CovSel

Variable Selection in highly multivariate and multi response cases

Application to NIR spectroscopy

JM Roger, B Palagos
E Fernandez and D Bertrand

jean-michel.roger@cemagref.fr

CovSel: Variable selection for highly multivariate and multi-response calibration: Application to IR spectroscopy
Chemometrics and Intelligent Laboratory Systems, Available online 21 October 2010
J.M. Roger, B. Palagos, D. Bertrand, E. Fernandez-Ahumada

Outline

- Introduction
- Theory
- Interpretation
- Implementation
- Examples
- Conclusion

Introduction

- For real industrial applications, multispectral sensors are designed :
  - INCA project aims at measuring outdoor the C contents in soils
  - SPECTRON is a portable sensor which monitors the maturity of grapes
  - TRI+ project develops waste sorting machines

Introduction

- The common way for determining the useful wavelengths :
  - To build a calibration database X, Y
  - To run a variable selection algorithm
- A lot of methods
  - Filters, Wrappers, Embedded

But no method for addressing the multi response case

Introduction

- CovSel is a new selection method, which explicitly addresses these specifications :
  - To yield meaningful variables
  - To avoid the correlation between the selected variables
  - To manage multi response cases

Theory

- Let **X** be a $n \times p$ matrix of predictor
- Let **Y** be a $n \times q$ matrix of responses

} centered

- Covsel principle :
  1. Select the variable $\mathbf{x}_i$ which :
     - carries variance
     - is close to **Y**
  2. Project **X** and **Y** orthogonally to $\mathbf{x}_i$
  3. GOTO 1

Theory

- What does "carries variance and is close to **Y**" mean ?
- For single response :
  - maximizes its absolute covariance with **y**

    $$i = \text{Argmax}(\text{cov}(\mathbf{x}_i, \mathbf{y})^2)$$

  - maximizes the norm of its projection onto **y**

    $$i = \text{Argmax}((\mathbf{x}_i^T \mathbf{y})^2) = \text{Argmax}(\mathbf{x}_i^T \mathbf{y}\, \mathbf{y}^T \mathbf{x}_i)$$

Theory

- For multiple responses ?
  - maximizes its projection onto **Y**

    $$i = \text{Argmax}(\mathbf{x}_i^T \mathbf{Y}\, \mathbf{Y}^T \mathbf{x}_i)$$

  - is the closest to **Yv**, for any **v**, $\mathbf{v}^2=1$

    $$i = \text{Argmax}(\text{Max}(\text{cov}(\mathbf{x}_i, \mathbf{Yv})^2)_{\mathbf{v}^2=1})$$

The two propositions are equivalent

Interpretation

Let $\mathbf{s}^i$ be the i^{th} vector of R^p basis (0 … 1 … 0)

1 … i … p

	PLS	CovSel
	$j=1$	$j=1$
1	$u_j=\text{ArgMax}_{\mathbf{u}}(\text{Max}_{\mathbf{v}}(\text{cov}(\mathbf{Xu},\mathbf{Yv})^2))_{u^2;v^2=1}$	$I_j=\text{ArgMax}_k(\text{Max}_{\mathbf{v}}(\text{cov}(\mathbf{Xs}^k,\mathbf{Yv})^2))_{v^2=1}$
2	$\mathbf{z}=\mathbf{Xu}_j$	$\mathbf{z}=\mathbf{Xs}^{I_j}=\mathbf{x}_{I_j}$
3	$\mathbf{X}\leftarrow(\mathbf{I}-\mathbf{z}(\mathbf{z}^T\mathbf{z})^{-1}\mathbf{z}^T)\,\mathbf{X}$	$\mathbf{X}\leftarrow(\mathbf{I}-\mathbf{z}(\mathbf{z}^T\mathbf{z})^{-1}\mathbf{z}^T)\,\mathbf{X}$
4	$\mathbf{Y}\leftarrow(\mathbf{I}-\mathbf{z}(\mathbf{z}^T\mathbf{z})^{-1}\mathbf{z}^T)\,\mathbf{Y}$	$\mathbf{Y}\leftarrow(\mathbf{I}-\mathbf{z}(\mathbf{z}^T\mathbf{z})^{-1}\mathbf{z}^T)\,\mathbf{Y}$
5	GOTO 2	GOTO 2
6		

CovSel is a particular case of PLS,
where the latent variables are constrained to be canonical vectors ($\mathbf{s}^i$)

