CovSel: Variable selection for highly multivariate and multi-response calibration. Application to IR spectroscopy

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

CovSel: Variable selection for highly multivariate and multi-response calibration. Application to IR spectroscopy.

J.M. ROGER*, B. PALAGOS*, D. BERTRAND** E. FERNANDEZ-AHUMADA*

*Cemagref BP 5095 - 34033 Montpellier Cedex1 France

**INRA BP 71627 - 44316 Nantes cedex 3 France

Abstract

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 . Variable selection for NIR spectroscopy must face two problems: (1) the huge number of variables yields a very large solution space; (2) variables are highly correlated, and if no special attention is paid the model built on the selection may be . This article presents a new method, CovSel, which tackles these two problems by following this procedure: (1) Variable selection step by step on the basis of their global covariance with all the responses; (2) Projection of the data orthogonally to the selected variable. CovSel was applied on three problems: the first one concerns a single response MIR calibration (Brix degree content in apricot), the second one concerns a multi-response NIR calibration (4 main constituents in corn)

and the last application concerns the NIR discrimination of 3 wine grape varieties.

Key words: Variable selection, Orthogonal projection

1 Introduction

Analytical chemistry and process monitoring involve more and more multivariate indirect sensors, such as spectrometers. For example, Near InfraRed (NIR) spectrometry is a powerful analytical tool, increasingly used in industry ([1],[2]). However these sensors require a calibration aiming at finding a relation between the measured spectra and the response to be estimated. A common practice involves collecting a sample set with spectral and response value information. If the response values are quantitative (e.g. concentrations) the usual method of calibration consists in a regression of the reference data on the spectral data. In the case of a qualitative response (e.g. an origin), tools for discrimination are used. For both model types, classical statistical methods are not efficient since the space carrying the useful information is much smaller than that of the spectra. Consequently a classical solution consists in using factorial methods. For quantitative responses, partial least squares regression (PLS) is the more commonly used method ([3]). In the case of qualitative responses, a similar procedure can be applied on binary variables (indicator variables) indicating the belonging of an observation to a given qualitative group. PLS can then be applied on these indicator variables, making it possible to carry out a discriminant analysis based on latent variables (PLS-DA) ([4]). Another solution involves choosing a restricted number of significant variables and then applying an ordinary least square (OLS) linear regression or a linear discriminant analysis (LDA). Moreover, numerous applications require

the conception of simplified instruments where only few variables are used. This is the case of spectrometry devoted to agricultural applications where practical specifications often impose conceiving robust and cheap filter instruments. All these reasons make variable selection an appropriate chemometric issue. Nevertheless, the nature of the data, i.e. NIR spectra, poses some particular problems because, on one side, variables are highly correlated and on the other side, the searching space is huge (if p is the number of variables there are 2^p-1 solutions). A supplementary problem occurs when a multi-response calibration is involved. The present paper addresses these problems.

There are numerous techniques of variable selection. In the context of PLS regression, a review can be found in ([5]). In the general domain of machine learning, the following taxonomy in three groups is commonly used ([6]):

- With *filter* methods, variable selection is done independently of the model that eventually makes use of them. Filter methods use the intrinsic characteristics of the whole data set in order to select some variables and/or eliminate others. This selection can be viewed as a pre-treatment of predictive variables. In the field of multivariate calibration, different filter criteria are used such as the absolute value of correlation or covariance between predictors and response ([7]). The theory of information is also used for selecting the predictive variables that maximise the mutual information with the variable to be predicted. However this method is difficult to implement when multi-responses are involved. An application in chemomectrics is found in ([8]). The UVE method ([9]) allows variable elimination by comparing them with noisy artificial variables.
- *Wrapper* methods scan the space of possible selections and use the prediction model as a black box to test the relevancy of selections. This is

often evaluated by means of a simple or cross validation. Depending on the strategies to perform the scan, there exist different wrapper methods (see [10], for a review). These are in most cases stochastic optimisation methods inspired by natural phenomena: Genetic algorithms ([11]) or simulated annealing ([12]). These methods are not repeatable due to their random nature. Moreover, their complex algorithms may pose a problem when the searching space is large and the relevancy of the selection is not easy to assess in the case of multiple responses.

- *Embedded* methods accomplish the variable selection during the calibration process. The subset of selected variables, optimising the training criterion, can be constructed by successive additions (forward), elimination (backward) or a combination of both approaches. Backward methods are not well adapted to the high multivariate cases because, at the beginning of the selection process, they take into account all the variables. Stepwise multiple linear regression (SMLR) ([13], pp 307-313) is one of the most popular examples of this kind of methods.

Successive Projection Algorithm (SPA, [14]) is a forward selection method that minimises colinearity between predictors by means of successive projections on interlinked sub-spaces. At each step, the selected variable is the one showing the maximum projection on the orthogonal sub-space generated by the already selected variables. SPA is a hybrid between filter and embedded methods. This paper proposes a new method of variable selection called CovSel (Covariance selection). It can be considered a hybrid method as SPA, from which it takes inspiration. CovSel is well adapted to multi-response calibration of spectrometers and can be applied to the problem of discrimination considering indicator variables as responses.

2 Theory

This section presents the theoretical aspects of CovSel and emphasizes its similarity with the construction of latent variables in PLS. Implementations for regression and discrimination will be successively presented.

