The successive projections algorithm

Sófacles Figueredo Carreiro Soares, Adriano A. Gomes,
Arlindo Rodrigues Galvão Filho, Mario Cesar Ugulino Araujo,
Roberto Kawakami Harrop Galvão

The successive projections algorithm (SPA) is a variable-selection technique that has attracted increasing interest in the analytical-chemistry community in the past 10 years. The present review presents the basic features of SPA for Multiple Linear Regression (MLR) and Linear Discriminant Analysis (LDA) and reports some variants that have been proposed for sample selection, calibration transfer and Quantitative Structure-Activity Relationship (QSAR) and Quantitative Structure-Property Relationship (QSPR) studies. We also discuss computational and pre-processing issues. By way of illustration we present two case studies involving near-infrared determination of protein in wheat and voltammetric classification of vegetable oils. The code employed in this article is freely available from us upon request.

Keywords: Calibration transfer; Classification; Linear Discriminant Analysis (LDA); Multiple Linear Regression (MLR); Multivariate calibration; Quantitative Structure-Activity Relationship (QSAR); Quantitative Structure-Property Relationship (QSPR); Sample selection; Successive projections algorithm (SPA); Variable selection

Sófacles Figueredo Carreiro Soares, Adriano A. Gomes, Mario Cesar Ugulino Araujo*
Universidade Federal da Paraíba, CCEN, Departamento de Química, Caixa Postal 5093, CEP 58051-970, João Pessoa, PB, Brazil

Arlindo Rodrigues Galvão Filho, Roberto Kawakami Harrop Galvão
Instituto Tecnológico de Aeronáutica, Divisão de Engenharia Eletrônica, 12228-900, São José dos Campos, SP, Brazil

1. Introduction

Modern analytical methods typically employ instrumental techniques to analyze solid, liquid or gaseous samples with fewer chemical treatments and reduced waste generation. The instruments employed for this purpose usually involve many analytical channels, so they generate data sets with a considerable number of variables. Examples include laser-induced breakdown spectroscopy (LIBS) [1] and near-infrared spectroscopy (NIR) [2], which deliver measurements over a large number of wavelengths for each sample. However, in many cases, the instrumental response exhibits strong correlation over different analytical channels, which leads to redundancy in the acquired data. Moreover, some channels may not provide relevant information for the problem under consideration and their use may even compromise the precision and the accuracy of the result [3]. For these reasons, the analytical method may benefit from the use of a reduced subset of channels, rather than the entire set of instrumental measurements obtained for each sample. In addition, the identification of an appropriate subset of channels facilitates the interpretation of the results and may be useful to guide the design of less costly instruments that are dedicated to the analytical application at hand [4]. In the chemometrics literature, this procedure is termed variable selection.

Variable selection generally benefits from *a priori* knowledge about the physical and chemical properties of the system under analysis and the technical features of the measurement instrument (e.g., in spectroscopy, the analyst should exclude wavelength regions in which the measured signal saturates, the signal-to-noise ratio of the detector is too small, or the analyte response is strongly overlapped by interferents). However in some cases, the decision is not so clear cut, which motivates the use of chemometrics techniques.

A pragmatic approach to variable selection involves using a computational method to search for the combination of variables that optimizes some performance index related to the analytical result [5]. This index is usually termed cost function when the optimization involves the search for a minimum value. Examples include minimization of the root-mean-square error of prediction (RMSEP) or cross-validation (RMSECV) in

*Corresponding author.
Tel.: +55 83 3216 7438;
Fax: +55 83 3216 7437;
E-mail: laqa@quimica.ufpb.br

0165-9936/$ - see front matter © 2012 Elsevier Ltd. All rights reserved. doi:http://dx.doi.org/10.1016/j.trac.2012.09.006

multivariate calibration [5,6] and the error rate in classification [7].

In principle, such optimization could be accomplished by an exhaustive search procedure (i.e. by evaluating the cost function for each and all combinations of variables). However, the computational effort involved in this process may be prohibitive, even for problems of modest dimensions. Indeed, if K variables are available, there are $(2^K - 1)$ possible subsets that can be formed with one up to K variables. For this reason, several methods have been proposed as alternatives to exhaustive search {e.g., genetic algorithms [8], generalized simulated annealing [9], tabu search [10], ant colonies [11], and the successive projections algorithm (SPA)}.

In SPA, the selection of variables is cast in the form of a combinatorial optimization problem with constraints. The optimization is said to be constrained because the search is restricted to a reduced number of variable subsets, which are formed according to a sequence of projection operations involving the matrix of instrumental responses. The number of cost-function evaluations is therefore much smaller than in an exhaustive search [4]. Moreover, the projection operations are used to choose subsets of variables with a small degree of multi-collinearity in order to minimize redundancy and ill-conditioning problems.

SPA was initially proposed for the construction of multivariate-calibration models [12], and it was subsequently extended to address classification problems [13]. Applications have involved different instrumental techniques and samples, as summarized in Tables 1 and 2, for multivariate-calibration and classification problems, respectively.

The present review presents the basic features of SPA for both multivariate calibration and classification. We also show that, through simple modifications to the original formulation, SPA has been used to handle sample selection and calibration-transfer problems, and non-linear Quantitative Structure-Activity Relationship (QSAR) and Quantitative Structure-Property Relationship (QSPR) studies. We also discuss the main computational and pre-processing issues related to the use of SPA. By way of illustration, we present two case studies involving NIR determination of protein in wheat and voltammetric classification of vegetable oils.

2. SPA for multivariate calibration

Within the scope of multivariate calibration, SPA is aimed at selecting variables for use in multiple linear regression (MLR) models. In this context, the collinearity-avoidance mechanism embedded in SPA is of value to reduce the propagation of measurement noise in the calibration process and to obtain MLR models that are statistically stable [65]. In what follows, the term SPA-MLR is employed with reference to the variable-selection algorithm and the final multivariate-calibration model.

As usual in multivariate-calibration problems, it is assumed that a set of modeling samples is available with the respective instrumental responses and reference values for the property of interest. Moreover, it is assumed that these samples have been split into a calibration set and a validation set with N_{cal} and N_{val} samples, respectively.

SPA-MLR comprises three phases [4,26]. In Phase 1, the instrumental responses of the calibration samples are disposed in a matrix $\mathbf{X_{cal}}$ of dimensions $(N_{cal} \times K)$, such that the k^{th} variable x_k is associated with the k^{th} column vector, $\mathbf{x}_k \in \Re^{N_{cal}}$. These column vectors are subjected to a sequence of projection operations that result in the creation of K chains of variables. The kth chain is initialized with variable x_k and is progressively augmented with variables that display the least collinearity with the previous ones. Such a collinearity is evaluated in terms of the associated column vectors, as depicted in Fig. 1a for a simple case in which $N_{cal} = 3$ samples and $K = 5$ variables. In this case, matrix $\mathbf{X_{cal}}$ comprises five column vectors $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_3$, $\mathbf{x}_4$, $\mathbf{x}_5$. Each of these vectors has three components, which correspond to the instrumental response values for the three calibration samples. This example illustrates the creation of a chain of variables starting from x_3. For this purpose, vectors $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_4$, $\mathbf{x}_5$ are projected onto the plane orthogonal to $\mathbf{x}_3$, which results in the projected vectors $\mathbf{Px}_1$, $\mathbf{Px}_2$, $\mathbf{Px}_4$, $\mathbf{Px}_5$. As can be seen, the largest projection corresponds to $\mathbf{Px}_1$, which indicates that $\mathbf{x}_1$ has the least collinearity with respect to $\mathbf{x}_3$. Therefore, variable x_1 is added to the chain. A third variable can be included in the chain by inspecting the projections of $\mathbf{Px}_2$, $\mathbf{Px}_4$, $\mathbf{Px}_5$ onto the line orthogonal to $\mathbf{Px}_1$, as illustrated in Fig. 1b. In this example, the largest projection is associated with $\mathbf{Px}_5$, so variable x_5 is added to the chain. The resulting chain of variables starting from x_3 is therefore (x_3, x_1, x_5).

By changing the initial vector employed in this procedure (i.e. by using $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_4$ or $\mathbf{x}_5$ to initialize the projections), four other chains of variables can be constructed. It is worth noting that no more than N_{cal} variables can be included in each chain. This number is reduced to $(N_{cal} - 1)$ if the data are mean-centered, which is usually the case in multivariate calibration, because one degree of freedom is lost as a result of this operation. A full description of the mathematical equations involved in the projections is described elsewhere [4,26].

