Building a regression vector for variable selection

Reinaldo F. Teófilo, Márcia M. C. Ferreira*
Universidade Estadual de Campinas, Campinas, SP, Brazil; Instituto de Química, P.O. Box 6154, 13084 – 971.
*marcia@iqm.unicamp.br

Keywords: Regression vector, OPS, variable selection.

Introduction

Although the multivariate calibration methods like PLS and PCR are able to deal with large quantities of highly correlated independent variables, and also with small sets of samples, it is well defined and accepted that better predictions are obtained when appropriate variables are selected [11,3]. The feature selection is a technique that aids to identify variables subsets that are, to a proposed problem, the most useful to obtain a more accurate regression model. Besides, the selected subsets can aid in the chemical interpretation of the regression model what is highly relevant for sensorial analysis and quantitative structure-activity relationships (QSAR), among other areas. [3,9].

In multivariate calibration, it is expected that regions with high signal intensity of some vector considered informative, are intuitively connected with those regions from original data that improve the predictions. Thus, it is a usual practice among chemometricians to visualize the plots of informative or prognostic vectors to localize desirable regions from multivariate data.

Several authors [7,4,2] advocate the regression vector as a potential informative tool to select variables in multivariate calibration. Variables with low regression coefficients do not contribute significantly for the prediction and, hence, can be eliminated. Thus, the regression vector can be considered yet, as a weighted sum of loadings included in model [8]. But, many times this vector does not provide improve the prediction and its use for this purpose has not been explored extensively.

Recently, a new method for variable selection based upon the use of informative vectors has been presented and named as Ordered Prediction Selection (OPS) [10,6,5]. The essence of this procedure is sorting the most important variables from an informative vector and to investigate these ordered variables from the most relevant.

The goal of this work is to propose the use of the regression vector for feature selection. The method OPS, a powerful tool for variable selection, is used to explore the potentialities of this informative vector.

Theory

The OPS algorithm

The following steps compose the working OPS algorithm (Figure 1). Initially, the vector that contains information about the location of the best independent variables for prediction is obtained. In the second step, the original independent variables (**X** matrix columns) are differentiated according to the corresponding values of the informative vector obtained previously The higher the absolute values, the more important the original independent variables, what enables their sorting in descending order of magnitude in the third step. In the next step, multivariate regression models are built and evaluated (using some validation strategy) over a window of variables, and further over the window extended by a fixed increment of variables. This procedure is repeated until all variables or a variable percentage is taken into account. Finally, the evaluated variable sets (the initial window and its extensions) are compared using the quality parameters calculated during validations. The model with the best quality parameters contains the variables that present the best prediction power and so, these are the selected variables.