Implementation

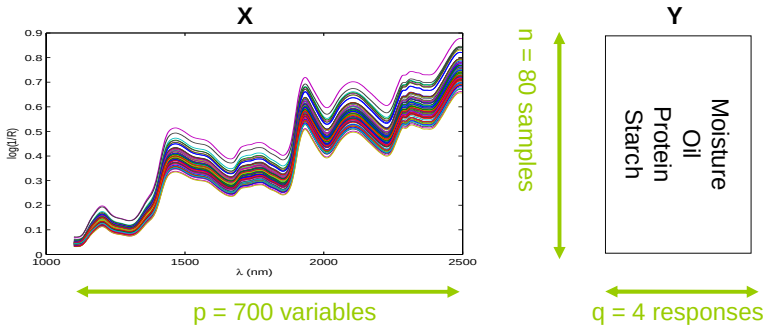
- In the case of regression
  1. Run CovSel on *k* steps between **X** and **Y**
     - Yields a global selection, for all responses
     - Watch to the explained variance
  2. Run CovSel between the k variables and each response
     - Yields specific sub-selections for each response
     - The optimization can rely on cross validation

Implementation

- In the case of discrimination
  1. Build **Y** with the membership degrees
     - $\mathbf{y}_i$=[0 0 … 1 … 0 0]
  2. Run CovSel on *k* steps between **X** and **Y**
     - Yields a global selection
  3. Run LDA (e.g.) on 1, 2, …, *k* variables
     - Examine the cross validation error
     - Watch to the explained variance

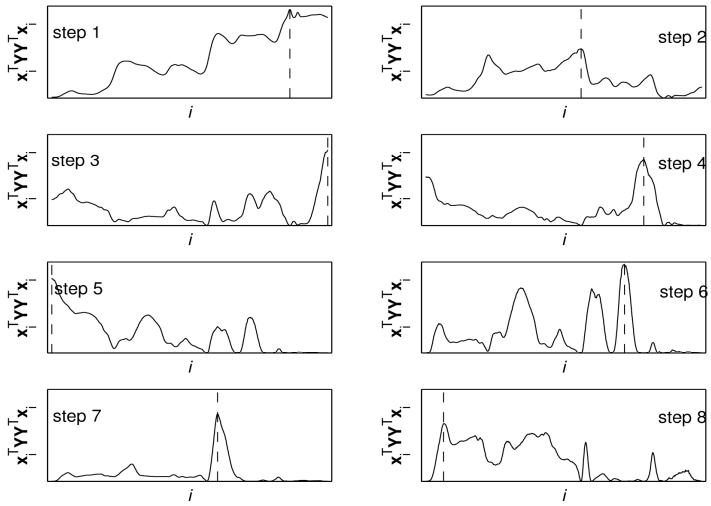
Example 1: Corn

- Data from Eigenvector web site :
  - http://software.eigenvector.com/Data/Corn/index.html

