Available online at www.sciencedirect.com
ScienceDirect
Chemometrics and Intelligent Laboratory Systems 92 (2008) 83–91
Chemometrics and intelligent laboratory systems
www.elsevier.com/locate/chemolab

A variable elimination method to improve the parsimony of MLR models using the successive projections algorithm

Roberto Kawakami Harrop Galvão [a], Mário César Ugulino Araújo [b,*], Wallace Duarte Fragoso [b], Edvan Cirino Silva [b], Gledson Emidio José [b], Sófacles Figueredo Carreiro Soares [b], Henrique Mohallem Paiva [c]

[a] Instituto Tecnológico de Aeronáutica, Divisão de Engenharia Eletrônica, 12228-900, São José dos Campos, SP, Brazil
[b] Universidade Federal da Paraíba, CCEN, Departamento de Química, Caixa Postal 5093, CEP 58051-970 — João Pessoa, PB, Brazil
[c] Empresa Brasileira de Aeronáutica (EMBRAER), Flight Control Systems, 12227-901, São José dos Campos, SP, Brazil

Received 16 May 2007; received in revised form 31 October 2007; accepted 21 December 2007
Available online 7 January 2008

Abstract

The successive projections algorithm (SPA) is a variable selection technique designed to minimize collinearity problems in multiple linear regression (MLR). This paper proposes a modification to the basic SPA formulation aimed at further improving the parsimony of the resulting MLR model. For this purpose, an elimination procedure is incorporated to the algorithm in order to remove variables that do not effectively contribute towards the prediction ability of the model as indicated by an *F*-test. The utility of the proposed modification is illustrated in a simulation study, as well as in two application examples involving the analysis of diesel and corn samples by near-infrared (NIR) spectroscopy. The results demonstrate that the number of variables selected by SPA can be reduced without significantly compromising prediction performance. In addition, SPA is favourably compared with classic Stepwise Regression and full-spectrum PLS. A graphical user interface for SPA is available at www.ele.ita.br/~kawakami/spa/.

Keywords: Multiple linear regression; Variable selection; Successive projections algorithm; Near-infrared spectrometry; Diesel analysis; Corn analysis

1. Introduction

The successive projections algorithm (SPA) is a variable selection technique designed to improve the conditioning of multiple linear regression (MLR) by minimizing collinearity effects in the calibration data set. For this purpose, candidate subsets of variables are constructed according to a sequence of projection operations involving the columns of the instrumental response matrix. These candidate subsets are then evaluated according to the prediction performance of the resulting MLR model. In several applications concerning the use of UV–VIS [1,2], ICP-OES [3], FT-IR [4] and NIR spectrometry [4–6], MLR-SPA has been shown to deliver models with good prediction ability when compared to conventional full-spectrum models obtained with partial-least-squares (PLS). SPA has also been successfully employed in other fields such as QSAR (quantitative structure–activity relationships) [7] and classification [8].

The present paper proposes a modification to the original SPA algorithm in order to reduce the number of selected variables without significantly compromising the prediction ability of the resulting MLR model. For this purpose, a final elimination step is incorporated to the algorithm to remove variables that do not effectively contribute towards the prediction ability of the model as indicated by an *F*-test.

A simulation study is presented to clarify the benefits of incorporating the proposed variable elimination procedure to SPA. Moreover, two case studies involving the analysis of diesel and corn samples by near-infrared (NIR) spectroscopy are discussed. In these cases, the results obtained by SPA before and after the elimination procedure are contrasted. The results are also compared with those obtained by classic Stepwise Regression [9]. Furthermore, a comparison with full-spectrum PLS is presented to demonstrate that the use of a reduced number of wavelengths

* Corresponding author. Tel.: +55 83 3216 7438; fax: +55 83 3216 7437.
E-mail address: laqa@quimica.ufpb.br (M.C.U. Araújo).

0169-7439/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2007.12.004

does not impair the prediction ability of the MLR-SPA model. A graphical user interface (GUI) for SPA was developed for the convenience of prospective users.

2. Background and theory

2.1. The successive projections algorithm

In what follows, the instrumental response data are disposed in a matrix $\boldsymbol{X}$ of dimensions $(N \times K)$ such that the kth variable x_k is associated to the kth column vector $\mathbf{x}_k \in \Re^N$. Let $M = \min(N-1, K)$ be the maximum number of variables that can be included in an MLR model with intercept term.

SPA comprises two phases. The first phase consists of projections carried out on the $\boldsymbol{X}$ matrix, which generate K chains of M variables each. Each element in a chain is selected in order to display the least collinearity with the previous ones. The construction of each chain starts from one of the variables x_k, $k = 1, \ldots, K$, and follows the operations described below:

- Step 1 (Initialization): Let

  $\mathbf{z}^1 = x_k$ (vector that defines the initial projection operations)
  $\mathbf{x}_j^1 = \mathbf{x}_j$, $j = 1, \ldots, K$
  $\boldsymbol{SEL}(1, k) = k$
  $i = 1$ (iteration counter)

- Step 2: Calculate the matrix $\boldsymbol{P}^i$ of projection onto the subspace orthogonal to $\mathbf{z}^i$ as

$$\boldsymbol{P}^i = \boldsymbol{I} - \frac{\mathbf{z}^i (\mathbf{z}^i)^T}{(\mathbf{z}^i)^T \mathbf{z}^i} \tag{1}$$

  where $\boldsymbol{I}$ is an $(N \times N)$ identity matrix.
- Step 3: Calculate the projected vectors $\mathbf{x}_j^{i+1}$ as

$$\mathbf{x}_j^{i+1} = \boldsymbol{P}^i \mathbf{x}_j^i \tag{2}$$

  for all $j = 1, \ldots, K$.
- Step 4: Determine the index j^* of the largest projected vector and store this index in element $(i+1, k)$ of the $\boldsymbol{SEL}$ matrix:

$$j^* = \arg \max_{j = 1, \ldots, K} ||\mathbf{x}_j^{i+1}|| \tag{3}$$

$$\boldsymbol{SEL}(i+1, k) = j^* \tag{4}$$

- Step 5: Let $\mathbf{z}^{i+1} = \mathbf{x}_{j^*}^{i+1}$ (vector that defines the projection operations for the next iteration)
- Step 6: Let $i = i+1$. If $i < M$ return to Step 2.