Upper case bold characters will be used for matrices, e.g. $\mathbf{X}$ will denote a matrix of n individuals (lines) by p variables (columns); lower case bold characters for column vectors, e.g. $\mathbf{x}$ will denote a simple individual (a spectrum); non-bold characters will be used for scalars, e.g. matrix elements x_{ij} or indices i. $\mathbf{I}_n$ will denote the identity matrix of $\mathbb{R}^n$. If $\mathbf{U}$ is a $(n \times k)$ matrix of rank k, $\mathbf{P_U}$ will represent the matrix of the projector on $\mathbf{U}$ in $\mathbb{R}^n$: $\mathbf{P_U} = \mathbf{U}(\mathbf{U}^{\mathrm{T}}\mathbf{U})^{-1}\mathbf{U}^{\mathrm{T}}$ and $\mathbf{P}_{\mathbf{U}}^{\perp}$ the matrix of the projector orthogonal to $\mathbf{U}$: $\mathbf{P}_{\mathbf{U}}^{\perp} = \mathbf{I}_n - \mathbf{P_U}$. The symbol $\mathbf{s}^i$ will denote a column vector containing null values, except the i^{th}, which is unitary: $s_j^i = 0$ for $i \neq j$ and $s_i^i = 1$.

Let $\mathbf{X}$ be a matrix of n objects described by p descriptors and $\mathbf{Y}$ a matrix of the same n objects described by q responses to be predicted. CovSel aims at classifying the k most useful variables of $\mathbf{X}$ in decreasing order of their interest. The procedure includes two main steps: (i) selecting the most useful variable, (ii) projecting the data orthogonally to this selected variable. In the same way as the Gram-Schmidt decomposition ([13], p 277) or as the SPA selection, CovSel approximates the $\mathbf{X}$ row space $\mathbb{R}^n$ as a sum of complementary subspaces. The difference with SPA lies in that CovSel carries out the variable selection on the basis of their global covariance with all the responses.

2.1 Algorithm

CovSel method performs variable selection by iterating the following two steps:

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

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

(2) :

$$\mathbf{X} \leftarrow \mathbf{P}^{\perp}_{\mathbf{x}_{I_1}} \mathbf{X} \tag{2}$$

$$\mathbf{Y} \leftarrow \mathbf{P}^{\perp}_{\mathbf{x}_{I_1}} \mathbf{Y} \tag{3}$$

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

2.2 Interpretation

Equation 1 can be written as:

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

Furthermore it can be demonstrated (Cf. annexes) that this equation is equivalent to:

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

Equation 4 is close to that of PLS where the first latent variable is given by the first eigenvector of: $\mathbf{X}^{\mathrm{T}} \mathbf{Y}\mathbf{Y}^{\mathrm{T}} \mathbf{X}$ ([15]). Equation 5 reminds the general objective of PLS as it is expressed in the basic algorithms such as NIPALS ([3]). To reach this objective, PLS allows any linear combinations of the columns of $\mathbf{X}$. CovSel aims at performing a similar optimisation, but by allowing only

linear combinations of the columns of $\mathbf{X}$ in the form [0,0,...,1,...0], since its role is the selection of variables. At last, as for the PLS algorithm, orthogonal projections accomplished by equations 2 and 3 ensure that variances of $\mathbf{X}$ and $\mathbf{Y}$ are captured in a cumulative way by every step of the algorithm. Therefore CovSel implements a PLS-like variable selection, as shown in table 1.

2.3 Implementation

The implementation of CovSel differs according to the objective of the user. Three cases are addressed here:

- **Data analysis**: Running CovSel between $\mathbf{X}$ and $\mathbf{Y}$ without any modelling phase makes it possible to identify the variables of $\mathbf{X}$ which explain $\mathbf{Y}$ at the most. This analysis will exploit the evolution of the variances explained by the successive steps of CovSel.
- **Regression**: If $\mathbf{Y}$ consists in continuous responses, like concentrations, CovSel could be used in a hierarchical process: (i) a first variable selection is made on the basis of all responses and (ii) this global selection is refined for each individual response in a second step.
- **Discrimination**: If $\mathbf{Y}$ contains the indicator variables, CovSel could use this multi-response for selecting variables prior to a LDA.

2.4 Evolution of variances explained by CovSel

In every iteration, during stages 4 and 5 as represented in table 1, the algorithm of CovSel erodes a part of the variance contained in $\mathbf{X}$ and $\mathbf{Y}$. Let $V_x(k)$ and $V_y(k)$ be the sum of these variances, according to k, expressed in percentage

of the whole variances of $\mathbf{X}$ and $\mathbf{Y}$. Curves V_x and V_y as a function of the iteration step are compulsorily increasing. Their shapes depend on the data configuration. If the rank of $\mathbf{X}$ is p and all variables of $\mathbf{X}$ are independent, $V_x(k)$ evolves linearly up to 100% for $k = p$, as illustrated on the two graphs on the left of figure 1. If $\mathbf{X}$ variables are correlated, the shape is different. The covariance maximized by CovSel is a compromise between $\mathbf{X}$ variance, $\mathbf{Y}$ variance and their correlation. For two variables with the same correlation with $\mathbf{Y}$, the one with the highest covariance will be chosen. Therefore curve V_x will show a convex shape, as illustrated on the two graphs of the right of figure 1. The shape of V_y thus depends on the relation between $\mathbf{X}$ and $\mathbf{Y}$. If, on one extreme, $\mathbf{Y}$ variables are orthogonal to $\mathbf{X}$, since the $\mathbf{Y}$ variance captured in every step is void, V_y is horizontal whereas V_x increases rapidly. On the other extreme, if the q variables of $\mathbf{Y}$ are completely determined by m variables of $\mathbf{X}$, V_y adopts a regular growing behaviour to attain 100% for $k = m$. Between these extreme situations, V_y should present a first step of fast increase, corresponding to the most important variables to be selected and then a step of slow increase, as illustrated on the bottom graphs of figure 1.

