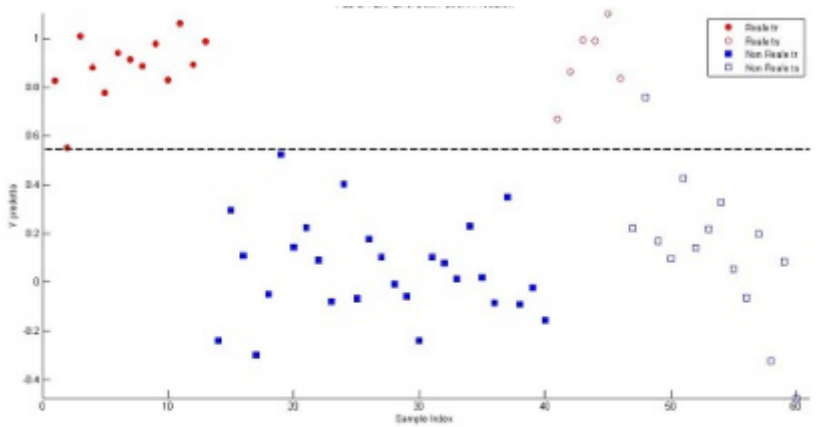
UV



Predictions		
Pretreatment	% Correct Class. (Pred)	
	"Reale"	"Not Reale"
Mean Centering	**100.0**	**85.7**

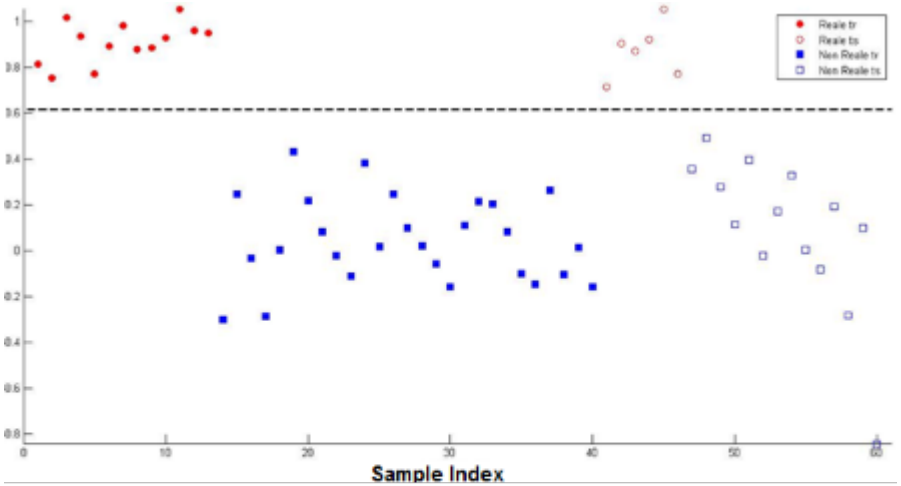
Vis

Data fusion



Predictions		
Pretreatment	% Correct Class. (Pred)	
	"Reale"	"Not Reale"
Without block scaling	**100.0**	**92.3**
With block scaling	**100.0**	**78.6**