The second phase of SPA consists of evaluating candidate subsets of variables extracted from the chains generated in the first phase. The candidate subset of m variables starting from x_k is defined by the index set $\{\boldsymbol{SEL}(1, k), \boldsymbol{SEL}(2, k), \ldots, \boldsymbol{SEL}(m, k)\}$. Since m ranges from one to M and k ranges from one to K, a total of $M \times K$ subsets of variables are tested. Different prediction performance metrics [10] could be used to choose the best variable subset. For this purpose, the present work adopts the RMSEV (root mean square error of validation) value obtained by applying the resulting MLR model to an independent validation set, as in previous papers [1–6].

2.2. Proposed variable elimination procedure

The proposed elimination procedure can be described as follows. Let $\{v_1, v_2, \ldots, v_L\}$ be the subset of L variables that were selected in Phase 2 described above and let $\{b_1, b_2, \ldots, b_L\}$ be the corresponding regression coefficients obtained by building an MLR model. It is worth noting that L is usually much smaller than the original number K of wavelengths in the full spectrum ($L \ll K$) [1–6]. A relevance index r_j is then calculated for each variable v_j as

$$r_j = s_{v_j} |b_j|, \quad j = 1, 2, \ldots, L \tag{5}$$

where s_{v_j} is the standard deviation of variable v_j obtained in the calibration data set.

It is worth noting that the idea of using the regression coefficients for variable selection has already been proposed in other settings. Williams, Swinkels and Maeder [11] employed the regression coefficients in the wavelength domain as a "prognostic vector" to identify spectral features more closely related to the property being modeled by PCR and PLS. Chong and Jun [12] stated that the magnitude of those coefficients can be used as a simple criterion for variable selection in PLS. A relevance metric similar to the index defined in Eq. (5) also exists in the context of classification models. More specifically, Altman [13] employed the product of the standard deviation of the variable by its linear discriminant coefficient to assess the relevance of financial ratios for prediction of corporate bankruptcy.

The rationale behind the relevance index r_j for an MLR model is the following. If the values of v_j display large variations from sample to sample (as indicated by a large standard deviation over the calibration set) and the corresponding regression coefficient b_j is large, variable v_j has a large contribution towards the overall "mean square due to regression" [9]. If the x-variables are auto-scaled prior to the regression, it can be easily shown that the value of r_j remains the same. Moreover, if the y-variable is auto-scaled, all regression coefficients would be divided by a common factor (the standard deviation of y) and thus the relative relevance of the x-variables would remain the same.

It is important to note that, in general, the relevance index in Eq. (5) could not be computed before the initial selection carried out by SPA because the determination of the regression coefficients by MLR would be ill-conditioned.

Let $\{v_1', v_2', \ldots, v_L'\}$ denote the sequence of variables disposed in decreasing order of relevance, as indicated by the metric defined in Eq. (5). A sequence of RMSEV values is then calculated in the following manner (cross-validation can be employed if a separate validation set is not available):

For $j = 1$ to L
  Build an MLR model with variables $\{v_1', v_2', \ldots, v_j'\}$
  Apply this model to a validation set
  Calculate the resulting RMSEV(j) value
Next j

Let $RMSEV_{min}$ be the minimum value of the RMSEV sequence thus obtained. Finally, the minimum number of variables for which the RMSEV value is not significantly larger than $RMSEV_{min}$ is adopted. For this purpose, an F-test is employed to compare the squared RMSEV values. This criterion is similar to the method of Haaland and Thomas [14], which is used to determine an appropriate number of latent variables for PLS. A significance level $\alpha = 0.25$ for the F-test can be adopted as suggested elsewhere [15].

3. Simulation study

In this section, a simulation study is presented to illustrate the potential benefits of the proposed variable elimination method. For this purpose, simulated spectra were generated by assuming a linear relation between the matrix $\boldsymbol{X}$ of instrumental responses and the matrix $\boldsymbol{Y}$ of concentrations for three analytes:

$$\boldsymbol{X} = \boldsymbol{YW} + \boldsymbol{N} \tag{6}$$

where $\boldsymbol{N}$ is a noise term. The element at line i and column k of matrix $\boldsymbol{W}$ corresponds to the proportionality coefficient for the ith analyte ($i = 1, 2, 3$) at the kth spectral bin ($k = 1, \ldots, 300$). The adopted $\boldsymbol{W}$-values for the three analytes, termed A, B, and C are presented in Fig. 1a.

A calibration data set comprising 27 samples was generated according to a full factorial design with three levels (1.0, 5.5, 10.0) for the concentration values ($\boldsymbol{Y}$ matrix). The validation set consisted of 27 samples specified according to another full factorial design with three levels (2.0, 5.5, 9.0). Finally, a prediction set of 100 samples was generated by using concentration values randomly distributed within the calibration range. White, zero-mean, homoscedastic Gaussian noise with standard deviation of 0.1 was added to all mixture spectra as in Eq. (6). The resulting spectra for the calibration, validation, and prediction sets are presented in Fig. 1b, c, and d, respectively.

It is worth noting that the validation set is employed to guide the selection of candidate subsets of variables in Phase 2 of SPA, as discussed in Section 2.1. It is also used in the proposed variable elimination procedure described in Section 2.2. The prediction set is employed in the final performance assessment of the resulting models. It is not used in any step of the calibration and validation procedures.