2.5 Regression case

If there is no technical interest in reducing the number of selected variables or if there is only one response, CovSel may be performed individually on each column of $\mathbf{Y}$, as in any classical selection method. However, CovSel addresses advantageously the other cases, where a unique common selection must be found to multiple responses. Let's assume that k is the maximal desired number of variables. The complete model building then relies on two steps:

- CovSel is first run on the centred $\mathbf{X}$ matrix and the autoscaled $\mathbf{Y}$ matrix, with a limit of k steps. This yields a selection $\{I_1, I_2, \cdots, I_k\}$.
- Secondly, CovSel is run between the submatrix $[\mathbf{x}_{I_1}, \mathbf{x}_{I_2}, \cdots, \mathbf{x}_{I_k}]$ centred and the columns $\mathbf{y}_i$ of $\mathbf{Y}$ also centred, for $i = 1, \cdots, q$.

This process gives q ordered choices of the same list of k variables, which can then be introduced stepwise in q classical mono-response OLS models. A cross validation of these q x k models produces q curves of SEC and q curves of $SECV$ which can guide the user to the choice of the best q selections. A set of q OLS models are then built between each of these selections of $\mathbf{X}$ and the corresponding column of $\mathbf{Y}$.

2.6 Discrimination case

Let $\mathbf{g}$ be a vector of n integers indicating the belonging of each observation of the calibration set to a given qualitative group. A value g_i gives the number of the group in which the observation of index i is *a priori* classified. Let q be the number of different groups. From $\mathbf{g}$, a matrix of indicators $\mathbf{Y}$, dimensioned $(n \times q)$ is constructed. In this matrix an element y_{ij} takes the value 1 if $j = g_i$, and 0 otherwise. A selection of k variables (sufficiently large number) is performed using CovSel between $\mathbf{X}$ and $\mathbf{Y}$, both centred. For each step i in selection, a LDA is tested by cross-validation between the current selection $\{I_1, I_2, \cdots, I_i\}$ and $\mathbf{g}$. The classification procedure aims at finding the minimal Mahalanobis distance to the centre of classes. Cross-validation results are expressed in terms of percentage of wrong classified samples. Two error curves are provided, one for calibration $(SEC(j)_{j=1\cdots k})$ and the other one for cross-validation $(SECV(j)_{j=1\cdots k})$ which can help the user to choose

the best selection. A model of discrimination by LDA is then developed on this selection.

3 Material and methods

CovSel was applied on several experimental data sets. A first example with an unique response was used to compare CovSel with a classical SMLR. A second one was used to illustrate the multi-response regression and the third one addressed the discrimination problem :

- **Set Apricots**: The **X** matrix consisted of 731 mid infrared spectra of apricots, acquired on $p = 292$ variables (a complete description of the collection can be found in [16]). The Brix degree, evaluating the soluble solid content, was measured on each fruit and was taken as the **y** single response. Calibration and validation sets were randomly drawn 100 times, with a proportion of 2/3 and 1/3, respectively. Each time, CovSel was applied on the calibration set with a number of variables $k = 30$. Then, 30 models were developed by OLS, introducing one after the other the variables previously chosen by CovSel. In parallel, two classical stepwise regressions (SMLR) were also performed with $P < 0.1$ and $P < 0.01$ as limits of probability for introducing the variables. All these models were then applied on the validation set, yielding 100 occurrences of 30 CovSel models and 100 occurrences of the two SMLR models. These occurrences were used to compute boxplots of the standard errors of validation (RMSEV) and of the norm of the models.
- **Set Corn**: The **X** data set, which can be found at http://software.eigenvector.com/Data/Corn, consisted of 80 near infrared spectra of corn samples. The wavelength range was 1100-2498 nm with a 2 nm step ($p = 700$

wavelengths). The moisture, oil, protein and starch contents of the samples were taken as the $\mathbf{Y}$ multi-response. A calibration and a validation set 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 $k = 15$. According to the implementation described in 2.5, CovSel was run a second time for each response to produce 4 sorting of the 15 selected variables. Four series of 15 OLS regressions were then calculated, using the variables in the order previously obtained, and cross-validated on the calibration set, with a leave-one-out splitting. The optimal models were then chosen by studying the evolution of the $SECV$, for each response independently. The four models were then applied to the validation set.

- **Set Wine grapes**: CovSel was applied to discriminate 3 varieties of wine grapes, by means of Visible/very Near Infrared spectrometry (310 - 1050 nm). The experimentation related to 3 varieties: *carignan* (crg), *grenache blanc* (grb) and *grenache noir* (grn). The $\mathbf{X}$ matrix contained 250 spectra measured on $p = 256$ variables. According to the procedure described in 2.6, the $q = 3$ class indicators were used as $\mathbf{Y}$ multi-response. The data set was cut randomly in two equal parts, each set containing 50 samples of crg, 50 samples of grb and 25 samples of grn. The selected variables as given by CovSel were then used as input of LDA. 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. The results were expressed with a prediction error ($PE(\%)$, percentage of wrongly classified samples) and a confusion matrix.

