SO-CovSel: A NEW METHOD FOR VARIABLE SELECTION IN MULTI-BLOCK DATA

Alessandra Biancolillo [1], Federico Marini [1], Jean-Michel Roger [2]

[1] University of Rome La Sapienza, Piazzale Aldo Moro 5, I-00185, Rome, Italy,
[2] ITAP, Irstea, Montpellier SupAgro, Univ Montpellier, Montpellier, France

Introduction

- More and more multi-block data:
  - same samples measured by different techniques
- A lot of methods for analyzing all the blocks simultaneously
  - CCSWA (ComDim), MBPCA, MBPLS, SO-PLS, Statis, …
- But there is a lack of dedicated variable selection methods

The idea

To mix SO-PLS [1] and CovSel [2] because they act similarly

[1]: T. Næs, O. Tomic, B.-H. Mevik, H. Martens, Path modelling by sequential PLS regression (2011), J. Chemometr. 25, 28–40

[2]: J.M.Roger, B.Palagos, D.Bertrand, E.Fernandez-Ahumada, (2011) Chemometr. Intel. Lab. Syst.106, 216-223

SO-PLS

- Principle: To extract complementary latent variables from each block, successively

- Algorithm:
  1. i=1
  2. Extract k_i PLS latent variables from (X_i, Y)
  3. Orthogonalize each block > i wrt already extracted LV
  4. Orthogonalize Y wrt already extracted LV
  5. i=i+1; GOTO 2

CovSel

- Principle: To extract original variables from one block X which explain the Covariance between X and Y

- Algorithm:
  1. Extract the variable i from (X, Y) / Max($cov^2(X_i,Y)$)
  2. Orthogonalize X wrt already extracted X_i
  3. Orthogonalize Y wrt already extracted X_i
  4. GOTO 1

CovSel = PLS with loadings as [0 0 … 1 … 0 0 0]

SO - CovSel

- Principle: To extract complementary original variables from each block, successively

- Algorithm:
  1. i=1
  2. Extract k_i original variables from (X_i, Y) by CovSel
  3. Orthogonalize each block > i wrt already extracted variables
  4. Orthogonalize Y wrt already extracted variables
  5. i=i+1; GOTO 2

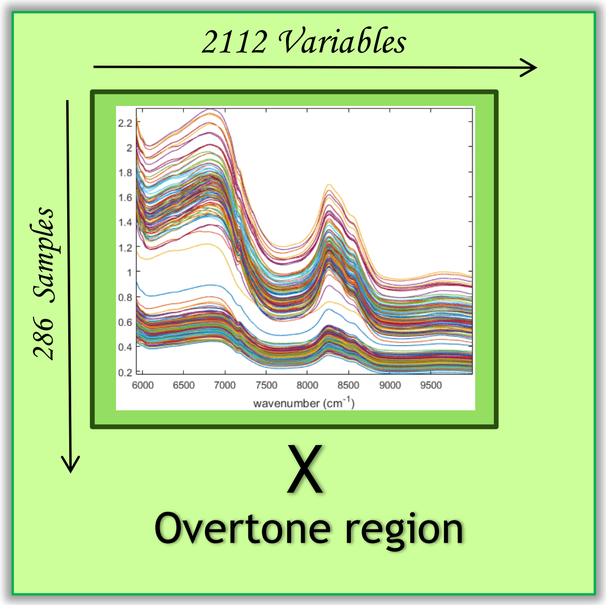
Under publication in Journal of Chemometrics