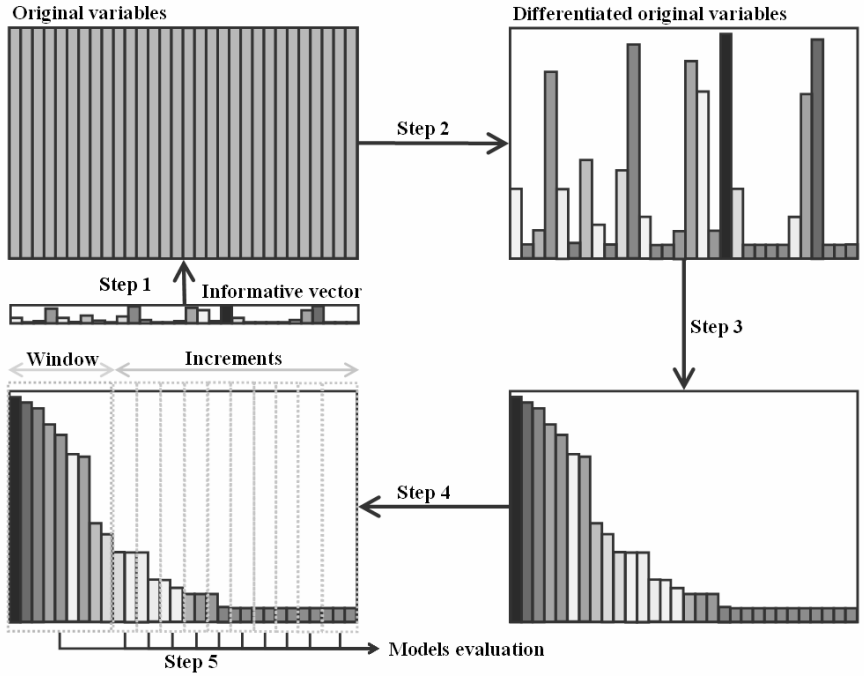

Figure 1. Variable selection steps using the OPS method.

The regression vector

The bidiagonal algorithm for the PLS1 method [1] was used to build the regression vector. This algorithm considers that any **X**(*I*×*J*) matrix can be written as

$$\mathbf{X} = \mathbf{U}\mathbf{R}\mathbf{V}^{t} \tag{1}$$

where **U**(*I*×*J*) and **V**(*I*×*J*) are matrices with orthonormal columns and **R**(*J*×*J*) is a bidiagonal matrix. Thus, considering **U, V** and **R** matrices computed with *h* components truncated in **R**, we can estimate the Moore-Penrose pseudo-inverse of **X** and be able to solve the least squares problem as

$$\mathbf{y} = \mathbf{X}\mathbf{b} \quad \rightarrow \quad \mathbf{y} = \mathbf{U}_h\mathbf{R}_h\mathbf{V}_h^{t}\mathbf{b} \quad \rightarrow \quad \hat{\mathbf{b}} = \mathbf{V}_h\mathbf{R}_h^{-1}\mathbf{U}_h^{t}\mathbf{y} \tag{2}$$

where **b** is the regression vector built with *h* components.

The first PLS model was built and validated, from which *h* = *hMod* was determined. A study using the OPS algorithm was performed by increasing the component number (*h* = *hOPS*) starting from *hMod*, just for building the regression vector which will be used in the first step of OPS algorithm. Two optimum component numbers are employed in this work, one representing the component number to model building (*hMod.*) and the other representing the component number employed to generate the informative vector in OPS method (*hOPS*).

Material and methods

Data set

NIR data set: The data set was composed by NIR spectra of diesel measured at the Southwest Research Institute (SWRI) on a project sponsored by the US Army. The data were obtained from the Eigenvector Research homepage at *http://www.eigenvector.com*. The parameters used were: bp50 - boiling point at 50% recovery/ ºC (ASTM D 86); d4052 - density, g/mL, 15 ºC, (ASTM D 4052); freeze - freezing fuel temperature/ ºC. In this work, splitting the data was considered in agreement with information obtained from the web site, excluding high leverage samples. Besides, the obtained spectra were preprocessed using first

derivative. The number of samples in the training/test sets for bp50, d4052 and freeze were 113/113, 122/121 and 116/115, respectively. Leave-*N*-out cross validation method, where *N* was 10 % of the total number of samples in the training set, was used to validate the models.

Model Evaluation

The calculated error was the root mean square error (RMSE) given in equation 3 and the correlation coefficient *R* was calculated using equation 4.

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{I_m}(y_i - \hat{y}_i)^2}{I_m}} \qquad (3)$$

$$R = \frac{\sum_{i=1}^{I_m}(\hat{y}_i - \bar{\hat{\mathbf{y}}})(y_i - \bar{\mathbf{y}})}{\sqrt{\sum_{i=1}^{I_m}(\hat{y}_i - \bar{\hat{\mathbf{y}}})^2 (y_i - \bar{\mathbf{y}})^2}} \qquad (4)$$

where $\hat{y}$ and $\hat{\mathbf{y}}$ are the scalar and vector estimated values, respectively, $\bar{\mathbf{y}}$ is a scalar of mean values in $\boldsymbol{y}$ and I_m is the number of samples and indicates the validation applied. When internal validation (cross validation) is applied, I_m is the number of samples not included in the model built during cross validation (CV), and the error and correlation are named RMSECV and R_{cv}, respectively. For external validation, I_m is the number of external samples used for prediction and the error and correlation are named RMSEP and R_p, respectively.

