Research Article

JOURNAL OF CHEMOMETRICS

Received: 30 May 2008, Revised: 1 August 2008, Accepted: 8 August 2008, Published online in Wiley InterScience: 29 October 2008

(www.interscience.wiley.com) DOI: 10.1002/cem.1192

Sorting variables by using informative vectors as a strategy for feature selection in multivariate regression

Reinaldo F. Teófilo[a], João Paulo A. Martins[a] and Márcia M. C. Ferreira[a]*

A new procedure with high ability to enhance prediction of multivariate calibration models with a small number of interpretable variables is presented. The core of this methodology is to sort the variables from an informative vector, followed by a systematic investigation of PLS regression models with the aim of finding the most relevant set of variables by comparing the cross-validation parameters of the models obtained. In this work, seven main informative vectors *i.e.* regression vector, correlation vector, residual vector, variable influence on projection (VIP), net analyte signal (NAS), covariance procedures vector (CovProc), signal-to-noise ratios vector (StN) and their combinations were automated and tested with the main purpose of feature selection. Six data sets from different sources were employed to validate this methodology. They originated from: near-Infrared (NIR) spectroscopy, Raman spectroscopy, gas chromatography (GC), fluorescence spectroscopy, quantitative structure-activity relationships (QSAR) and computer simulation. The results indicate that all vectors and their combinations were able to enhance prediction capability with respect to the full data sets. However, regression and NAS informative vectors from partial least squares (PLS) regression, both built using more latent variables than when building the model presented in most of tested data sets, were the best informative vectors for variable selection. In all the applications, the selected variables were quite effective and useful for interpretation. Copyright © 2008 John Wiley & Sons, Ltd.

Keywords: variable selection; informative vectors; OPS; partial least squares; chemometrics

1. INTRODUCTION

Multivariate regression is a widely established method for conducting multivariate chemical calibration (determination of a chemical quantity from measured physical quantities), most frequently using the inverse model

$$\mathbf{y} = \mathbf{Xb} + \mathbf{e} \quad (1)$$

where the rows of $\mathbf{X}$ ($I \times J$) matrix contain the values measured or calculated at J response variables (*e.g.* wavelengths, potentials, sensory attributes and molecular descriptors) for I individual samples. In the above equation, $\mathbf{y}$ is the dependent variable, an $I \times 1$ vector containing values of the property of interest (*e.g.* concentrations, panel scores and biological activity, among others) determined from a reference method. The $J \times 1$ vector $\mathbf{b}$ contains the unknown calibration regression coefficients and $\mathbf{e}$ represents an $I \times 1$ error vector normally distributed with mean zero and covariance matrix $\sigma^2\mathbf{I}$ [1,2].

For each sample, a large number of physical quantities are measured, frequently several hundreds. So, the classical multiple linear regression (MLR) cannot be performed and techniques such as principal components regression (PCR) or partial least squares (PLS) regression must be applied [1,3].

Although the multivariate calibration methods as PLS and PCR are able to deal with a large number of highly correlated response variables (predictors or descriptors) and with small sets of samples, there are several situations in which better predictions are obtained when a subset from a larger number of variables is selected [4–8]. This occurs mainly because in a set of hundreds or thousands of variables, most of them enclose noise and irrelevant and/or redundant information. Feature selection is a way to identify variable subsets that in fact reproduce the observed values of a dependent variable, *i.e.* those subsets that are, for a proposed problem, the most useful to obtain a more accurate regression model. Although the main emphasis is upon the prediction, it is desirable that the selected subsets should aid the chemical interpretation of the regression model, which is highly relevant for sensorial analysis and quantitative structure-activity relationships (QSAR), among other areas [3,6,9,10]. Thus, the aim of variable selection is to reduce significantly the number of variables to obtain simple, robust and interpretable models [11].

Nowadays, there is considerable interest for variable selection and the subject has been intensively studied [12,13]. A great number of procedures for variable selection are available in the literature [3,14]; the majority is focused on wavelength selection from spectroscopic data. These procedures can be distinguished from each other by the searching criteria to locate an optimum subset [9], but the majority of them is not generalizable and

* *Correspondence to: M. M. C. Ferreira, Instituto de Química, Universidade Estadual de Campinas, P.O. Box 6154, 13084—971 Campinas, São Paulo, Brazil.*
E-mail: marcia@iqm.unicamp.br

a *R. F. Teófilo, J. P. A. Martins, M. M. C. Ferreira*
Instituto de Química, Universidade Estadual de Campinas, P.O. Box 6154, 13084—971 Campinas, São Paulo, Brazil

efficient for all types of variables. For example, in chemistry, diverse analytical techniques provide different types of predictors and each one presents peculiar and well-defined characteristics.

It is a usual practice among chemometricians to visualize the plots of informative or prognostic vectors to identify desirable features from multivariate data. The informative vectors are those obtained from some mathematical data treatment using the predictors and/or dependent variables. These vectors make the standard behavior of the variables more visible.

In multivariate calibration, the elements of an informative vector that have high absolute values are intuitively connected with the regions from original data that improve the predictions. In fact, if the visualized vector contains the needed information, the feature selection can be performed from this vector. The informative vector can be generated from different types of response variables, because the information they contain is inherent to variables, no matter their nature.

Several authors [14–16] advocate the use of the regression vector as a potential tool to select variables in multivariate calibration. Variables with low regression coefficients would not contribute significantly for prediction and, hence, should be eliminated.

Other informative vectors that could be used for variable selection are those that, in some way, relate responsive variables with the dependent variable as, for example, the correlation coefficients that compose the correlation vector [3,12,17,18]. Usually, a poorly correlated variable is not informative and can be excluded. However, one must pay special attention to such types of vectors since they bear univariate and not multivariate information. Besides the regression and correlation vectors, other types of vectors occurring in the literature [14,19,20] have the same goal. The strategy of using informative vectors is simple, intuitive and, in general, leads directly to the interpretation of the selected variables and also to better predictions of unknowns.

However, in spite of the successful but relatively rare use of informative or prognostic vectors for selecting variables, little has been found in the literature about algorithm automation, definition of cut off criterion and vectors combination.

Hoskuldsson and Reinikainen [12,13,21] have proposed informative vectors and strategies for variable selection. However, these authors did not compare or combine them with other well known informative vectors such as regression and/or correlation vectors.

In this work, a new informative vector is proposed and a new strategy for automatic multiple variable selection/elimination is presented. The methodology developed is based upon several informative vectors and their combinations, introducing a simple and intuitive automatic procedure for variable selection.

2. THEORY

2.1. Notation

Scalars are defined as italic lower case characters (*a, b* and *c*), vectors are typed in bold lower case characters (**a**, **b** and **c**) and matrices as bold upper case characters (**A**, **B** and **C**). Matrix elements are represented by corresponding italic lower case characters with row and column index subscripts (x_{ij} is an element of **X**). In some cases, matrices will be written explicitly as $\mathbf{X}$ ($I \times J$) to emphasize their dimensions (I rows and J columns). The identity matrix is represented as **I** with its proper dimensions indicated.

Superscripts t, $+$ and -1 represent transpose, pseudo-inverse and inverse operations, respectively. The symbol ^, *e.g.* $\hat{y}$, represents an estimated matrix, vector or scalar.

2.2. Method: ordered predictors selection (OPS®)

In general, the essence of the method is to obtain a vector (informative vector) that contains information about the location of the best response variables for prediction (see Figure 1A). Such vectors can be obtained directly from calculations performed with response and dependent variables, or from combinations of

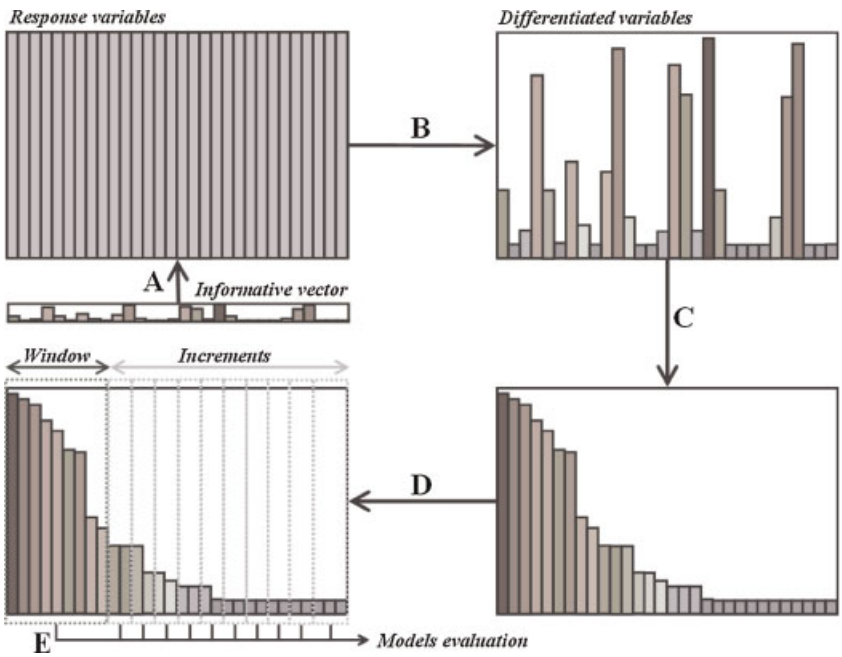

Figure 1. Variable selection steps using the OPS method. This figure is available in color online at www.interscience.wiley.com/journal/cem

different vectors obtained with the same purpose. Obviously, the vector length must be equal to the number of response variables and each position in the vector must be aligned to the corresponding response. In the second step (Figure 1, B), the original response variables (**X** matrix columns) are differentiated according to the corresponding absolute values of the informative vector elements obtained previously in step A. The higher the absolute value, the more important the response variable, which enables their sorting in descending order of magnitude in the third step (Figure 1C).

In the fourth step (Figure 1D), multivariate regression models are built and evaluated using a cross validation strategy. An initial subset of variables (window) is selected to build and evaluate the first model. Then, this matrix is expanded by the addition of a fixed number of variables (increment) and a new model is built and evaluated. New increments are added until all or some percentage of variables are taken into account. Quality parameters of the models are obtained for every evaluation and stored for future comparison.

In the last step (Figure 1E), the evaluated variable sets (initial window and its extensions) are compared using the quality parameters calculated during validations. The model with the best quality parameters should contain variables with the best prediction capability and so these are the selected variables.

2.3. The OPS-PLS algorithm

Specifically, the algorithm used in this work consists of the following steps:

(i) Obtaining informative vectors or their combinations from **X** and **y**;
(ii) Building PLS regression models;
(iii) Calculating quality parameters by leave-*N*-out cross validation and
(iv) Comparing the quality parameters for the obtained models.

The algorithm has many distinguishing features: (1) it is computationally efficient when compared to other variable selection algorithms (*e.g.* genetic algorithm); (2) it may be completely automated with different informative vectors and their combinations; (3) it can be adapted for variable selection when treating multiway data sets [22] and also (4) it can be used in methods of variable selection by intervals (*e.g. i*PLS) or in discriminant analysis [23].