Fig. 2(a–c) presents the scree plots obtained for each analyte by applying SPA and using the relevance index described in Section 2.2. As can be seen, the initial part of the RMSEV curve displays a sharp fall as the number of variables is increased from one to three because at least three variables are required to resolve the spectral overlapping features of the analytes. The RMSEV values continue to decrease after that point but the improvement becomes marginal as the number of variables is further increased

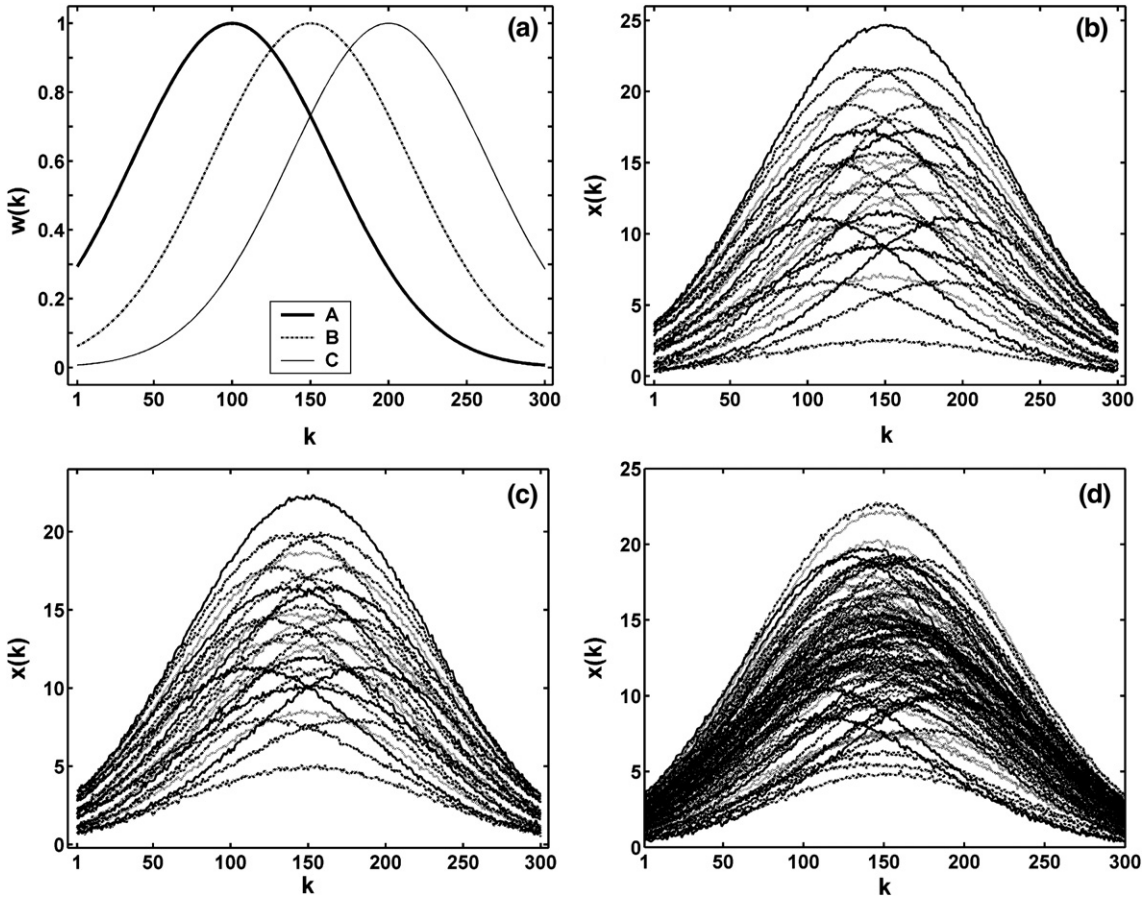

Fig. 1. (a) Pure spectra for analytes A, B, C and mixture spectra for (b) calibration, (c) validation, and (d) prediction.