Low Level



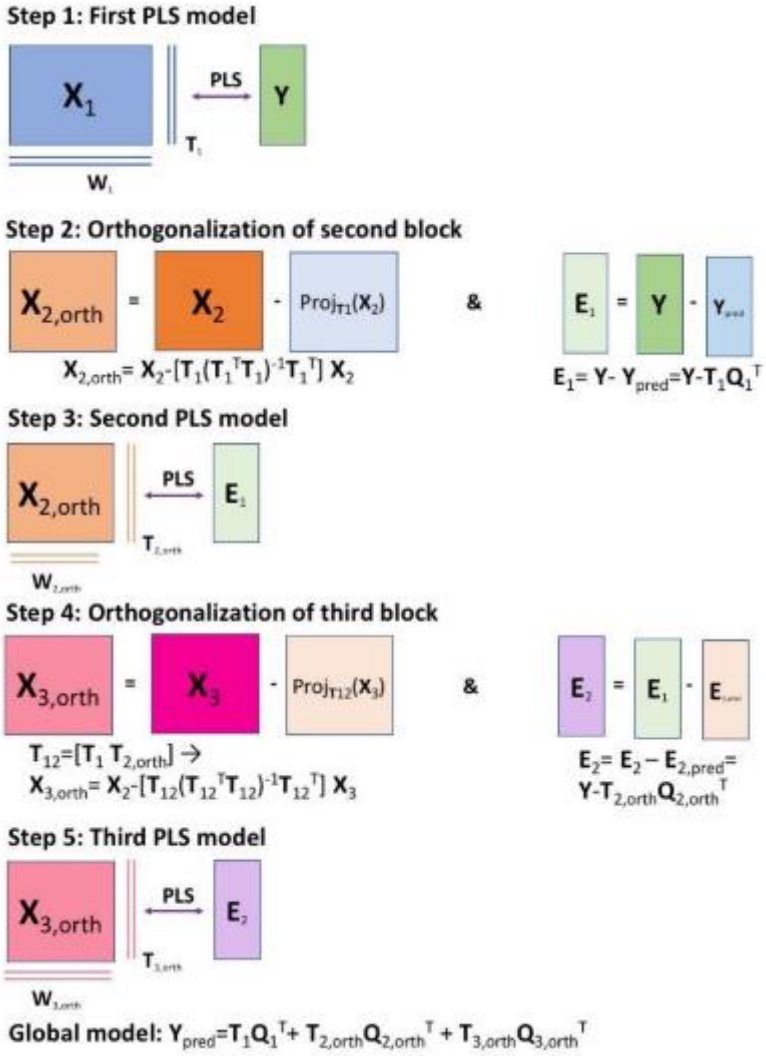
Predictions		
Pretreatment	% Correct Class. (Pred)	
	"Reale"	"Not Reale"
Mean Centering	**100.0**	**100.0**

Mid Level

Other multi-block paradigms

SO-PLS

- Sequential modeling after orthogonalization wrt scores of previous blocks



Example 1: Discrimination

Chemometrics and Intelligent Laboratory Systems 215 (2021) 104348

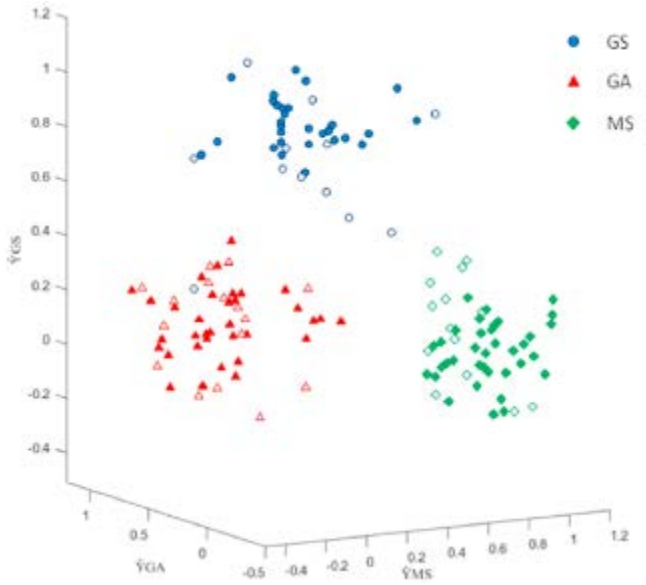
Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Spectroscopic fingerprinting and chemometrics for the discrimination of Italian Emmer landraces

Martina Foschi [a,*], Alessandra Biancolillo [a], Simona Vellozzi [b], Federico Marini [b], Angelo Antonio D'Archivio [a], Ricard Boqué [c]