The PLS method to obtain all informative vectors and the bidiagonal algorithm for PLS1, were used in this work.

Manne [24] has shown that PLS1 is equivalent to an algorithm developed by Golub and Kahan [25] for matrix bidiagonalization. Matrix bidiagonalization is a useful decomposition often employed as a fast initialization in algorithms for singular values decomposition [26].

This method considers that any matrix $\mathbf{X}(I \times J)$ can be written as:

$$\mathbf{X} = \mathbf{URV}^{\mathrm{t}} \tag{2}$$

where $\mathbf{U}(I \times J)$ and $\mathbf{V}(I \times J)$ are matrices with orthonormal columns, *i.e.* they satisfy $\mathbf{U}^{\mathrm{t}}\mathbf{U} = \mathbf{V}^{\mathrm{t}}\mathbf{V} = \mathbf{I}$, and $\mathbf{R}(J \times J)$ is a bidiagonal matrix.

Several papers in the literature describe the interesting relation between PLS1 and bidiagonal decomposition [24,27–30]. The bidiagonal decomposition algorithm (PLSBdg), similar to NIPALS and SIMPLS, considers the **y** information during computations. The PLSBdg algorithm can be summarized as follows [27,29].

(1) initialize the algorithm for the first component, $v_1 = \mathbf{X}^{\mathrm{t}}\mathbf{y}/\|\mathbf{X}^{\mathrm{t}}\mathbf{y}\|$; $\alpha_1 u_1 = \mathbf{X}v_1$
(2) for i = 2, . . . , *h* components
  2.1. $\gamma_{i-1}v_i = X^t u_{i-1} - \alpha_{i-1}v_{i-1}$
  2.2. $\alpha_i u_i = Xv_i - \gamma_{i-1}u_{i-1}$

with,
$\mathbf{V}_h = (v_i, ..., v_h)$, $\mathbf{U}_h = (u_i, ..., u_h)$ and $\mathbf{R}_h$

$$= \begin{pmatrix} \alpha_1 & \gamma_1 & & \\ & \alpha_2 & \gamma_2 & 0 \\ & & \ddots & \ddots \\ & 0 & & \alpha_{k-1} & \gamma_{k-1} \\ & & & & \alpha_{k-1} \end{pmatrix}$$

It can be proved that

$$\mathbf{XV}_h = \mathbf{U}_h\mathbf{R}_h \rightarrow \mathbf{R}_h = \mathbf{U}_h^{\mathrm{t}}\mathbf{XV}_h. \tag{3}$$

The bidiagonal matrices are analogous but not identical to those derived by singular values decomposition (SVD) and, therefore, the PCR regression method is slightly different from PLS.

An important conceptual detail is that the scores calculated by the bidiagonalization algorithm are different by a normalization factor from those produced in NIPALS and SIMPLS algorithms [30]. Ergon [31] and Pell *et al*. [30] have shown that there is a difference between the reconstructed matrix obtained by the PLSBdg method ($\hat{\mathbf{X}}_h = \mathbf{U}_h\mathbf{R}_h\mathbf{V}_h^{\mathrm{t}}$) and those reconstructed by NIPALS and SIMPLS methods, becoming smaller with higher numbers of latent variables.

In this work, one of the informative vectors proposed makes use of the reconstructed matrix and, therefore, PLSBdg is fundamental to obtain a consistent reconstructed matrix.

2.4. The informative vectors

2.4.1. *Regression vector (Reg.).*

The regression vector obtained when multivariate regression is performed can be defined as the expected change in the response, per unit change in the variable, if all the variables and responses are linearly related [32]. In this work, the regression vector was considered as an informative vector for variable selection.

Once the matrices **U**, **V** and **R** are computed with *h* components truncated in **R**, according to Equation 3, the regression vector can be estimated directly by solving the least squares problem as shown in Equation 4.

$$\mathbf{y} = \mathbf{Xb} \quad \rightarrow \quad \mathbf{y} = \mathbf{U}_h\mathbf{R}_h\mathbf{V}_h^{\mathrm{t}}\mathbf{b} \quad \rightarrow \quad \hat{\mathbf{b}} = \mathbf{V}_h\mathbf{R}_h^{-1}\mathbf{U}_h^{\mathrm{t}}\mathbf{y} \tag{4}$$

Although the regression vectors obtained by NIPALS or PLSBdg are the same [29,30], the reconstructed **X** matrix from the bidiagonal method with *h* components is more consistent to calculate the *residual vector*. Besides, the PLSBdg algorithm is significantly more efficient computationally when compared to NIPALS algorithm.

When using the regression vector for variable selection in the OPS method, the first question asked is about the number of components, *h*, to be used when obtaining this informative

vector. Firstly, it is proposed in this work to build and validate a PLS model, from which $h = hMod$ is determined. But maybe *hMod* cannot generate a regression vector sufficiently informative for variable selection. To find the best informative vector, a study was then performed on the full data set by increasing the number of components in the model, starting from $h = hMod$ and carrying out the variable selection using the OPS algorithm. By varying the h value, different informative vectors are generated, from which the best component number ($h = hOPS$) is selected. Therefore, two optimum numbers of components are employed in this work, one representing the component number for model building (*hMod.*) and the other representing the component number employed to generate the best informative vector in OPS method (*hOPS*).

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.

2.4.2. Correlation vector between columns of X and y (Corr)

The Pearson correlation coefficient (R) is a natural population parameter for bivariate normal distribution used for assessing the degree of linear association between two variables **x** and **y**. It is a dimensionless measure and lies in the interval from -1 to $+1$, with zero indicating the absence of correlation (but not necessarily the independence of the two variables). Data falling exactly on a straight line (sloped upwards or downwards) indicates that $|R| = 1$ [33].

The informative correlation vector contains the correlation coefficients between each predictor $\mathbf{x}_j$ and the dependent variable **y**. This vector shows how each predictor in **X** is correlated to **y**, and high correlation indicates that the corresponding variable should contain important information for the model. A disadvantage of this methodology is that the correlation between a combination of predictors and **y** is not taken into account. An expression for the calculation of the correlation vector is shown in Equation 5

$$\mathbf{r} = \frac{{}^{a}\mathbf{X}^{t}\,{}^{a}\mathbf{y}}{I-1} \tag{5}$$

where ${}^{a}\mathbf{X}$ and ${}^{a}\mathbf{y}$, with superscript a, are the autoscaled matrix and vector for predictor and dependent variables, respectively.

2.4.3. Residual vector (SqRes)

In data compression methods such as PLS and PCR, when the matrix is truncated, an estimate for the original matrix (reconstructed matrix) can be obtained, *i.e.* $\hat{\mathbf{X}}_h$. The reconstructed matrix with h components should contain the relevant information needed while the eliminated information is considered as residuals and defined by $\mathbf{E}_h = \mathbf{X} - \hat{\mathbf{X}}_h$, where **X** is the original data matrix. The residual matrix can give information about important variables in **X**. When relevant information for the regression is being transferred to $\hat{\mathbf{X}}_h$, each component carries information from important columns of **X** into $\hat{\mathbf{X}}_h$ and the sum of squared residuals of the corresponding columns in $\mathbf{E}_h$ approaches zero. So, the columns in $\mathbf{E}_h$ with low values of squared sum indicate the more effective variables for the regression. The informative vector proposed in this case is the inverse of the sum of squared residuals defined as **q**, according to Equation 6

$$\mathbf{E}_h = \mathbf{X} - \hat{\mathbf{X}}_h \quad \rightarrow \quad q_j = \frac{1}{\mathbf{e}_j^{t}\mathbf{e}_j} \tag{6}$$

where $\mathbf{e}_j$ is the j-th column of $\mathbf{E}_h$.

2.4.4. Covariance procedures vector (CovProc)

Reinikainen and Hoskuldsson [21] have proposed a vector, named CovProc, to rank and then select the variables. Its calculation is based on ranking the variables according to their covariance and selecting 'an optimal' number of variables to use in sequential dynamic systems. The *H*-principle suggests using $\mathbf{X}^{t}\mathbf{y}\mathbf{y}^{t}\mathbf{X}$ as the measure of strength between **X** and **y**. The method CovProc suggests sorting the variables according to weights derived from $\mathbf{X}^{t}\mathbf{y}\mathbf{y}^{t}\mathbf{X}$ and judging the subset selection based on expanding **X** according to the sorting obtained. The informative vector is the diagonal of $\mathbf{X}^{t}\mathbf{y}\mathbf{y}^{t}\mathbf{X}$.

$$\mathbf{CovProc} = diag(\mathbf{X}^{t}\mathbf{y}\mathbf{y}^{t}\mathbf{X}) \tag{7}$$

2.4.5. Variable influence on projection (VIP)

The VIP score of a predictor, first published by Wold *et al.* [34], is a summary of the importance for the projections to find h latent variables. The VIP score for the j-th variable, which is a measure based on the weighted PLS coefficients, can be calculated by Equation (8). On the other hand, since the average of squared VIP scores equals 1, the 'greater than one rule' is generally used as a criterion for variable selection [14].

$$VIP_j = \frac{J \times \sum_{k=1}^{h} t_k^2 v_{jk}^2}{\sum_{k=1}^{h} t_k^2} \tag{8}$$

where $\mathbf{t} = \mathbf{R}\mathbf{V}^{t}\mathbf{b}$

Attention should be called to the fact that the *VIP* vector calculated from Equation 8 using the loadings and scores from the PLSBdg algorithm is equivalent to those calculated by using **w** and **t** from NIPALS (the weights are the same and the scores differ by a normalization factor).

2.4.6. Net analyte signal (NAS) vector

For inverse calibration (PLS, PCR, etc.) the multivariate vector net analyte signal (NAS) is defined as a part of the mixture's signal that is useful for prediction [35].

Faber [36] has presented an efficient way to perform the calculation of the NAS vector. This method is based on the use of the regression vector since this vector is orthogonal to the interference signal.

The dot product in the equation $y_i = \mathbf{b}^{t}\mathbf{x}_i + e$ will give only the contribution of that part of the mixture spectrum **x** which is orthogonal to the interference spectra: this part is the NAS vector.

According to Ferre *et al*. [35] and Bro *et al*. [37], the NAS vector can be calculated using the following equation

$$\mathbf{x}_i^{nas} = \hat{y}_i(\mathbf{b}^t)^+ = (\hat{y}_i/\mathbf{b}^t\mathbf{b})\mathbf{b} \quad (9)$$

Since the NAS vector is obtained for each sample, an average of the columns of the matrix that contains all vectors is a good estimate for an informative vector to be employed in the OPS method.

As in the regression vector, the NAS vector was also built using the component number $h = hOPS$.

