CovSel, Variable selection for multivariate and multi-response calibration.

J.M. Roger, B. Palagos, E. Fernandez, D. Bertrand

CovSel : Variable selection for highly multivariate and multi-response calibration

J.M. ROGER[1] B. PALAGOS[1] D. BERTRAND[2] E. FERNANDEZ-AHUMADA[1]

[1]**ITAP Mixed research Unit** (Cemagref SupAgro), BP 5095 - 34033 Montpellier Cedex 1 France
[2]**UR1268 BIA** (INRA), BP 71627 - 44316 Nantes Cedex 3 France

I theory

1. algorithm

1. Searching index I_1 corresponding to the predictor *closest* to the responses, by :

$$I_1 = \text{ArgMax}_i \left(\mathbf{x}_i^T \mathbf{Y}\mathbf{Y}^T \mathbf{x}_i\right) \quad (1)$$

2. Collinear information is removed from $\mathbf{X}$ and $\mathbf{Y}$ by orthogonal projection :

$$\mathbf{X} \leftarrow \mathbf{P}^{\perp}_{\mathbf{x}_{I_1}} \mathbf{X} \quad (2)$$

$$\mathbf{Y} \leftarrow \mathbf{P}^{\perp}_{\mathbf{x}_{I_1}} \mathbf{Y} \quad (3)$$

This process is then repeated for I_2, I_3, $\cdots$, I_k.

2. properties

Equation 1 can be written as :

$$I_1 = \text{ArgMax} \left(\text{diag}\left(\mathbf{X}^T \mathbf{Y}\mathbf{Y}^T \mathbf{X}\right)\right) \quad (4)$$

Furthermore it can be demonstrated that this equation is equivalent to :

$$I_1 = \text{ArgMax}_i \left(\text{Max}_{\mathbf{v},\mathbf{v}^2=1} \left(\text{cov}\left(\mathbf{x}_i, \mathbf{Y}\mathbf{v}\right)^2\right)\right) \quad (5)$$

CovSel is a PLS like selection process

II material and methods

– **Set Corn** (http ://software.eigenvector.com/Data/Corn) :
  – $\mathbf{X}$ = 80 NIR spectra of corn samples × 700 wavelengths (1100 to 2498 nm).
  – $\mathbf{Y}$ = 80 reference values of 4 responses : moisture, oil, protein and starch

A calibration and a validation sets were randomly drawn in the proportion of 2/3 and 1/3, respectively. CovSel was applied on the calibration set, with a predefined number of variables $k = 15$. Four models were then optimized individually on each response and applied to the validation set.

– **Set Wine Grapes** :
  – $\mathbf{X}$ = 250 Vis/VNIR spectra of wine grain samples × 256 wavelengths (310 to 1100 nm).
  – $\mathbf{Y}$ = 250 membership degrees to 3 varieties : *carignan* (crg), *grenache blanc* (grb) and *grenache noir* (grn) ; e.g. $\mathbf{y} = [0, 1, 0]$ for a sample belonging to class 2.

A calibration and a validation sets were randomly and equally drawn. The variables selected by CovSel were used as input of a Linear Discriminant Analysis. The observation of the leave-one-out cross-validation results allowed the determination of the optimal number of selected variables. The discriminant model calibrated on this subset was applied on the test set.

III results on corn dataset

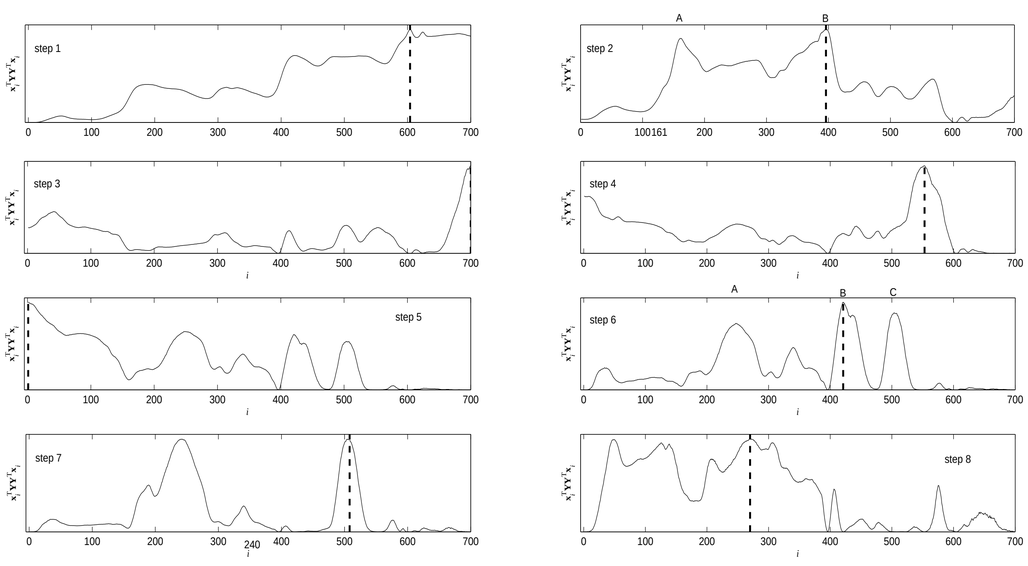
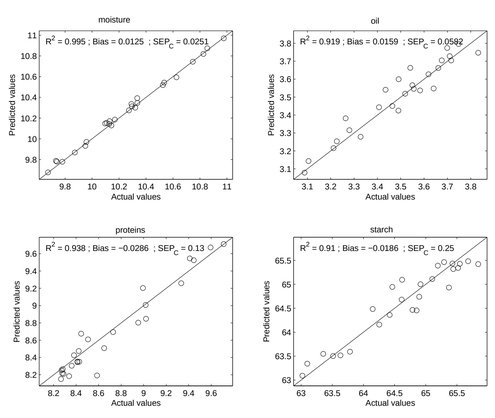