SO-PLS vs SO-CovSel

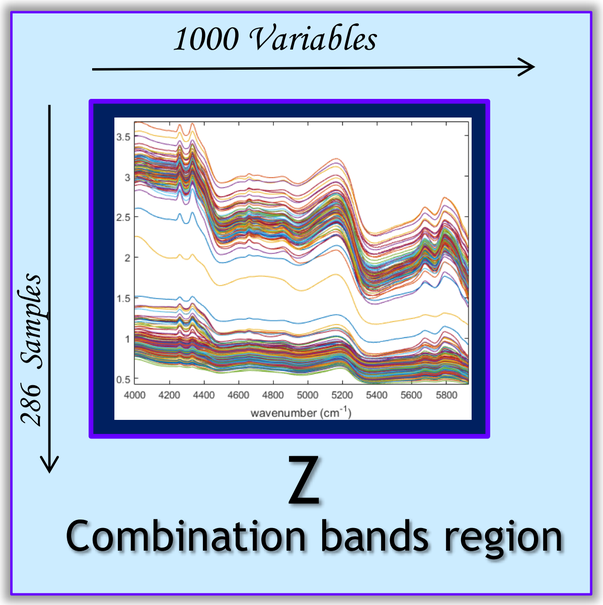
- From two blocks X1 and X2
- Let K1, K2 be the number of (latent) variables to extract from X1, X2
- SO-PLS provides scores $T1_{(NxK1)}$ $T2_{(NxK2)}$, as linear combination of the original variables of X1 and X2
- SO-CovSel provides scores $T1_{(NxK1)}$ $T2_{(NxK2)}$, as subsets of original variables of X1 and X2
- Depending on the Y, the scores of both methods can be inputted into:
  - a linear regression -> SO-PLS-R & SO-CovSel-R
  - a discriminant analysis -> SO-PLS-DA & SO-CovSel-DA