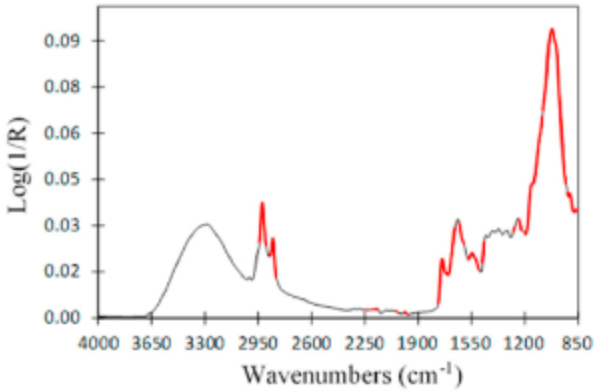


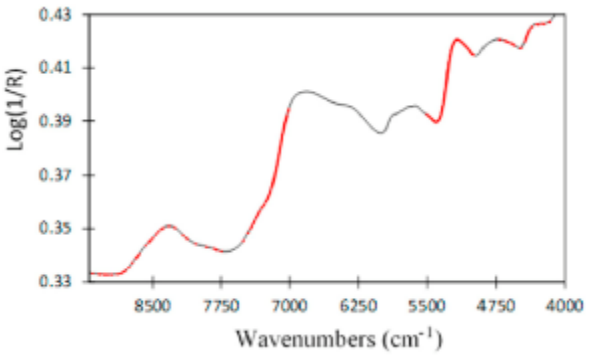
GS: 91.7%

GA: 100.0%

MS: 100.0%

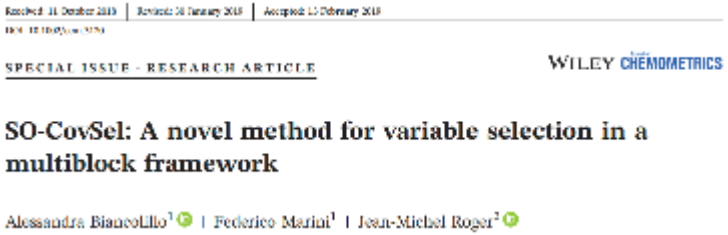
Tot: 97.2%





Introducing variable selection: SO-CovSel

- Couples SO-PLS with CovSel variable selection
- Extremely parsimonious variable selection



1. Selection of features from the first input block by CovSel → $\boldsymbol{X}_{1,sel}$
2. Prediction of the response, by ordinary least squares regression

$$\boldsymbol{Y} = \boldsymbol{X}_{1,sel}\boldsymbol{B}_{1,sel} + \boldsymbol{E}_1$$

3. Orthogonalization of the second input block wrt selected vars

$$\boldsymbol{X}_{2,orth} = \boldsymbol{X}_2 - \boldsymbol{X}_{1,sel}\left(\boldsymbol{X}_{1,sel}^T\boldsymbol{X}_{1,sel}\right)^{-1}\boldsymbol{X}_{1,sel}^T\boldsymbol{X}_{1,sel}$$

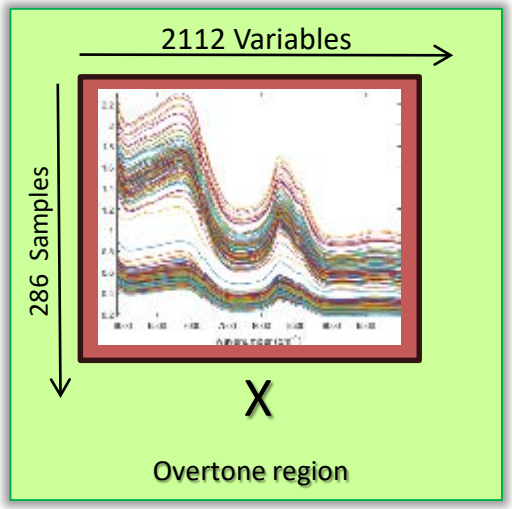
4. Selection of features from the orthogonalized second input block by CovSel → $\boldsymbol{X}_{2,sel}$
5. Prediction of the Y residuals by ordinary least squares regression

