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

Outline

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

Introduction

- Variable selection for multivariate calibration :
  - For extracting meaningful features
  - For designing multispectral devices
- A lot of methods
  - Filters, Wrappers, Embedded

But none addresses explicitly the multi response case

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 "<u>carries variance and is close to **Y**</u>" 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**

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

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

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

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

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

Implementation

- In the regression case :

Run CovSel on *k* steps between **X** and **Y**

- Yields a global selection, for all responses
- Watch to the explained variance

Run CovSel between the k variables and each response

- Yields specific sub-selections for each response
- The optimization can rely on cross validation

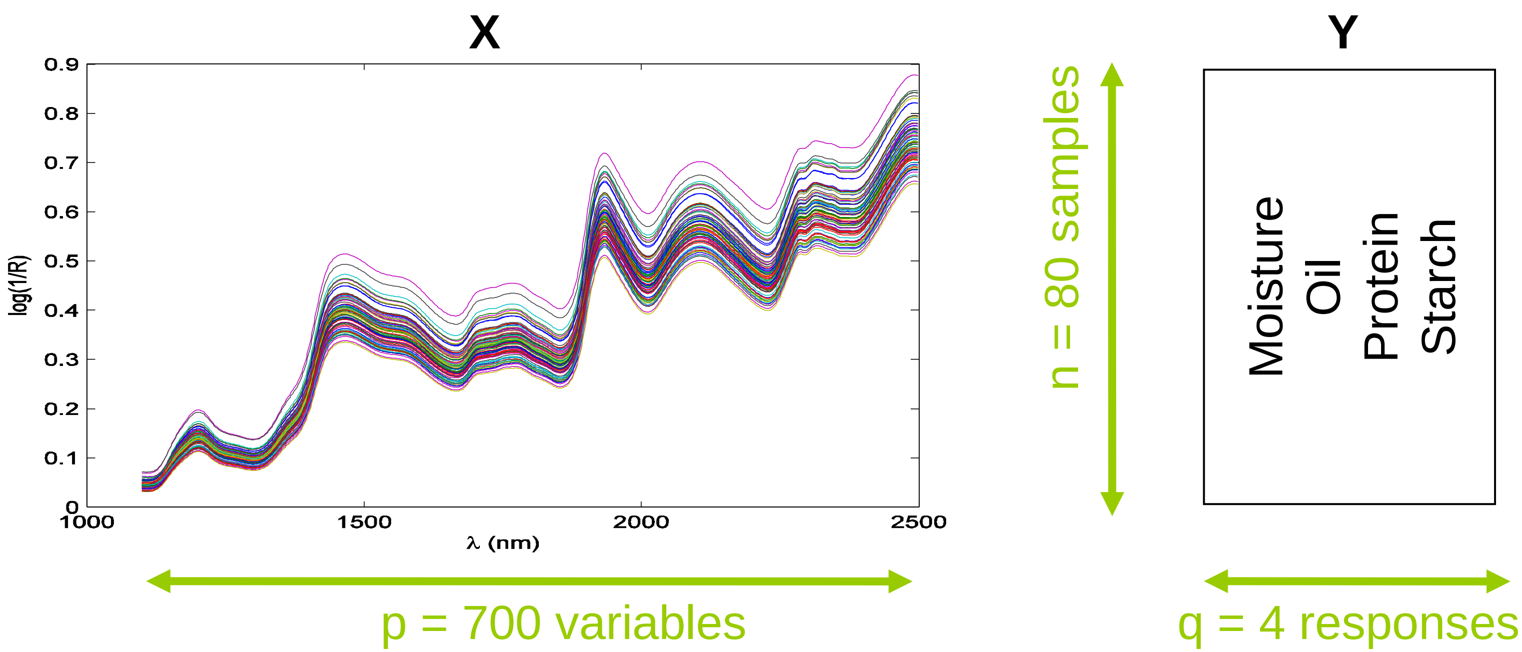
Implementation

- In the discrimination case :

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 on 1, 2, …, *k* variables
   - Examine the cross validation error
   - Watch to the explained variance