It is interesting to note that the selection of each new variable in a chain amounts to maximizing the determinant of the $\mathbf{X}^T\mathbf{X}$ matrix, which needs to be inverted in the construction of the associated MLR model. A demonstration of this mathematical property is provided in the Appendix. In this sense, the first phase of SPA has similarities with classical algorithms for constructing

Table 1. Applications of the Successive Projections Algorithm (SPA) involving multivariate calibration

Ref.	Year	Technique	Sample	Remarks
[12]	2001	UV-Vis	Synthetic mixtures	Original SPA formulation
[14]	2001	ICP-OES	Simulated data and Steel	Determination of Mn, Mo, Cr, Ni and Fe using a low-resolution plasma spectrometer/diode array detection system
[15]	2003	ICP-OES	Steel	Selection of wavelet coefficients
[16]	2003	NIR	Diesel	Determination of total sulfur
[17]	2004	UV-Vis and NIR	Diesel and Synthetic mixtures	Selection of samples
[18]	2004	NIR	Simulated data and Diesel	Selection of wavelet coefficients
[19]	2005	MIR and NIR	Gasoline and Corn	Selection of variables for calibration transfer
[20]	2005	UV-Vis	Polyvitamin/Polymineral Drug	Determination of Cu^{2+}, Mn^{2+} and Zn^{2+}
[21]	2006	NIR	Diesel	Use of subagging to improve prediction ability
[22]	2006	Molecular descriptors	HEPT Derivatives	Nonlinear QSAR study
[7]	2006	MIR	Lubricating Oils	Prediction of viscosity
[23]	2007	NIR	Diesel and Corn	Use of cross-validation
[24]	2007	Molecular descriptors	HEPT Derivatives	Correlation-weighted SPA
[25]	2007	UV-Vis	Sea Water	Determination of five phenolic compounds
[26]	2008	NIR	Diesel and Corn	Introduction of a backward elimination phase in SPA to improve the parsimony of the resulting model
[27]	2008	UV	Antibiotic enzymatic synthesis	Determination of the concentrations of the components present in the enzymatic synthesis of ampicillin
[28]	2008	UV-Vis	Pharmaceutical preparations	Determination of levodopa and carbidopa
[29]	2008	NIR	Vegetable oils	Combination of iPLS and SPA for determination of acidity, refractive index and viscosity
[30]	2008	NIR	Tobacco and pharmaceutical tablet	Combination of UVE and SPA for determination of nicotine and active pharmaceutical ingredient
[31]	2009	Vis-NIR	Plum vinegar	Determination of acetic, tartaric and lactic acids: variables selected by SPA were used in PLS and LS-SVM models
[32]	2009	Vis-NIR	Beer	Determination of soluble solids content: variables selected by SPA were used in LS-SVM models
[33]	2009	MIR	Yogurt	Determination of protein content: variables selected by SPA were used in BP-ANN models
[34]	2009	NIR	*Auricularia auricula*	Determination of protein content: variables selected by SPA were used in PLS and LS-SVM models
[35]	2009	Molecular descriptors	Series of Glycogen Synthase Kinase-3β Inhibitors	Prediction of bioactivity: variables selected by SPA were used in PLS, ANN and LS-SVM models
[36]	2009	Molecular descriptors	Pesticides	Prediction of adsorption coefficients: variables selected by SPA were used in ANN models
[37]	2010	Vis-NIR	Grape juice	Combination of UVE and SPA for non-invasive quantitative determination of soluble solids content and pH: variables selected by UVE-SPA were used in PLS models
[38]	2010	NIR	Wheat	Use of sequential regressions to reduce computation time in SPA
[39]	2010	NIR	Wheat	Use of parallel processing to reduce computation time in SPA
[40]	2010	MIR and NIR	Gasoline and Corn	Use of SPA and subagging for calibration transfer
[41]	2010	NIR and MIR	Biodiesel	Prediction of oxidative stability index, acid number and water content
[42]	2010	UV	Dehydrated broths	Determination of monosodium glutamate, guanosine-5'-monophosphate and inosine-5'-monophosphate
[43]	2010	NIR and MIR	Wheat and Gasoline	Selection of wavelet coefficients with different wavelet functions, followed by combination of the resulting models
[44]	2010	Spectrofluorimetric	Air	Determination of hydroquinone, resorcinol, phenol, *m*-cresol and *p*-cresol
[45]	2011	UV-Vis and NIR	Colorants and Gasoline	Modification of SPA to handle the presence of unknown interferents in new samples to be analyzed
[46]	2011	NIR	Oilseed Rape Leaves	Determination of total amino acids under a new herbicide stress: variables selected by SPA were used in PLS and LS-SVM models
[47]	2011	Vis-NIR	Biodiesel/diesel blends	Determination of biodiesel content

Table 1 (*continued*)

Ref.	Year	Technique	Sample	Remarks
[48]	2011	NIR	Insulating oils	Determination of interfacial tension and relative density
[49]	2011	NIR	Diesel and Wheat	Choice of subsampling ratio in the SPA-subagging approach
[50]	2011	NIR	Water	Determination of benzene, toluene and xylenes
[51]	2011	Thermogravimetry	Drugs in pharmaceutical formulations	Determination of L-ascorbic acid
[52]	2011	NIR	Biodiesel	Prediction of density, viscosity, methanol content, and water concentration
[53]	2012	Thermogravimetry	Pharmaceutical tablet	Determination of paracetamol and codeine phosphate
[54]	2012	NIR- Overtone regions	Diesel/Biodiesel	Determination of biodiesel content and adulterations
[55]	2012	NIR	Injectable formulations	Determination of dipyrone in closed ampoules
[56]	2012	Imaging spectroscopy	Dehydrated prawns	Determination of moisture content: variables selected by SPA were used in PLS and LS-SVM models

UV–Vis (UV–Vis spectrophotometry), UV (UV spectrometry), ICP-OES (Inductively coupled plasma - optical emission spectrometry), MIR (middle-infrared spectrometry), NIR (Near-infrared spectrometry), HEPT (1-[2-hydroxyethoxy-methyl]-6-(phenylthio) thymine), QSAR (quantitative structure-activity relationship), UVE (uninformative variable elimination), PLS (partial least squares), iPLS (interval PLS), ANN (artificial neural network), LS-SVM (least squares-support vector machines).

D-optimal designs [66,67]. It is worth noting that the search procedure adopted in Phase 1 does not guarantee that the overall maximum of the determinant for a given number of variables will be achieved. However, such a procedure is a reasonable alternative to an exhaustive search, which is usually impractical in terms of computational workload. Moreover, the variable-selection process in SPA-MLR is guided by not only the maximization of the determinant of $\mathbf{X}^T\mathbf{X}$, but also the relation of the x-variables with the y-property of interest, which is assessed in Phase 2, as described below.

The second phase of SPA involves evaluating candidate subsets of variables extracted from the chains created in Phase 1. Candidate subsets with L variables are obtained by taking the L first variables of each chain. The best subset of variables is selected on the basis of the smallest root-mean-square error for the validation set (RMSEV). This performance index is calculated as

$$RMSEV = \sqrt{\frac{1}{N_{val}}\sum\nolimits_{i=1}^{N_{val}}\left(y_{val,i} - \hat{y}_{val,i}\right)^2} \tag{1}$$

where $y_{val,i}$ and $\hat{y}_{val,i}$ denote the reference value and the predicted value of the y-property of interest for the ith validation sample, respectively.

The third phase involves a backward elimination procedure aimed at discarding uninformative variables and thus improving the parsimony of the model. For this purpose, a relevance index is defined for each variable belonging to the subset selected at the end of Phase 2. This index is calculated by multiplying the standard deviation of the variable (calculated over the calibration set) by the absolute value of its regression coefficient. The variables are then sorted according to this relevance index, and a scree plot of $RMSEV$ against the number of variables included in the MLR model is generated. The final solution is then obtained by using the smallest number of variables so that $RMSEV$ is not significantly larger than the minimum value observed in the scree plot, according to an F-test. A significance level $\alpha = 0.25$ for the F-test is typically adopted, as suggested in [26].

It is worth noting that the final SPA-MLR model provides a direct relation between the measurements in the original analytical channels (x-variables) and the y-property to be predicted. An alternative involves working with the orthogonalized x-variables resulting from the projection operations carried out in Phase 1, as proposed elsewhere [68]. However, the interpretation of the resulting model may not be so straightforward, as the orthogonalized x-variables no longer correspond to the original measurements.

The basic SPA-MLR formulation described above requires the division of the available samples into calibration and validation data sets. For this purpose, the calibration set could be extracted in a random fashion from the pool of available samples. However, such a

Table 2. Applications of the Successive Projections Algorithm (SPA) involving classification

Ref.	Year	Technique	Sample	Remarks
[13]	2005	NIR	Vegetable oils and diesel	Adaptation of the original SPA formulation to address classification problems
[57]	2009	LIBS	Soils	Use of wavelet compression; classification of the samples into three different soil orders
[58]	2009	SWV	Edible vegetable oils	Classification with respect to type and conservation state.
[59]	2009	NIRR	Cigarettes	Classification into four brands of different chemical composition
[39]	2010	SWV	Vegetable oils	Use of parallel processing to reduce computation time in SPA
[60]	2010	UV-Vis	Aqueous extracts of coffee	Classification with respect to type and conservation state
[61]	2011	NIR	Diesel/Biodiesel blends	Detection of adulterations
[62]	2012	NIR	Beer	Classification with respect to ageing state
[63]	In Press	MIR	Blue pen ink	Classification according to types and brands
[64]	2012	NIR and MIR	Ethanol fuel	Detection of adulterations with water and methanol

UV–Vis (UV–Vis spectrophotometry), MIR (middle-infrared spectrometry), NIR (Near-infrared spectrometry), SWV (Square wave voltammetry).