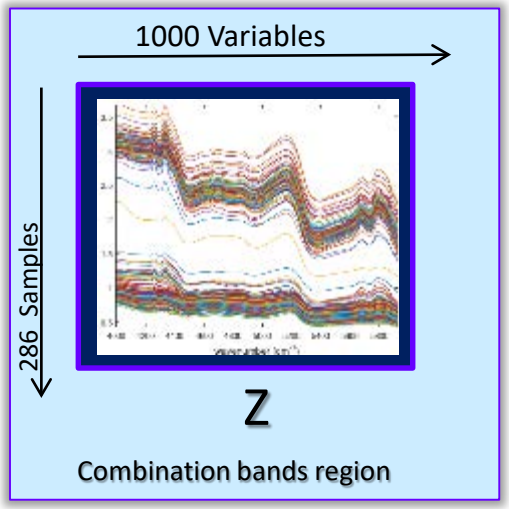
$$\boldsymbol{E}_1 = \boldsymbol{X}_{2,sel}\boldsymbol{B}_{2,sel} + \boldsymbol{E}_2$$

6. Repeat steps 3-5 for the remaining input blocks
7. Calculate overall model by OLS (or SO-PLS) on selected vars

$$\widehat{\boldsymbol{Y}} = \boldsymbol{X}_{1,sel}\boldsymbol{B}_{1,sel} + \boldsymbol{X}_{2,sel}\boldsymbol{B}_{2,sel} + \cdots + \boldsymbol{X}_{B,sel}\boldsymbol{B}_{B,sel}$$