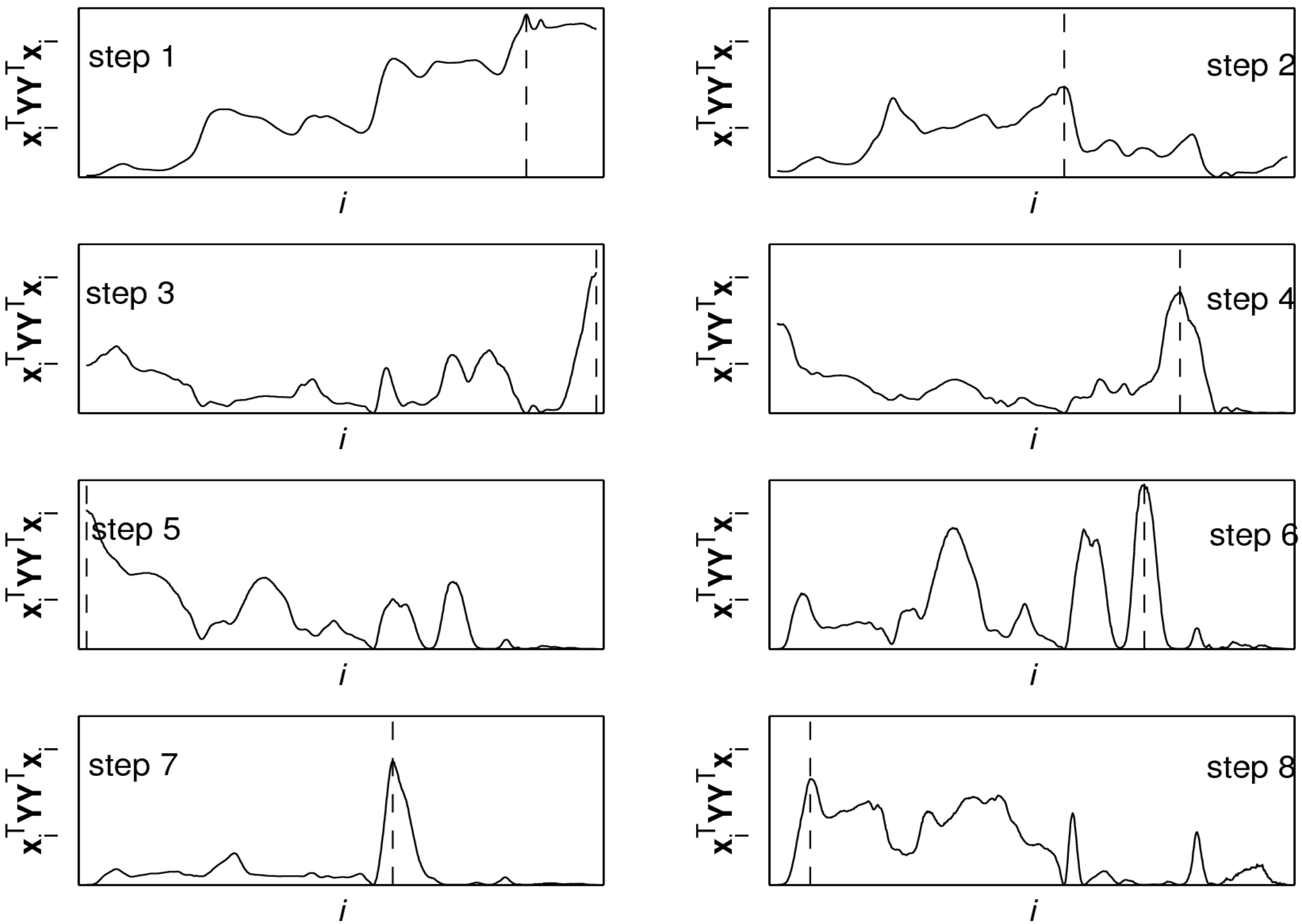
Example 1: Corn

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

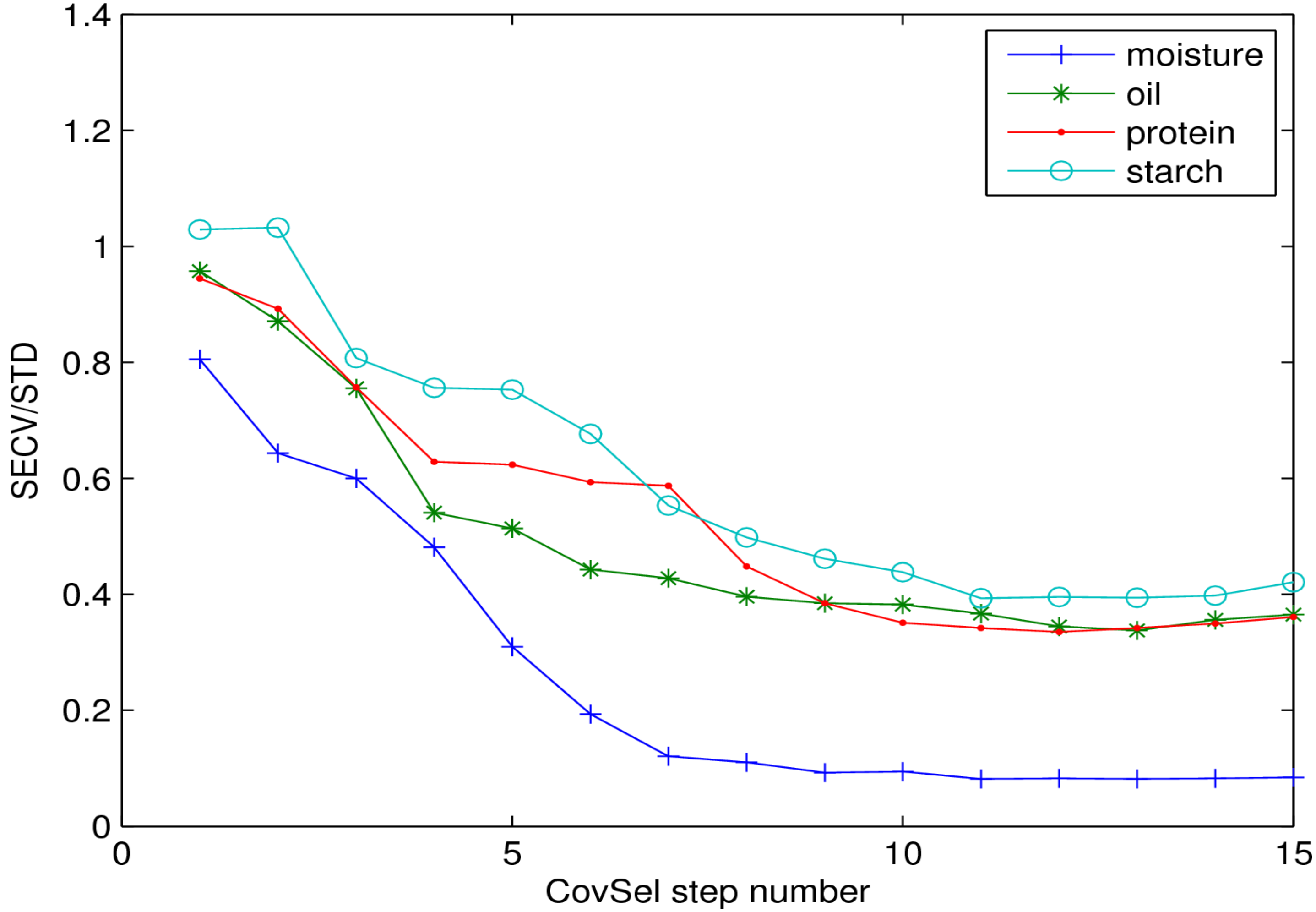


2/3 in the learning set, 1/3 in the test set

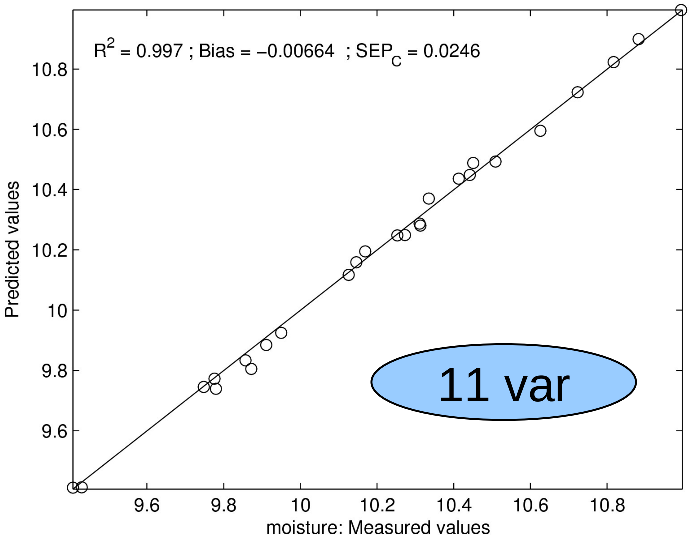
Example 1: Corn