2.4.7. Signal-to-noise (StN) vector

This method was described by Brown [17] and consists in calculating a signal-to-noise statistic for each variable. As presented by Miller [18], parameters of a least squares fit between each intensity variable ($\mathbf{x}_j$) to the constituent concentrations ($\mathbf{y}$) are calculated according to Equation 10,

$$\mathbf{y} = b_0\mathbf{1}_j + b_1\mathbf{x}_j + \mathbf{e}_{\mathbf{y},j} \quad (10)$$

where $\mathbf{x}_j$ is the j-th column in the data matrix $\mathbf{X}$, $\mathbf{1}_j$ is a vector of ones and $\mathbf{e}_{\mathbf{y},j}$ is the residual. Once the least-squares fit is made, the signal-to-noise (StN) vector for variable j is then calculated

$$StN_j = \frac{\hat{b}_1}{(\hat{\mathbf{e}}_{\mathbf{y},j}^t, \hat{\mathbf{e}}_{\mathbf{y},j})} \quad (11)$$

where $\hat{\mathbf{e}}_{\mathbf{y},j} = \mathbf{y} - \hat{b}_0\mathbf{1}_j - \hat{b}_1\mathbf{x}_j$. This procedure is repeated for each variable used in the data matrix $\mathbf{X}$.

2.4.8. Vectors' combinations

Besides the vectors Reg, Corr, SqRes, CovProc, VIP, NAS and StN, their combination also can be used to search the most predictive variables. This combination is obtained by performing the product of the absolute value of each element in one vector times the corresponding element in the other vector. Before doing that, the vectors are normalized. In the present work, pairs of these vectors were investigated.

To make the graphical representation straightforward, an abbreviation for vectors pairs was used, *i.e.* Reg-Corr (*RC*); Reg-SqRes (*RS*); Reg-CovProc (*RCP*); Reg-VIP (*RV*); Reg-NAS (*RN*); Reg-StN (*RStN*) and so on for other pairs.

2.4.9. Model evaluation

The quality of the models is assessed by the root mean square error (RMSE) calculated according to Equation 12 and the correlation coefficient R given by Equation 13,

$$RMSE = \sqrt{\frac{\sum_{i=1}^{I_m}(y_i - \hat{y}_i)^2}{I_m}} \quad (12)$$

$$R = \frac{\sum_{i=1}^{I_m}(\hat{y}_i - \bar{\hat{\mathbf{y}}})(y_i - \bar{\mathbf{y}})}{\sqrt{\sum_{i=1}^{I_m}(\hat{y}_i - \bar{\hat{\mathbf{y}}})^2(y_i - \bar{\mathbf{y}})^2}} \quad (13)$$

where $\hat{y}$ and $\hat{\mathbf{y}}$ are the scalar and vector of estimated values, respectively, $\bar{\mathbf{y}}$ is a scalar of mean values in $\mathbf{y}$ and I_m is the number of samples. When internal validation (cross validation—CV) is applied, I_m is the number of samples in the calibration set (training), and the error and correlation coefficients are named RMSECV and R_{cv}, respectively. For external validation (a new set of samples), I_m is the number of predicting samples (P) and in this case, the error and correlation coefficients are named RMSEP and R_p, respectively.

3. EXPERIMENTAL

3.1. Data sets

Six data sets were used in this work. They were obtained from different sources, *i.e.* near-Infrared (NIR) spectroscopy, Raman spectroscopy, fluorescence spectroscopy, gas chromatography (GC), quantitative structure activity relationship (QSAR) data and finally, one simulated data set.

The full data sets were split into training and external validation sets. Approximately 30% of samples were selected by the algorithm of Kennard and Stone [38], available on the Internet at http://www.vub.ac.be/fabi/publiek/index.html, to be the external validation set.

For cross validation, the method leave-*N*-out was applied, where N was set as 10% of total sample number in the training set.

3.1.1. NIR data set

The first data set was composed by NIR spectra of diesel samples measured at the Southwest Research Institute (SWRI) in a project sponsored by the US Army. The data set was taken from the Eigenvector Research homepage at http://www.eigenvector.com. The following physical properties were modeled: bp50, boiling point at 50% recovery/°C (ASTM D 86); CN, cetane number (equivalent to octane number for gasoline, ASTM D 613); d4052, density, g mL^{-1}, 15°C, (ASTM D 4052); freeze, freezing temperature of the fuel/°C; Total, total aromatics, mass % (ASTM D 5186) and Visc., viscosity, cSt/40°C. In this work, the data was split according to that on the web site without high leverage samples. Also, the spectra obtained were preprocessed by taking the first derivative. The number of samples in the training set/test set for bp50, CN, d4052, freeze, Total and Visc. were 113/113; 113/112; 122/121; 116/115; 118/118 and 116/116, respectively.

Further information about this data can be found at http://www.idrc-chambersburg.org/shootout_2002.htm or in Reference [39].

3.1.2. Raman data set

This data set from literature is available at http://www.models.kvl.dk/research/data/ and presented by Dyrby *et al*. [40]. The data set consisted of Raman scattering for 120 samples using 3401 wave numbers in the range of 200–3600 cm^{-1} and using the average of 64 scans for each sample. The dependent variable referred to the amount of active substance (containing a C=N group) in *Escitalopram*® tablets in %w/w units. More experimental details can be obtained in the literature [40].

Second derivative pre-treatment was applied to the samples before building the model.

3.1.3. Fluorescence data set

This data set was designed by Bro *et al*. [41] for the study of several topics in fluorescence spectroscopy and can be found at http://www.models.kvl.dk/research/data/. The selected analytes have very similar excitation and emission spectra. Consequently, the calibration problem is rather complicated.

Six different analytes were used: catechol (CATE), hydroquinone (HYDR), indole (INDO), resorcinol (RESO), *L*-tryptophane (TRYP) and *DL*-tyrosine (TYRO). The set used in this work contained 404 mixtures of 2–4 fluorophores. The concentration ranges of the fluorophores in the samples were: CATE (0–87 μmol L^{-1}), HYDR (0–22 μmol L^{-1}), INDO (0–5.46 μmol L^{-1}), RESO (0–39.96 μmol L^{-1}), TRYP (0–7.44 μmol L^{-1}) and TYRO (0–12.14 μmol L^{-1}). Only the first replicate measurement of each sample was used in the present work.

The scan settings emission wavelengths were 230–500 nm (recorded every 2 nm) and excitation wavelengths 230–320 nm (recorded every 5 nm). For further experimental details see Reference [41].

Prior to analysis, part of the recorded data was removed (Raman and Rayleigh scattering) and the region removed was interpolated using the algorithm from Bahram *et al*. [42] available at http://www.models.kvl.dk/source/. This procedure was carried out in order to avoid the presence of any scattering effects. Besides, the excitation wavelengths 230–240 and 300–320 nm were excluded, together with the emission wavelengths 230–296 and 422–500 nm. The 3D array generated for each sample with dimensions 19×136 (excitation $\times$ emission) was cut down to 11×62, corresponding to the 11 emission spectra recorded from each experiment. An unfolding was performed to apply PLS regression. Thus a total of 682 variables were investigated for each sample.

3.1.4. GC data set

The data GC was selected from the examples available in Pirouette software [43]. This set is formed by 35 responses consisting of peak areas for a set of gas chromatographic runs on fuel samples and 3 dependent variables made up from measurements of the following physical properties: flash point, specific gravity and freeze point. A total of 16 samples were measured and one was detected as outlier and removed. Due to the small data set, the OPS method was carried out using all samples and after feature selection the training set and test set were split into 11 and 4 samples, respectively.

3.1.5. Data set QSAR

The data set QSAR is from our research group [44] and available at http://pcserever.iqm.unicamp.br/~marcia//hiv1qsardata.html. The 14 molecular descriptors for 48 HIV-1 protease inhibitors were generated on the basis of 1D and 2D formulas and the dependent variable was the *in vitro* inhibition activity [45], $pIC_{50} = -\log IC_{50}$. The data set was split into 32 compounds for the training set and the other 16 were for external validation.

3.1.6. Simulated data set

The simulated data set was proposed by one of reviewers of this paper and consisted of 20 mixtures simulated by using UV-type spectra from 4 analytes and their respective concentrations randomly generated. The data set was split into 10 samples for the training set and 10 samples for external validation.

3.2. Programs

All data analyses were performed using home-built functions written for Matlab 7 (MathWorks, Natick, USA). The OPS® Toolbox routines, implemented in MATLAB 7, were registered and are available on the Internet at http://lqta.iqm.unicamp.br

4. RESULTS AND DISCUSSION

4.1. NIR data set

Recent advances in process instrumentation (using NIR spectroscopy, in particular) and chemometric methods have led to the popularization of NIR spectroscopy.

For oil fractions and diesel fuels, the NIR spectroscopic region (750–1550 nm) is especially attractive because most absorption bands observed in this region arise from overtones and combination bands of carbon-hydrogen (C-H) stretching vibrations of hydrocarbon molecules.

Figure 2 presents the minimum RMSECV obtained by leave-eleven-out cross validation. Besides the full data set (first bar from bottom and named Full), when no variables were excluded, results for the seven informative vectors and their combinations introduced in the previous session are shown. Vertical dot lines indicate the RMSECV value for full data set (with no variable selection) and for the minimum RMSECV obtained by the OPS method. A significant decrease in RMSECV values occurred when other informative vectors were combined with Reg. The subset of variables selected by the following informative vectors: Corr, SqRes, CovProc, VIP and StN presented reasonable improvement in the prediction quality of the models principally when combined with Reg (RC, RS, RCP, RV and RStN), but the results are not still comparable to Reg and NAS vectors and their combination RN.

It was observed that when the regression vector was used as an informative vector in the OPS algorithm, the number of components to build this informative vector played a primordial role to obtain good results. The study performed by increasing the number of components to build the informative vector while keeping constant the number of components (*hMod.*) to build the model was fundamental. It was found that the optimum number of components (*hOPS*) to build a regression vector with the purpose of variable selection was, in most cases, significantly higher than *hMod*.

Typical variations of RMSECV as the number of components is increased are presented in Figure 3. Three replicates were carried out and the mean value of minimum errors (RMSECV) and its standard deviation bar (error bar) are shown for each selection. The number of components for minimum RMSECV indicates $h = hOPS$. Figure 3 indicates that the number of components with minimum error is significantly higher than those for building the model. In Figure 2, for comparison, each result was obtained with *hMod* equal to 9 for Visc. and 10 for Total. This number of components for each specific physical property was kept constant for the window and each added increment.