First example: Hazelnut origin discrimination

Overtone region



Combination bands region



221 PDO Romana Hazelnut

155 Common Hazelnut

2 Classes

Training Set of 286 samples

Test Set of 90 samples

49 Romana PDO 41 Common

Hazelnuts: Results

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

Latent V = 3, 5

6 Misclassified

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

Selected V = 2, 10

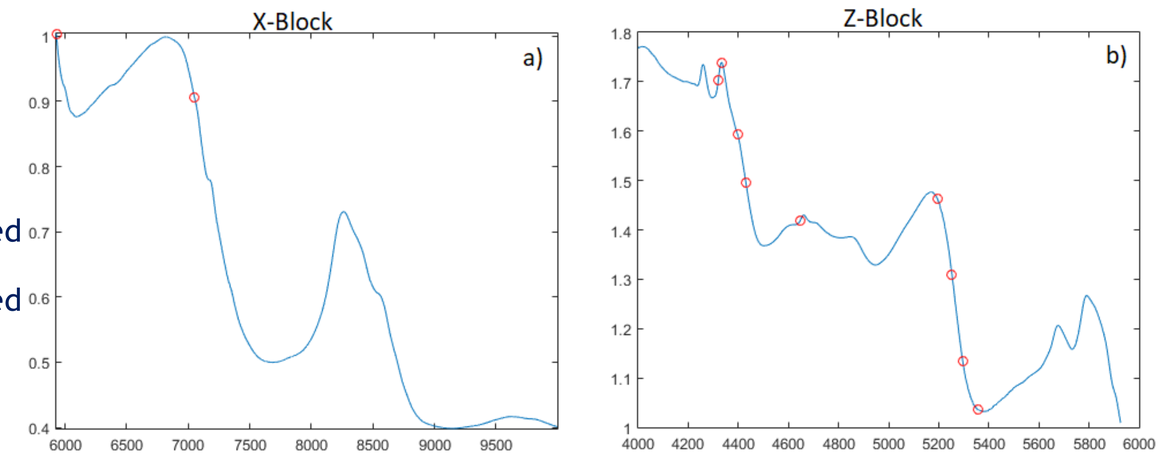
5 Misclassified

Hazelnuts: Results

CovSel

X: 2 variables selected

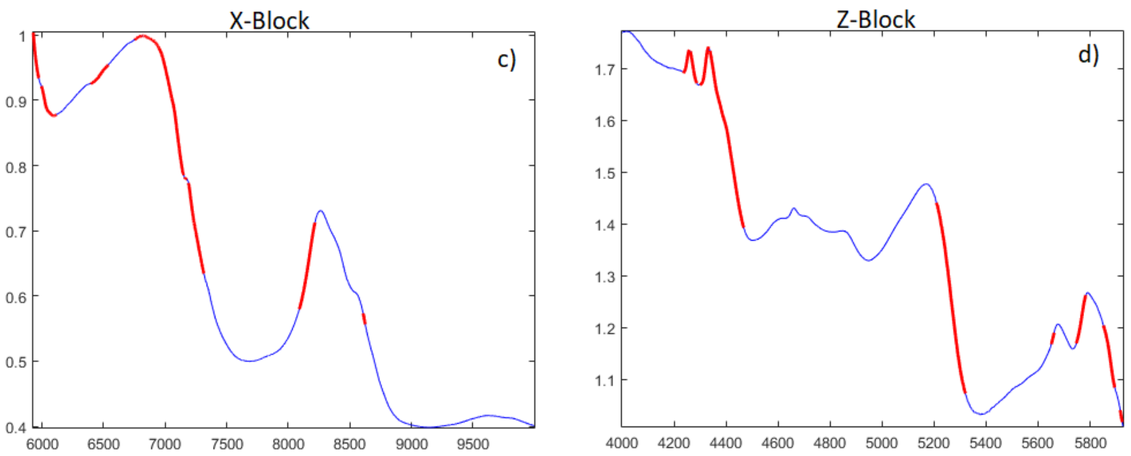
Z: 9 variables selected



VIP

X: 526 variables selected