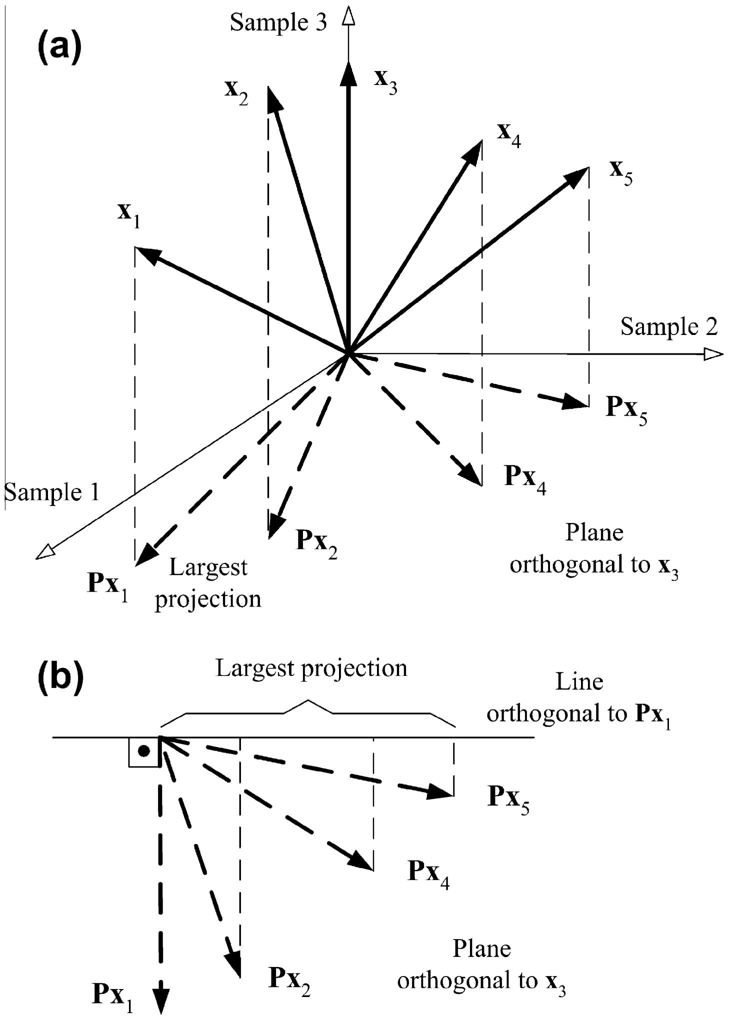

Figure 1. Generation of a chain of variables in Phase 1 of SPA-MLR with $N_{cal} = 3$ samples and $K = 5$ variables, starting from x_3. The projection operations in (a) and (b) lead to the selection of x_1 and x_5, respectively. The resulting chain of variables is (x_3, x_1, x_5).

simple procedure does not guarantee that the resulting set will be adequately distributed over the multidimensional sample space. A possibly better approach involves using the well-known Kennard-Stone (KS) algorithm [69], which is typically initialized by using the pair of samples separated by the largest distance in the instrumental response (**x**-vector) space. Each of the subsequent calibration samples is then chosen in order to maximize the minimum distance with respect to the samples that have already been selected. Alternatively, the KS algorithm can be applied to the values of the y-property of interest, instead of the **x**-vectors. In this manner, the samples with the smallest and largest values of y are guaranteed to be included in the calibration set, thus avoiding extrapolation problems [70]. It is also possible to take into account the **x** and y distances simultaneously, as proposed elsewhere [71].

Another option involves evaluating the candidate subsets of variables by using cross-validation instead of employing an independent validation set, as proposed in [23]. Such an option may be especially useful if the data set contains a small number of samples. The simplest cross-validation procedure involves removing one sample at a time from the calibration set and applying the resulting model to the prediction of this sample (the leave-one-out method). It is also possible to remove and to predict a larger number of samples at each time (leave-many-out cross-validation), which could possibly be more efficient to avoid overfitting problems [72]. However, a systematic comparison of different cross-validation procedures within the scope of SPA-MLR remains to be carried out. A variant of this procedure termed ''subagging'' (**sub**sample **agg**regat**ing**) was employed in [21,49] to improve the prediction performance of the resulting calibration model. In this case, the SPA-MLR models obtained for different calibration/validation divisions were combined to generate an ''ensemble'' model.

2.1. Application example: NIR spectrometric determination of protein in wheat

This illustrative example involves the application of SPA-MLR to a publicly available data set containing 882 VIS-NIR spectra of whole-kernel wheat samples, which were harvested over a period of eight years [73]. Protein content was chosen as the y-property of interest. The SPA-MLR model was built by using 775 samples corresponding to seven of the eight years. The 107 samples corresponding to the remaining year were employed as an external prediction set to evaluate the performance of the resulting model.

The spectra were acquired in the range 400–2500 nm with a resolution of 2 nm. In the present example, only the NIR region in the range 1100–2500 nm was employed, as shown in Fig. 2a. In order to remove undesirable baseline features, first derivative spectra were calculated by using a Savitzky-Golay filter with a 2nd-order polynomial and an 11-point window (Fig. 2b). The KS algorithm was applied to the derivative spectra in order to divide the 775 modeling samples into calibration and validation sets with $N_{cal} = 517$ and $N_{val} = 258$ samples, respectively.

Fig. 3a presents the smallest RMSEV values obtained as a function of the number of variables in the candidate subsets evaluated in Phase 2 of SPA-MLR. In order to save computational time, the RMSEV calculations could have been interrupted when the curve started to rise (e.g., for candidate subsets with more than 300 variables). However, the entire curve is shown here for the purpose of illustration. Overall, the smallest RMSEV was achieved by using a subset of 117 variables. In Phase 3, these variables were sorted according to their relevance index and the RMSEV was recalculated by progressively including the sorted variables in the MLR model. The new RMSEV values thus obtained are presented in Fig. 3b. Finally, by using an F-test as described above, the number of variables was greatly reduced from 117 to 16. The white and black markers in Fig. 3b indicate the variables discarded and retained as a result of this rocedure.

It may be argued that the F-test employed in Phase 3 could also be employed in Phase 2 to carry out a preliminary reduction in the number of variables, before the relevance sorting procedure. So far, such an alternative has not been investigated in the literature.

Finally, the performance of the resulting SPA-MLR model was evaluated by using the external set of prediction samples. The results are presented in Fig. 4. As can be seen, the samples are randomly distributed on both sides of the bisecting line, which indicates the absence of systematic error. The root-mean-square error of prediction (RMSEP) was 0.2% m/m and the correlation between predicted and reference values was 0.99.

2.2. Variants of SPA-MLR: correlation-weighted SPA and adaptive SPA

Some variants of the basic SPA-MLR formulation have been proposed in the literature through modifications in one or more phases of the algorithm. In correlation-weighted SPA (CWSPA), a modification in Phase 1 was introduced to favor the selection of variables with a larger correlation with the y-property of interest [24]. For this purpose, each of the projected vectors was multiplied by a factor ρ^{α}, where ρ is the correlation coefficient between the y-property of interest and the x-variable associated with the projected vector. The exponent α can be chosen by the analyst in order to adjust the importance ascribed to this correlation in the formation of the chains of variables. It is worth noting that CWSPA with $\alpha = 0$ is equivalent to the original SPA-MLR formulation. In [24], seven values for α were tested ($\alpha = 0, 1,\ldots, 6$) and the best value was chosen on the basis of the smallest relative standard error.

In adaptive SPA, the cost function in Phase 2 was modified to favor the selection of variables in which the effect of interferents is less pronounced over a set of $N_{unknown}$ unknown samples to be analyzed [45]. To this end, the proposed cost function was defined as:

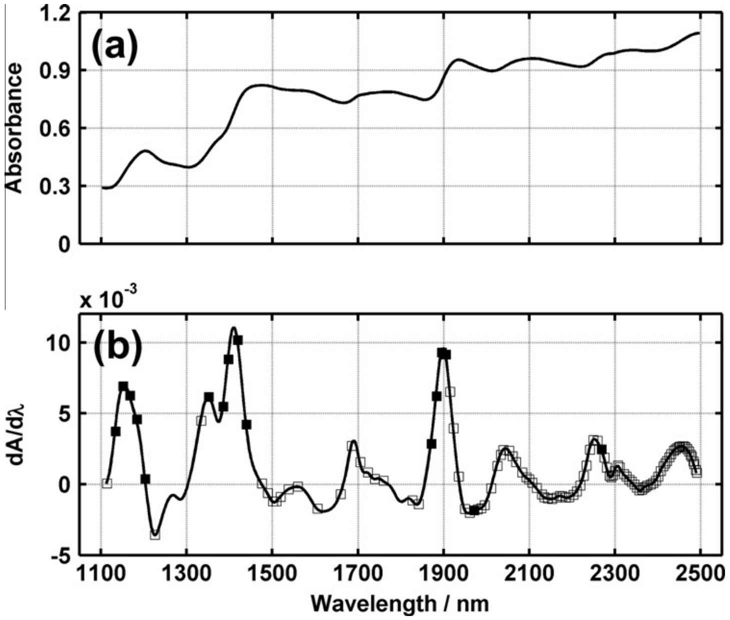

Figure 2. (a) Original and (b) derivative NIR mean spectrum of the 775 whole-kernel wheat samples used for modeling purposes. The wavelengths discarded and retained by the elimination procedure in Phase 3 of SPA-MLR are indicated by white and black square markers, respectively.