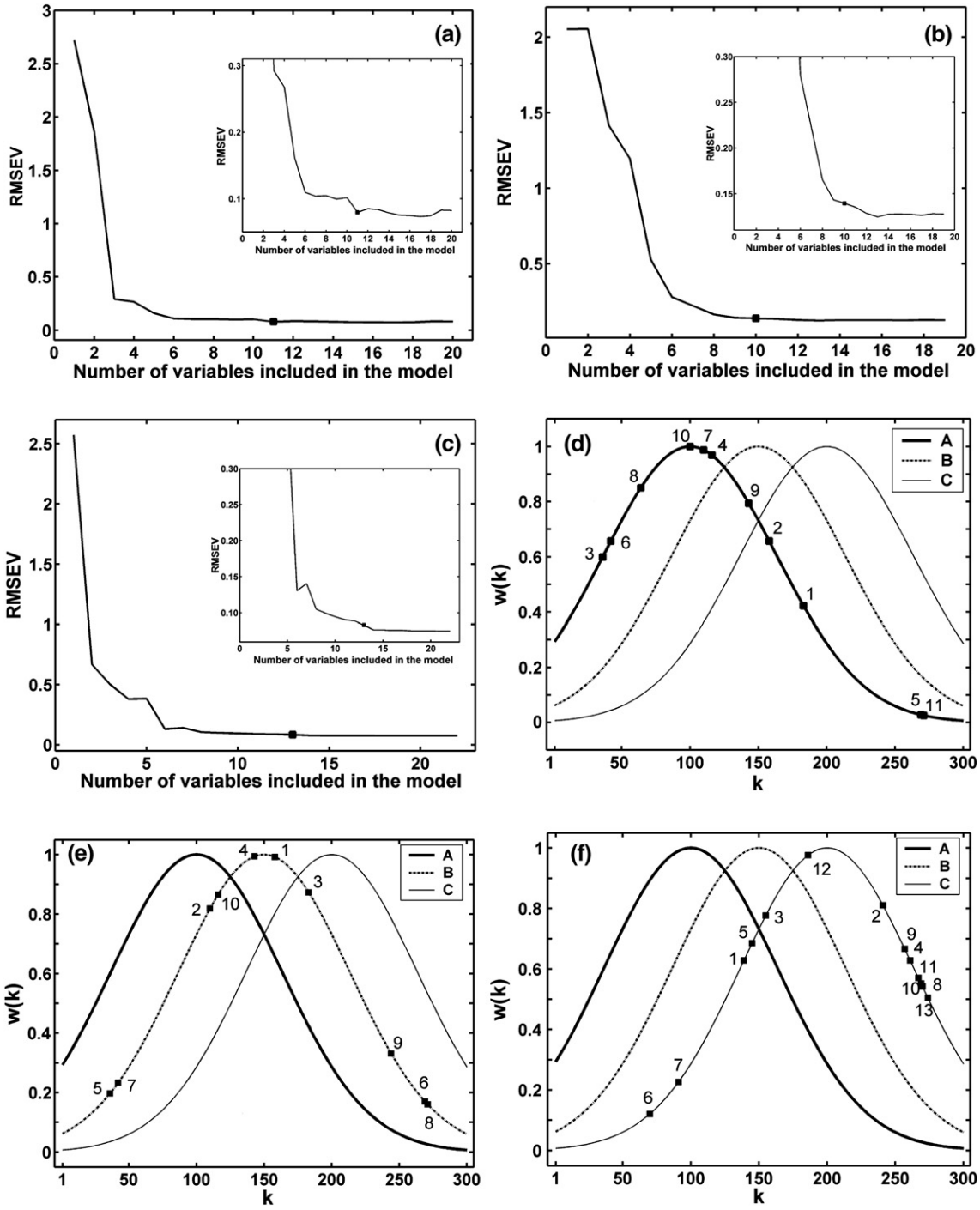

Fig. 2. RMSEV scree plots obtained for analytes A, B, C (graphs (a), (b), (c), respectively). In each plot, the inset presents an expanded view of the region around the selected point (black square). The selected variables for each analyte are presented in (d), (e), (f).

and thus the curve tends to level off. It is worth noting that the validation set employed to generate the scree plots is the same that was used in Phase 2 of SPA to choose a suitable subset of variables. That explains why the RMSEV curve levels off instead of progressively increasing after a certain point.

For this reason, some criterion is required to determine a point in the scree plot beyond which the minor RMSEV improvements do not justify the loss of model parsimony. The *F*-test criterion described in Section 2.2 (with $\alpha = 0.25$) leads to the number of variables indicated in Fig. 2a–c by a black square symbol. The resulting variables are indicated in Fig. 2d–f in order of relevance (i.e., number 1 corresponds to the first variable in the scree plot).

An inspection of Fig. 2d–f sheds additional light on the interpretation of the scree plots. In fact, the sharp fall from two to three variables in Fig. 2a (analyte A) occurs because the third variable (indicated by number 3 in Fig. 2d) is located in a region in which the interference of analytes B and C is relatively smaller. For the same reason, the largest fall in Fig. 2c (analyte C) occurs when the second variable is included. On the other hand, Fig. 2b

does not display such a feature because there is no region in which the spectral signature of analyte B is considerably dominant over the interferents. For this reason, the scree plot for analyte B has a more gradual descent as more variables are introduced.

In addition, it is worth noting that the scree plots tend to level off when adjacent variables start to be included in the model. For instance, in Fig. 2a, the decrease in RMSEV is noticeably smaller after the sixth variable is included. In fact, as can be seen in Fig. 2d, the seventh variable is very close to the fourth. Of course, a non-subjective criterion, such as the *F*-test adopted in this paper, is needed to automate the choice of an appropriate point in the scree plot. The significance level of the *F*-test criterion could be adjusted to change the number of selected variables. However, since the best value of such parameter will depend on the nature of the data set, a more detailed study of this aspect was not carried out in the present simulated study. The value $\alpha=0.25$ suggested in Section 2.2 led to appropriate results for the experimental data sets, as will be seen in Section 5.

For comparison purposes, Table 1 shows the number of variables selected by SPA for each analyte before and after the proposed elimination procedure. As can be seen, a considerable reduction (almost one-half) in the number of selected variables was attained for all three analytes. This table also presents the RMSEP (root mean square error of prediction) values obtained when the resulting MLR models were applied to the independent prediction set, which was not employed in the variable selection procedures. The results show that the gains in parsimony (reduction in the number of variables) were obtained without compromising the prediction ability of the resulting models. In fact, according to an *F*-test at a confidence level of 95%, the small increase in the RMSEP value for each analyte is not significant. The worst prediction results (both before and after the variable elimination procedure) were obtained for analyte B. Such a finding can be ascribed to the fact that the signature of this analyte is considerably overlapped by analytes A and C along the entire spectral range employed in the simulation. This problem is less severe for analytes A and C, which are less affected by interferences in the left and right-hand sides of the spectral interval, respectively.

4. Experimental

4.1. Diesel data set

One hundred and seventy diesel samples were collected from gas stations in the city of Recife (Pernambuco State, Brazil). NIR spectra were obtained using an FT-NIR/MIR spectrometer Perkin Elmer GX equipped with a Hellma® 130-QS quartz flow-through cell presenting an optical path length of 1.0 cm. A spectral resolution of 2 cm^{-1} and 16 scans were used. The NIR region in the range 880–1675 nm was adopted for this study.

The reference values for sulphur content and distillation temperatures (initial point IP, *T*10% and *T*90%) were obtained according to the ASTM (American Society for Testing and Materials) 4294-90 and D86 standards, respectively.

The reference values for sulphur content were determined by using energy-dispersive X-ray fluorescence. For this purpose, a Spectro Titan spectrophotometer (current of 400 μA, tube voltage of 5.5 kV and irradiation time of 300 s) was employed. The reference values for distillation temperatures were determined by using a Herzog HDA 628 automatic distiller.

Table 1
Number of selected variables (*m*) and RMSEP values obtained in the prediction set before and after the variable elimination procedure

Analyte	A		B		C	
	RMSEP	*m*	RMSEP	*m*	RMSEP	*m*
Before	0.14	20	0.22	19	0.11	22
After	0.15	11	0.23	10	0.12	13

4.2. Corn data set

This data set is publicly available at www.eigenvector.com/Data/Corn/. It consists of NIR spectra from 80 corn samples, which were acquired in the range 1100–2498 nm by using three spectrometers. In this study, only the data from spectrometer "m5" were employed. The data set also includes moisture, oil, protein and starch content values for each sample.