Z: 210 variables selected



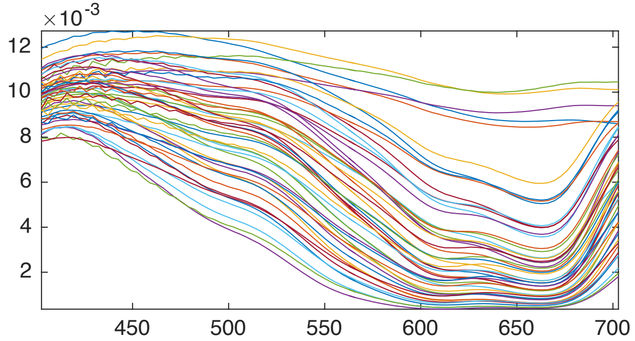
Second example: Polarisation spectroscopy

- Four complementary settings to measure the reflectance according to four polarisation states :
  - none, perpendicular, at 45 degrees, circular
- Measurement of visible spectra of a complete mixture design of:
  - 10 concentrations of Blue Methylene
  - 5 concentrations of pure scattering medium

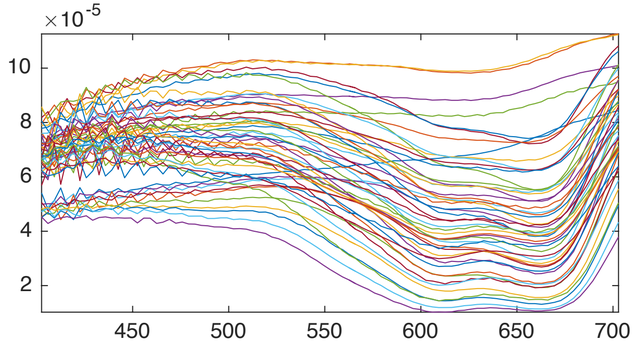
Second example:
Polarisation spectroscopy

DATA:

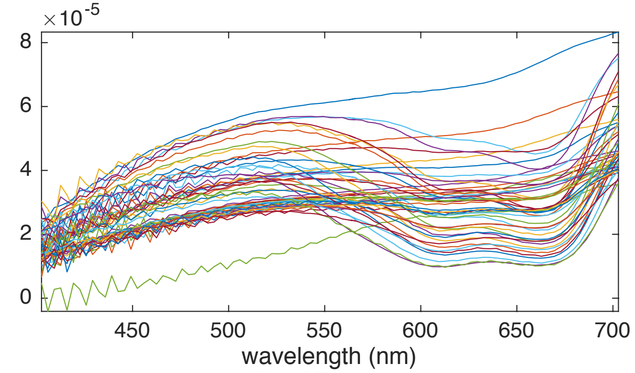
X1, X2, X3, X4 :
4 blocs of 50 spectra x 91 wavelengths