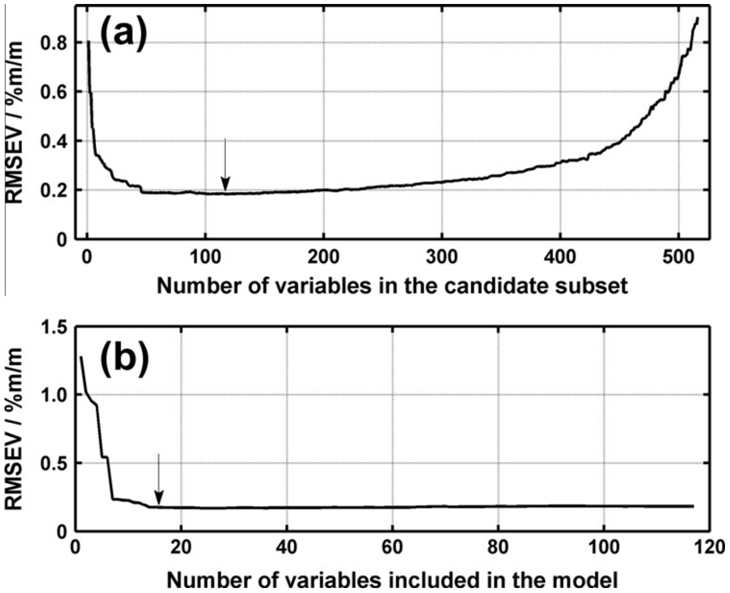

Figure 3. RMSEV values obtained in (a) Phase 2 and (b) Phase 3 of SPA-MLR. The arrows indicate the position corresponding to the number of selected variables in each phase.

$$J_{cost} = \begin{cases} \infty, & \gamma > 1 \\ \gamma \times RMSEV, & \gamma \leqslant 1 \end{cases} \quad (2)$$

where γ is given by

$$\gamma = \overline{MD}_{unknown} / \overline{MD}_{cal} \quad (3)$$

with

$$\overline{MD}_{cal} = \frac{1}{N_{cal}} \sum_{i=1}^{N_{cal}} MD_{cal,i} \quad (4)$$

$$\overline{MD}_{unknown} = \frac{1}{N_{unknown}} \sum_{j=1}^{N_{unknown}} MD_{unknown,j} \quad (5)$$

where $MD_{cal,i}$ and $MD_{unknown,j}$ denote the Mahalanobis distance [74] of the ith calibration sample ($i = 1, 2, \ldots,$

N_{cal}) and the jth ($j = 1, 2, \ldots, N_{unknown}$) unknown sample with respect to the center of the calibration set. The reason for using the cost function defined in Equation (2) can be explained as follows. If a candidate subset of variables is strongly affected by the interferents over the unknown samples, the resulting Mahalanobis distance with respect to the center of the calibration set will be larger than usual. Therefore, $\overline{MD}_{unknown}$ will be larger than the corresponding value for the calibration samples ($\overline{MD}_{cal}$), resulting in a large value of the γ parameter defined in Equation (3). If the interference effects are so strong that $\gamma > 1$, this subset of variables will not be considered a suitable candidate for use in the MLR model ($J_{cost} = \infty$). If $\gamma \leqslant 1$, the adopted expression for the cost ($J_{cost} = \gamma \times RMSE$) will favor subsets of variables with small γ (i.e. with small interference effects).

2.3. QSAR/QSPR applications

In addition to analytical applications involving different instrumental techniques, SPA has been employed to select molecular descriptors in Quantitative Structure-Activity Relationship (QSAR) and Quantitative Structure-Property Relationship (QSPR) studies. Examples include the prediction of anti-HIV-1 activity [22,24], bioactivity of glycogen synthase kinase (GSK-3β) inhibitors [35] and adsorption coefficients of pesticides (Koc) [36].

It is worth noting that QSAR and QSPR often require the use of non-linear models. Within this scope, two modifications to SPA-MLR modeling have been proposed. The simplest modification involves applying the standard SPA-MLR formulation to select the molecular descriptors and then using these descriptors to build a non-linear model [35,36]. However, in this approach, the selection process is still aimed at minimizing the prediction error of an MLR model, so there is no guarantee that the resulting descriptors will be appropriate for use in a non-linear model. A more elaborate modification involves building non-linear models in Phase 2 of SPA in order to evaluate the candidate subsets of variables [22,24].

2.4. Calibration transfer using SPA-MLR

The prediction accuracy of a model can be compromised by the presence of sources of variation that were not taken into account in the calibration process {e.g., changes in the physical or chemical features of the samples, and alterations in the instrumental response caused by ageing or maintenance interventions}. In these cases, calibration transfer procedures may be of value to avoid the need for a full recalibration of the model [75,76].

A typical application of calibration transfer involves the use of a model calibrated for one instrument (primary) to generate predictions on the basis of measurements from a different instrument (secondary). For this purpose, a common approach involves using standardization methods that transform the measurements of the secondary instrument to resemble those of the primary [77]. Alternatively, the reverse transformation can be applied to the original set of calibration measurements in order to build a new model for the secondary instrument. In both cases, a set of representative samples needs to be measured using both instruments in order to construct the mathematical transformation involved in the standardization process [78].

The need to measure the same samples at primary and secondary instruments is inconvenient if the instruments are located at different laboratories or the primary instrument is no longer available. This limitation motivated the development of a variant of SPA-MLR to address calibration-transfer problems [19]. For this purpose, the following cost function was adopted in Phase 2 of SPA-MLR:

$$J_{cost} = 1/2(\mathrm{RMSEV} + \mathrm{RMSET}) \qquad (6)$$

In this cost function, RMSET is the root-mean-square error for a set of ''transfer samples'' measured at the secondary instrument. The purpose of introducing this new term involves taking into account not only the predictive ability of the SPA-MLR model (measured by RMSEV), but also its robustness with respect to differences between the instruments (measured by RMSET). In this sense, the robustness is not directly assessed in terms of changes in the recorded spectrum, but rather in terms of the resulting error in the prediction of the y-property of interest for the transfer samples. Such a prediction is accomplished by using the spectra recorded on the secondary instrument, so it is not necessary to measure the transfer samples on the primary instrument.

2.5. Selection of samples using SPA-MLR

Through a simple modification, SPA-MLR can be used to select suitable calibration samples instead of variables [17]. The goal in this case involves selecting a subset of samples that are minimally redundant but still representative of the data set. For this purpose, it is sufficient to apply the standard SPA-MLR algorithm to the transposed matrix $(\mathbf{X_{cal}})^T$ instead of $\mathbf{X_{cal}}$. In this manner, SPA-MLR will select columns of $(\mathbf{X_{cal}})^T$, which correspond to calibration samples. It is worth noting that the number of selected samples must be larger than the number of variables to allow the use of MLR, so a preliminary selection of variables may be required.

This variant of SPA-MLR may be of value to reduce the experimental workload and the cost involved in calibrating a group of instruments. To this end, an initial set of samples needs to be measured using the first instrument. Subsequently, only the selected samples need to be measured using the remaining instruments. Moreover, this sample-selection algorithm can be useful to choose representative samples for calibration-transfer purposes, as reported elsewhere [19].

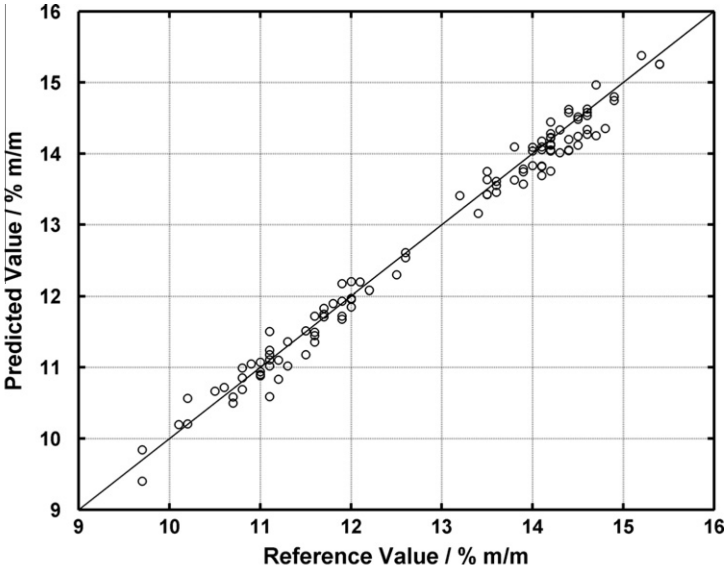

Figure 4. Predicted *versus* reference values of protein content for the prediction set.

2.6. Computational improvements

Typically, Phase 2 is considerably more demanding than the other two phases of SPA-MLR in terms of computational workload [38,39]. Such a workload is mainly associated with construction and validation of an MLR model for each candidate subset of variables. This bottleneck has motivated the development of computational improvements to reduce the time required for completion of Phase 2.

Soares et al. [38] proposed the use of a sequential regression procedure in Phase 2 to simplify the matrix inverse computations involved in MLR. The idea involves using a model with i variables as a starting point to obtain a model with $(i + 1)$ variables without the need to repeat all MLR calculations. In a case study involving NIR spectrometric determination of protein in wheat, five-fold computational gains were obtained by using this strategy [38].