4.3. Sample set partitioning

The SPXY algorithm [16] was used to divide the available samples into calibration, validation, and prediction sets. The diesel data were divided into 70 (calibration), 50 (validation), and 50 (prediction) samples. The corn data were divided into 40 (calibration), 20 (validation), and 20 (prediction) samples. These sets are used for model-building and performance evaluation purposes as in the simulated example presented in Section 3.

4.4. Stepwise regression

The stepwise regression (SR) method adopted for comparison with SPA is a classic formulation, which combines forward inclusion and backward elimination procedures, as described in [9]. The algorithm starts from the *x*-variable with the largest correlation with the dependent variable *y*. Each subsequent iteration of the selection procedure comprises an inclusion phase followed by an exclusion phase, which are guided by partial *F*-tests for the *x*-variables [9].

Seven different *F*-test significance levels ($\alpha=0.01$, 0.02, 0.05, 0.10, 0.15, 0.20, 0.25) were tested for each property *y* under consideration. In each case, the best value of α was selected according to the same RMSEV criterion employed in SPA.

4.5. Software

Savitzky–Golay differentiation and PLS modelling were performed by using The Unscrambler® 9.6 (CAMO AS, Oslo, Norway). The number of latent variables for PLS was determined on the basis of the validation error by using the default settings of the software. SPA variable selection, Stepwise Regression, MLR modelling and SPXY sample selection were implemented in MATLAB® 6.5. Function finv.m from the Matlab Statistics Toolbox is required for the *F*-test

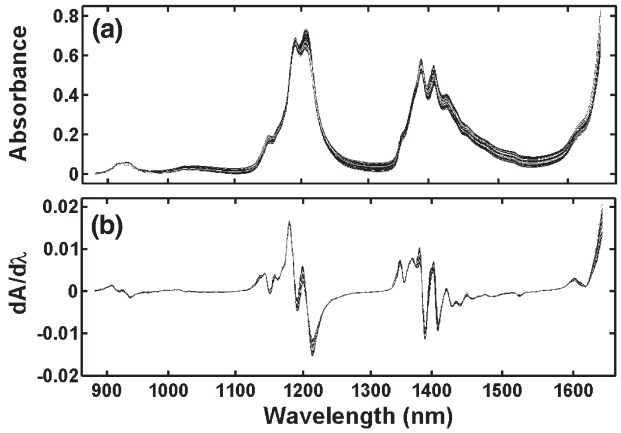

Fig. 3. (a) Original and (b) derivative NIR spectra from the 170 diesel samples.

criterion in the variable elimination procedure, as well as in the Stepwise Regression method.

5. Results and discussion

5.1. Diesel analysis

Fig. 3a presents the raw spectra from the 170 diesel samples. As can be seen, the spectra display undesirable baseline features, which are often found in FT-NIR instruments and are not correlated to the parameters under analysis. For this reason, first derivative spectra were calculated with a Savitzky–Golay filter [17] using a 2nd-order polynomial and a 13-point window, as shown in Fig. 3b. The resulting derivative spectra comprised 1579 variables and were employed throughout the work.

Fig. 4 presents the RMSEV scree plots obtained according to the proposed variable elimination procedure for the four diesel parameters under consideration (sulphur content, IP, *T*10%, *T*90%). As in the simulation example, the scree plots tend to level off after a certain number of variables is added to the model. The result of applying the *F*-test criterion with $\alpha = 0.25$ is indicated in each graph by a square symbol.

The middle graph in Fig. 5 indicates the variables selected by SPA for *T*10%. The variables discarded by the elimination procedure and the retained variables are marked by white and black squares, respectively. Similar results were obtained for sulphur content, IP and *T*90%. As can be seen, all retained variables are associated to absorbance bands, that is, no variables were selected in uninformative regions. It is worth noting that the elimination procedure removed variables in the short NIR region that had been selected by SPA. This region contains absorption features, but the signal-to-noise ratio is relatively poor. Therefore, the elimination is justifiable. For comparison, the variables selected by SR are presented at the