It seems that when an excessive number of components is included to calculate the informative vector, more information from each variable and its real contribution to the model seems to be better represented. This is an empirical observation and it

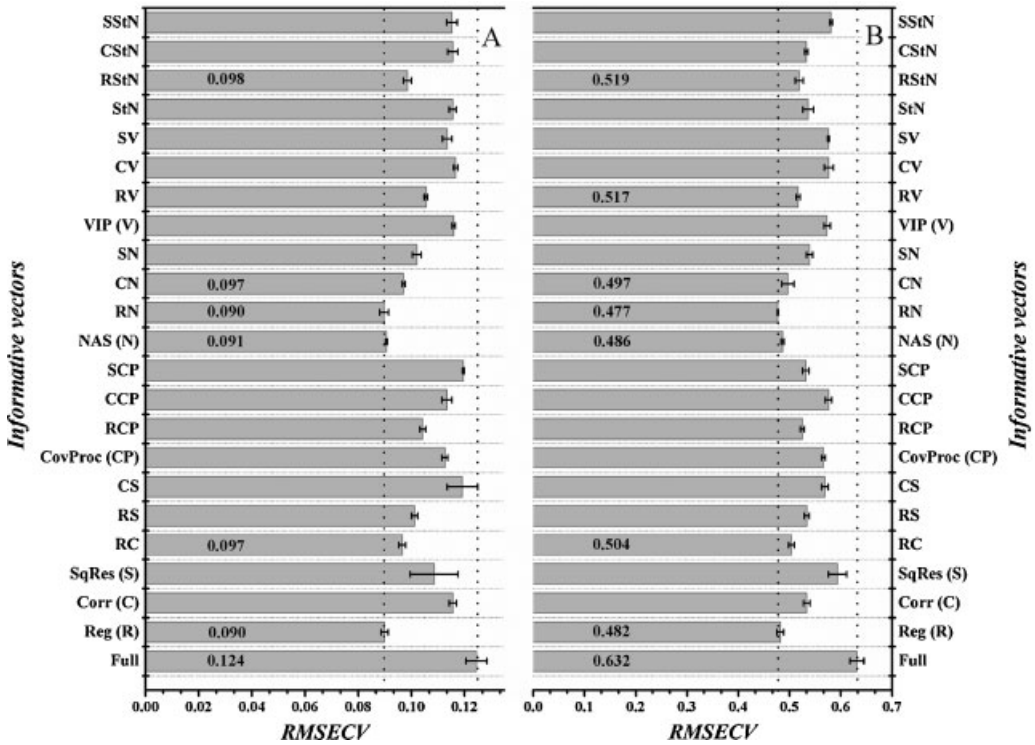

Figure 2. Minimum RMSECV obtained for the seven informative vectors and their combinations used in the OPS algorithm. The standard deviation of three replicates is represented by a horizontal error bars. These results were obtained for two physical properties Visc. (A) and Total (B). The full bar RMSECV obtained (right hand dotted line) were 0.124 and 0.632 for Visc. and Total, respectively. The best results (left hand dotted line) obtained were 0.090 and 0.477 for Visc. and Total, respectively. Other good values are indicated inside the bars for comparison.

depends upon the data structure. More statistical studies are necessary, however these are out of the scope of the present work.

NAS vector, which is similar to the regression vector for inverse calibration, was also very efficient to improve the prediction when using $h = hOPS$. The vectors SqRes and VIP, which are also dependent upon the number of components, did not show better results when using $h = hOPS$, as obtained for the regression and NAS vectors. Thus, regression and NAS vectors were built with a number of components equal to *hOPS* while SqRes and VIP vectors were built with a number of components equal to *hMod*.

Figure 4 illustrates both regression vectors applied in different situations in this context. The absolute values of regression coefficients are larger for *hOPS* than for *hMod*.

Figure 5 shows the OPS plots for detection of the best points for selection/elimination of variables. When the variables are arranged into descending order as indicated by the informative vector and the first window is selected, a decrease in RMSECV and increase in R_{cv} is expected when expanding this first variable subset by adding new variables. When the optimal number of variables is reached, the error increases and the R_{cv} decreases with the inclusion of noninformative variables. The region where values of minimum RMSECV and maximum R_{cv} occur is the

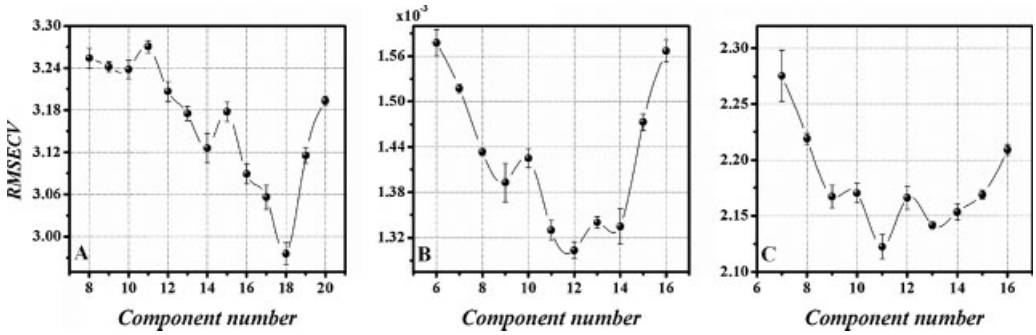

Figure 3. Plots of RMSECV versus the number of components for building the regression vector used as informative vector. The bars are the standard deviations of three replicates. Three physical properties are considered: (A) BP50 with $hOPS = 18$; (B) D4052 with $hOPS = 12$ and (C) Freeze with $hOPS = 11$

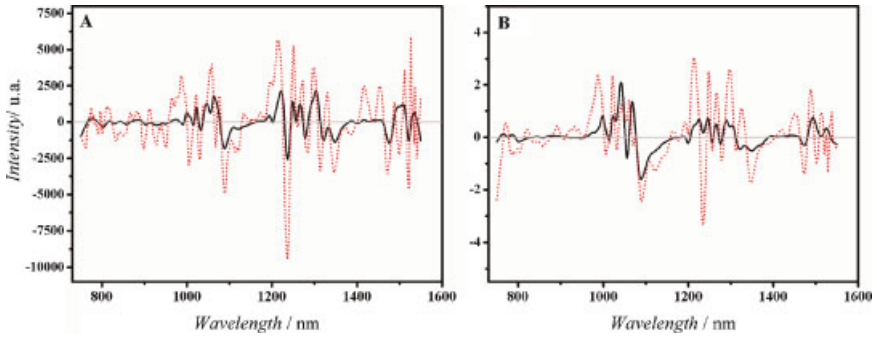

Figure 4. Regression vectors built with different numbers of components, *hMod* (black solid line) and *hOPS* (red doted line). (A) Physical property BP50 with *hMod* = 8 and *hOPS* = 18; (B) physical property D4052 with *hMod* = 6 and *hOPS* = 12. This figure is available in color online at www.interscience.wiley.com/journal/cem

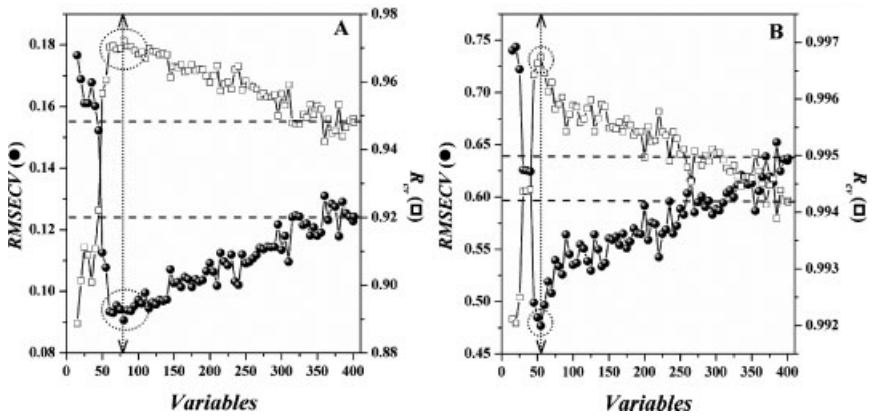

Figure 5. Typical OPS plots. The vertical arrows indicate the set of variables with better prediction ability. The doted circles indicate the optimum regions. The horizontal dashed lines indicate the statistical parameters using the full data set. These results were obtained for the physical properties (A) Visc. and (B) Total. For comparison, each result was obtained with the maximum number of *hMod* equal to 9 for Visc. and 10 for Total. The maximum number of components was fixed in the window and each added increment.

optimum (vertical arrows with dotted circles) for the selection/elimination of variables. The horizontal dashed lines in the plots indicate the RMSECV or R_{cv} for full data.

For all physical properties of this data set, the results obtained when using the variables indicated by the optimum regions are significantly better than those obtained when all variables (full data set) were used. These results suggest that the proposed methodology is in fact very efficient for variable selection.

Table I shows results obtained for the models built with all variables (full model) and selected variables. A meaningful reduction of the number of variables with improvement in statistical parameters was obtained. Note that the vectors Reg and NAS are important vectors either alone or in combination.

The selected variables are shown in Figure 6 where well defined regions are clearly observed. Notice that baseline regions were not selected and peaks with rather clear physical meaning were selected. This is a great advantage of the OPS method compared to the other methods.

4.2. Raman data set

Raman is a rapidly developing spectroscopic technique that, unlike infrared spectroscopy (IR), does not require special sampling techniques. It gives a measure of the weak inelastic scattering created by interaction between the incoming light and the species present in the sample [40]. Comparing the raw spectrum from the pure active substance containing the cyanide group (C=N) [40], with that from the investigated tablets, it is visible that the cyanide group is easily detected in this region. The distinct peak at 2233 nm seen in the tablet spectrum originated from the C=N group. Several others peaks from the active substance can be seen in the tablet spectrum (*e.g.* at 1614 and 3075 nm), but none are as specific as the cyanide peak. The tablet spectra contain several peaks from the excipients including three intense peaks from titanium dioxide in the coating material (395, 511 and 634 nm) and several peaks from cellulose (1095, 1121, 1380 and 2900 nm). There is a strong fluorescence background originating from an unknown residual in the primary microcrystalline cellulose excipient. Besides, cellulose contains trace amounts of organic substances that can be highly fluorescent. To eliminate this fluorescence background the second derivative of the spectra was calculated.