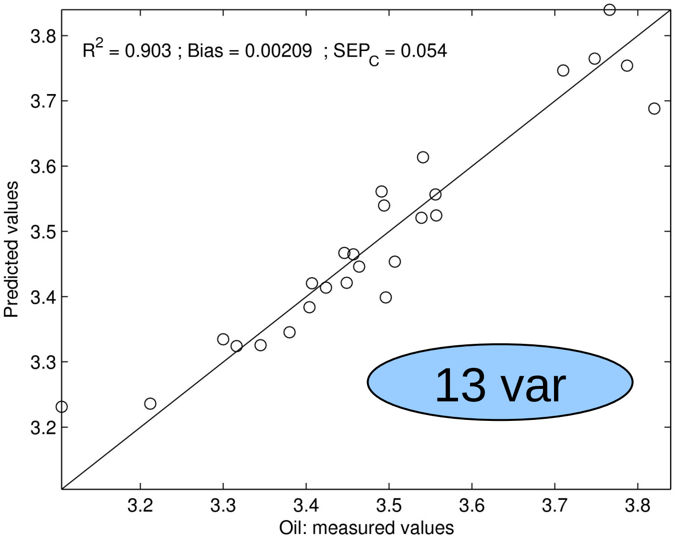


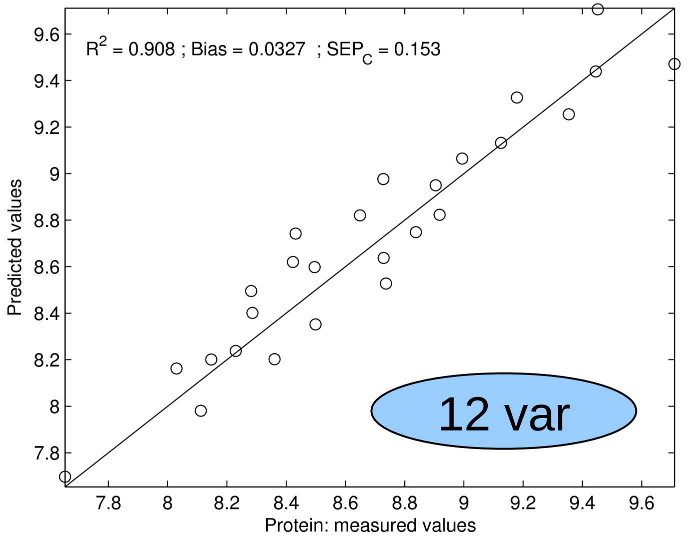
Example 1: Corn

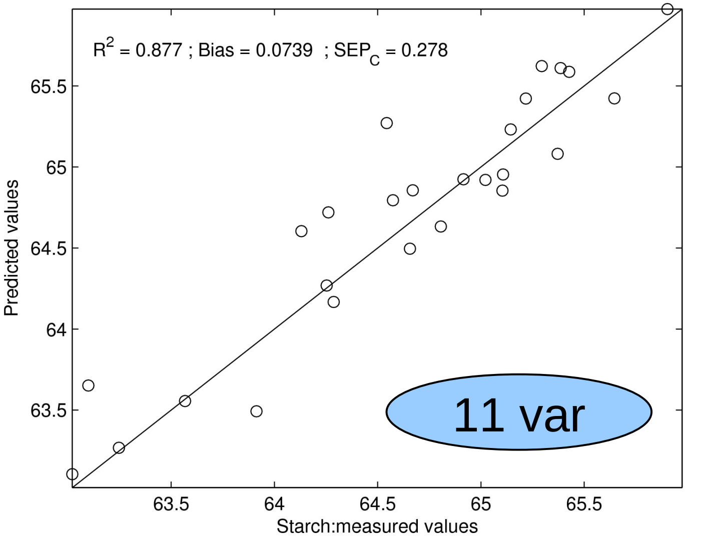


Example 1: Corn

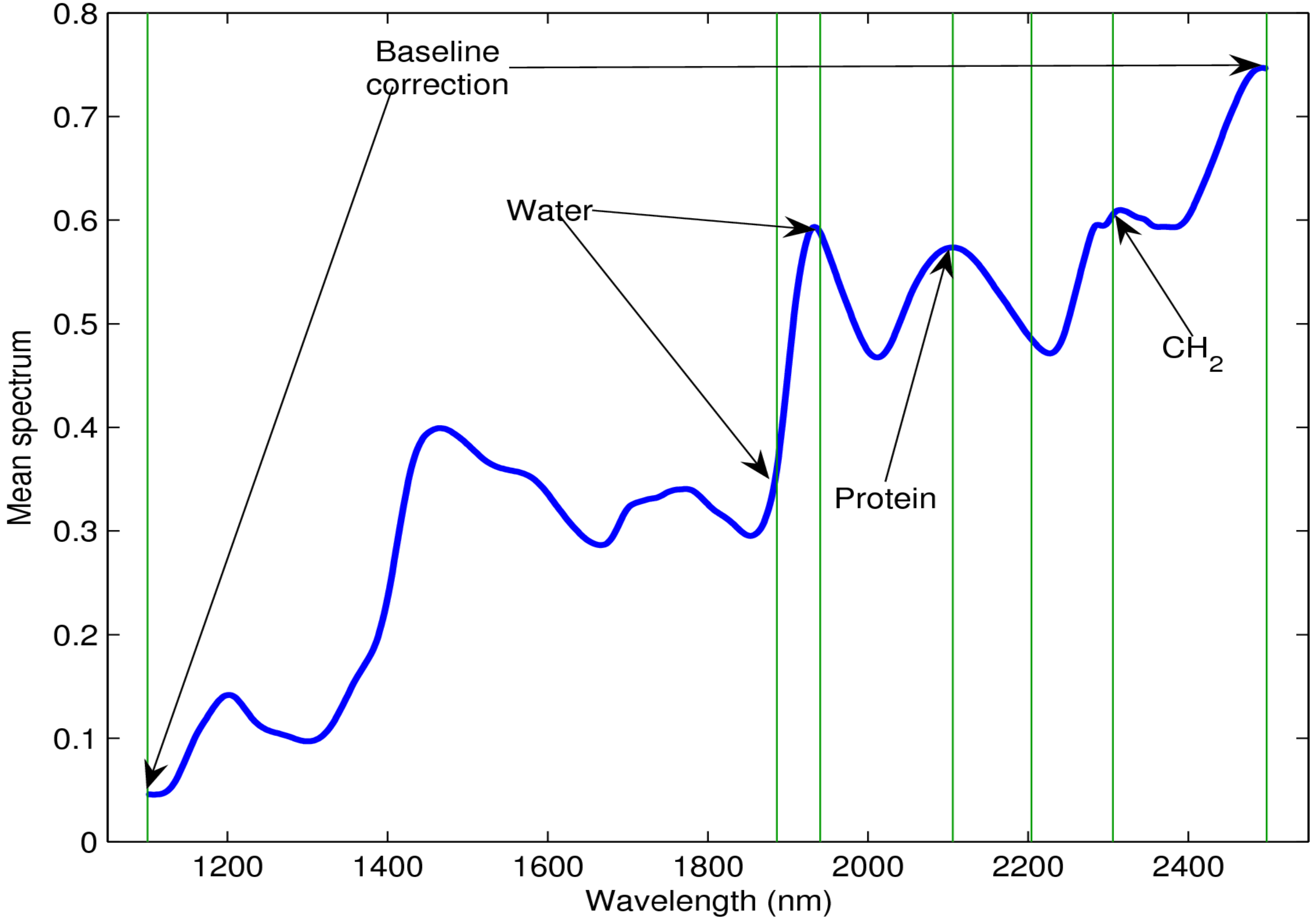








Example 1: Corn



Example 2 : Apricots

X : MIR spectra of apricots (*n*=731 x *p*=292)