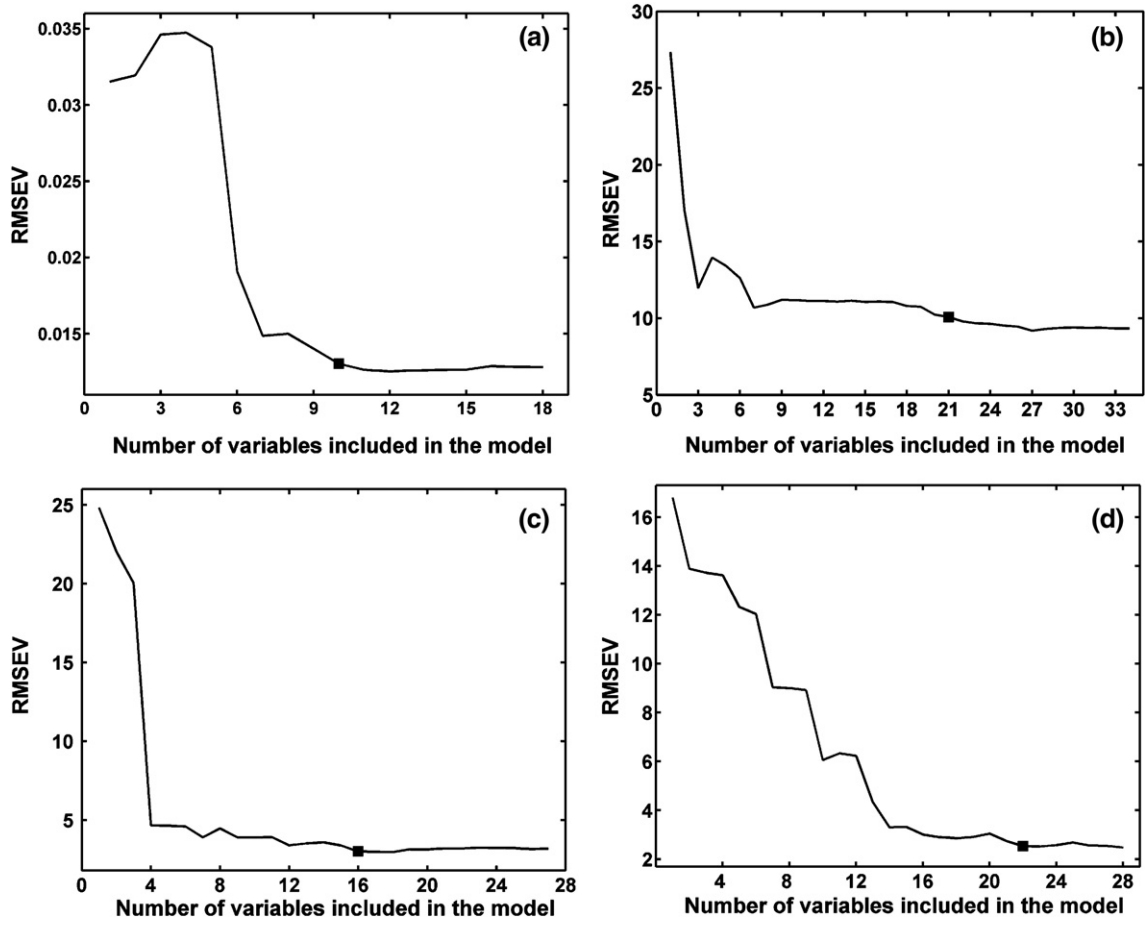

Fig. 4. RMSEV scree plots of SPA for (a) sulphur content, (b) IP, (c) *T*10%, and (d) *T*90%.

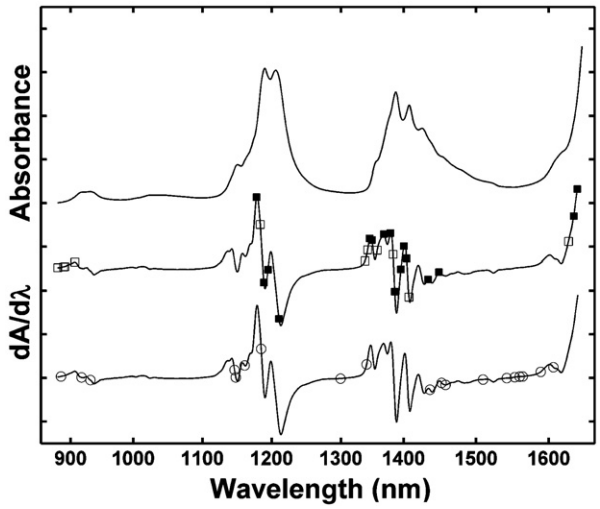

Fig. 5. Wavelengths selected by SPA (middle graph, square markers) and SR (bottom graph, circle markers) for *T*10%. The wavelengths discarded and retained by the elimination procedure in SPA are indicated by white and black square markers, respectively. An original spectrum is also presented for comparison (top graph).

bottom graph of Fig. 5 (circle markers). As can be seen, most of these variables are different from those obtained with SPA.

Table 2 presents the RMSEP values obtained when the resulting MLR-SPA, MLR-SR and PLS models were applied to the independent prediction set, which was not employed in the variable selection procedures. On the overall, the use of MLR-SPA yielded slightly better results as compared to the other two techniques. The results in Table 2 also reveal that the proposed elimination procedure did not compromise the prediction ability of the MLR model, which corroborates the findings of the simulation study. For IP and *T*90%, the use of fewer variables even provided a small reduction in RMSEP, which shows that the proposed modification in SPA may be advantageous not only to obtain simpler models, but also to improve prediction.

5.2. Corn analysis

Fig. 6a presents the raw spectra from the 80 corn samples. As in the diesel case study, the spectral baseline shifts were removed by using a first derivative procedure, as shown in Fig. 6b. For this purpose, a Savitzky–Golay filter using a 2nd-order polynomial and a 21-point window was employed. The resulting derivative spectra comprised 680 variables and were employed throughout the work.

Table 2
RMSEP results for sulphur content, IP, *T*10%, and *T*90%

Parameter	Range	MLR-SPA		PLS	MLR-SR
		Before elimination	After elimination		
Sulphur	0.03–0.31 w/w	0.01 (16)	0.01 (10)	0.02 (05)	0.03 (10)
IP	142.2–240.7 °C	10.3 (34)	9.2 (21)	9.5 (06)	12.4 (06)
*T*10%	186.6–269.9 °C	2.8 (27)	3.0 (16)	4.9 (06)	5.0 (19)
*T*90%	317.2–385.5 °C	3.7 (28)	3.5 (22)	5.3 (04)	3.4 (08)

The values in parentheses correspond to the number of latent variables in PLS and wavelengths in MLR-SPA and MLR-SR. In all cases, the best value of α for MLR-SR was 0.02.

Fig. 7 presents the RMSEV scree plots obtained by applying the variable elimination procedure for the parameters moisture, oil, protein, and starch. As in all previous cases, the scree plots tend to level off as more variables are added to the model. Again, the *F*-test criterion with $\alpha = 0.25$ was employed to obtain a suitable number of variables, which is indicated by a square symbol in each graph. As can be seen in the middle graph of Fig. 8, the retained variables for starch content are associated to absorption bands distributed along the entire NIR spectrum. Similar results were obtained for moisture, oil, and protein. For comparison, the variables selected by SR are presented at the bottom graph of Fig. 5 (circle markers). Again, there is not a clear correspondence between these variables and those obtained with SPA.