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

221 PDO Romana Hazelnut

155 Other Hazelnut

2 Classes

Training Set of 286 samples

Test Set of 90 samples

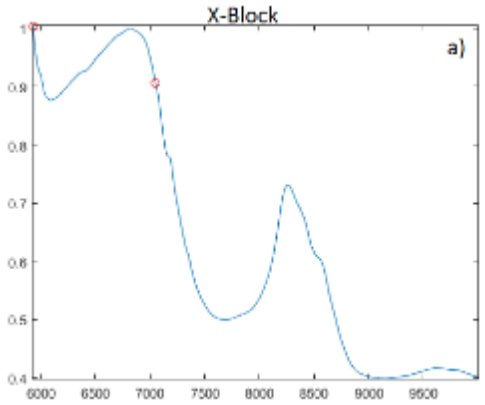
49 Romana PDO 41 Others

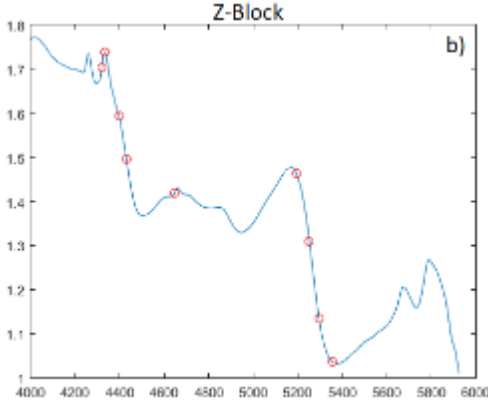
Hazelnuts data set: Interpretation

CovSel

X: 2 variables selected

Z: 9 variables selected

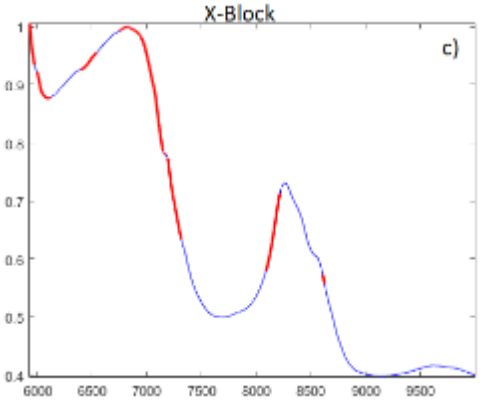


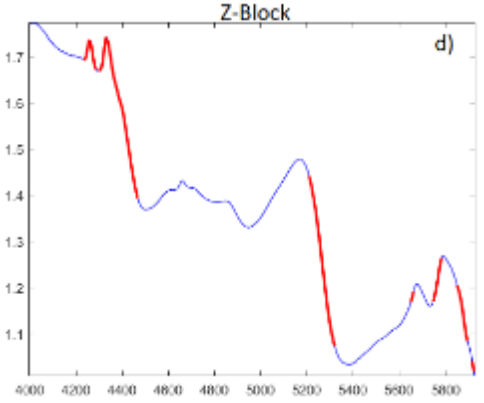


VIP

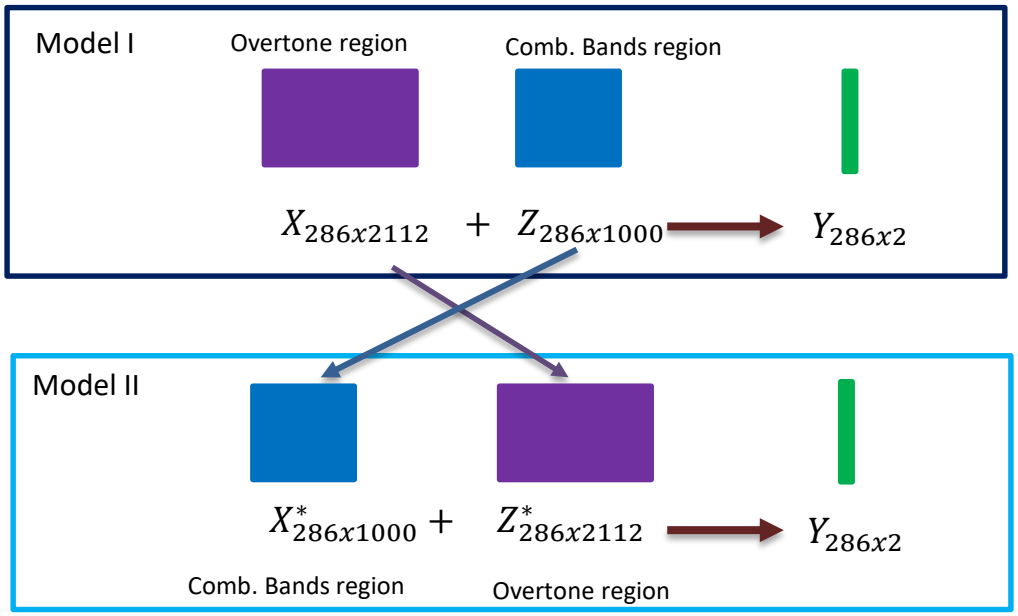
X: 526 variables selected

Z: 210 variables selected





Interpretation – Common and distinct information



- 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

Example 2: Calibration – EE prediction

	RMSEP	R^2	bias
MIR	10.9	0.89	0.1
NIR	8.8	0.93	4.7
ML-PLS	7.6	0.95	3.1
SO-PLS	7.8	0.95	1.3
SO-CovSel	12.0	0.87	1.7