Figure 7A shows the trend in RMSECV with increasing the number of components to generate the informative vectors (regression vector and NAS, SqRes and VIP). The leave-twelve-out cross-validation method was used to calculate the RMSECV. The value *hOPS* = 9 was used to obtain the informative vectors and *hMod* = 3 for model building. The vectors Reg, NAS and the combination RS, RN and SN, which are shown in Figure 7B yielded the best results (RMSECV ≅ 0.47). Hence, these vectors could be used to select variables. The vertical lines in Figure 7B indicate the

Table I. Statistical results obtained from NIR data set for all physical properties

	Full model	OPS model	Full model	OPS model	Full model	OPS model
	BP50		CN		D4052	
Vector	—	Reg.	—	Reg.-Corr.	—	Reg.
hOPS	—	18	—	11	—	12
hMod.	8		6		6	
nVars	401	60	401	105	401	40
RMSECV	4.51	2.97	1.99	1.91	2.5×10^{-3}	1.3×10^{-3}
R_{cv}	0.953	0.979	0.802	0.820	0.977	0.994
RMSEP	4.60	3.90	2.09	2.02	2.5×10^{-3}	1.3×10^{-3}
R_p	0.963	0.976	0.811	0.826	0.975	0.992
	Full model	OPS model	Full model	OPS model	Full model	OPS model
	Freeze		Total		Visc.	
Vector	—	Reg.	—	Reg.	—	NAS
hOPS	—	11	—	18	—	15
hMod.	7		10		9	
nVars	401	55	401	45	401	80
RMSECV	2.76	2.12	0.63	0.48	0.12	0.09
R_{cv}	0.743	0.828	0.994	0.995	0.952	0.972
RMSEP	3.24	2.65	0.66	0.64	0.13	0.10
R_p	0.630	0.771	0.994	0.995	0.944	0.968

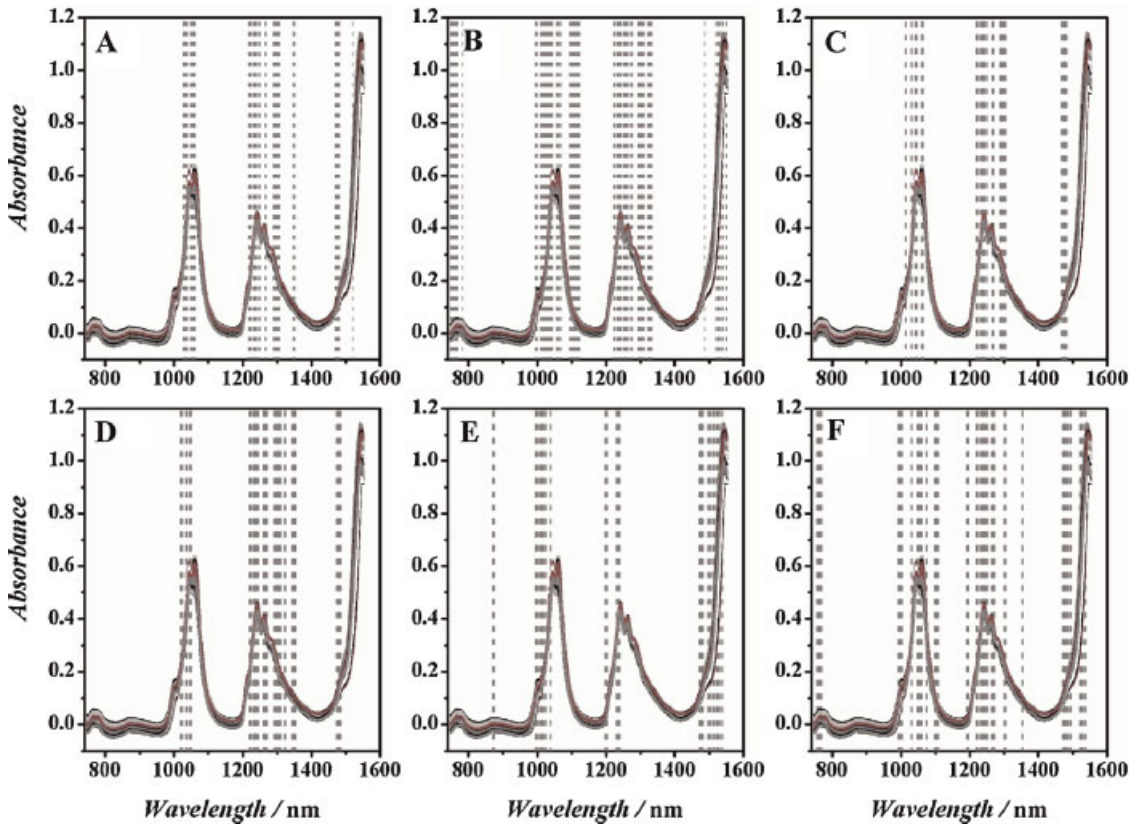

Figure 6. Selected variables from the original spectra for each physical property. (A) BP50, (B) CN, (C) D4052, (D) Freeze, (E) Total and (F) Visc. All calculations were performed on the first derivative spectra. This figure is available in color online at www.interscience.wiley.com/journal/cem

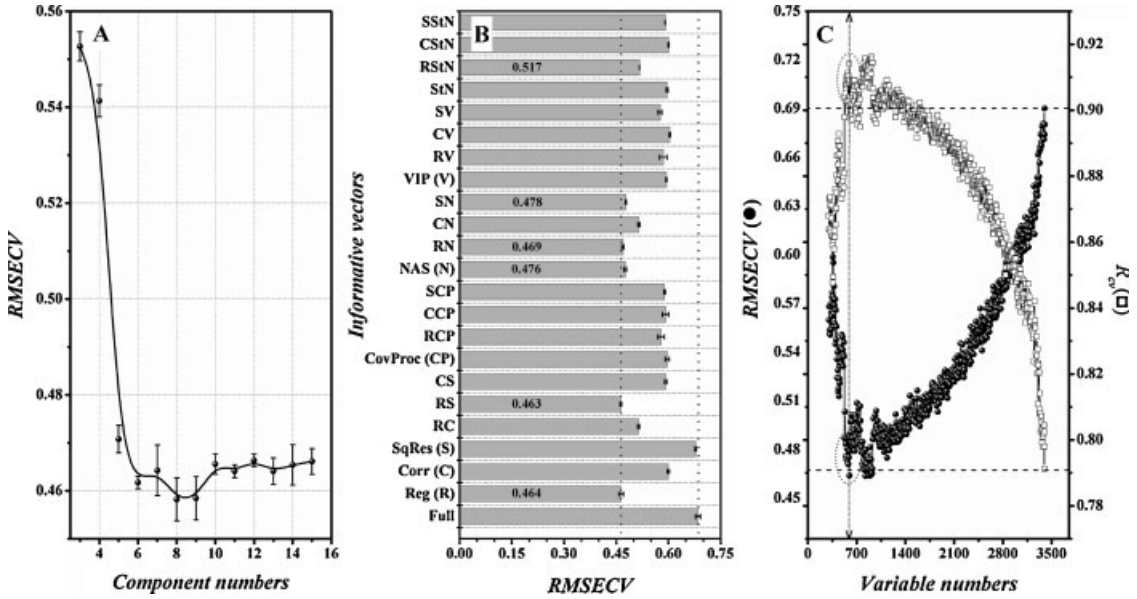

Figure 7. (A) Typical decrease of RMSECV with an increasing number of components to build the regression vector used as informative vector; (B) minimum RMSECV obtained for various informative vectors used in the OPS algorithm and for the full data set (other good values are indicated inside the bars for comparison); (C) OPS plot indicating the optimum region for selection/elimination of variables (dotted ellipses). The vertical arrow indicates the set of variables (left of the arrow) with better prediction ability. The horizontal dashed lines indicate the statistical parameters using the full data set. The vertical and horizontal error bars are, respectively, the standard deviations of three replicates in the graphs A and B.

maximum RMSECV (0.685, right vertical line) and the minimum RMSECV (0.463, left vertical line). The combination Reg.-SqRes showed the least RMSECV value and so it was selected as the informative vector. Note that in Figure 7B the error of the SqRes vector is very close to that of the full data set. However, when SqRes was combined with Reg., (RS), a significant decrease in RMSECV was observed. This result justifies the study of the informative vector combinations.

Figure 7C shows the OPS plot indicating the optimum region for selection/elimination of variables (vertical arrow and ellipses). From 3401 original variables, 595 were selected with significant improvement of the prediction ability of the model. The first window and increment values in the OPS algorithm were 300 and 5, respectively.

Table II. Statistical results for active substance for data set Raman

	Full model	OPS model
	Activity	
Vector	—	Reg.-SqRes
hOPS	—	9
hMod.	3	
nVars	3401	595
RMSECV	0.69	0.44
R_{cv}	0.800	0.924
RMSEP	0.66	0.49
R_p	0.792	0.890

After variable selection, the data set was split into 85 samples for modeling and 35 samples for external validation.

The statistical parameters for the final model are shown in Table II. The results for the full model (no variables excluded) are included for comparison. Approximately 18% of total variables were selected and a meaningful improvement in the prediction capacity was obtained. The selected variables are marked in the spectra of Figure 8. Note that selected regions are coherent with peaks obtained from the active substance. Several variables were selected in the region around of 2233 nm originating from the

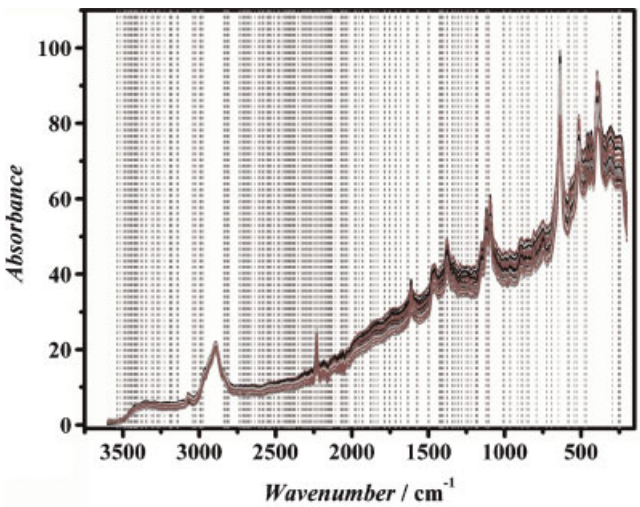

Figure 8. Selected variables for the active substance containing the cyanide group (C=N). Variable selection was performed on the second derivative spectra. This figure is available in color online at www.interscience.wiley.com/journal/cem

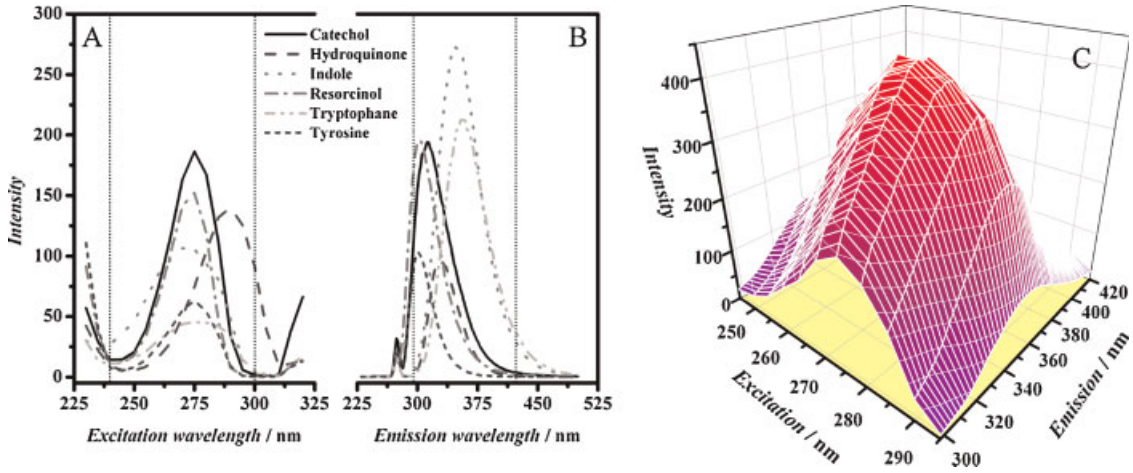

Figure 9. (A and B) Pure excitation and emission spectra for the six fluorophores. The regions inside the vertical dotted lines in plots A and B were used for data analysis. (C) Representative landscape spectrum from one mixture of fluorophores. This figure is available in color online at www.interscience.wiley.com/journal/cem

cyanide group. The spread of selected variables along the spectra is justified due to the presence of several other peaks in the tablet spectrum derived from the active substance [40] and also to the complexity of samples. Dyrby *et al*. [40] have tried to select variables by using the *i*PLS method, but no improvement was obtained. The use of a genetic algorithm for variable selection in this data set is also not appropriate because it is time consuming due to the high number of variables. On the other hand, the OPS method was computationally efficient, and its high potential to select interpretative variables has been shown again in this example.