4 Results and discussion

Figure 2 shows the results of the tests done on the apricot dataset. For each value of k between 1 and 30, a boxplot summarizes the distribution of the RMSEV obtained by CovSel in each of the 100 validation tests. The two boxplots on the right are devoted to SMLR results, with $P < 0.1$ (left) and $P < 0.01$ (right). The dispersion is very similar for all the values of k. The median value of RMSEV decreases rapidly from $k = 1$ to $k = 12$ and reaches a value close to the one of SMLR (about 0.75 Brix) and then decreases more slowly down to 0.7 Brix, for $k = 20$. The median values of the number of variables selected by the SMLR models was 13 and 28, respectively for $P < 0.01$ and $P < 0.1$. Figure 3 shows the evolution of the norm of the regression coefficients in the same way as previously. Contrarily to what was observed with RMSEV, the dispersion of these norms increases with k. The regularity of this increasing confirms the above conclusions about the insensitivity of CovSel to overfitting. Moreover, for a same value of the norm of the regression coefficients, CovSel generally gives smallest RMSEV than SMLR. Like PLS, Covsel indeed presents the advantage of maximizing the covariance between $\mathbf{X}$ and $\mathbf{Y}$ rather than the correlation. The consequence of such maximization is that the variables showing high variances play a large role in the regression model, which is not compulsorily the case in SMLR. The norm of the SMLR models is much more variable than those produced by CovSel. This is probably due (i) to the variability of the number of variables chosen by the SMLR (ii) to the management of the variable colinearity, not explicitly performed in SMLR method. This advantage of CovSel is clearly illustrated by the figure 4, showing the selections produced by SMLR ($P < 0.1$) and by CovSel on the

whole data set. The variables selected by CovSel are well spread on the whole spectrum and then obviously less correlated than those selected by SMLR.

Figure 5 illustrates the functioning of CovSel, on the corn dataset, without any preprocessing. Each graph of this figure shows the quantity that is maximized by CovSel, i.e. $\mathbf{x}_i^{\mathrm{T}}\mathbf{Y}\mathbf{Y}^{\mathrm{T}}\mathbf{x}_i$ as a function of the variable index i. The $k = 8$ first steps of CovSel are represented here. Vertical dashed lines indicate the selected variables, located at the curve maximum. It is noticeable that each curve (except the first one) presents a wide depression around the variable that has been selected at the previous step. Two reasons can be put forward for that: (i) the orthogonal projection carried out between two consecutive steps (according to equations 2 and 3) removes the information which is correlated to the selected variable, thus drastically decreases the variance of the neighboring variables in the further steps; (ii) the criterion used by CovSel is based on the covariance, so implicitely on the variance. This depression would not be observed if the correlation was used in place of the covariance because high correlation can be observed even if the variance is low. It is also noticeable that the curves of figure 5 look like peak-shaped spectra that are very different from one step to another. This clearly shows that the deflation achieved by the orthogonal projections allows CovSel to deal with complementary and structured information. Concerning steps 1, 3, 4 and 5, the position of the maximum is neat and unambiguous. Contrarily, in step 2, two high peaks (A and B on the figure) appear. The highest one (B) is chosen and the two peaks totally disappear at the following step. That is explained by the high correlation ($r = 0.9$) existing between the two variables associated with these peaks. Once one peak is selected, all what is correlated to it disappears by means of the orthogonal projection. A contrary situation can be observed in

step 6. Three peaks (A, B and C) can be observed here. The highest one (B) is selected and, at step 7, the peaks A and C remain. This is due to the poor correlation existing between the variables of (A,B) and (B,C) ($r = 0.2$ in both cases). Hence, the peaks A and C bring information that is complementary to the one of peak B and are thus not affected by its selection. These two examples show that, if two peaks have similar height, the choice of one peak in place of the other is not a critical point of the method. At last, one can also notice that in steps 3 and 5 extreme variables were selected. This is probably due to the presence of a baseline, which must appear in the regression model.

Figure 6 shows the evolution of the variance captured by CovSel. It is noticeable that the evolution of these variances complies with the shape illustrated in figure 1, bottom right. This indicates that a model should exist between $\mathbf{X}$ and $\mathbf{Y}$. The curves drawn on figure 7 report the evolution of the $SECV$s as a function of k for the four models (each $SECV$ was divided by the standard deviation of the response, in order to produce comparable curves). Each curve corresponds to a re-ordering of the $k = 15$ variables previously chosen at the first run of CovSel. The best model is the one addressing moisture, for which a $SECV/\sigma$ of about 0.1 is reached for 11 variables. The other models reach a $SECV/\sigma$ of about 0.4, with 13, 12 and 12 variables for oil, protein and starch, respectively. Applying the corresponding models to the test set yielded the results reported in figure 8. Considering the predictions, the results are very satisfactory for moisture ($R^2 > 0.99$), quite good for oil and protein ($R^2 \simeq 0.90$) and less good for starch ($R^2 \simeq 0.88$). The same hierarchy can be observed for the performances of individual PLS regressions calculated on the whole spectra (not shown). Table 2 summarizes the wavelength selections for the 4 models and proposes some assignments. Globally, the selection

seems coherent with the spectrocopic knowledge. However, some wavelengths actually assigned to specific compounds are used for all the responses, like for example the water at 1940 nm or the oil at 2306 nm. This clearly demonstrates that CovSel performs a compromise among the responses. Some bands are not directly assigned to chemical absorptions and are certainly useful for geometrical features, like the baseline that is probably taken into account by the two extreme wavelengths.

Figure 9 reports the results concerning the wine grapes discrimination problem. It shows the evolution of the calibration and cross-validation errors of the linear discriminant model built with the variables selected by CovSel, as a function of the number of steps (k). Both errors decrease very rapidly from about 35% for $k = 1$ to less than 5% for $k = 5$, and then more slowly, down to less than 2% for $k = 8$. The discriminant model built with 8 variables and applied to the test set yielded the errors reported in table 3. The performances are quite satisfactory, in comparison with the ones obtained with a PLS-DA model (not shown here, but published in [17]), which led to the same level of prediction error. This example shows the potential of CovSel to process variable selection in the framework of discriminant problems.