Another approach [39] exploits the availability of multi-core processors in modern off-the-shelf computers. The proposed approach involves the distribution of the MLR calculations among the processor cores, which would otherwise be idle. By applying this method to the same data set employed in [38], computation gains of up to 204% were obtained for SPA-MLR.

3. SPA for classification

The extension of SPA to classification problems was proposed in [13] with the purpose of improving the performance of linear discriminant analysis (LDA) models, which are also known to be adversely affected by multi-collinearity among the input variables [65]. In what follows, the term SPA-LDA is employed with reference to the variable-selection algorithm, as well as the final classification model. Moreover, for consistency with the terminology usually employed in the classification literature, the expression ‘‘calibration set’’ will be replaced with ‘‘training set’’. It is still assumed that a separate validation set will be available to guide the variable-selection process, and that an external set will be used to assess the performance of the final LDA model. This external set will be termed ‘‘test set’’.

In Phase 1, the only difference between SPA-LDA and SPA-MLR involves the mean-centering procedure. In SPA-MLR, the calibration data are centered in the overall mean of the set, whereas, in SPA-LDA, the training data are centered in the mean of each class. Due to the loss of degrees of freedom associated with the calculation of the class means, the length of the chains of variables constructed in Phase 1 of SPA-LDA is limited by $N - C$, where N is the number of training samples and C is the number of classes involved in the problem. The mean-centered training data are employed in Phase 2 to calculate a pooled covariance matrix, as described below.

In Phase 2, the candidate subsets of variables are evaluated according to a cost function related to the average risk of incorrect classification over the validation set. This cost function is defined as:

$$J_{\cos t} = \frac{1}{N_{val}} \sum_{n=1}^{N_{val}} g_n \tag{7}$$

where

$$g_n = \frac{MD^2[\mathbf{x}_{val,n}, \bar{\mathbf{x}}(I_n)]}{\min\limits_{I_j \neq I_n} MD^2[\mathbf{x}_{val,n}, \bar{\mathbf{x}}(I_j)]} \tag{8}$$

Table 3. Oil classes and number of training, validation and test samples employed in the example

Class	Set		
	Training	Validation	Test
Canola*	9	3	3
Sunflower*	10	3	3
Corn*	10	3	4
Soybean*	10	4	4
Expired	20	10	18
Total	59	23	32

*Non-expired samples.

In Equation (8), numerator $MD^2[\mathbf{x}_{val,n}, \bar{\mathbf{x}}(I_n)]$ is the squared Mahalanobis distance [74] between the nth validation sample $\mathbf{x}_{val,n}$ (of class index I_n) and the mean $\bar{\mathbf{x}}(I_n)$ of its true class (both row vectors) calculated over the training set. This distance is given by

$$MD^2[\mathbf{x}_{val,n}, \bar{\mathbf{x}}(I_n)] = [\mathbf{x}_{val,n}, \bar{\mathbf{x}}(I_n)]\mathbf{S}^{-1}[\mathbf{x}_{val,n}, \bar{\mathbf{x}}(I_n)]^T \tag{9}$$

where $\mathbf{S}$ is a pooled covariance matrix calculated over the training set [79,80]. The denominator in Equation (8) corresponds to the squared Mahalanobis distance between $\mathbf{x}_{val,n}$ and the center of the nearest wrong class. A small value of g_n indicates that $\mathbf{x}_{val,n}$ is close to the center of its true class and distant from the centers of the remaining classes. The cost function J_{cost} is defined as the average value of g_n over all validation samples (n = 1, 2, ..., N_{val}), so minimization of J_{cost} results in better separation of the samples according to their true classes.

After the variables have been selected, the classification of a new sample, $\mathbf{x}_{new}$, can be carried out by calculating the Mahalanobis distance of $\mathbf{x}_{new}$ with respect to the mean vector of each class. The new sample is then assigned to the class for which the Mahalanobis distance is the smallest. It should be noted that the mean vectors and pooled covariance matrix are calculated over the training set by using the selected variables.

As in SPA-MLR, computational gains can be obtained by distributing the calculations of Phase 2 over different processor cores [39]. However, an equivalent to the sequential regressions procedure employed in [38] has not yet been proposed for SPA-LDA. Moreover, an equivalent to Phase 3 of SPA-MLR has not been developed for SPA-LDA. For this purpose, the individual discriminability [80] of the selected variables with respect to the classes under consideration could be used as a relevance index, as in Phase 3 of SPA-MLR. However, a (possibly non-parametric) hypothesis test would need to be devised in order to choose an appropriate point in the resulting scree plot of J_{cost}, as the F-test employed to compare RMSEV values in SPA-MLR cannot be directly applied to compare values of J_{cost} in SPA-LDA.

3.1. Application example: voltammetric classification of vegetable oils

This illustrative example involves the application of SPA-LDA for classification of vegetable oils by using square-wave voltammetry. The apparatus and experimental procedure are detailed elsewhere [58]. A total of 114 samples of canola, sunflower, corn and soybean oil

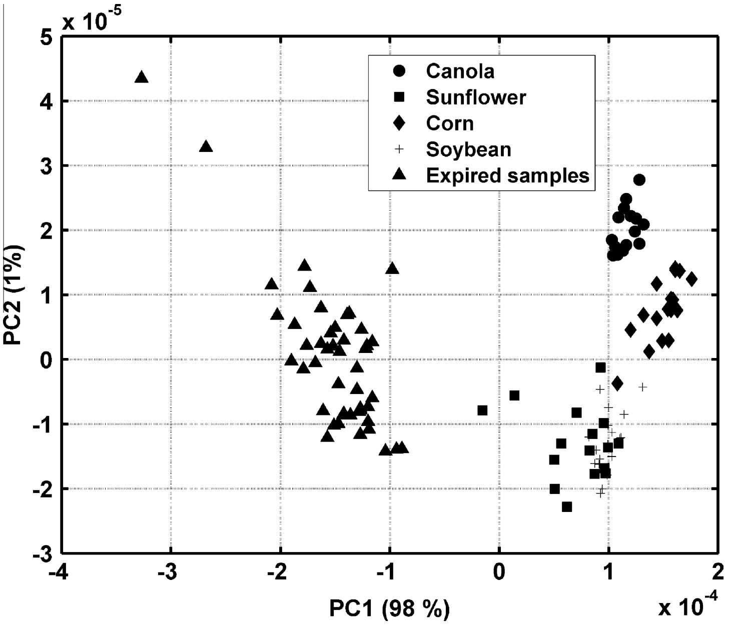

Figure 5. PCA score plot of the vegetable oil data set. The variance explained by each principal component is indicated at the corresponding axis.

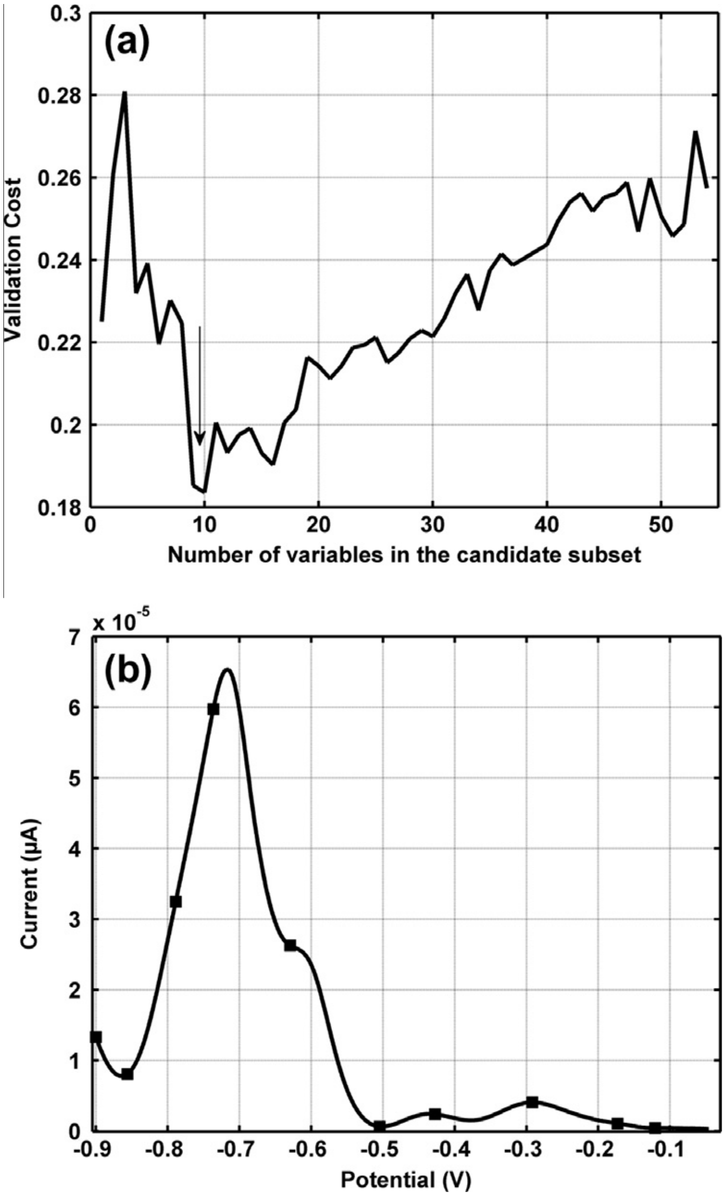

Figure 6. (a) Validation cost (J_{cost}) values obtained in Phase 2 of SPA-LDA. The arrow indicates the minimum of the curve. (b) Mean voltammogram of the 114 vegetable oil samples with indication of the selected variables.

were employed. The problem involved discriminating ''expired'' samples (i.e. samples that were analyzed several months past the expiry date) from ''non-expired'' ones. In addition, the non-expired samples were to be classified according to the oil type. As a result, five classes were considered, as shown in Table 3. The samples within each class were separated in training, validation and test sets by using the KS algorithm.