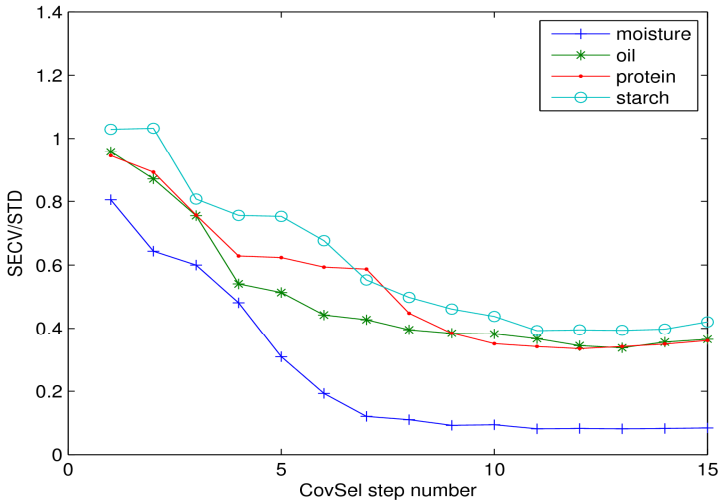


2/3 in the learning set, 1/3 in the test set

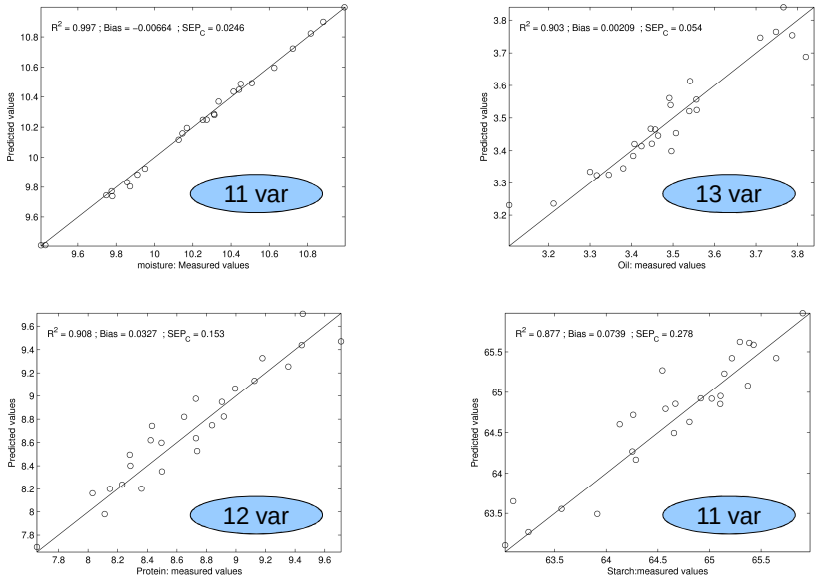
Example 1: Corn



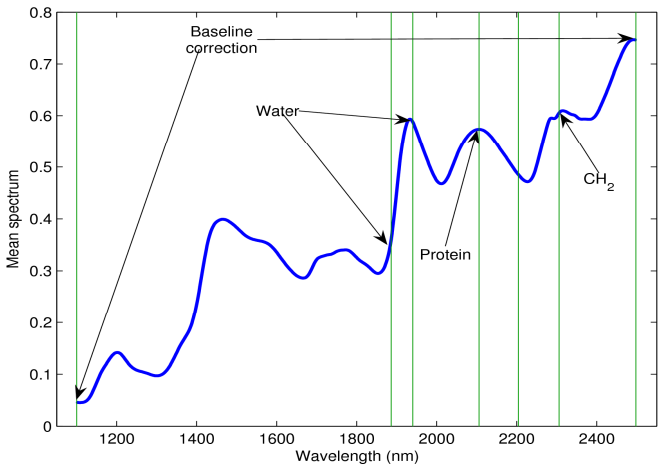
Example 1: Corn



Example 1: Corn



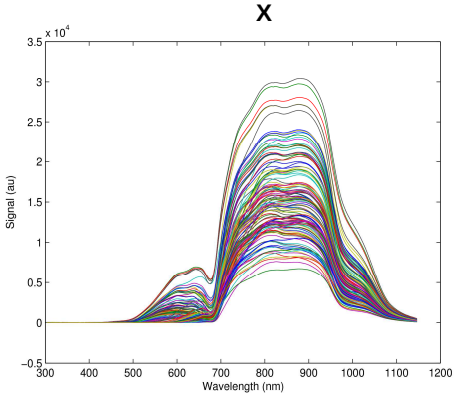
Example 1: Corn



Example 2: Grape variety

Vis/VNIR spectra of wine grape berries (Zeiss MMS1 spectrometer)