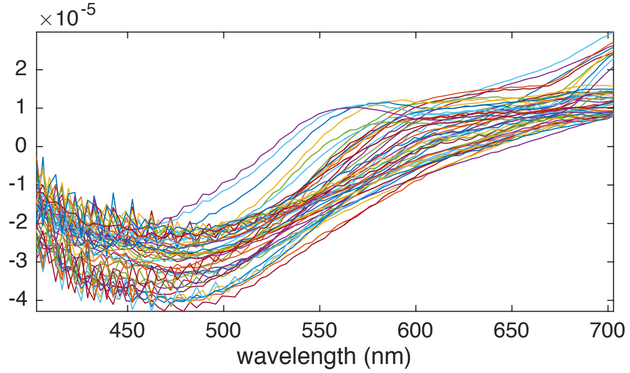
X1 : no pola



X2 : perpendicular pola

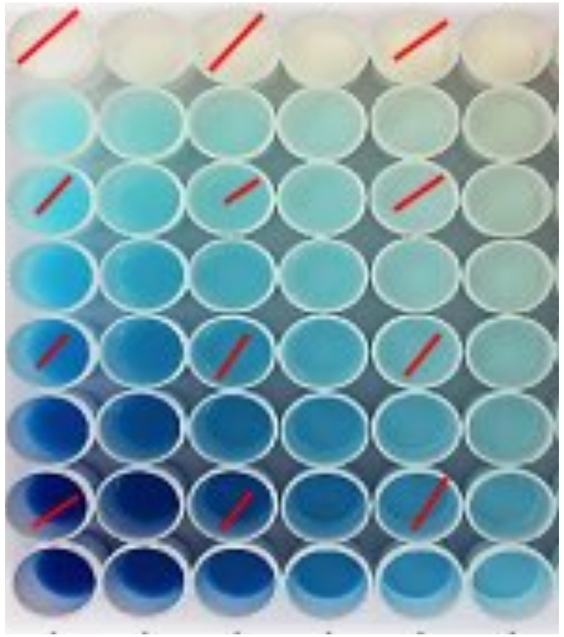


X3 : 45° pola



X4 : circular pola

Y: 1 bloc of 50 samples
x 2 concentrations Cabs, Cdif



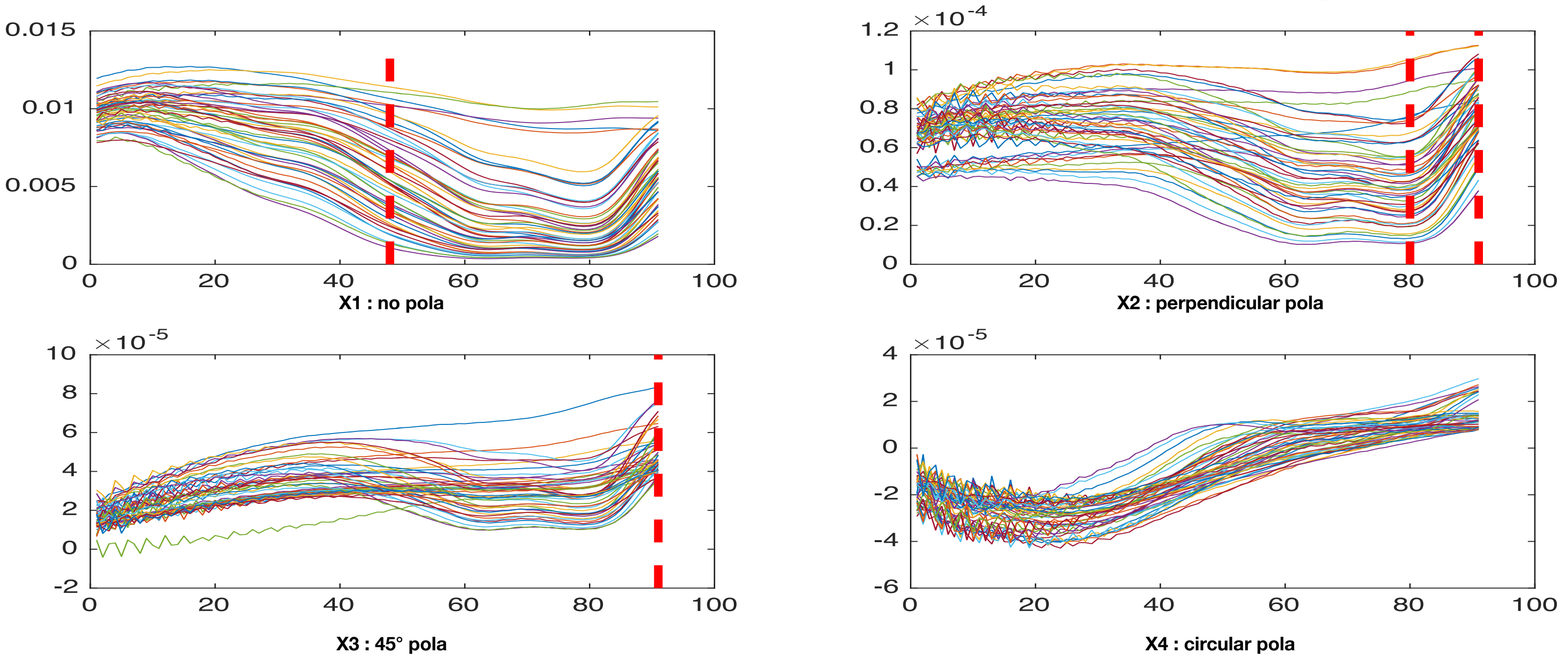
Second example: Polarisation spectroscopy

RESULTS

	Cabs		Cdif	
	SECV	R^2	SECV	R^2
PLS on global reflectance	20.7	0.71	0.92	0.54
SO-PLS on 4 blocks #LV = 6, 7, 2, 0	13.6	0.88	0.58	0.82
SO-CovSel on 4 blocks 7, 3, 2, 3	15.6	0.84	0.82	0.72
SO-CovSel on 4 blocks 1, 2, 1, 0	15.1	0.85	1.3	0.2