Conclusion

This paper proposes a new method (CovSel), dedicated to the problem of variable selection for highly multivariate data related to single or multiple responses. CovSel consists in an iterative procedure that looks like PLS-NIPALS algorithm. Thanks to the deflation operated at each step of the CovSel algorithm, it produces selections that can be relevantly used in classical multivari-

ate modeling methods. The comparison of CovSel with stepwise multilinear regression in a mono-response case showed a better performance and a better stability for the proposed method. An application to a multi-response case dealing with Near Infrared spectrometry showed that CovSel performed well and that the variable selection was meaningful according to spectrocopy knowledge. A second application on wine variety discrimination from the spectra of berries showed that CovSel is also relevantly applicable to discrimination problems.

Acknowledgements

The authors are grateful to all assistance provided by Sylvie BUREAU from INRA "Sécurité et Qualité des Produits d'Origine Végétale" INRA, Avignon (France)

References

[1] C. W. Huck, R. Maurer, G. K. Bonn, Quality control of liquid plant extracts in the phytopharmaceutical industry in near infrared spectroscopy, in: A. M. C. Davis, R. Giangiacomo (Eds.), Near Infrared Spectroscopy : Proceedings of the 9th International Conference, NIR Publications, 2000, pp. 487–491.

[2] G. Lachenal, Structural investigations and monitoring of polymerisation by nir spectroscopy, Journal of Near Infrared Spectroscopy 1-4 (1998) 299–306.

[3] H. Martens, T. Naes, Multivariate Calibration, Wiley, New York, 1989.

[4] M. Barker, W. Rayens, Partial least squares for discrimination., Journal of Chemometrics 17 (2002) 166–173.

[5] P. C. JP. Gauchi, Comparaison of selection methods of explanatory variables in pls regression with application to manufacturing process data, Chemometrics and Intelligent Laboratory Systems 58 (2001) 171–193.

[6] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, Journal of Machine Learning Research 3 (2003) 1157–1182.

[7] D. Jouan-Rimbaud, B. Walczack, D. Massart, I. Last, K. Prebble, Comparison of multivariate methods based on latent vectors and methods based on wavelength selection for the analysis of near-infrared spectroscopic data, Analytica Chimica Acta 304 (1995) 285–295.

[8] N. Benoudjit, D. François, M. Meurens, M. Verleysen, Spectrophotometric variable selection by mutual information, Chemometrics and Intelligent Laboratory Systems 74 (2004) 243–251.

[9] V. Centner, D. luc Massart, O. E. Noord, S. de Jong, B. M. Vandeginste, C. Sterna, Elimination of uniformative variables for multivariate calibration, Anal. Chem. 68 (1996) 3851–3858.

[10] R. Kohavi, G. H. John, Wrappers for feature subset selection, Artificial Intelligence 97 (1-2) (1997) 273–324.

[11] R. Leardi, A. Lupiáñez, Genetic algorithms applied to feature selection in pls regression, Chemometrics and Intelligent Laboratory Systems 41 (1998) 195Ű207.

[12] E. Llobetnext, O. Gualdrona, M. Vinaixaa, N. El-Barbrib, J. Brezmesa, X. Vilanovaa, B. Bouchikhib, R. Gomezc, J. Carrascoc, X. Correiga, Efficient feature selection for mass spectrometry based electronic nose applications selection, Chemometrics and Intelligent Laboratory Systems 85 (2007) 253–261.

[13] N. R. Draper, H. Smith, Applied Regression Analysis,3rd.ed, Wiley, New York, 1998.

[14] M. C. U. Araujo, T. C. B. Saldanha, R. K. H. Galvao, T. Yoneyama, H. C. Chame, V. Visani, The successive projections algorithm for variable selection in spectroscopic multicomponent analysis, Chemometrics and Intelligent Laboratory Systems 57 (2) (2001) 65–73.

[15] A. Phatak, S. DeJong, The geometry of partial least squares, Journal of Chemometrics 11 (1997) 311–338.

[16] S. Bureau, D. Ruiz, M. Reich, B. Gouble, D. Bertrand, J.-M. Audergon, C. M. Renard, Application of atr-ftir for a rapid and simultaneous determination of sugars and organic acids in apricot fruit, Food Chemistry 115 (3) (2009) 1133 – 1140.

[17] J. M. Roger, B. Palagos, S. Guillaume, V. Bellon-Maurel, Discriminating from highly multivariate data by focal eigen function discriminant analysis. application to nir spectra., Chemometrics and Intelligent Laboratory Systems 79/1-2 (2005) 31–41.

[18] B. G. Osborne, T. Fearn, Near infrared spectroscopy in food analysis, John Wiley and Sons, N.Y., 1986.

[19] P. C. Williams, K. Norris, Near infrared technology in the agricultural and food industries, American association of cereal chemistry Inc., St Paul - Minnesota, 1987.