y : brix degree of apricots

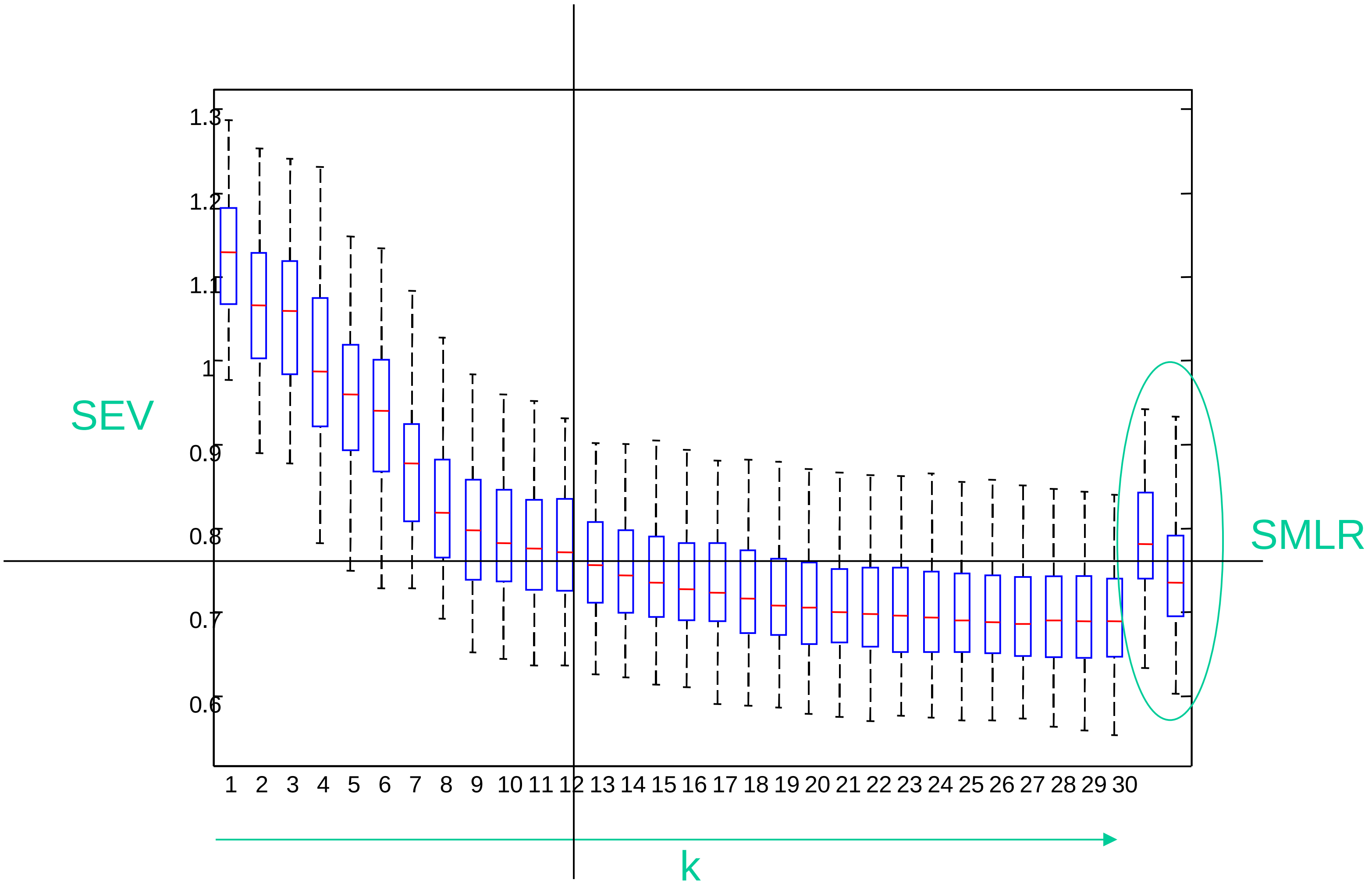
2/3 for calibration

1/3 for validation

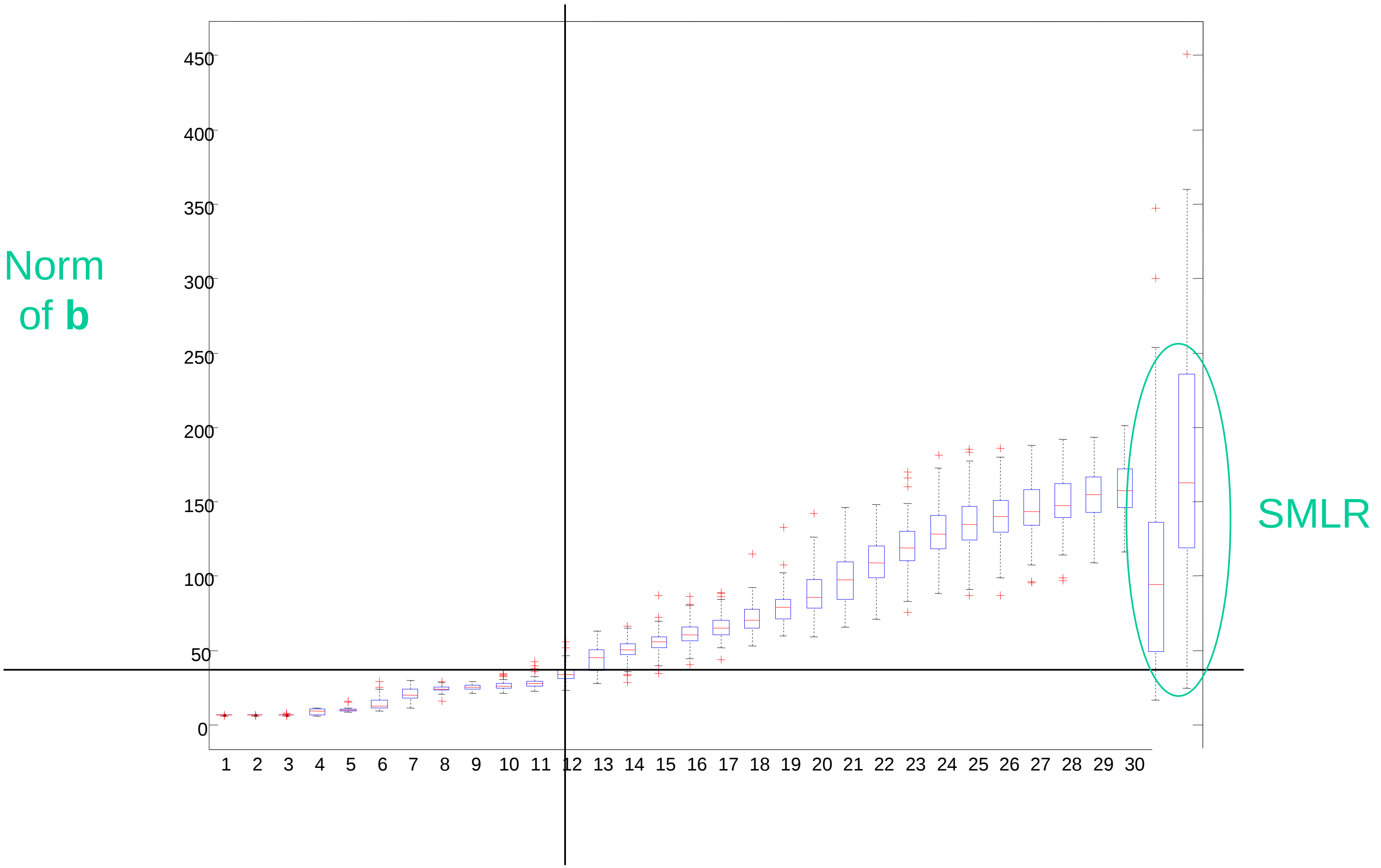
} Repeated 100 times randomly