Fig. 5 presents a PCA score plot of the entire data set. The first two principal components (PC1 and PC2) account for 99% of the data variability. As can be seen, the expired samples can be clearly separated from the non-expired ones along the PC1 axis. In addition, some separation among the types of non-expired oils can be observed along PC2. However, the soybean and sunflower classes cannot be discriminated from each other.

Fig. 6a presents the smallest J_{cost} values obtained as a function of the number of variables in the candidate subsets evaluated in Phase 2 of SPA-LDA. As in the case of SPA-MLR, the cost calculations could have been interrupted when the curve started to rise (e.g., for candidate subsets with more than 20 variables). Overall, the smallest cost was achieved by using a subset of 10 variables, as indicated in Fig. 6b. By applying the

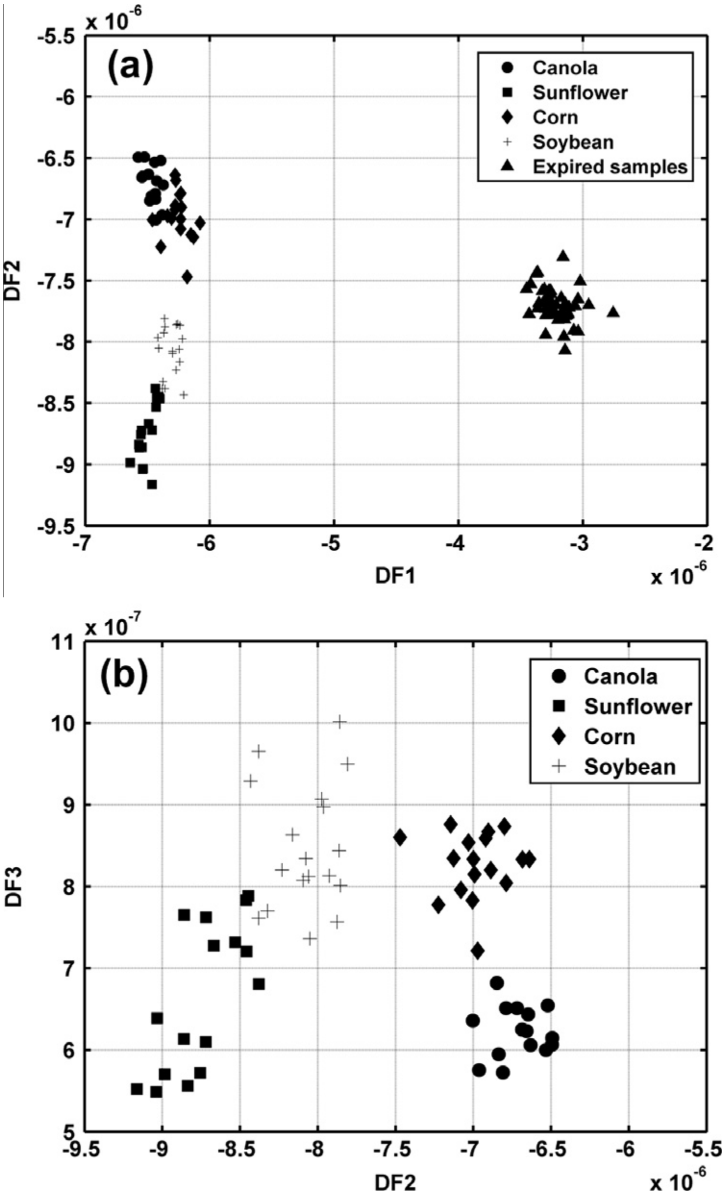

Figure 7. (a) DF1 × DF2 and (b) DF2 × DF3 discriminant function values calculated by using the variables selected by SPA-LDA.

resulting SPA-LDA model to the test set, all samples were correctly classified.

Fig. 7a presents the first two (non-standardized) discriminant functions (DF1, DF2) for the overall data set [81]. The coefficients of these functions were calculated by using the training-set statistics (class means and pooled covariance matrix) for the 10 selected variables. As can be seen, DF1 discriminates between expired and non-expired samples, whereas DF2 separates two groups of non-expired samples (canola/corn and sunflower/soybean). Such groups are further separated along the third discriminant function (DF3), as shown in Fig. 7b. It is worth noting that the sunflower and soybean classes, which could not be distinguished from each other in the PCA plot (Fig. 5), are now clearly separated. Such a result illustrates the convenience of using the variables selected by SPA-LDA compared to entire voltammograms.

4. Pre-processing issues

Among the variety of pre-processing methods available in the chemometrics literature, three procedures are of particular value for SPA-MLR and SPA-LDA, namely range selection, signal smoothing and data compression.

The pre-selection of ranges of variables is useful to exclude analytical channels that do not convey useful information, due to the physical and chemical properties of the system, or technical features of the measurement instrument. Such a selection can be carried out on the basis of *a priori* knowledge or using chemometrics techniques {e.g., interval-PLS (iPLS) [29]}. This procedure may be important to avoid the possible selection of variables that exhibit chance correlation with the$y$-property of interest in SPA-MLR or the class index in SPA-LDA. Moreover, a reduction in the overall number of variables is useful to decrease the computational workload of SPA [12].

Signal smoothing is often employed in chemometrics modeling to improve the signal-to-noise ratio of the measurements. This is particularly important in SPA because noisy variables tend to be included in the chains generated in Phase 1 due to their small collinearity with informative variables. It is expected that candidate subsets with too many noisy variables will result in large values for the cost function and thus will not be selected in Phase 2. Moreover, if some of these variables are still present in the selected subset, they can be eliminated in Phase 3 of SPA-MLR. However, despite these safeguards, smoothing is usually of value to reduce the possibility that noisy variables still remain in the SPA outcome [7].

Data compression involves extracting a reduced number of informative features from the data set by means of a convenient mathematical transformation, which usually provides an improvement in signal-to-noise ratio, because the signal content is concentrated in the resulting features. This pre-processing procedure is therefore useful in SPA to reduce computational workload and to avoid the selection of noisy variables. In this context, the wavelet transform (WT) has been the most commonly used method for data compression [15,18,43] In a problem involving the classification of soil samples on the basis of LIBS spectra, the use of a wavelet-compression procedure provided a 100-fold reduction in computational workload without significantly compromising the classification accuracy of the resulting SPA-LDA model [57].

5. Conclusions

This article reviewed the basic formulation of SPA for multivariate-calibration and classification problems, and several recently proposed variants. In particular, we discussed extensions to calibration transfer, sample selection and non-linear QSAR/QSPR modeling. We also analyzed the main computational and pre-processing issues related to the use of SPA.

The value of SPA can be gauged from the growing number of applications reported in the literature with different techniques and samples, as summarized in Tables 1 and 2. Arguably, the main advantage of SPA involves the simplicity of the algorithm, which can be easily modified to incorporate contributions from the chemometrics community. It should also be emphasized that the algorithm is deterministic (i.e. it does not involve random factors), so the results obtained with a given data set can be easily reproduced and double-checked by different research groups.

Future research avenues may include the applications of SPA to higher-order calibration methods and process analytical technology. Within the scope of classification, it may be of value to investigate the use of SPA together with recent methods {e.g., support vector machines and random forests [80,82]}.

The code employed in this article is freely available from us upon request. In addition, a graphical user interface for SPA-MLR can be obtained at www.ele.ita.br/kawakami/~spa/. A detailed description of this interface is presented in [83].

Acknowledgments
We acknowledge the support of CNPq, Capes and FAPESP (research fellowships, scholarships and INCTAA grant).

Appendix A

Consider the problem of selecting the $(n + 1)$st variable to be included in a given chain during Phase 1 of SPA. As illustrated in Fig. 1, this variable will be selected in order to maximize the projection of the associated column vector in the subspace orthogonal to the column vectors of the previous n variables of the chain. An interpretation for the result of this procedure can be derived as follows.