Appendix

Proof of property 1 *:*

$$\text{ArgMax}\left(\text{diag}\left(\mathbf{X}^{\text{T}}\mathbf{Y}\mathbf{Y}^{\text{T}}\mathbf{X}\right)\right) = \text{ArgMax}_i\left(\text{Max}_{\mathbf{v},\mathbf{v}^2=1}\left(\text{cov}\left(\mathbf{x}_i, \mathbf{Y}\mathbf{v}\right)^2\right)\right)$$

- Proof 1 : Let $m = \text{Max}_{\mathbf{v},\mathbf{v}^2=1}\left(\text{cov}\left(\mathbf{x}, \mathbf{Y}\mathbf{v}\right)^2\right)$

  Applying the Lagrange multipliers on $F(\mathbf{v}) = \text{cov}\left(\mathbf{x}, \mathbf{Y}\mathbf{v}\right)^2$ yields :

$$\frac{\partial}{\partial \mathbf{v}}\left((\mathbf{x}^{\text{T}}\mathbf{Y}\mathbf{v})^2 - \lambda\left(\mathbf{v}^2 - 1\right)\right) = 0$$

$$2\mathbf{Y}^{\text{T}}\mathbf{x}\left(\mathbf{x}^{\text{T}}\mathbf{Y}\mathbf{v}\right) - 2\lambda\mathbf{v} = 0$$

$$\left(\mathbf{Y}^{\text{T}}\mathbf{x}\mathbf{x}^{\text{T}}\mathbf{Y}\right)\mathbf{v} = \lambda\mathbf{v}$$

$$(\mathbf{Y}^{\text{T}}\mathbf{x})(\mathbf{Y}^{\text{T}}\mathbf{x})^{\text{T}}\mathbf{v} = \lambda\mathbf{v}$$

  Then, m is the largest eigenvalue of the $q-$square matrix $(\mathbf{Y}^{\text{T}}\mathbf{x})(\mathbf{Y}^{\text{T}}\mathbf{x})^{\text{T}}$.

- Proof 2 : Let $\mathbf{u}$ be a non nul vector. The matrix $\mathbf{u}\mathbf{u}^{\text{T}}$ has only one non nul eigenvalue $\lambda = \mathbf{u}^{\text{T}}\mathbf{u}$

  We have : $\text{rank}(\mathbf{u}\mathbf{u}^{\text{T}}) = 1$, then $\mathbf{u}\mathbf{u}^{\text{T}}$ has only one non nul eigenvalue.

  Moreover, the trace of a matrix equals the sum of its eigenvalues. Then, we have :

$$\lambda = \text{trace}(\mathbf{u}\mathbf{u}^{\text{T}})$$

$$\lambda = \sum_i u_i^2 = \mathbf{u}^{\text{T}}\mathbf{u}$$

- Finally, combining proof 1 and 2, with $\mathbf{u} = \mathbf{Y}^{\text{T}}\mathbf{x}$, yields :

$$\mathbf{x}^{\text{T}}\mathbf{Y}\mathbf{Y}^{\text{T}}\mathbf{x} = \text{Max}_{\mathbf{v},\mathbf{v}^2=1}\left(\text{cov}\left(\mathbf{x}, \mathbf{Y}\mathbf{v}\right)^2\right)$$

  And consequently :

$$\text{ArgMax}\left(\text{diag}\left(\mathbf{X}^{\text{T}}\mathbf{Y}\mathbf{Y}^{\text{T}}\mathbf{X}\right)\right) = \text{ArgMax}_i\left(\text{Max}_{\mathbf{v},\mathbf{v}^2=1}\left(\text{cov}\left(\mathbf{x}_i, \mathbf{Y}\mathbf{v}\right)^2\right)\right)$$

Table 1

Analogy between PLS and CovSelmethod.

	PLS	CovSel
1	$j = 1$	$j = 1$
2	$\mathbf{u}_j$=ArgMax$_{\mathbf{u}}$(Max$_{\mathbf{v}}$(cov($\mathbf{Xu}$,$\mathbf{Yv}$)^2))$_{\mathbf{u}^2,\mathbf{v}^2=1}$	I_j=ArgMax$_m$(Max$_{\mathbf{v}}$(cov($\mathbf{Xs}^m$,$\mathbf{Yv}$)^2)$_{\mathbf{v}^2=1}$)
3	$\mathbf{z} = \mathbf{X}\mathbf{u}_j$	$\mathbf{z} = \mathbf{X}\mathbf{s}^{I_j} = \mathbf{x}_{I_j}$
4	$\mathbf{X} \leftarrow \mathbf{P}_{\mathbf{z}}^{\perp}\mathbf{X}$	$\mathbf{X} \leftarrow \mathbf{P}_{\mathbf{z}}^{\perp}\mathbf{X}$
5	$\mathbf{Y} \leftarrow \mathbf{P}_{\mathbf{z}}^{\perp}\mathbf{Y}$	$\mathbf{Y} \leftarrow \mathbf{P}_{\mathbf{z}}^{\perp}\mathbf{Y}$
6	$j \leftarrow j + 1$; goto 2	$j \leftarrow j + 1$; goto 2

Table 2

Corn: Summary of the selected wavelengths for the 4 models.

λ (nm)	moisture	oil	protein	starch	assignement
1100	×	×	×	×	baseline
1190		×	×		oil ([18])
1306		×		×	
1428	×		×	×	starch ([19])
1500		×			NH ([18])
1592	×	×	×	×	
1718	×	×	×	×	oil ([19])
1886	×	×	×	×	
1940	×	×	×	×	water
2106	×		×	×	starch ([18], [19])
2204	×	×	×	×	
2250	×	×	×		starch ([18])
2306	×	×	×	×	oil ([19])
2388		×		×	
2498	×	×	×	×	baseline

Table 3

Wine grapes: confusion matrix of the model built with 8 variables and applied to the test set.