molecules

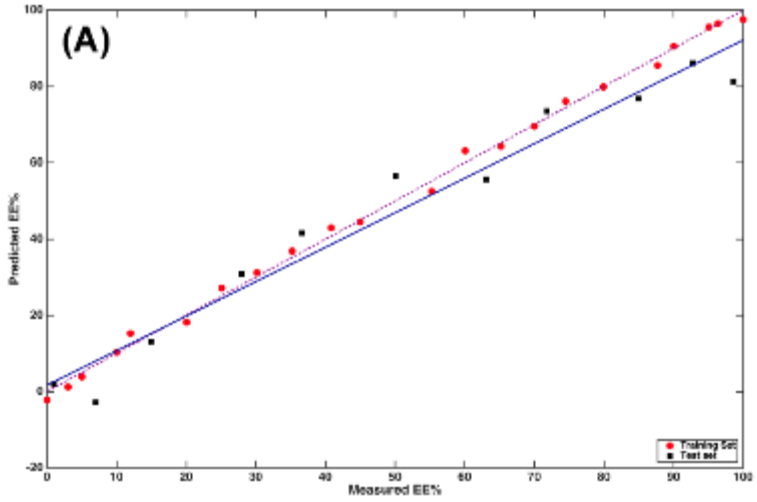
MDPI

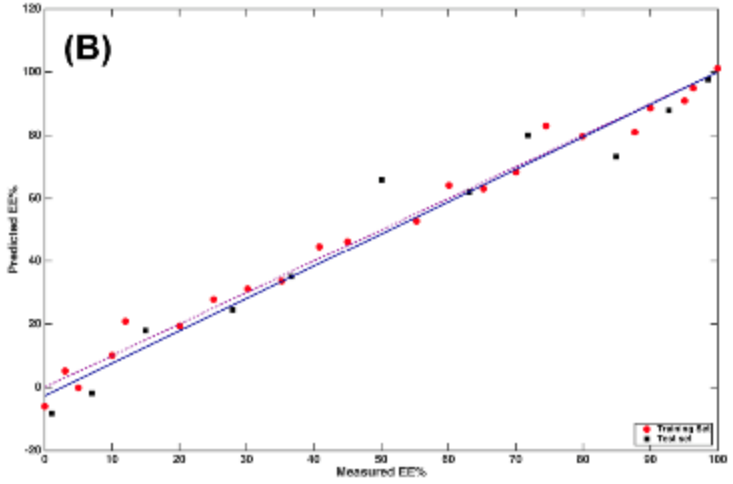
Article

Green multi-platform solution for the quantification of levodopa enantiomeric excess in solid-state mixtures for pharmacological formulations

Alessandra Biancolillo*, Stefano Battistoni, Regina Presutto, Federico Marini*

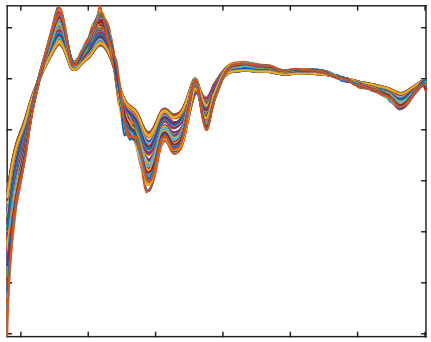
→ 5 wavenumbers





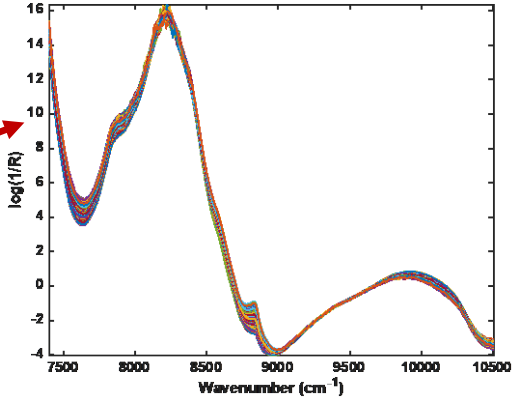
Preprocessing → Multi-block data

- The same data matrix pre-processed with different approaches



1st Der.

VSN

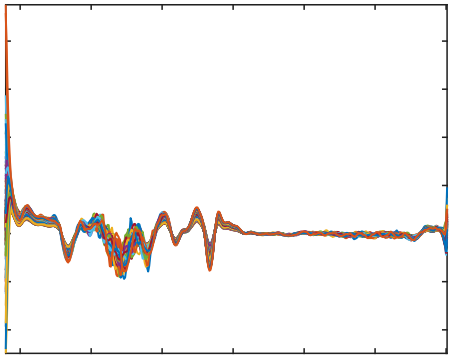


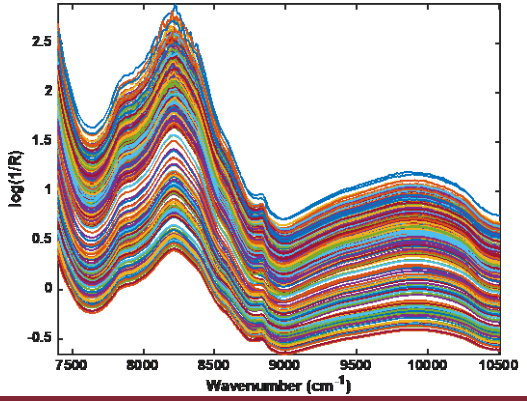
X

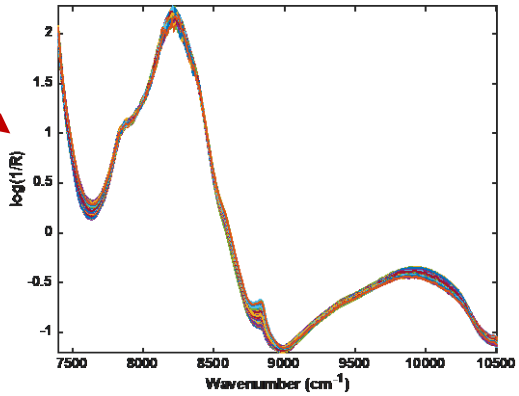
2nd Der.