Let $\mathbf{X}_n$ $(N_{cal} \times n)$ be a matrix formed with the columns of $\mathbf{X_{cal}}$ associated with the n variables of the chain under consideration. It is worth noting that these columns are linearly independent, because they have been selected as the result of previous projection operations to minimize collinearity, so matrix $\mathbf{X}_n^T\mathbf{X}_n$ is invertible. Moreover, let $\mathbf{z}$ be one of the remaining columns of $\mathbf{X_{cal}}$, which corresponds to one of the variables still available for inclusion in the chain. The projection of $\mathbf{z}$ onto the subspace orthogonal to the columns of $\mathbf{X}_n$ is given by

$$\mathbf{z_P} = (\mathbf{I} - \mathbf{X}_n(\mathbf{X}_n^T\mathbf{X}_n)^{-1}\mathbf{X}_n^T)\mathbf{z} \tag{10}$$

where $\mathbf{I}$ is an identity matrix and $\mathbf{z_P}$ denotes the projected column vector. It follows that

$$\begin{aligned} ||\mathbf{z_P}||^2 &= \mathbf{z_P}^T\mathbf{z_P} = \mathbf{z}^T(\mathbf{I} - \mathbf{X}_n(\mathbf{X}_n^T\mathbf{X}_n)^{-1}\mathbf{X}_n^T)(\mathbf{I} - \mathbf{X}_n(\mathbf{X}_n^T\mathbf{X}_n)^{-1}\mathbf{X}_n^T)\mathbf{z} \\ &= \mathbf{z}^T(\mathbf{I} - \mathbf{X}_n(\mathbf{X}_n^T\mathbf{X}_n)^{-1}\mathbf{X}_n^T - \mathbf{X}_n(\mathbf{X}_n^T\mathbf{X}_n)^{-1}\mathbf{X}_n^T \\ &\quad + \mathbf{X}_n(\mathbf{X}_n^T\mathbf{X}_n)^{-1}\mathbf{X}_n^T\mathbf{X}_n(\mathbf{X}_n^T\mathbf{X}_n)^{-1}\mathbf{X}_n^T)\mathbf{z} \\ &= \mathbf{z}^T(\mathbf{I} - 2\mathbf{X}_n(\mathbf{X}_n^T\mathbf{X}_n)^{-1}\mathbf{X}_n^T + \mathbf{X}_n(\mathbf{X}_n^T\mathbf{X}_n)^{-1}\mathbf{X}_n^T)\mathbf{z} \\ &= \mathbf{z}^T(\mathbf{I} - \mathbf{X}_n(\mathbf{X}_n^T\mathbf{X}_n)^{-1}\mathbf{X}_n^T)\mathbf{z} \end{aligned} \tag{11}$$

Now, let $\mathbf{X}_{n+1}$ be a matrix formed by augmenting $\mathbf{X}_n$ with the additional column $\mathbf{z}$ {i.e.$\mathbf{X}_{n+1} = [\mathbf{X}_n\ \mathbf{z}]$}. The determinant of $\mathbf{X}_{n+1}^T\mathbf{X}_{n+1}$ can be calculated as

$$\det(\mathbf{X}_{n+1}^T\mathbf{X}_{n+1}) = \det\left(\begin{bmatrix}\mathbf{X}_n^T\\ \mathbf{z}^T\end{bmatrix}[\mathbf{X}_n \quad \mathbf{z}]\right) = \det\begin{bmatrix}\mathbf{X}_n^T\mathbf{X}_n & \mathbf{X}_n^T\mathbf{z}\\ \mathbf{z}^T\mathbf{X}_n & \mathbf{z}^T\mathbf{z}\end{bmatrix} \tag{12}$$

which can be re-written as

$$\det(\mathbf{X}_{n+1}^T\mathbf{X}_{n+1}) = [\det(\mathbf{X}_n^T\mathbf{X}_n)][\mathbf{z}^T\mathbf{z} - \mathbf{z}^T\mathbf{X}_n(\mathbf{X}_n^T\mathbf{X}_n)^{-1}\mathbf{X}_n^T\mathbf{z}] \tag{13}$$

(see, e.g., [84] for a demonstration). It follows from Equation (13) that

$$\det(\mathbf{X}_{n+1}^T\mathbf{X}_{n+1}) = [\det(\mathbf{X}_n^T\mathbf{X}_n)]\mathbf{z}^T[\mathbf{I} - \mathbf{X}_n(\mathbf{X}_n^T\mathbf{X}_n)^{-1}\mathbf{X}_n^T]\mathbf{z} \tag{14}$$

Finally, from Equations (11) and (14), one arrives at

$$\det(\mathbf{X}_{n+1}^T\mathbf{X}_{n+1}) = [\det(\mathbf{X}_n^T\mathbf{X}_n)]||\mathbf{z}_\mathbf{P}||^2 \tag{15}$$

so, by choosing $\mathbf{z}$ according to the criterion of largest projection onto the subspace orthogonal to the columns of $\mathbf{X}_n$, the determinant of $\mathbf{X}_{n+1}^T\mathbf{X}_{n+1}$ is maximized.

References

[1] C. Pasquini, J. Cortez, L.M.C. Silva, F.B. Gonzaga, J. Braz. Chem. Soc. 18 (2007) 463.

[2] P. Williams, K. Norris, Near-Infrared Technology in the Agricultural and Food Industries. Amer. Assoc. Cereal Chem., St. Paul, Minnesota, USA, 2001.

[3] Z. Xiaobo, Z. Jiewen, M.J.W. Povey, M. Holmes, M. Hanpin, Anal. Chim. Acta 667 (2010) 14.

[4] R.K.H. Galvão, M.C.U. Araújo, in: B. Walczak, R. Tauler, S. Brown (Editors), Comprehensive Chemometrics, Elsevier, Oxford, UK, 2009, Vol. 3, pp. 233–283.

[5] K. Baumann, Trends Anal. Chem. 22 (2003) 395.

[6] A. Alexandridis, P. Patrinos, H. Sarimveis, G. Tsekouras, Chemom. Intell. Lab. Syst. 75 (2005) 149.

[7] A.R. Caneca, M.F. Pimentel, R.K.H. Galvão, C.E. Matta, F.R. Carvalho, I.M. Raimundo Jr., C. Pasquini, J.J.R. Rohwedder, Talanta 70 (2006) 344.

[8] R. Leardi, J. Chemom. 15 (2001) 559.

[9] O. Bohachevsky, M.E. Johnson, M.L. Stein, Technometrics 28 (1986) 209.

[10] F. Glover, J. Comput. 1 (1989) 190.

[11] M. Shamsipur, V. Zare-Shahabadi, B. Hemmateenejad, M. Akhond, J. Chemom. 20 (2006) 146.

[12] M.C.U. Araújo, T.C.B. Saldanha, R.K.H. Galvão, T. Yoneyama, H.C. Chame, V. Visani, Chemom. Intell. Lab. Syst. 57 (2001) 65.

[13] M.J.C. Pontes, R.K.H. Galvão, M.C.U. Araújo, P.N.T. Moreira, O.D.P. Neto, G.E. José, T.C.B. Saldanha, Chemom. Intell. Lab. Syst. 78 (2005) 11.

[14] R.K.H. Galvão, M.F. Pimentel, M.C.U. Araújo, T. Yoneyama, V. Visani, Anal. Chim. Acta 443 (2001) 107.

[15] C.J. Coelho, R.K.H. Galvão, M.C.U. Araújo, M.F. Pimentel, E.C. Silva, Chemom. Intell. Lab. Syst. 66 (2003) 205.

[16] M.C. Breitkreitz, I.M. Raimundo, J.J.R. Rohwedder, C. Pasquini, H.A.D Filho, G.E. José, M.C.U. Araújo, Analyst (Cambridge, UK) 128 (2003) 1204.

[17] H.A.D. 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. 72 (2004) 83.

[18] R.K.H. Galvão, G.E. José, H.A.D. Filho, M.C.U. Araújo, E.C. Silva, H.M. Paiva, T.C.B. Saldanha, E.S.O.N. Souza, Chemom. Intell. Lab. Syst. 70 (2004) 1.

[19] F.A. Honorato, R.K.H. Galvão, M.F. Pimentel, B.B. Neto, M.C.U. Araújo, F.R. Carvalho, Chemom. Intell. Lab. Syst. 76 (2005) 65.

[20] H.A.D. 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.

[21] R.K.H. Galvão, M.C.U. Araújo, M.N. Martins, G.E. José, M.J.C. Pontes, E.C. Silva, T.C.B. Saldanha, Chemom. Intell. Lab. Syst. 81 (2006) 60.

[22] Y. Akhlaghi, M. Kompany-Zareh, J. Chemom. 20 (2006) 1.

[23] R.K.H. Galvão, M.C.U. Araújo, E.C. Silva, G.E. José, S.F.C. Soares, H.M. Paiva, J. Braz. Chem. Soc. 18 (2007) 1580.

[24] M. Kompany-Zareh, Y. Akhlaghi, J. Chemom. 21 (2007) 239.

[25] M.S. Di Nezio, M.F. Pistonesi, W.D. Fragoso, M.J.C. Pontes, H.C. Goicoechea, M.C.U. Araújo, B.S.F. Band, Microchem. J. 85 (2007) 194.

[26] R.K.H. Galvão, M.C.U. Araújo, W.D. Fragoso, E.C. Silva, G.E. José, S.F.C. Soares, H.M. Paiva, Chemom. Intell. Lab. Syst. 92 (2008) 83.

[27] M.P.A. Ribeiro, T.F. Pádua, O.D. Leite, R.L.C. Giordano, R.C. Giordano, Chemom. Intell. Lab. Syst. 90 (2008) 169.

[28] M. Grunhut, M.E. Centuríon, W.D. Fragoso, L.F. Almeida, M.C.U. Araújo, B.S.F. Band, Talanta 75 (2008) 950.

[29] A.F.C. Pereira, M.J.C. Pontes, F.F.G. Neto, S.R.B. Santos, R.K.H. Galvão, M.C.U. Araújo, Food Res. Int. 41 (2008) 341.