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

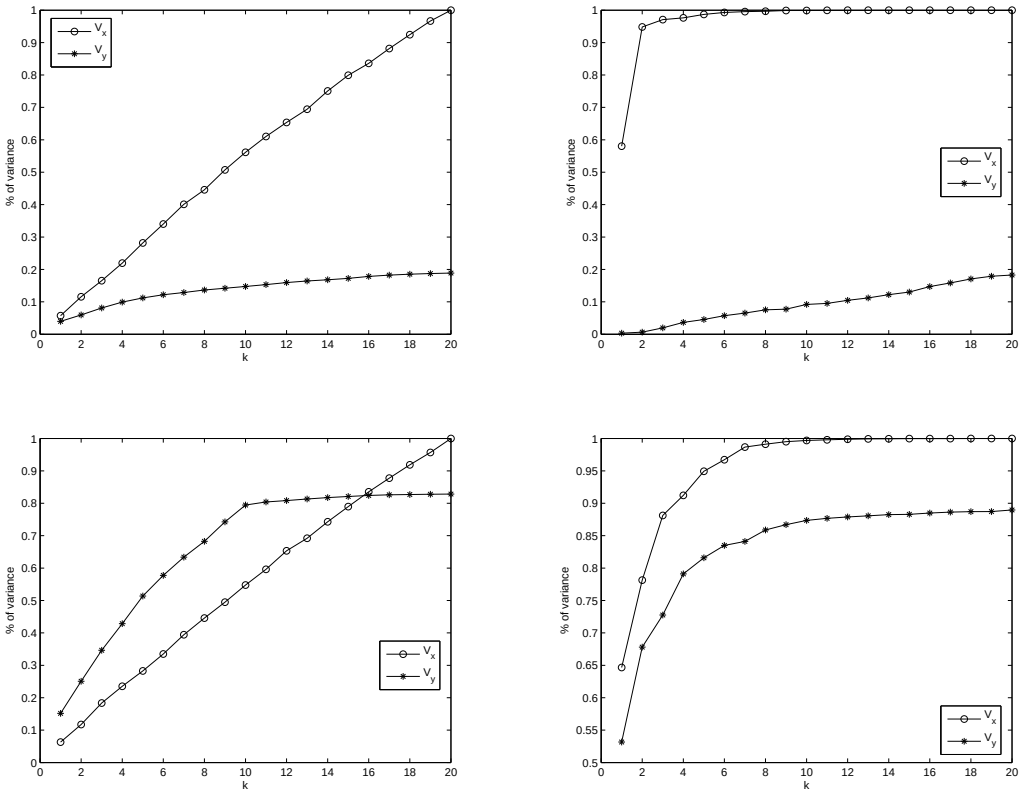

Fig. 1. Typical curves of the evolution of variance explained by CovSel applied to simulated data. $\mathbf{X}$ is made up of 100 lines and 20 columns; $\mathbf{Y}$ is made up of 100 lines and 3 columns. Left: $\mathbf{X}$ variables are independent. Right: $\mathbf{X}$ variables are dependent. Top: no relationship between $\mathbf{X}$ and $\mathbf{Y}$. Bottom: $\mathbf{Y}$ is built by a linear combination of 10 variables de $\mathbf{X}$ and noise addition.

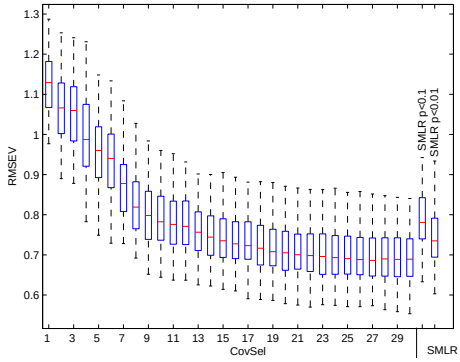

Fig. 2. Apricot: Evolution of the RMSEV distribution according to the dimension of the model based on CovSel selection ($k = 1 \cdots 30$) and RMSEV distribution of SLMR models ($p < 0.1$ and $p < 0.01$). Each boxplot represents the distribution for the 100 trials

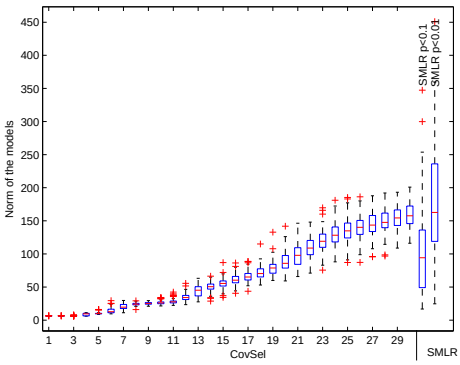

Fig. 3. Apricot: Evolution of the distribution of the model norm according to the dimension of the model based on CovSel selection ($k = 1 \cdots 30$) and distribution of the SLMR model norm ($p < 0.1$ and $p < 0.01$). Each boxplot represents the distribution for the 100 trials

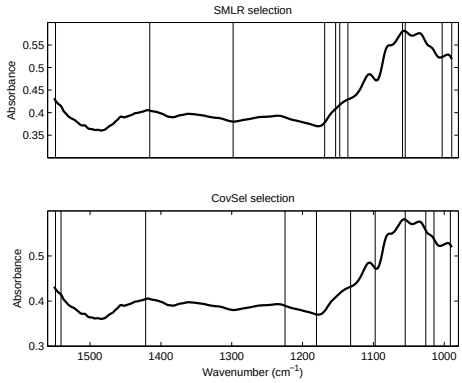

Fig. 4. Apricot: Selections performed by SMLR, p<0.1 (top) and by CovSel (bottom) on the whole data set, superimposed to the mean spectrum.