Illustration of the iterative erosion of $\mathbf{X}^T\mathbf{Y}\mathbf{Y}^T\mathbf{X}$



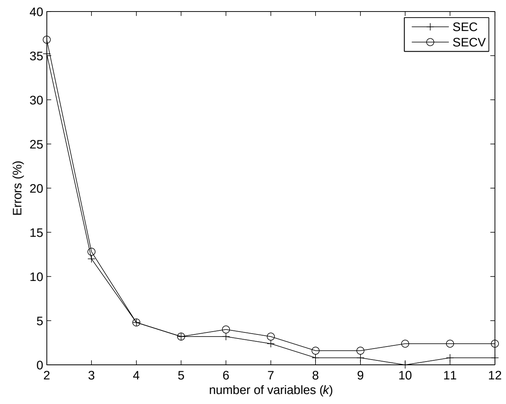
Test results on the four responses

λ (nm)	moisture	oil	protein	starch	assignement
1100	×	×	×	×	baseline
1190		×	×		oil
1306		×		×	
1428	×		×	×	starch
1500		×			NH
1592	×	×	×	×	
1718	×	×	×	×	oil
1886	×	×	×	×	
1940	×	×	×	×	water
2106	×		×	×	starch
2204	×	×	×	×	
2250	×	×	×		starch
2306	×	×	×	×	oil
2388		×		×	
2498	×	×	×	×	baseline

Spectral assignment of the selected variables

CovSel yields parcimonious, meaningful and little correlated selections

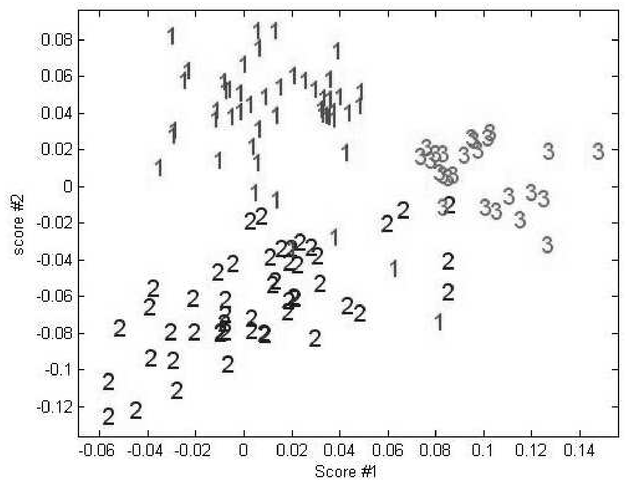
IV results on wine discrimination



Evolution of calibration and cross-validation errors of the LDA model, as a function of the number of CovSel steps. $k = 8$ variables were retained.

$\widehat{\mathbf{Y}}^T\mathbf{Y}$	crg	grb	grn
crg	43	-	-
grb	4	46	-
grn	3	4	25
PE = 8.8 %			

Confusion matrix of the LDA model when applied on the test set



Scores of the LDA calculated on the selected variables.

CovSel is efficient on discrimination problems

V conclusion

– CovSel is a variable selection method well suited to highly multivariate and multi-response calibration

– CovSel acts as the PLS, splitting the covariance between $\mathbf{X}$ and $\mathbf{Y}$

– Covsel can be used on discrimination problems

– CovSel is particularly adapted to the design of multispectral devices

XII CHEMOMETRICS IN ANALYTICAL CHEMISTRY
18-21 October 2010
Antwerp, Belgium