4.3. Fluorescence data set

Fluorescence spectroscopy has been used in many scientific fields such as chemistry, medicine, environmental and food science. However, the fluorescence signal can be rather complex and, therefore, the analysis may become complicated owing to interferences, scattering, overlapping signals, etc. [46].

When the fluorescence spectrum of a sample is measured at several emission and excitation wavelengths, the data analysis can be carried out by using multi-way methods, or applying two-way methods on the unfolded matrices. In this work the data was unfolded and the PLS regression method employed. Such types of data contain several peaks with different intensities and with a relatively symmetric behavior. Besides, the excitation and emission bands are highly overlapped as shown in Figure 9. These two factors make this data set very complicated for the selection of interpretable variables from multiple peaks.

Plots for determination of the number of components, *hOPS*, to build the informative regression vectors that will be used for variable selection are shown in Figure 10. As expected, *hOPS* values [12 for CATE (Figure 10A), 9 for INDO (Figure 10B), 9 for TRYP (Figure 10C) and 10 for TYRO (Figure 10D)], were higher than *hMod*, which corresponds to the first point in each plot (*hMod* = 6, except for HYDR and INDO where *hMod* = 5).

The OPS plots in Figure 11 show distinct trends of RMSECV for different analytes. However, for all analytes, a few variables were selected and the improvements were significant, as observed in the plots with a wide gap between values for the full data set (horizontal dashed lines) and optimum region for selection/elimination of variables (indicated by dotted circles).

Table III shows significant improvement in the prediction ability of models with a few variables, except for HYDR. An interesting fact observed was that the vector Corr has potential as an informative vector for variable selection, especially when combined with the Reg vector. However, the vector Reg is once more the best informative vector for variable selection.

Figure 12 shows the variables selected by the OPS method. Note that specific peaks were selected, where each peak corresponded to the emission spectra of a given excitation wavelength. Within each peak selected specific wavelengths were chosen. In all cases an interpretation of selected variables is possible when comparing with the respective excitation and emission wavelengths from pure spectra (see Figure 9).

The variables selected for CATE are approximately located in the center of excitation band where most of the peaks are superimposed and at short wavelengths from emission spectra (Figures 12A and 12Az). For RESO (Figures 12B and 12Bz) and TYRO (Figures 12C and 12Cz), similar spectral regions are expected. It is very clear that the selected variables are situated mainly in the central part of excitation spectra and at short wavelengths of the emission band. Finally, for TRYP, though the position in excitation and emission spectra are dislocated to longer wavelengths (see pure spectra in Figures 9A and B), the selected variables follow exactly the same tendency, where the center to right part of excitation band and final part of emission spectra were indicated by the OPS method (Figures 12F and 12Fz). For other phenols (not shown), the behavior follows the expected trends.

This conclusion reinforces the high performance of the OPS method in selecting meaningful variables.

4.4. GC data set

A constant issue in industry is the analysis of finished products for quality control and chromatography has shown to be an extremely versatile technique in this context. The data set used

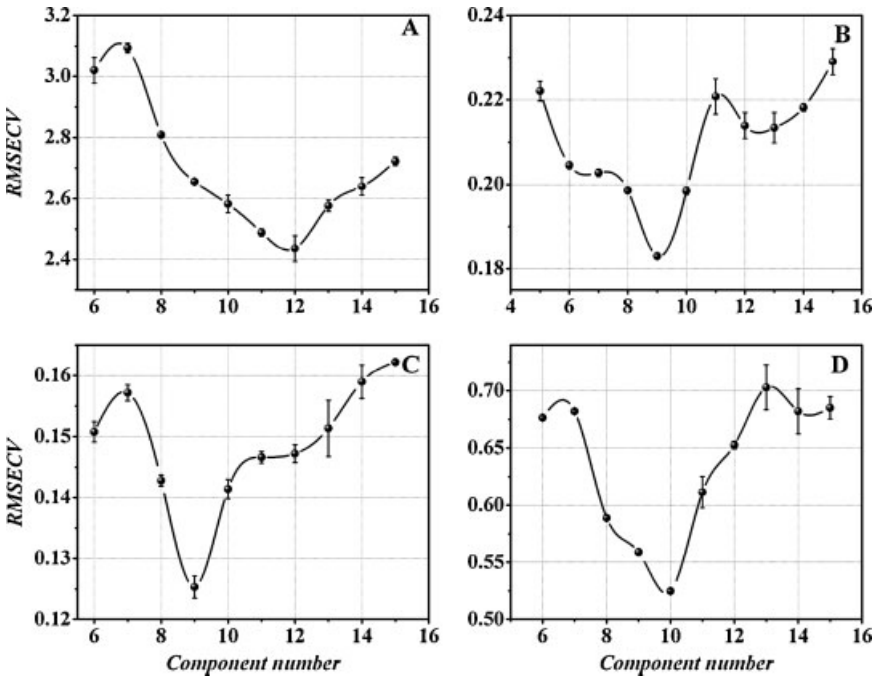

Figure 10. Decrease of RMSECV with increasing number of components to build the regression vector used as an informative vector. The error bars are the standard deviations of three replicates. This behavior is illustrated for (A) CATE, (B) INDO, (C) TRYP and (D) TYRO.

here was extracted from Pirouette™ package and contains peak areas from gas chromatography applied to fuel. This example is a challenge to improve the prediction power by variable selection. In this case, the OPS algorithm was applied to the autoscaled data using leave-one-out cross-validation.

Table IV shows the improvement in the models obtained by the selected peak areas using the OPS method. The variables selected by importance in decreasing order were: flash point [12, 16 and 32]; specific gravity [8, 31 32, 27, 26, 2, 28 and 25]; freezing point [18, 29, 24 and 16].

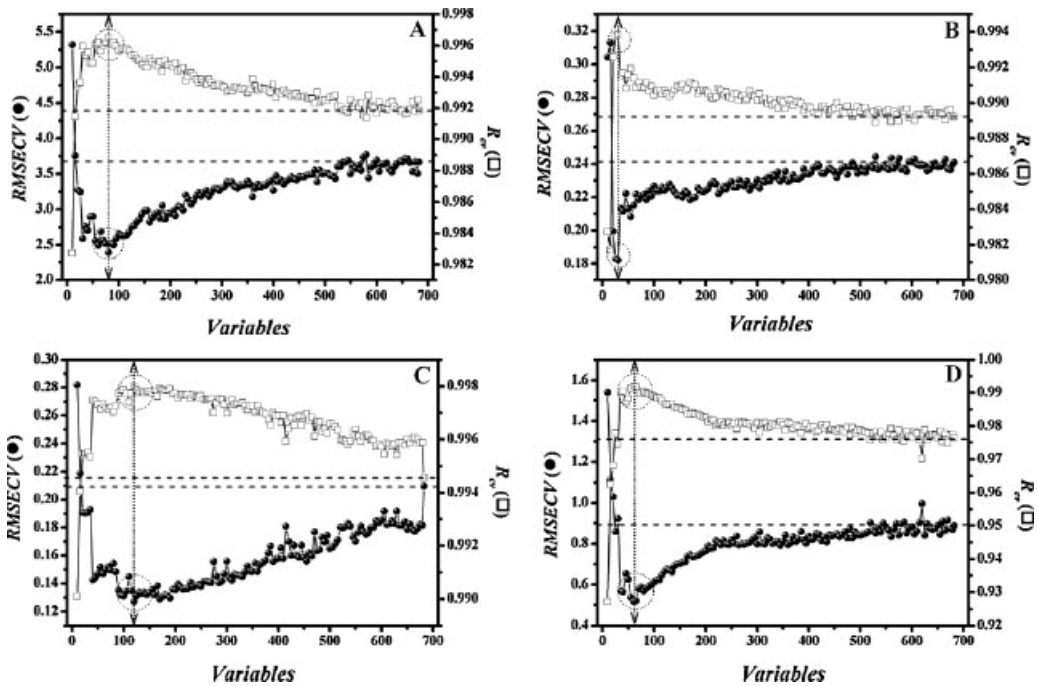

Figure 11. OPS plots for (A) CATE, (B) INDO, (C) TRYP and (D) TYRO. The horizontal dashed lines show statistical parameters for the full data set. The arrows indicate the set of variables (left from the arrow) with better prediction ability. The dotted circles indicate the optimum region for selection/elimination of variables. For comparison, each result was obtained with a fixed number *hMod* equal to 6 for the majority of analytes, except HYDR and INDO where *hMod* = 5. The cross validation method used was leave-twenty five-out.

Table III. Statistical results for all analytes from the fluorescence data set

	Full model	OPS model	Full model	OPS model	Full model	OPS model
	Catechol		Hydroquinone		Indole	
Vector	—	Reg.	—	Reg.-SqRes	—	Reg.
hOPS	—	12	—	10	—	9
hMod.	6		5		5	
nVars	682	80	682	30	682	30
RMSECV	3.66	2.40	1.37	1.31	0.24	0.18
R_{cv}	0.992	0.996	0.983	0.985	0.989	0.994
RMSEP	4.55	3.59	1.27	1.25	0.34	0.26
R_p	0.990	0.995	0.989	0.989	0.986	0.992
	Full model	OPS model	Full model	OPS model	Full model	OPS model
	Resorcinol		L-Tryptophane		DL-Tyrosine	
Vector	—	Reg.-Corr.	—	Reg.-Corr.	—	Reg.
hOPS	—	14	—	9	—	10
hMod.	6		6		6	
nVars	682	20	682	120	682	60
RMSECV	1.96	1.09	0.21	0.13	0.89	0.52
R_{cv}	0.949	0.984	0.994	0.998	0.976	0.992
RMSEP	3.29	1.84	0.23	0.22	1.54	0.91
R_p	0.956	0.987	0.997	0.997	0.951	0.983

4.5. QSAR data set

Variable selection is essential in QSAR studies. A good and useful QSAR model for prediction and interpretation has low errors and must be understandable from the chemical and biological points of view. Thus, variable selection is essential to obtain a well interpretable model.

The OPS algorithm was applied to the autoscaled data using leave-one-out cross-validation and the results can be seen in Table V.

