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

J.M. Roger,[1]* B. Palagos,[1] D. Bertrand,[2]

[1]Cemagref-ITAP, Montpellier, France, jean-michel.roger@cemagref.fr ; [2]INRA, Nantes, France

[Introduction] Variable selection is of major interest for NIR calibration, either as a feature selection or for the design of multi-wavelength devices. Some dedicated methods have been developed in chemometrics, but none of them addresses the case of multi response calibration. The present paper reports a new method devoted to this task and presents two applications on spectrometry.

[Theory] Variable selection for NIR spectroscopy must face two problems : (1) the huge number of variables yield a very large solution space ; (2) the variables are highly correlated, and if no precautions are taken the model built on the selection may be inconsistent. CovSel tackles these two problems: (1) by selecting variables step by step on the basis of their global covariance with all the responses ; (2) by projecting the data orthogonally to the selected variable. Thus, one step of CovSel is :

(1) Select the variable i which maximizes the quantity $\mathbf{x}_i^T\mathbf{Y}\mathbf{Y}^T\mathbf{x}_i$, with **X** spectra and **Y** response matrices

(2) Replace **X** by $(\mathbf{I}-\mathbf{x}_i(\mathbf{x}_i^T\mathbf{x}_i)^{-1}\mathbf{x}_i^T)\,\mathbf{X}$ and **Y** by $(\mathbf{I}-\mathbf{x}_i(\mathbf{x}_i^T\mathbf{x}_i)^{-1}\mathbf{x}_i^T)\,\mathbf{Y}$ with **I** the identity matrix

A predefined number of rounds are performed. At each one, some figures of merit are calculated, like the cumulated variance of **X** and **Y** captured by the selection, or the cross validation error. The best selection can then be adopted regarding the evolution of these criteria.

[Materials and Methods] CovSel was applied on two problems. The first one concerns the calibration of 4 responses : moisture, oil, protein and starch contents in corn with regards to NIR spectra. The data set (available at http://software.eigenvector.com) contained 80 samples. The wavelength range was 1100-2498 nm over 700 variables (2 nm step). Ten variables were selected by CovSel and introduced in a multi-linear regression, one by one. The best selection was determined on the basis of leave one out cross validation. The second application concerned the discrimination of 3 wine grape varieties. A set of 300 Vis/VNIR spectra was acquired with a ZEISS MMS1 spectrometer (256 wavelengths from 310 to 1100 nm ; 3,27 nm step). On one half of the set, CovSel was run against a matrix of 3 responses, encoding the membership of each sample ([1 0 0] for class 1, etc.). The best selection was determined on the basis of the explained variance, and entered in a discriminant analysis. The resulting model was applied on the other half of the set.

[Results and Discussion] Table 1 shows the performance of the cross-validation for the corn example, which is quite satisfactory. Figure 1 shows the selected variables on the mean corn spectrum. It is noticeable that the selected variables are far each from the others, i.e. not too much correlated, giving some guaranty of a well-conditioned model. Table 2 gives the confusion matrix for the second application and Figure 2 shows the 3 classes in the discriminant space calculated on the 5 variables selected by CovSel. It appears that CovSel is an efficient method for selecting variables for addressing the multiple response cases, including the discrimination one.

Table 1. Results for the corn set (10 variables)

Response	RPD
Moisture	2,7
Oil	3,3
Protein	3,3
Starch	2,8

Table 2. Discrimination of the 3 grape varieties (5 var)

		Actual		
		C1	C2	C3
Predicted	C1	44	5	0
	C2	6	44	0
	C3	0	1	25

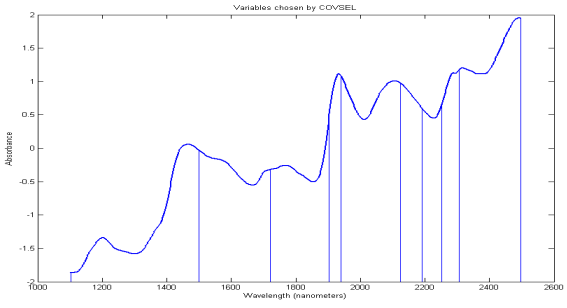

Figure 1. Selection for corn set

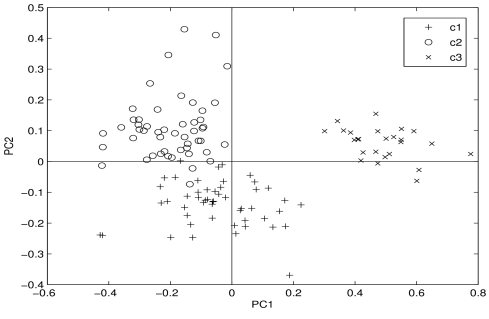

Figure 2. Discrimination of the 3 grape classes