Algorithm

The algorithm employed to study the regression vector consists of the following steps:

for *h* = *hMod* to *n* components

generate the regression vector for variable selection with *hOPS* = *h*;

run the OPS algorithm using the previously generated vector (use *hMod* for model building);

store the minimum RMSECV obtained in the selection for all *h*;

end

plot the component numbers versus RMSECV.

Programs

The OPS® Toolbox [5] routines and all data analysis were performed using made in-house functions of Matlab™ 7 (MathWorks, Natick, USA).

Results and discussion

NIR data set

The results indicate that when the regression vector is used as informative vector in OPS algorithm, the number of components necessary to build this vector plays a crucial role for improving the results. It was observed that by increasing the number of components, the RMSECV value dropped until a suitable *hOPS* value is reached. If the process continues, the RMSECV values increase, *i.e.*, there exists an optimum *hOPS* number, which can be significantly higher than *hMod*.

Typical results are presented in Figure 2. Three replicates were done and the RMSECV is presented with standard deviation bar for each selection. Notice that the number of components for which the error is minimum, is significantly higher than the *hMod* used for model building. By the results obtained, we

suppose that when new components are added for building the informative vector, more information is introduced, and consequently, important variables can be better differentiated by this vector.

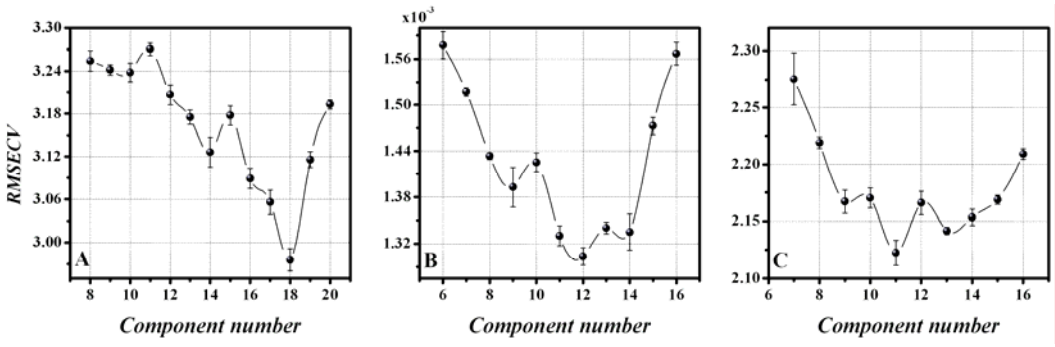

Figure 2. Typical decrease of RMSECV when more components are included for building the regression vector used as informative vector. The bars are the standard deviations of three replicates. This behavior is illustrated for parameters (A) BP50, (B) D4052 and (C) Freeze.

Figure 3 presents both regression vectors for *h* equal to *hMod* and *hOPS* in different context (parameters DP50 and D4052). Comparing the absolute values of elements from both regression vectors, they are always greater in the OPS. It is likely that the regions more important for increasing the prediction power are significantly discriminated from those employed for model building.

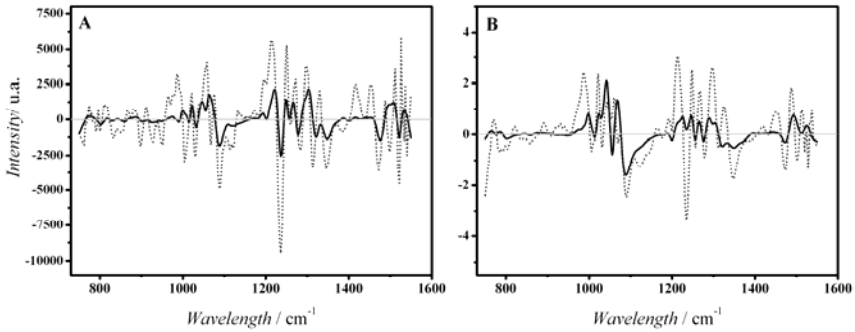

Figure 3. Regression vectors built with different number of components, i.e., hMod (solid line) and hOPS (short dot line). (A) Parameter BP50 with hMod = 8 and hOPS =18; (B) parameter D4052 with hMod = 11 and hOPS = 21.