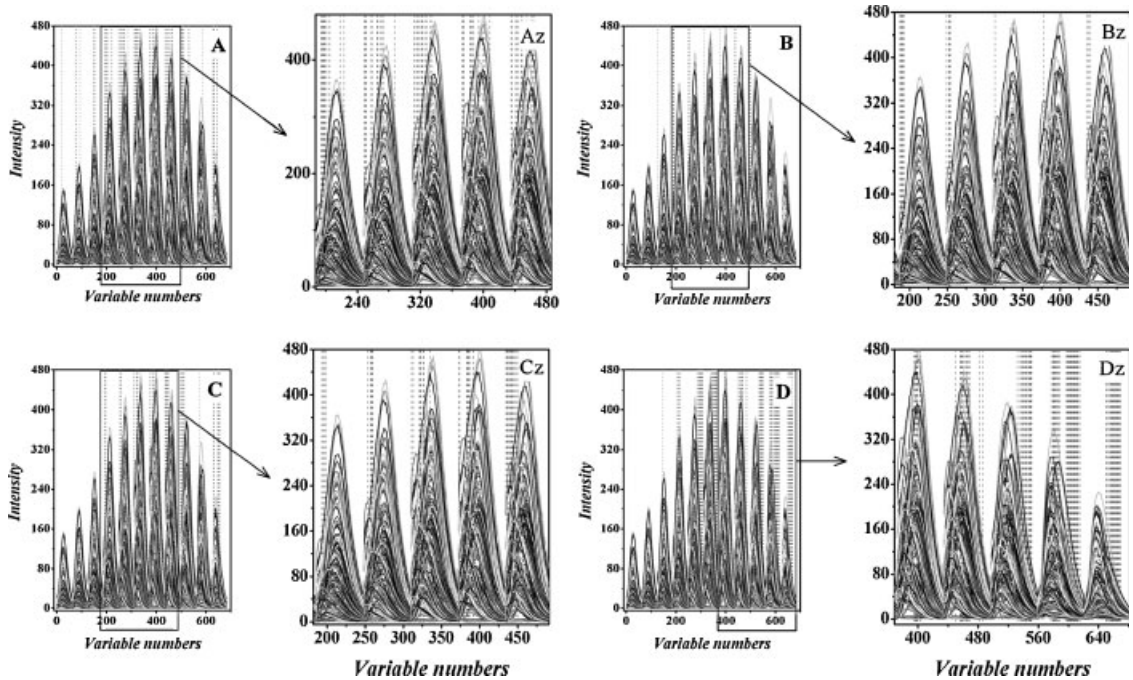

Figure 12. Selected variables for the four phenolic compounds. Each peak corresponds to one of the 11 emission spectra recorded. (A) CATE, (Az) zoom of the selected CATE emission peaks, (B) RESO, (Bz) zoom of the selected RESO spectral region, (C) TYRO, (Cz) zoom of the selectd TYRO emission peaks and (D) TRYP, (Dz) zoom of the selected TRYP emission peaks.

Table IV. Statistical results from different models* of physical properties obtained from GC peak areas

	Full model	OPS model	Full model	OPS model	Full model	OPS model
	Flash		Spec grd		Freeze	
Vector	—	Reg.	—	Reg.	—	NAS
hOPS	—	2	—	4	—	9
hMod.	1		3		3	
nVars	35	3	35	8	35	4
RMSECV	8.03	7.28	0.005	0.003	2.99	1.29
R_{cv}	0.450	0.520	0.316	0.724	0.361	0.910
RMSEP	4.92	2.48	0.009	0.002	3.65	1.34
R_p	0.683	0.993	−0.044	0.950	0.241	0.841

*The cross validation method used was leave-one-out.

Table V. Statistical results for the QSAR data set

	Full model	OPS model
Vector	—	Reg
hOPS	—	3
hMod	3	1
nVars	14	3
RMSECV	0.53	0.47
R_{cv}	0.92	0.94
RMSEP	1.12	1.09
R_p	0.90	0.91

The results obtained after variable selection are slightly better than those obtained with all variables (Table V). The three variables selected were: X9, X10 and X11 (see Reference [44]). X9 is the effective number of substituents on the central chain (a steric-geometrical descriptor). X10 is the number of potential hydrogen bonds (an electronic-geometrical descriptor) and X11 is the effective number of ring substituents (a steric-geometrical descriptor). The selection of these three descriptors (variables) can be well understood from both the chemical and the chemometric points of view: (1) the descriptors are of complex natures, including steric, geometrical, electronic, hydrogen bonding and even hydrophobic features of the treated molecules; (2) the descriptors are good representatives of clusters that can be seen in exploratory analysis; (3) the absolute values of the regression vector elements are the greatest for these descriptors and (4) among correlation coefficients between the biological activity and molecular descriptors, the greatest absolute values are observed for X9 and X11. The model with the three selected variables, as has been just mentioned, is justified in terms of chemical concepts, and it presents an interesting alternative to chemical interpretation of interactions between the HIV-1 protease and inhibitors when compared to the model with fourteen variables reported in literature [44].

In this example, the PLS model was further validated by leave-*N*-out crossvalidation, where *N* varied from 1 to 5, to check the robustness of the model. Chance correlation was tested by performing ten y-randomizations according to Wold and Eriksson [47]. These are standard validation procedures applied to QSAR

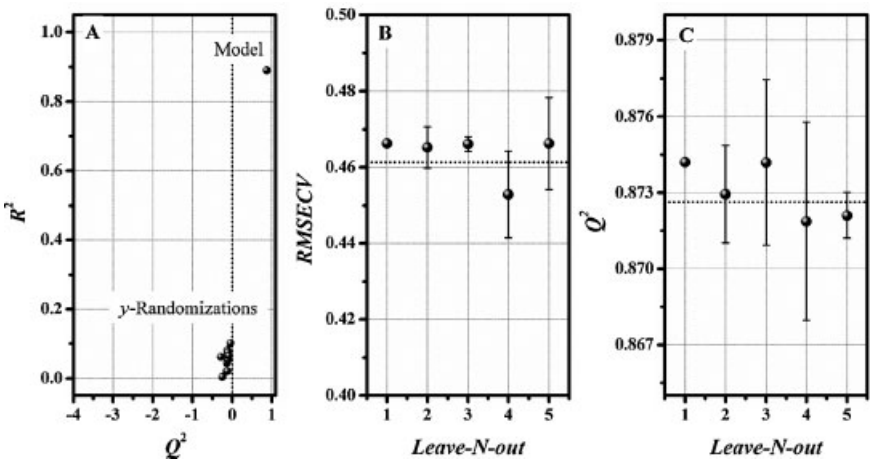

Figure 13. Plots for model validation. Chance correlation plot (A) and leave-*N*-out crossvalidation (*N* varied from 1 to 5) plots with *N* versus RMSECV (B) and Q^2 (C).

Table VI. Statistical results for all analytes from the simulated data sets*

	Full model	OPS model	Full model	OPS model
	Analyte 1		Analyte 2	
Vector	—	Reg-StN	—	Reg.
hOPS	—	6	—	7
hMod.	4		4	
nVars	100	42	100	28
RMSECV	0.014	0.0028	0.0136	0.0047
R_{cv}	0.999	1.000	1.000	1.000
RMSEP	0.0123	0.0028	0.0126	0.0037
R_p	1.000	1.000	1.000	1.000
	Full model	OPS model	Full model	OPS model
	Analyte 3		Analyte 4	
Vector	—	Reg-Corr	—	Reg.
hOPS	—	7	—	7
hMod	4		4	
nVars	100	57	100	40
RMSECV	0.0173	0.0137	0.0154	0.0062
R_{cv}	0.999	1.000	0.999	1.000
RMSEP	0.0039	0.0028	0.0100	0.0051
R_p	1.000	1.000	1.000	1.000

*The cross validation method used was leave-one-out.

data as strongly suggested by the QSAR scientific community. The original data was randomized prior to leave-*N*-out cross-validation and y-randomization. These results are shown in Figure 13.

According to Figure 13A, it is possible to conclude that there is no chance correlation using the selected variables. Besides that, a robust model was obtained as can be observed from results of leave-*N*-out crossvalidation (Figure 13B). This application reaffirms the great potential of the OPS method for interpretable variable selection.

Althought not applied in this work, a strategy of using the OPS method for obtaining a substantially reduced number of variables (which are important in QSAR/QSPR modeling) can be carried out through feedback of the selected variables. This consists in performing the OPS method several times and, in each run, only the selected variables are used as input data in the next run. The procedure is repeated until satisfatory RMSECV and variable numbers are obtained.

4.6. Simulated data

A study using the simulated data described in the experimental session was performed to verify the OPS ability in selecting interpretable regions. Informative regions were selected for all four components. The improvement in the statistical parameters for all models can be confirmed in Table VI when a significant decrease in RMSEP values was observed after feature selection. In all performed selections the informative vector selected was built with *hOPS* > *hMOD*. For analytes 2 and 4 the regression vector was the best informative vector. However, for analytes 1 and 3, its

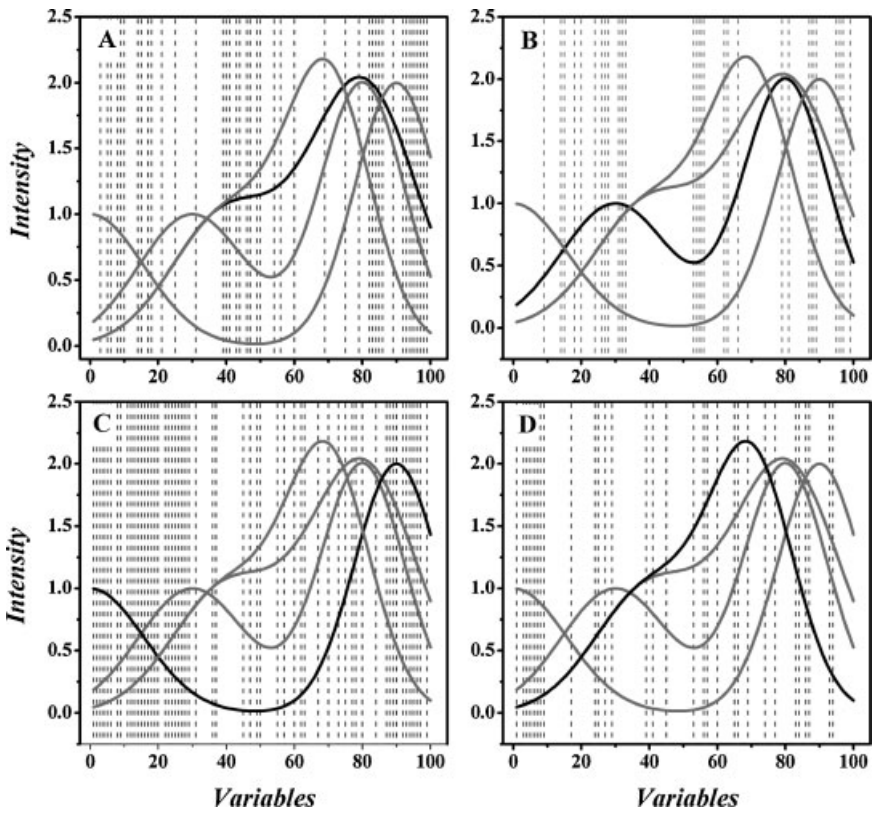

Figure 14. Selected variables for the four-component simulated data. (A) analyte 1, (B) analyte 2, (C) analyte 3 and (D) analyte 4. The black solid line indicates the analyte being studied.

combination with other vectors was necessary for obtaining better results. This makes clear the necessity of studying other related vectors and the combinations among themselves. Figure 14 shows the selected regions and the pure spectrum of the four components. The densest regions for analyte 1 are in the middle and far right side of the pure spectrum and for analyte 3, two informative regions one in the beginning and another at the end were selected, indicating that in both cases the selected variables are indeed informative. For these two analytes, the informative vector was composed by a combination including the regression vector. Variable selection was also very effective for analyte 2. In the case of analyte 4, the best results were given when using the regression vector *per se* and $hOPS = 7$ and $hMOD = 4$. For this data set, although there are several variables selected in the far left side, the results can be considered reasonable for interpretation, since the relevant variables were also selected.