Raw

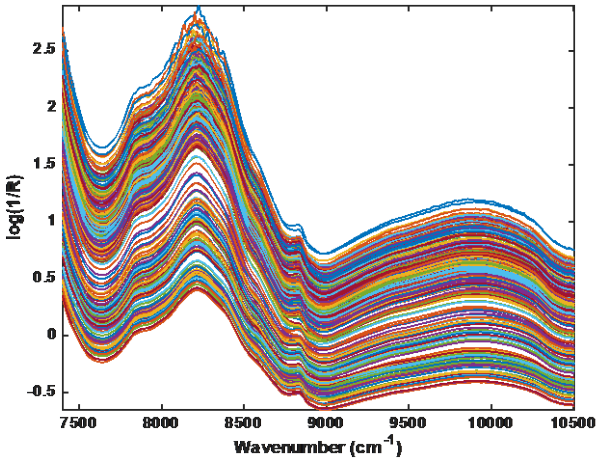
SNV





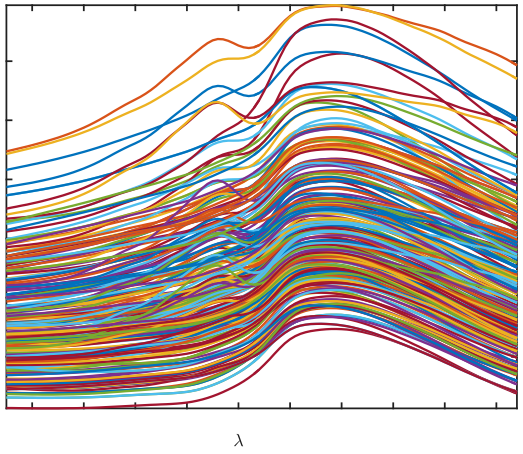


Data sets



Tablets

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



Meat

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

Chemometrics and Intelligent Laboratory Systems 199 (2020) 103975

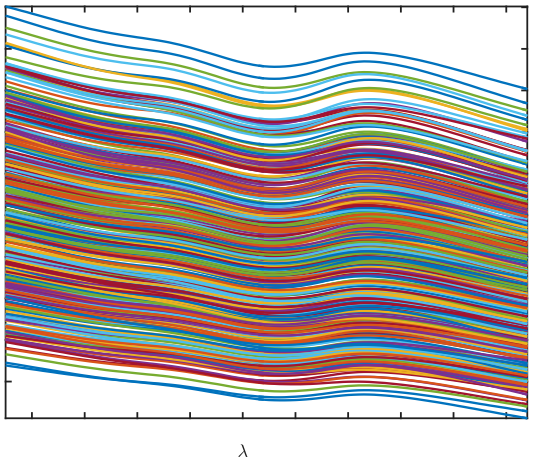
Contents lists available at ScienceDirect

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]