Comparison with stepwise MLR

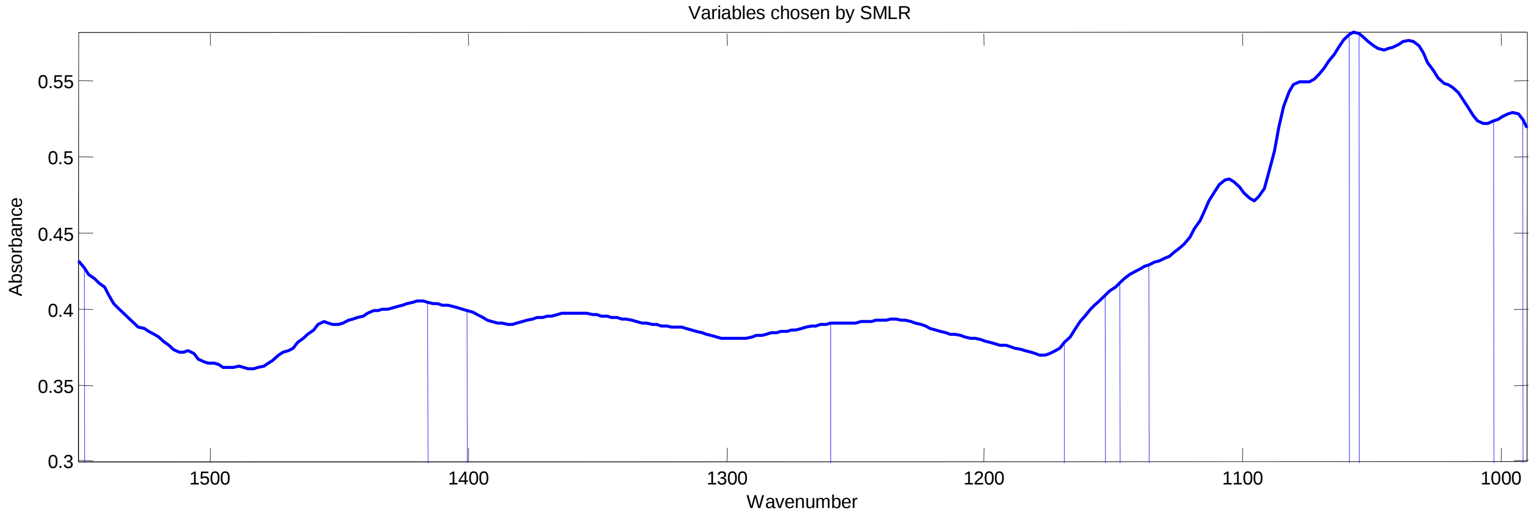
Example 2 : Apricots

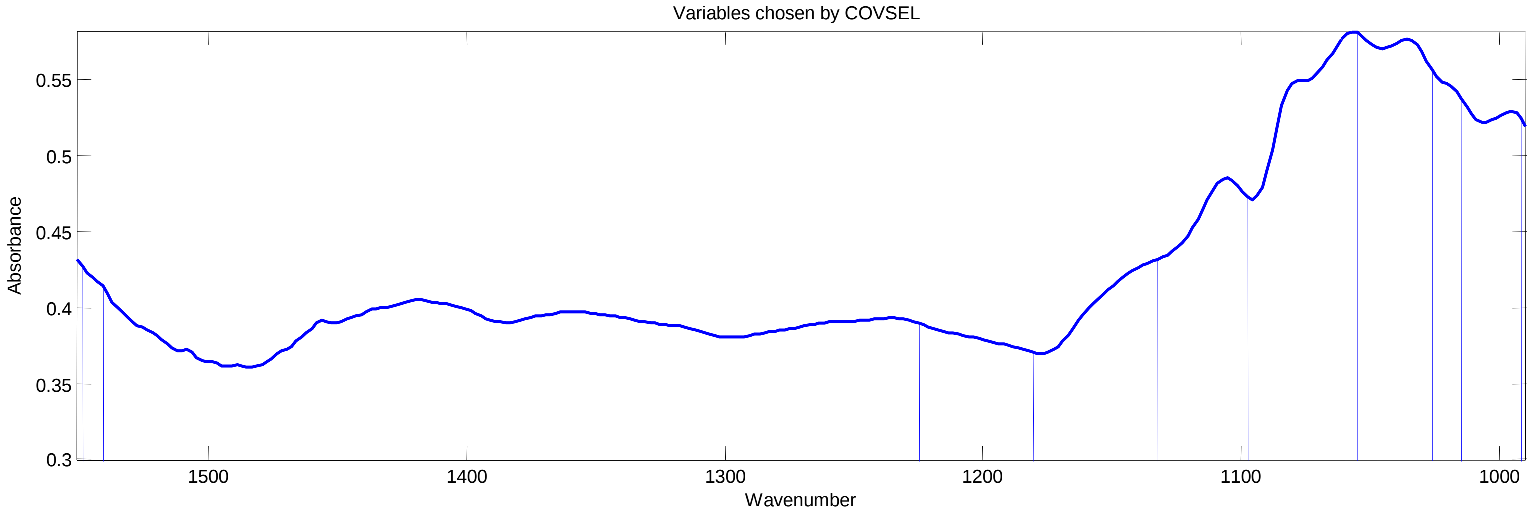


Example 2 : Apricots



Example 2 : Apricots

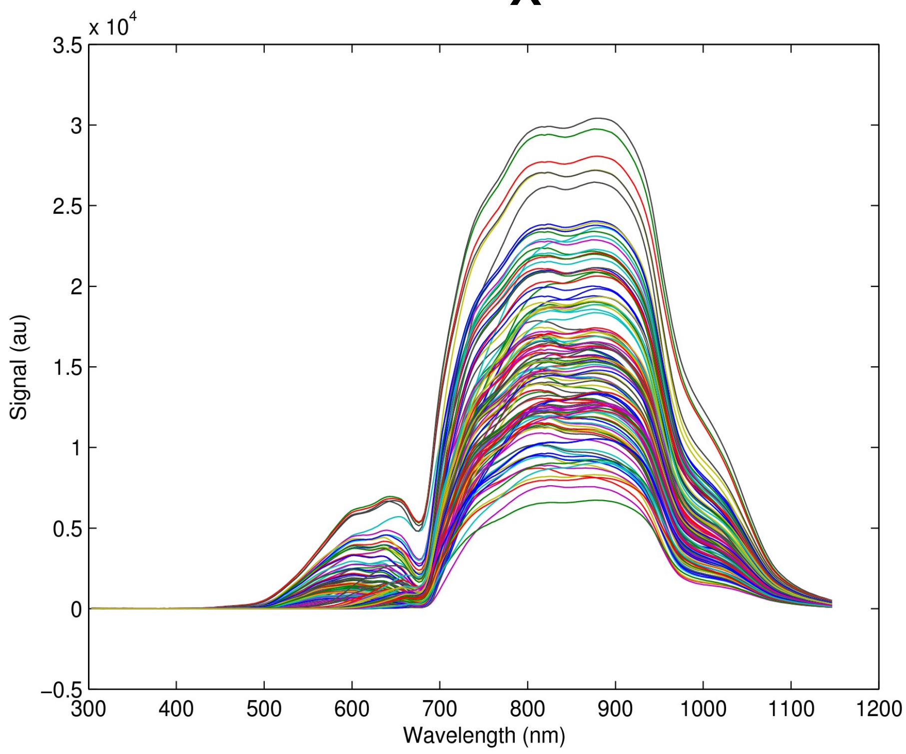