[30] S. Ye, D. Wang, S. Min, Chemom. Intell. Lab. Syst. 91 (2008) 194.

[31] F. Liu, Y. He, Food Chem. 115 (2009) 1430.

[32] F. Liu, Y. Jiang, Y. He, Anal. Chim. Acta 635 (2009) 45.

[33] M. Khanmohammadi, A.B. Garmarudi, K. Ghasemi, S. Garrigues, M. Guardia, Microchem. J. 91 (2009) 47.

[34] F. Liu, Y. He, G. Sun, J. Agric. Food Chem. 57 (2009) 4520.

[35] M. Goodarzi, M.P. Freitas, R. Jensen, J. Chem. Inf. Model. 49 (2009) 824.

[36] N. Goudarzi, M. Goodarzi, M.C.U. Araújo, R.K.H. Galvão, J. Agric. Food Chem. 57 (2009) 7153.

[37] D. Wu, Y. He, P. Nie, F. Cao, Y. Bao, Anal. Chim. Acta 659 (2010) 229.

[38] A.S. Soares, A.R.G. Filho, R.K.H. Galvão, M.C.U. Araújo, J. Braz. Chem. Soc. 21 (2010) 760.

[39] A.S. Soares, R.K.H. Galvão, M.C.U. Araújo, S.F.C. Soares, L.A. Pinto, J. Braz. Chem. Soc. 21 (2010) 1626.

[40] M.N. Martins, R.K.H. Galvão, M.F. Pimentel, J. Braz. Chem. Soc. 21 (2010) 127.

[41] L.F.B. Lira, M.S. Albuquerque, J.G.A. Pacheco, T.M. Fonseca, E.H.S. Cavalcanti, L. Stragevitch, M.F. Pimentel, Microchem. J. 96 (2010) 126.

[42] C.C. Acebal, M. Grünhut, A.G. Lista, B.S.F. Band, Talanta 82 (2010) 222.

[43] L.A. Pinto, R.K.H. Galvão, M.C.U. Araújo, Anal. Chim. Acta 682 (2010) 37.

[44] M.F. Pistonesi, M.S. Di Nezio, M.E. Centurión, A.G. Lista, W.D. Fragoso, M.J.C. Pontes, M.C.U. Araújo, B.S.F. Band, Talanta 83 (2010) 320.

[45] S.F.C. Soares, R.K.H. Galvão, M.C.U. Araújo, E.C. Silva, C.F. Pereira, S.I.E. Andrade, F.C. Leite, Anal. Chim. Acta 689 (2011) 22.

[46] F. Liu, Z.L. Jin, M.S. Naeem, T. Tian, F. Zhang, Y. He, H. Fang, Q.F. Ye, W.J. Zhou, Food Bioprocess. Technol. 4 (2011) 1314.

[47] D.D.S. Fernandes, A.A. Gomes, G.B. Costa, G.W.B. Silva, G. Véras, Talanta 87 (2011) 30.

[48] M.J.C. Pontes, A.M.J. Rocha, M.F. Pimentel, C.F. Pereira, Microchem. J. 98 (2011) 254.

[49] A.R.G. Filho, R.K.H. Galvão, M.C.U. Araújo, J. Braz. Chem. Soc. 22 (2011) 2225.

[50] K.M.G. Lima, I.M. Raimundo, M.F. Pimentel, Sens Actuators, B 160 (2011) 691.

[51] R.L.S. Otero, R.K.H. Galvão, M.C.U. Araújo, E.T.G. Cavalheiro, Thermochim. Acta 526 (2011) 200.

[52] R.M. Balabin, S.V. Smirnov, Anal. Chim. Acta 692 (2011) 63.
[53] M. Khanmohammadi, M. Soleimani, F. Morovvat, A.B. Garmarudi, M. Khalafbeigi, K. Ghasemi, Thermochim. Acta 530 (2012) 128.
[54] F.V.C. Vasconcelos, P.F.B. Souza, M.F. Pimentel, M.J.C. Pontes, C.F. Pereira, Anal. Chim. Acta 716 (2012) 101.
[55] F.A.C. Sanches, R.B. Abreu, M.J.C. Pontes, F.C. Leite, D.J.E. Costa, R.K.H. Galvão, M.C.U. Araújo, Talanta 92 (2012) 84.
[56] D. Wu, H. Shi, S. Wang, Y. He, Y. Bao, K. Liu, Anal. Chim. Acta 726 (2012) 57.
[57] M.J.C. Pontes, J. Cortez, R.K.H. Galvão, C. Pasquini, M.C.U. Araújo, R.M. Coelho, M.K. Chiba, M.F. Abreu, B.E. Madari, Anal. Chim. Acta 642 (2009) 12.
[58] F.F. Gambarra-Neto, G. Marino, M.C.U. Araújo, R.K.H. Galvão, M.J.C. Pontes, E.P. Medeiros, R.S. Lima, Talanta 77 (2009) 1660.
[59] E.D.T. Moreira, M.J.C. Pontes, R.K.H. Galvão, M.C.U. Araújo, Talanta 79 (2009) 1260.
[60] U.T.C.P. Souto, M.J.C. Pontes, E.C. Silva, R.K.H. Galvão, M.C.U. Araújo, F.A.C. Sanches, F.A.S. Cunha, M.S.R. Oliveira, Food Chem. 119 (2010) 368.
[61] M.J.C. Pontes, C.F. Pereira, M.F. Pimentel, F.V.C. Vasconcelos, A.G.B. Silva, Talanta 85 (2011) 2159.
[62] M. Ghasemi-Varnamkhasti, S.S. Mohtasebi, M.L. Rodriguez-Mendeza, A.A. Gomes, M.C.U. Araújo, R.K.H. Galvão, Talanta 89 (2012) 286.
[63] C.S. Silva, F.S.L. Borba, M.F. Pimentel, M.J.C. Pontes, R.S. Honorato, C. Pasquini, Microchem. J. 104 (2012) 49.
[64] A.C. Silva, L.F.B.L. Pontes, M.F. Pimentel, M.J.C. Pontes, Talanta 93 (2012) 129.
[65] T. Naes, B.H. Mevik, J. Chemom. 15 (2001) 413.
[66] J.Y. Tsay, J. Amer. Statist. Assoc. 71 (1976) 671.
[67] T.J. Mitchell, J. Amer. Statist. Assoc. 16 (1974) 203.
[68] M. Kompany-Zareh, N. Omidikia, J. Chem. Inf. Model. 50 (2010) 2055.
[69] R.W. Kennard, L.A. Stone, Technometrics 11 (1969) 137.
[70] A.C. Sousa, M.M.L.M. Lucio, O.F.B. Neto, G.P.S. Marcone, A.F.C. Pereira, E.O. Dantas, W.D. Fragoso, M.C.U. Araújo, R.K.H. Galvão, Anal. Chim. Acta 588 (2007) 231.
[71] R.K.H. Galvão, M.C.U. Araujo, G.E. José, M.J.C. Pontes, E.C. Silva, T.C.B. Saldanha, Talanta 67 (2005) 736.
[72] J. Shao, D. Tu, The Jackknife and Bootstrap, Springer, New York, USA, 1995.
[73] www.idrc-chambersburg.org/shootout.html (accessed 2 May 2012).
[74] R. Maesschalck, D. Jouan-Rimbaud, D.L. Massart, Chemom. Intell. Lab. Syst. 50 (2000) 1.
[75] R.N. Feudale, N.A. Woody, H. Tan, A.J. Myles, S.D. Brown, J. Ferré, Chemom. Intell. Lab. Syst. 64 (2002) 181.
[76] H. Swierenga, P.J. Groot, A.P. Weijer, M.W.J. Derksen, L.M.C. Buydens, Chemom. Intell. Lab. Syst. 41 (1998) 237.
[77] S.D. Brown, in: B. Walczak, R. Tauler, S. Brown (Editors), Comprehensive Chemometrics, Vol. 3, Elsevier, Oxford, UK, 2009.
[78] C.F. Pereira, M.F. Pimentel, R.K.H. Galvão, F.A. Honorato, L. Stragevitch, M.N. Martins, Anal. Chim. Acta 611 (2008) 41.
[79] W. Wu, Y. Mallet, B. Walczak, W. Penninckx, D.L. Massart, S. Heuerding, F. Erni, Anal. Chim. Acta 329 (1996) 257.
[80] R.O. Duda, P.E. Hart, D.G. Stork, Pattern Classification, Second Edition., Wiley, New York, USA, 2001.
[81] J.F. Hair, B. Black, B. Babin, R.E. Anderson, R.L. Tatham, Multivariate Data Analysis, Sixth Edition., Prentice Hall, Englewood Cliffs, NJ, USA, 2005.
[82] V. Kovalishyn, J. Aires-de-Sousa, C. Ventura, R.E. Leitão, F. Martins, Chemom. Intell. Lab. Syst. 107 (2011) 69.
[83] H.M. Paiva, S.F.C. Soares, R.K.H. Galvão, M.C.U. Araújo, Chemom. Intell. Lab. Syst. (2012) (accepted for publication) (DOI: http://dx.doi.org/10.1016/j.chemolab.2012.05.014).
[84] C.D. Meyer, Matrix Analysis and Applied Linear Algebra, SIAM, Philadelphia, PA, USA, 2000.