Table 3 presents the RMSEP values obtained when the resulting MLR-SPA, MLR-SR and PLS models were applied to the independent prediction set. As can be seen, MLR-SPA noticeably outperformed PLS for moisture, protein, and starch. The oil results were similar for both modeling techniques. MLR-SPA also yielded better results than MLR-SR for three of the four parameters (moisture, oil, starch). The variable elimination procedure provided a considerable reduction in the number of variables. As in the diesel case study, such a reduction did not compromise the prediction ability of the resulting MLR-SPA models and even granted a small reduction in RMSEP for protein and starch.

6. Conclusions

This paper presented a variable elimination method intended to improve the parsimony of MLR models obtained by the successive projections algorithm. The results obtained in a simulation study, as well as in application examples involving NIR data, demonstrate that the number of variables selected by SPA can indeed be reduced without compromising prediction performance. Therefore, the proposed method can be considered a valid and useful improvement to the basic SPA formulation. It

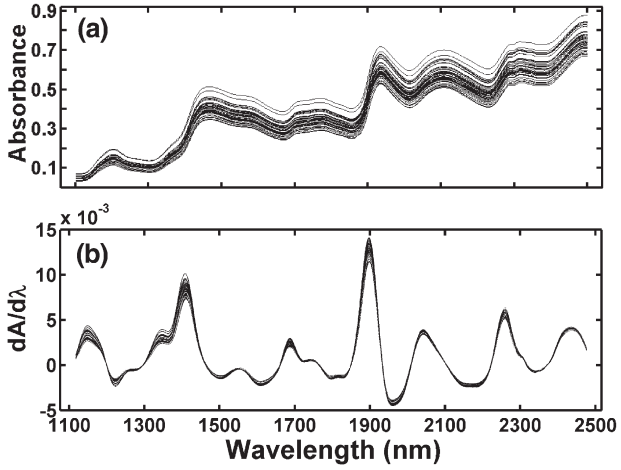

Fig. 6. (a) Raw and (b) derivative NIR spectra from the 80 corn samples.

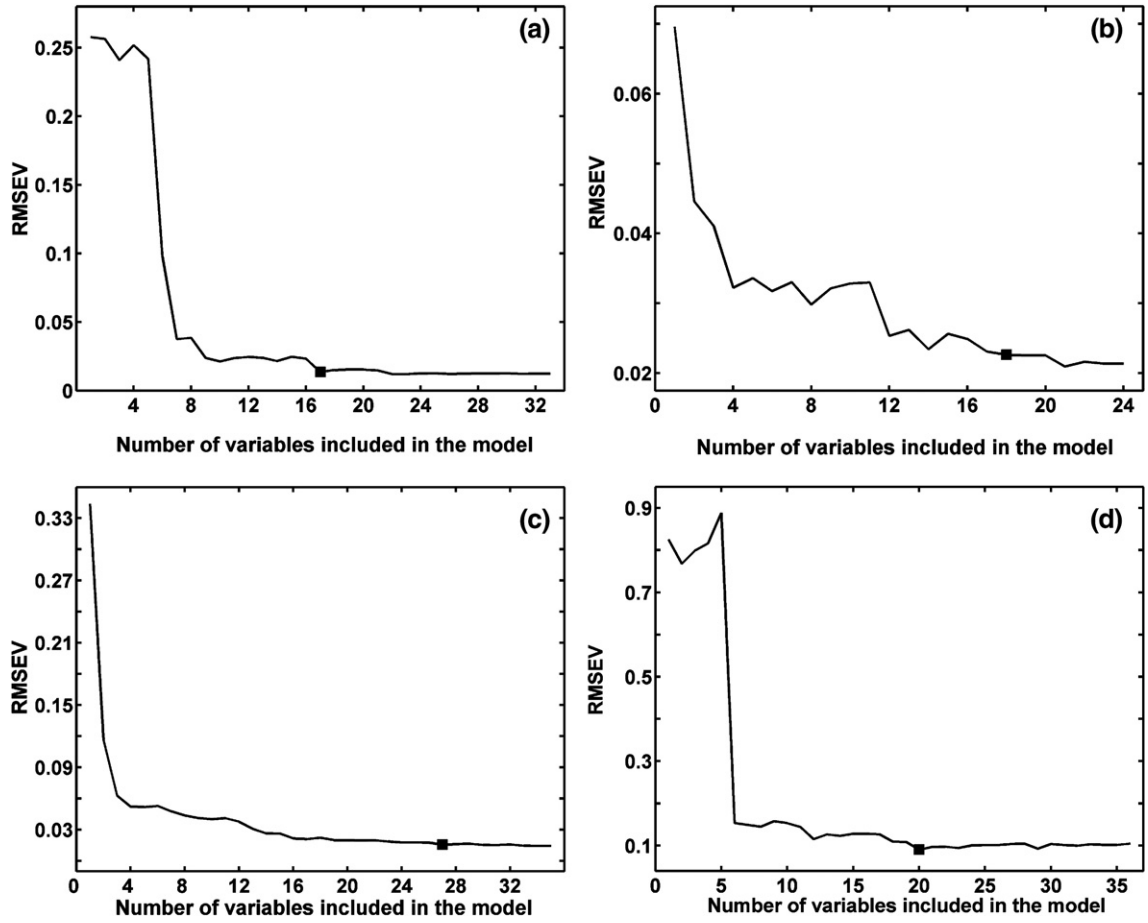

Fig. 7. RMSEV scree plots of SPA for (a) moisture, (b) oil, (c) protein, and (d) starch contents.

is worth noting that the resulting MLR-SPA models were comparable (and in some cases superior) to PLS in terms of prediction ability, as measured by the RMSEP values obtained in an independent set. In most cases, the results of MLR-SPA were also better than those obtained with stepwise regression.

The results showed that a significance level $\alpha = 0.25$ in the proposed elimination procedure was appropriate for the analytical problems under consideration. Future works may carry out a more extensive investigation concerning the choice of α for different data sets. In addition, the use of other relevance criteria to rank the variables may also be explored. The t statistic defined as the ratio between the regression coefficient and the corresponding standard error of estimation could be particularly useful for this purpose. Finally, it may be interesting to investigate direct statistical tests for the relevance index instead of using an F-test based on RMSEV values.