Example 3: Grape variety

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

X



p = 256 variables

n = 250 samples

Y

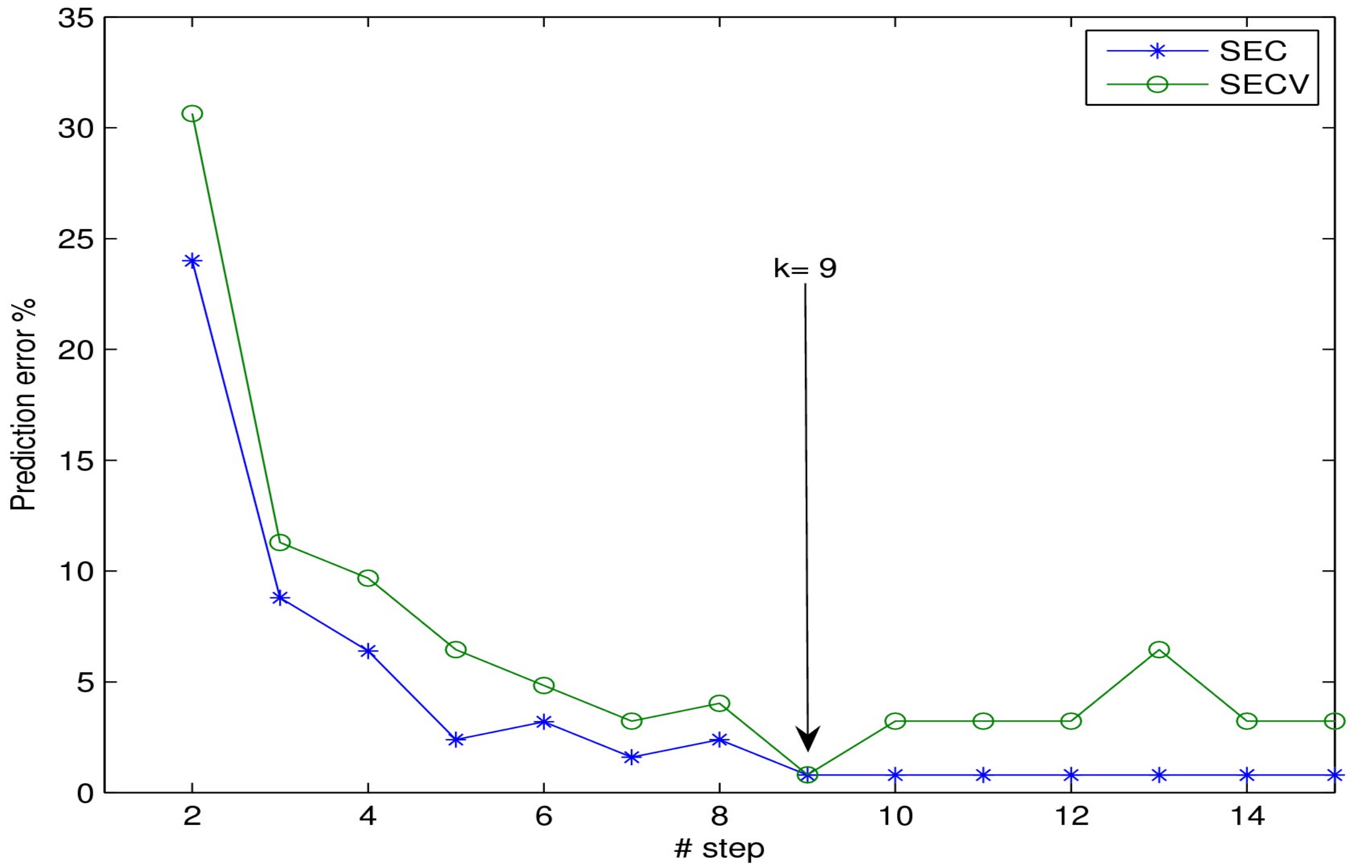