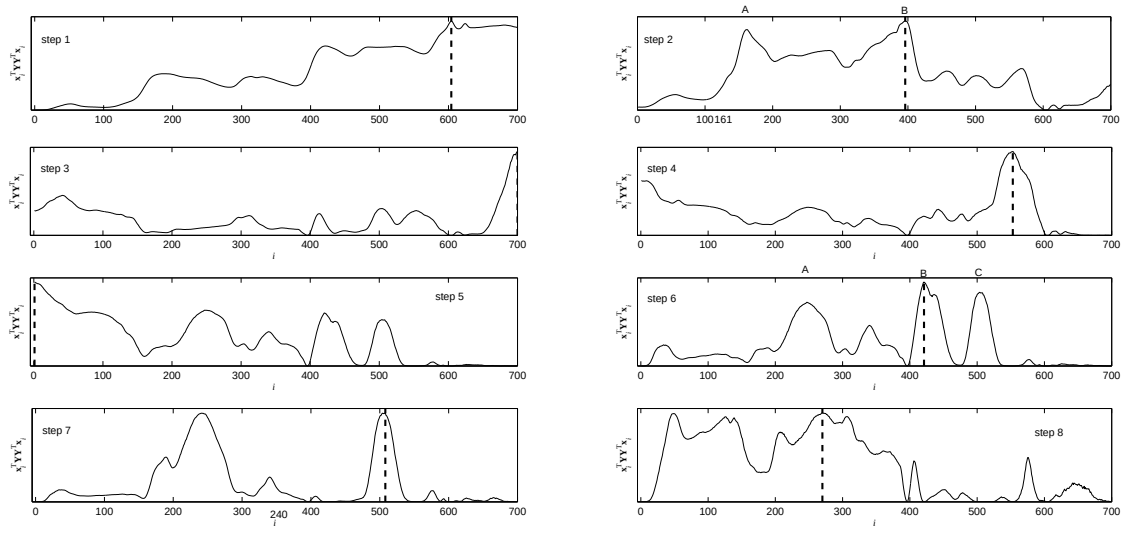

Fig. 5. Illustration of CovSel functioning on the Corn dataset.

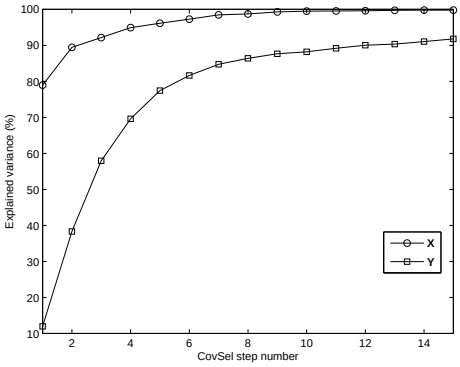

Fig. 6. Corn: Evolution of the cumulated sum of square (explained variance) as a function of the number of variables introduced by the covsel procedure

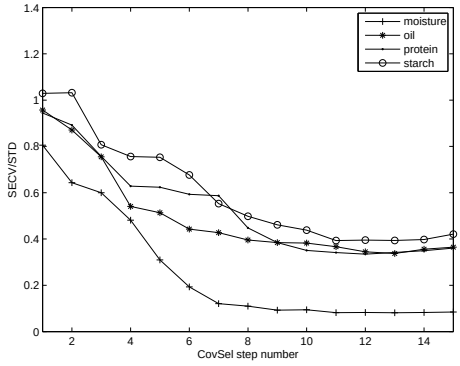

Fig. 7. Corn: Evolution of the SECV according to the number of CovSel steps.

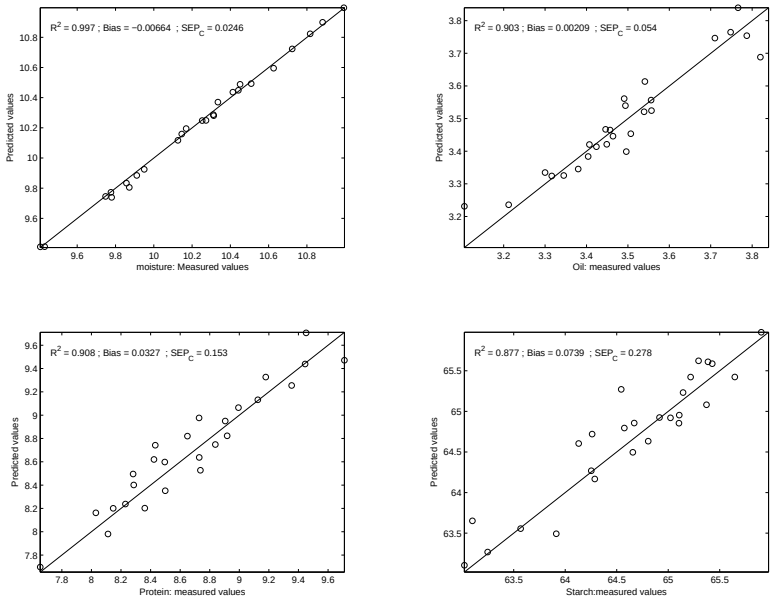

Fig. 8. Corn: Prediction of moisture, oil, protein and starch contents.

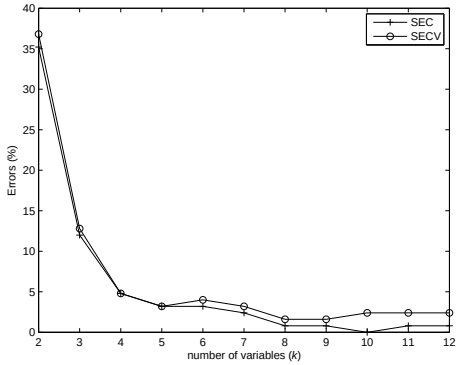

Fig. 9. Wine grapes: evolution of calibration and cross-validation errors of the linear discriminant model, as a function of the number of CovSel steps.