X



n = 250 samples

p = 256 variables

Y

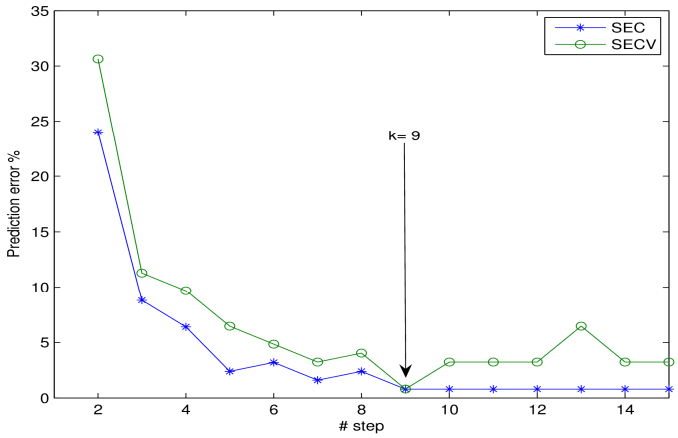
1 0 0
1 0 0
0 1 0
0 0 1
…

q = 3 classes

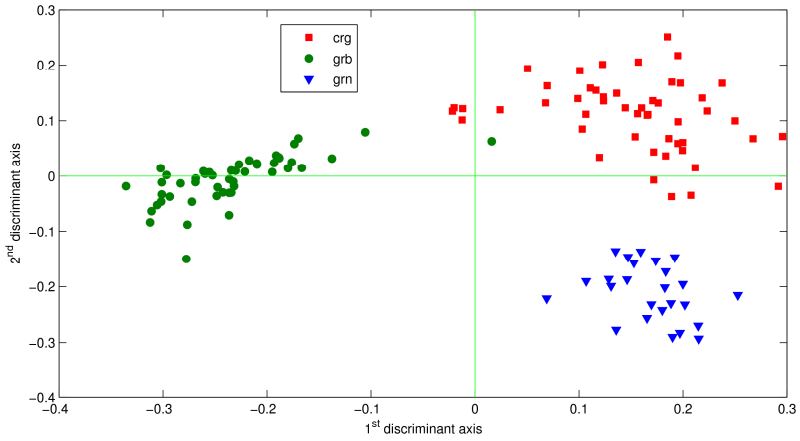
crg : *carignan*
grb : *grenache blanc*
grn : *grenache noir*

½ for the calibration ; ½ for the test

Example 2: Grape variety

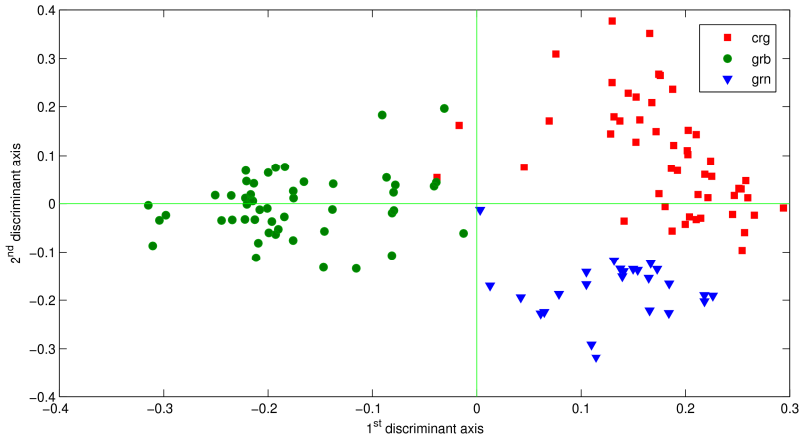


Example 2: Grape variety



Calibration Error : 0.8%

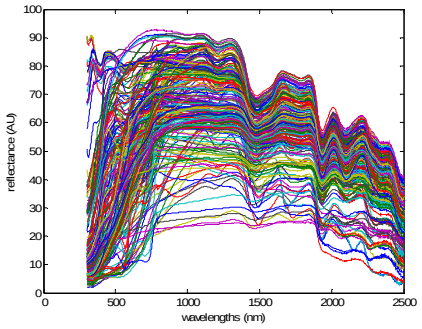
Example 2: Grape variety



Prediction Error : 6.4%

Example 3: Waste sorting

X **Y**



n = 230 samples

1 0 0
1 0 0
0 1 0
0 0 1
…

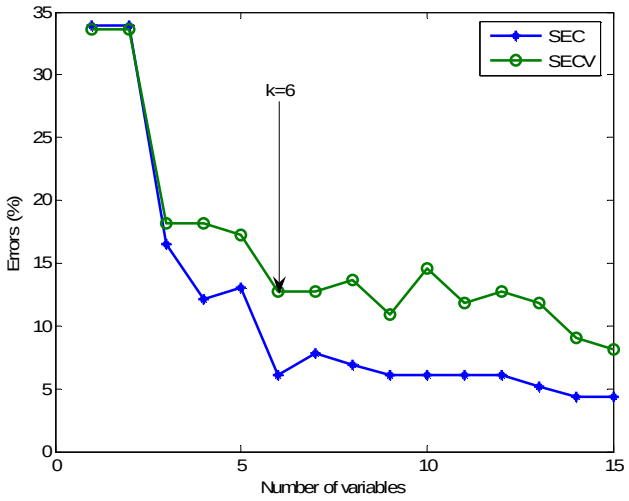
p = 1101 variables

q = 3 classes

½ for the calibration ; ½ for the test

A : *wood A*
B : *wood B*
P : *paper*

Example 3: Waste sorting



Example 3: Waste sorting

	Wood A	**Wood B**	**Paper**
Wood A	10	2	4
Wood B	1	23	0
Paper	0	0	75

Confusion matrix on the test set : 6% of errors

Conclusion

- CovSel is a new method that:
  - implements a PLS-like variable selection
  - handles multiple responses
  - can be applied on discrimination problems
  - produces well separated selections
  - is very little time consuming

Thanks for your attention