$$\begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ & \cdots & \end{pmatrix}$$

q = 3 classes

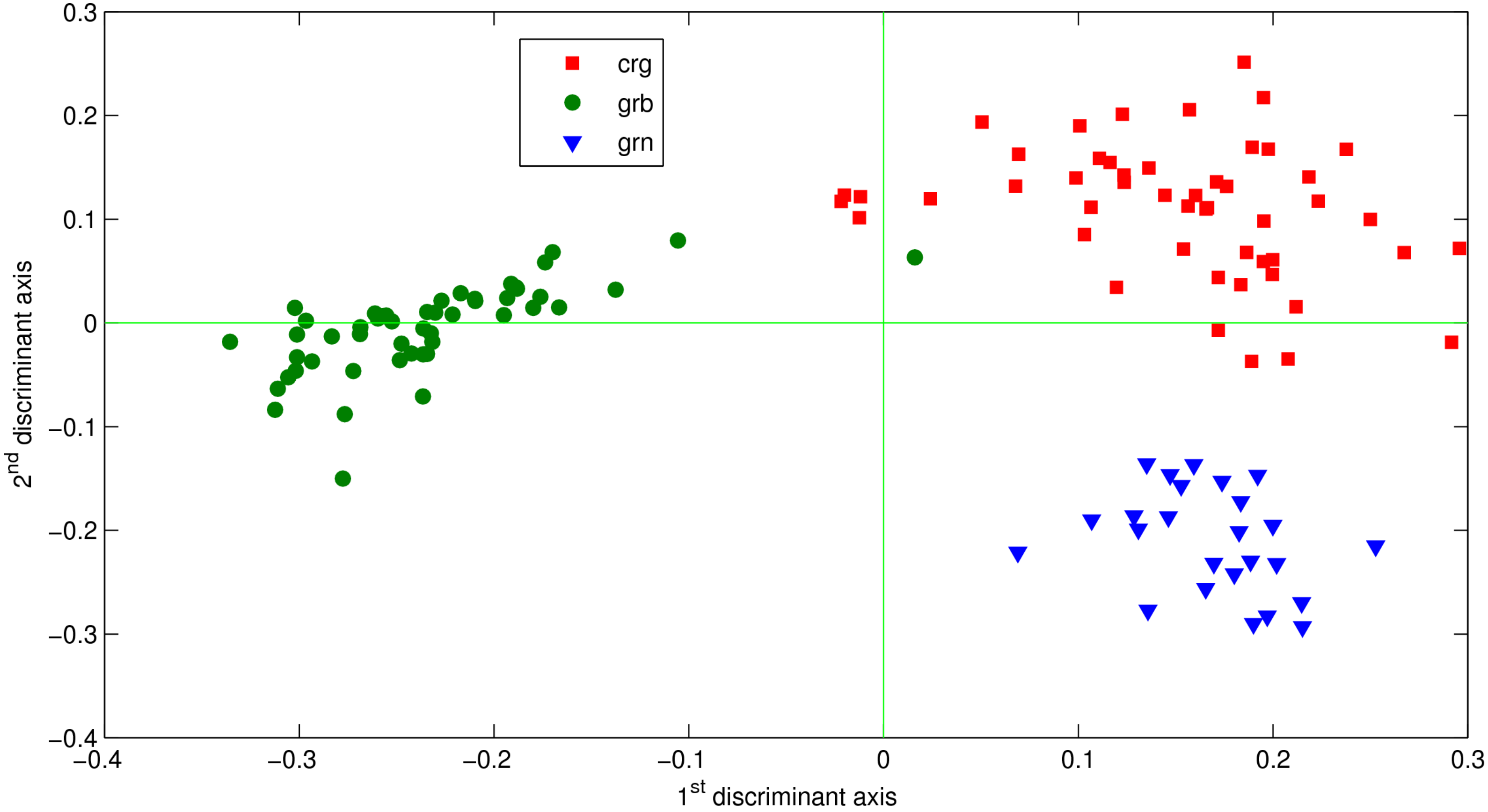
crg : *carignan*
grb : *grenache blanc*
grn : *grenache noir*