Table 1 shows the results obtained with all variables (full model) and the selected variables (OPS model). A significant improvement of the statistic parameters was obtained with a meaningful reduction in the number of variables (below 15 % of the initial number).

	Full model	OPS model	Full model	OPS model	Full model	OPS model
	BP50		D4052		Freeze	
hOPS	-	18	-	12	-	11
hMod	8		6		7	
nVars	401	55	401	40	401	55
RMSECV	4.65	3.33	0.00250	0.00130	2.76	2.29
R_{cv}	0.950	0.975	0.977	0.994	0.743	0.824
RMSEP	4.60	3.52	0.00250	0.00130	3.24	2.65
R_p	0.963	0.979	0.975	0.992	0.630	0.771

Table 1. Statistical results for all parameters calibrated for NIR data set.

The improvement achieved by the statistical results shows that the regression vector built with the optimum number of components certainly brings more information about the relevant variables for model building.

Conclusion

A new approach for variable selection is presented by using the regression vector. This vector, with higher number of components with respect to that required by the model, is substantially better to sort the most informative variables. This study was carried out by the powerful OPS method. The use of this new criterion to build the regression vector was applied successfully to data sets from other areas, such as fluorescence spectroscopy, voltammetry, gas chromatography and can be recommended as a new strategy for variable selection associated with the OPS method.

References

[1] G.H. Golub, W. Kahan : Calculating the singular values and pseudo-inverse of a matrix. *SIAM J. Num. Anal. Ser. B.*, 2205-224, 1965.

[2] I.G. Chong, C.H. Jun : Performance of some variable selection methods when multicollinearity is present. *Chemometrics Intell. Lab. Syst.*, 78(1-2):103-112, 2005.

[3] I.S. Helland : On the structure of partial least squares regression. *Commun. Stat.-Simul. Comput.*, 17581-607, 1988.

[4] L. Xu, I. Schechter : Wavelength selection for simultaneous spectroscopic analysis. Experimental and theoretical study. *Anal. Chem.*, 68(14):2392-2400, 1996.

[5] M. Forina, S. Lanteri, M. Oliveros, C.P. Millan : Selection of useful predictors in multivariate calibration. *Anal. Bioanal. Chem.*, 380(3):397-418, 2004.

[6] M.M.C. Ferreira : Multivariate QSAR. *J. Braz. Chem. Soc.*, 13(6):742-753, 2002.

[7] R.F. Teófilo, J.P.A. Martins, M.M.C. Ferreira : *OPS Toolbox 1.0*. INPI - 0000270703255138 , 2007.

[8] R.F. Teófilo, J.P.A. Martins, M.M.C. Ferreira : Sorting variables using informative vectors as strategy for variables select in multivariate regression. *In preparation*, 2008.

[9] R.F. Teófilo, J.P.A. Martins, M.M.C. Ferreira Ordered predictors Selection: an intuitive method to find the most relevant variable in multivariate calibration. *10th International Conference on Chemometrics in Analytical Chemistry (CAC'2006), 2006*, pp. P066

[10] R.P. Williams, A.J. Swinkels, M. Maeder : Identification and application of a prognostic vector for use in multivariate calibration and prediction. *Chemometrics Intell. Lab. Syst.*, 15185-193, 1992.

[11] T. Pirouette : *Pirouette: Multivariate data analysis*. Infometrix , version 3.11, 2003.