5. CONCLUSIONS

The application of the OPS method enables variable selection in analysis of multivariate data sets with abilities to

(1) improve the model's prediction power;
(2) improve the interpretability of the selected variables;
(3) reduce significantly the number of variables in the final model;
(4) be useful for different data set types and
(5) be a feature selection method, simple and effective.

Due to the criteria presented to select variables, this methodology has shown to be robust and avoid overfitting and chance correlation.

Among all vectors investigated, the regression vector Reg. and the correlated vector NAS are the most promising to sort the best variables. Moreover, a new criterion to choose the number of components to define the Reg. and NAS vectors is presented. Applying this criterion, the number of components selected to define Reg. and NAS is higher than that used to build the model.

The other vectors, VIP, CovProc, SqRes, StN and their combinations, although having lower performance with respect to Reg. and NAS vectors in the examples presented, could be more appropriate for some other specific data types. Besides, other vectors can be introduced in the future.

Although OPS was used in tandem with PLS regression, it can be combined with other regression methods.

Acknowledgements

The authors acknowledge CNPq for financial support and Dr Carol H. Collins for English revision. They thank Dr. Rudolf Kiralj for his significant suggestions and contributions.

REFERENCES

1. Martens H, Naes T. *Multivariate Calibration*. Wiley: New York, 1929; 523–550.
2. Fairchild SZ, Kalivas JH. PCR eigenvector selection based on correlation relative standard deviations. *J. Chemometr.* 2001; **15**: 615–625.
3. Forina M, Lanteri S, Oliveros M, Millan CP. Selection of useful predictors in multivariate calibration. *Anal. Bioanal. Chem.* 2004; **380**: 397–418.
4. Xu L, Schechter I. Wavelength selection for simultaneous spectroscopic analysis: experimental and theoretical study. *Anal. Chem.* 1996; **68**: 2392–2400.
5. Nadler B, Coifman RR. The prediction error in CLS and PLS: the importance of feature selection prior to multivariate calibration. *J. Chemometr.* 2005; **19**: 107–118.
6. Ferreira MMC, Multivariate QSAR. *J. Braz. Chem. Soc.* 2002; **13**: 742–753.
7. Jouanrimbaud D, Walczak B, Massart DL, Last IR, Prebble KA. Comparison of multivariate methods based on latent vectors and methods based on wavelength selection for the analysis of near-infrared spectroscopic data. *Anal. Chim. Acta* 1995; **304**: 285–295.
8. Henrique CM, Teófilo RF, Sabino L, Ferreira MMC, Cereda MP. Classification of cassava starch films by physicochemical properties and water vapor permeability quantification by FTIR and PLS. *J. Food Sci.* 2007; **72**: E184–E189.
9. Jiang JH, Berry RJ, Siesler HW, Ozaki Y. Wavelength interval selection in multicomponent spectral analysis by moving window partial least-squares regression with applications to mid-infrared and near-infrared spectroscopic data. *Anal. Chem.* 2002; **74**: 3555–3565.
10. Centner V, Massart DL, deNoord OE, De Jong S, Vandeginste BM, Sterna C. Elimination of uninformative variables for multivariate calibration. *Anal. Chem.* 1996; **68**: 3851–3858.
11. Herrero A, Ortiz MC. Qualitative and quantitative aspects of the application of genetic algorithm-based variable selection in polarography and stripping voltammetry. *Anal. Chim. Acta* 1999; **378**: 245–259.
12. Hoskuldsson A. Variable and subset selection in PLS regression. *Chemometr. Intell. Lab. Syst.* 2001; **55**: 23–38.
13. Hoskuldsson A. Analysis of latent structures in linear models. *J. Chemometr.* 2003; **17**: 630–645.
14. Chong IG, Jun CH. Performance of some variable selection methods when multicollinearity is present. *Chemometr. Intell. Lab. Syst.* 2005; **78**: 103–112.
15. Helland IS. On the structure of partial least squares regression. *Commun. Stat.: Simul. Comput.* 1988; **17**: 581–607.
16. Williams RP, Swinkels AJ, Maeder M. Identification and application of a prognostic vector for use in multivariate calibration and prediction. *Chemometr. Intell. Lab. Syst.* 1992; **15**: 185–193.
17. Brown PJ. Wavelength selection in multicomponent near-infrared calibration. *J. Chemometr.* 1992; **6**: 151–161.
18. Miller CE. The use of chemometric techniques in process analytical method development and operation. *Chemometr. Intell. Lab. Syst.* 1995; **30**: 11–22.
19. Wold S, Johansson E, Cocchi M. 3D QSAR. In *Drug Design: Theory, Methods, and Applications*, Escom T (ed.). Holland: Leiden, 1993; 523–550.
20. Dodds SA, Heath WP. Construction of an online reduced-spectrum NIR calibration model from full-spectrum data. *Chemometr. Intell. Lab. Syst.* 2005; **76**: 37–43.v
21. Reinikainen SP, Hoskuldsson A. COVPROC method: strategy in modeling dynamic systems. *J. Chemometr.* 2003; **17**: 130–139.
22. Teófilo RF. *Chemometric Methods in the Electrochemical Studies of Phenols on Boron-Doped Diamond Films*. PhD Thesis, Universidade Estadual de Campinas, Campinas, 2007; 1–109.
23. Ribeiro JS, Teófilo RF, Martins JPA, Ferreira MMC. *Melhorias na Análise Discriminatória de Cafés Comerciais Brasileiros Aplicando o Algoritmo de Seleção de Variáveis OPS® Sobre Dados Cromatográficos*. 14° ENQA, João Pessoa, PB, 2007, Abstract 104.
24. Manne R. Analysis of 2 partial-least-squares algorithms for multivariate calibration. *Chemometr. Intell. Lab. Syst.* 1987; **2**: 187–197.
25. Golub GH, Kahan W. Calculating the singular values and pseudo-inverse of a matrix. *SIAM J. Num. Anal. Ser. B.* 1965; **2**: 205–224.
26. Golub GH, Van Loan CF. *Matrix Computation*, (2nd edn). Johns Hopkins University Press: Baltimore, 1989; 498–499.
27. Lorber A, Kowalski BR. A note on the use of the partial least-squares method for multivariate calibration. *Appl. Spectrosc.* 1988; **42**: 1572–1574.
28. Phatak A. *Evaluation of Some Multivariate Methods and Their Applications in Chemical Engineering*, PhD Thesis, University of Waterloo, Ontario, 1993; 1–58.

29. Elden L. Partial least-squares vs. Lanczos bidiagonalization—I: analysis of a projection method for multiple regression. *Comput. Stat. Data Anal.* 2004; **46**: 11–31.
30. Pell RJ, Ramos LS, Manne R. The model space in partial least square regression. *J. Chemometr.* 2007; **21**: 165–172.
31. Ergon R. PLS score-loading correspondence and a bi-orthogonal factorization. *J. Chemometr.* 2002; **16**: 368–373.
32. Mosteller F, Tukey JW. *Data Analysis and Regression: A Second Course in Statistics*. Addison Wesley: London, 1977; 299–331.
33. Montgomery DC, Runger GC. *Applied Statistics and Probability for Engineers*, (3nd edn). Wiley: New York, 2003; 506–555.
34. Wold S, Johansson E, Cocchi M. PLS-partial least squares projections to latent structures. In *3D QSAR in Drug Design*, Kubinyi H (ed.). ESCOM Science Publishers: Leiden, 1993; 523–548.
35. Ferre J, Faber NM. Net analyte signal calculation for multivariate calibration. *Chemometr. Intell. Lab. Syst.* 2003; **69**: 123–136.
36. Faber NM. Efficient computation of net analyte signal vector in inverse multivariate calibration models. *Anal. Chem.* 1998; **70**: 5108–5110.
37. Bro R, Andersen CM. Theory of net analyte signal vectors in inverse regression. *J. Chemometr.* 2003; **17**: 646–652.
38. Kennard RW, Stone LA. Computer aided design of experiments. *Technometrics* 1969; **11**: 137–148.
39. Soyemi OO, Busch MA, Busch KW. Multivariate analysis of near-infrared spectra using the G-programming language. *J. Chem Inf. Comput. Sci.* 2000; **40**: 1093–1100.
40. Dyrby M, Engelsen SB, Norgaard L, Bruhn M, Lundsberg-Nielsen L. Chemometric quantitation of the active substance (containing CN) in a pharmaceutical tablet using near-infrared (NIR) transmittance and NIR FT-Raman spectra. *Appl. Spectrosc.* 2002; **56**: 579–585.
41. Bro R, Rinnan A, Faber NM. Standard error of prediction for multilinear PLS—2: practical implementation in fluorescence spectroscopy. *Chemometr. Intell. Lab. Syst.* 2005; **75**: 69–76.
42. Bahram M, Bro R, Stedmon C, Afkhami A. Handling of Rayleigh and Raman scatter for PARAFAC modeling of fluorescence data using interpolation. *J. Chemometr.* 2006; **20**: 99–105.
43. Pirouette 4.0, Infometrix, Inc., Bothwell, WA, 2007.
44. Kiralj R, Ferreira MMC. A priori molecular descriptors in QSAR: a case of HIV-1 protease inhibitors. I. The chemometric approach. *J. Mol. Graph.* 2003; **21**: 435–448.
45. Holloway M, Wai J, Halgren T, Fitzgerald P, Vacca J, Dorsey B, Levin R, Thompson W, Chen L, Desolms S, Gaffin N, Ghosh A, Giuliani E, Graham S, Guare J, Hungate R, Lyle T, Sanders W. A priori prediction of activity for HIV-1 protease inhibitors employing energy minimization in the active site. *J. Med. Chem.* 1995; **38**: 305–317.
46. Andersen CM, Bro R. Practical aspects of PARAFAC modeling of fluorescence excitation-emission data. *J. Chemometr.* 2003; **17**: 200–215.
47. Wold S, Eriksson L. *Statistical Validation of QSAR Results*. VCH: Weinheim, 1995; 309–318.