½ for the calibration ; ½ for the test

Example 3: Grape variety

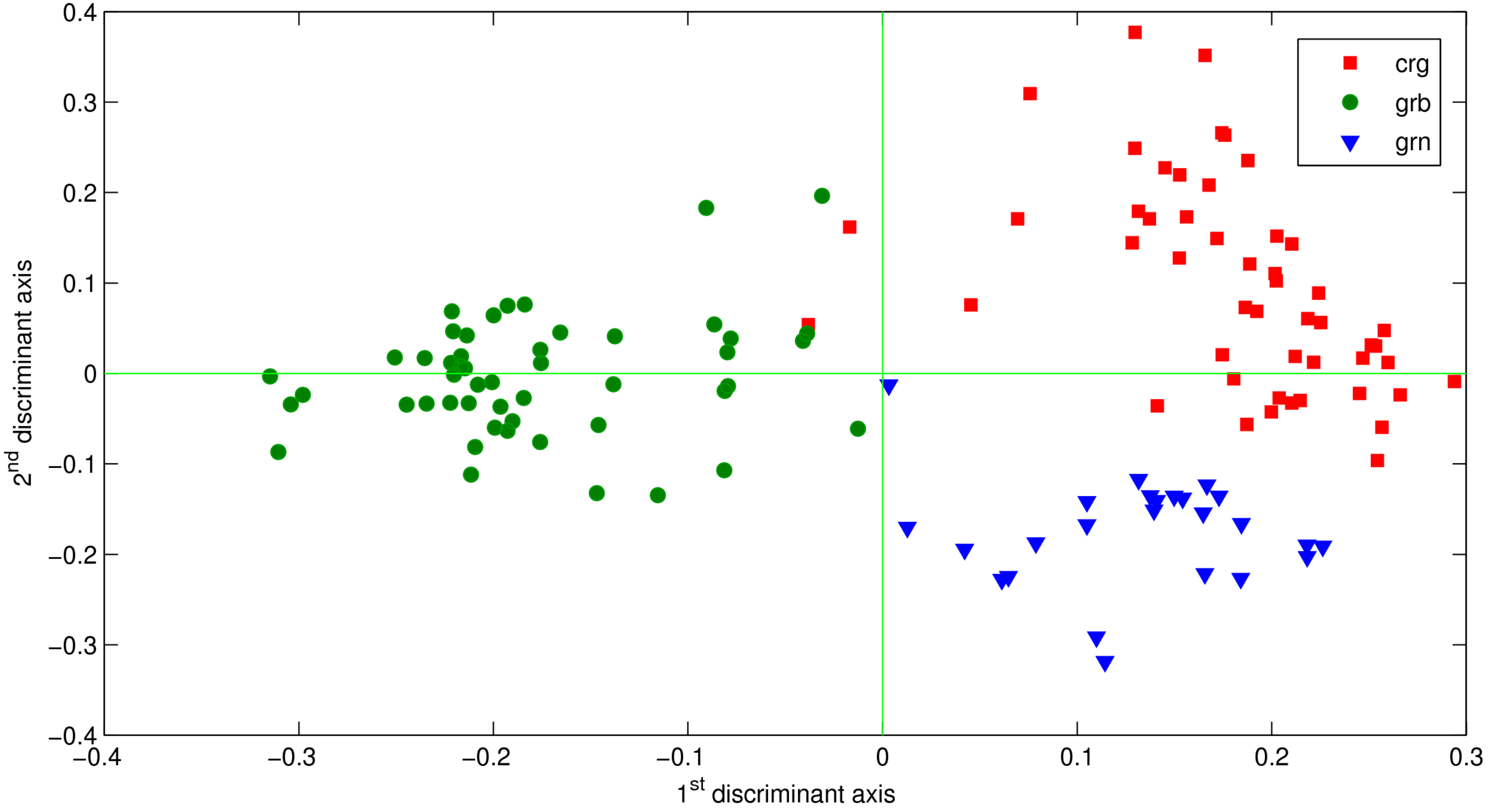


Example 3: Grape variety



Calibration Error : 0.8%

Example 3: Grape variety



Prediction Error : 6.4%

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