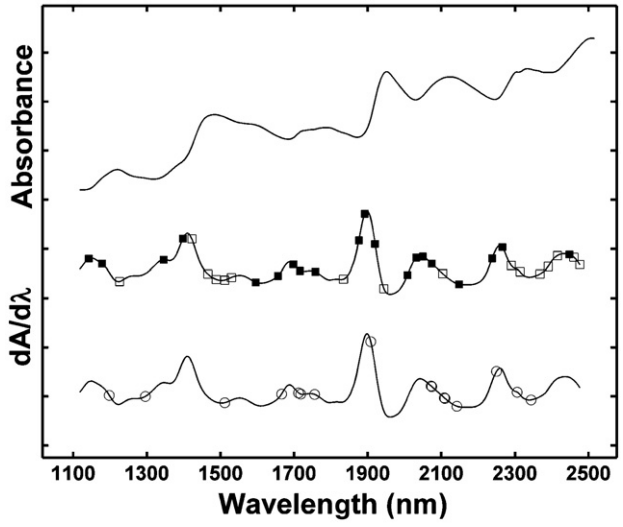

Fig. 8. Wavelengths selected by SPA (middle graph, square markers) and SR (bottom graph, circle markers) for starch content. The wavelengths discarded and retained by the elimination procedure in SPA are indicated by white and black square markers, respectively. An original spectrum is also presented for comparison (top graph).

Table 3
RMSEP results for moisture, oil, protein and starch contents

Parameter	Range (%)	MLR-SPA		PLS	MLR-SR
		Before elimination	After elimination		
Moisture	9.377–10.993	0.019 (33)	0.019 (17)	0.045 (06)	0.026 (09)
Oil	3.088–3.832	0.029 (24)	0.030 (18)	0.028 (10)	0.039 (13)
Protein	7.654–9.711	0.037 (35)	0.033 (27)	0.110 (07)	0.012 (18)
Starch	62.826–66.472	0.127 (36)	0.101 (20)	0.228 (05)	0.129 (16)

The values in parentheses corresponds to the number of latent variables in PLS and wavelengths in MLR-SPA and MLR-SR. The best values of α for MLR-SR were 0.05 (moisture), 0.10 (protein, starch) and 0.15 (oil).

Acknowledgments

The authors thank PROCAD/CAPES (Grant 0081/05-1) and CNPq (Grant 475204/2004-2) for partial financial support. The research fellowships and scholarships granted by CNPq are also gratefully acknowledged. The authors are also indebted to Mr. Cláudio Vicente Ferreira (Laboratório de Combustíveis, Departamento de Engenharia Química, Universidade Federal de Pernambuco) for providing the diesel samples and the reference values of the quality parameters employed in this study.

References

[1] M.C.U. Araujo, T.C.B. Saldanha, R.K.H. Galvão, T. Yoneyama, H.C. Chame, V. Visani, Chemom. Intell. Lab. Syst., Lab. Inf. Manag. 57 (2001) 65–73.
[2] H.A. Dantas Filho, E.S.O.N. Souza, V. Visani, S.R.R.C. Barros, T.C.B. Saldanha, M.C.U. Araújo, R.K.H. Galvão, J. Braz. Chem. Soc. 16 (2005) 58–61.
[3] R.K.H. Galvão, M.F. Pimentel, M.C.U. Araujo, T. Yoneyama, V. Visani, Anal. Chim. Acta 443 (2001) 107–115.
[4] F.A. Honorato, R.K.H. Galvão, M.F. Pimentel, B.B. Neto, M.C.U. Araújo, F.R. Carvalho, Chemom. Intell. Lab. Syst., Lab. Inf. Manag. 76 (2005) 65–72.
[5] M.C. Breitkreitz, I.M. Raimundo Jr., J.J.R. Rohwedder, C. Pasquini, H.A. Dantas Filho, G.E. José, M.C.U. Araújo, Analyst 128 (2003) 1204–1208.
[6] H.A.D. Dantas Filho, R.K.H. Galvão, M.C.U. Araújo, E.C. Silva, T.C.B. Saldanha, G.E. José, C. Pasquini, I.M. Raimundo Jr., J.J.R. Rohwedder, Chemom. Intell. Lab. Syst., Lab. Inf. Manag. 72 (2004) 83–91.
[7] Y. Akhlaghi, M. Kompany-Zareh, J. Chemom. 20 (2006) 1–12.
[8] M.J.C. Pontes, R.K.H. Galvão, M.C.U. Araújo, P.N.T. Moreira, O.D. Pessoa Neto, G.E. José, T.C.B. Saldanha, Chemom. Intell. Lab. Syst., Lab. Inf. Manag. 78 (2005) 11–18.
[9] N.R. Draper, H. Smith, Applied Regression Analysis, 3rd ed.Wiley, New York, 1998.
[10] T. Naes, T. Isaksson, T. Fearn, T. Davies, A User-friendly Guide to Multivariate Calibration and Classification, NIR Publications, Chichester, 2002.
[11] R.P. Williams, D.A.J. Swinkels, M. Maeder, Chemom. Intell. Lab. Syst., Lab. Inf. Manag. 15 (1992) 185–193.
[12] I.G. Chong, C.H. Jun, Chemom. Intell. Lab. Syst., Lab. Inf. Manag. 78 (2005) 103–112.
[13] E. Altman, J. Finance 23 (1968) 589–609.
[14] D.M. Haaland, E.V. Thomas, Anal. Chem. 60 (1988) 1193–1202.
[15] B.X. Li, D.L. Wang, C.L. Xu, Z.J. Zhang, Microchem. J. 149 (2005) 205–212.
[16] R.K.H. Galvão, M.C.U. Araújo, G.E. José, M.J.C. Pontes, E.C. Silva, T.C. B. Saldanha, Talanta 67 (2005) 736–740.
[17] K.R. Beebe, R.J. Pell, B. Seasholtz, Chemometrics — A Practical Guide, Wiley, New York, 1998.