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

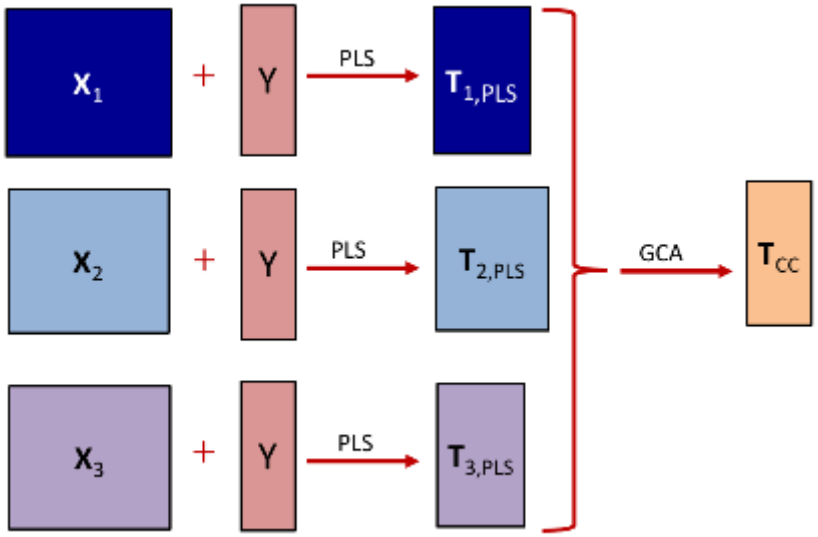
Tablets

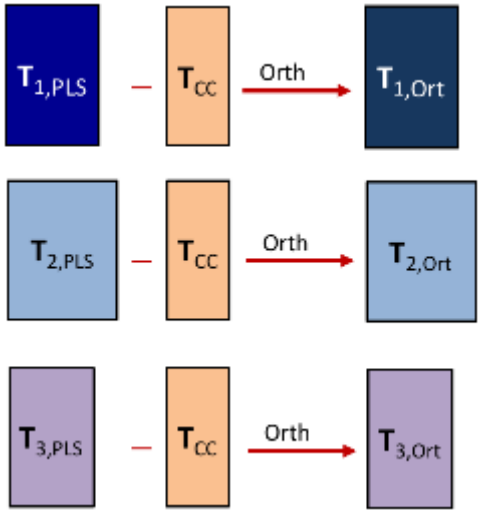
- By exchanging the order of the blocks, it is possible to explore common and distinctive information

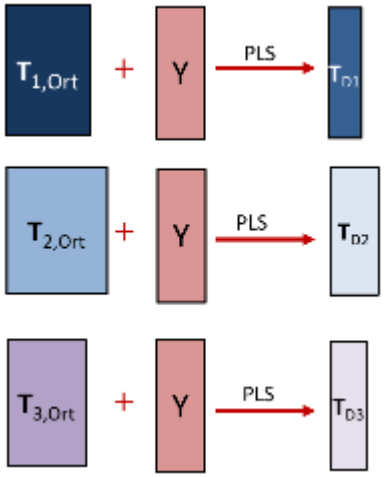
block number	Boosting 1	Boosting 2	Boosting 3
1	raw data	SNV	SG-15-3-2
2	SG-15-2-1	raw data	SNV
3	SG-15-3-2	SG-15-3-2	raw data
4	SNV	VSN, tol 0.0067, Npar 2	VSN, tol 0.0067, Npar 2
5	VSN, tol 0.0067, Npar 2	SG-15-2-1	SG-15-2-1
#LV	0,3,0,0,4	0,5,0,2,0	0,0,5,2,0
RMSEC	0.27	0.28	0.28
RMSEP	0.33	0.34	0.34

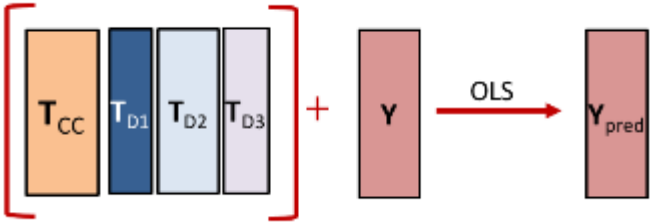
- Exchanging the order of the blocks has little effect on the predictivity, but impacts the selected pre-processings

PO-PLS


X1
+
Y
PLS
T1,PLS
X2
+
Y
PLS
T2,PLS
GCA
TCC
X3
+
Y
PLS
T3,PLS


T1,PLS
–
TCC
Orth
T1,Ort
T2,PLS
–
TCC
Orth
T2,Ort
T3,PLS
–
TCC
Orth
T3,Ort


T1,Ort
+
Y
PLS
TD1
T2,Ort
+
Y
PLS
TD2
T3,Ort
+
Y
PLS
TD3


TCC
TD1
TD2
TD3
+
Y
OLS
Ypred

Conclusions

- Fusing multiple data matrices can improve prediction or interpretation or lead to more robust/parsimonious models
- Use of sequential or parallel approaches + orthogonalization improves interpretation (and may embed selection)
- More and more strategies and algorithms are emerging

Your attention thanks for!