Cross Validation by blocks of identical Cabs

Second example: Polarisation spectroscopy



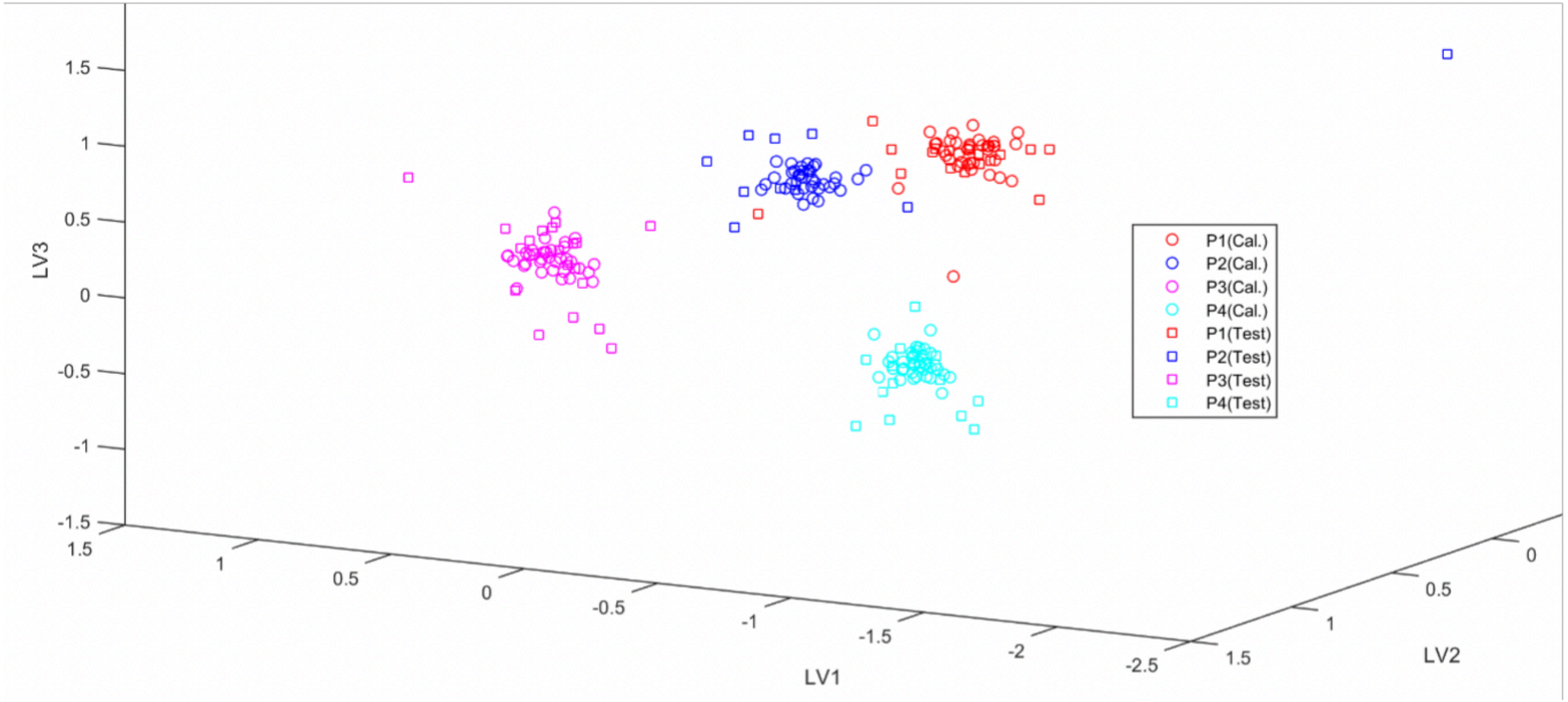
- Selected variables :
  - 559 nm on back scattering
  - 666 & 700 nm on perpendicular polarisation
  - 700 nm on 45° polarisation

Should be feasible to build a sensor with 3 wavelengths and 3 polarizers
Possible application: chlorophyll content in vegetation

Third example: Sensory analysis of chocolates

- 208 chocolates samples belonging to 4 sensory poles
- Measurement of global fingerprints by:
  - NIRS, PTRMS, SPME, 3D-Fluorescence, 8 organic acids by HPLC
- 144 samples for calibration
- 62 samples for validation

Third example:
Sensory analysis of chocolates



1 sample misclassified over 62

Third example: Sensory analysis of chocolates

Technique	#Sel	Selection
Vis-NIRS	3	408; 598; 1116 nm
PRTMS	20	41.054;43.055; 43.068; 44.995; 61.064; 68.055; 68.075; 69.086;82.984; 85.064; 89.084; 93.041; 101.094; 104.046112.150; 114.145; 115.111; 119.068; 125.058; 127.029
SPME	1	AClacet
3D-Fluo	0	
Acids	1	Citric

Conclusion

- SO-CovSel is an adaptation of SO-PLS multi-block PLS
- It is able to select non redundant variables in each block
- The selection is optimal in terms of covariance
- The selection is parsimonious
- SO-CovSel can be used on any type of blocks, for regression or discrimination

Acknowledgements

- Arnaud Ducanchez for the data on polarized spectroscopy

- Agropolis Fondation for supporting the Chaman project on chocolates and Zoé Deuscher and Jean-Luc Le Quéré from the Centre des Sciences du Goût et de l’Alimentation, Dijon, for the PTRMS data