Analytica Chimica Acta 689 (2011) 22–28

Contents lists available at ScienceDirect

Analytica Chimica Acta

journal homepage: www.elsevier.com/locate/aca

A modification of the successive projections algorithm for spectral variable selection in the presence of unknown interferents

Sófacles Figueredo Carreiro Soares[a], Roberto Kawakami Harrop Galvão[b], Mário César Ugulino Araújo[a,*], Edvan Cirino da Silva[a], Claudete Fernandes Pereira[a], Stéfani Iury Evangelista de Andrade[a], Flaviano Carvalho Leite[a]

[a] Universidade Federal da Paraíba, CCEN, Departamento de Química, Caixa Postal 5093, CEP 58051-970 - João Pessoa, PB, Brazil
[b] Instituto Tecnológico de Aeronáutica, Divisão de Engenharia Eletrônica, 12228-900, São José dos Campos, SP, Brazil

ARTICLE INFO

Article history:
Received 19 October 2010
Received in revised form
24 December 2010
Accepted 12 January 2011
Available online 19 January 2011

Keywords:
Variable selection
Successive projections algorithm
Interferents
Ultraviolet–visible and near-infrared spectrometry
Colorants
Gasoline

ABSTRACT

This work proposes a modification to the successive projections algorithm (SPA) aimed at selecting spectral variables for multiple linear regression (MLR) in the presence of unknown interferents not included in the calibration data set. The modified algorithm favours the selection of variables in which the effect of the interferent is less pronounced. The proposed procedure can be regarded as an adaptive modelling technique, because the spectral features of the samples to be analyzed are considered in the variable selection process. The advantages of this new approach are demonstrated in two analytical problems, namely (1) ultraviolet–visible spectrometric determination of tartrazine, allure red and sunset yellow in aqueous solutions under the interference of erythrosine, and (2) near-infrared spectrometric determination of ethanol in gasoline under the interference of toluene. In these case studies, the performance of conventional MLR-SPA models is substantially degraded by the presence of the interferent. This problem is circumvented by applying the proposed Adaptive MLR-SPA approach, which results in prediction errors smaller than those obtained by three other multivariate calibration techniques, namely stepwise regression, full-spectrum partial-least-squares (PLS) and PLS with variables selected by a genetic algorithm. An inspection of the variable selection results reveals that the Adaptive approach successfully avoids spectral regions in which the interference is more intense.

1. Introduction

In general, multiple linear regression models are simpler and more amenable to chemical interpretation as compared to latent variable models obtained by principal component regression (PCR) or partial least squares (PLS). However, MLR usually requires the selection of a suitable subset of variables in order to ensure proper numerical conditioning and to minimize the propagation of random errors [1]. In this context, several variable selection techniques have been proposed in the literature [2]. Examples include Genetic Algorithms [3], Forward Selection, Backward Elimination and Stepwise Regression [4], Generalized Simulated Annealing [5], Mallow's Cp statistic [6], use of branch-and-bound for combinatorial optimization [7], minimization of the condition number of the calibration matrix [8], analysis of MLR weights [9], and the successive projections algorithm [10,11]. The selection of variables in the wavelet domain has also been discussed as an alternative to modelling in the original domain [12–14].

Usually, a variable selection algorithm is expected to choose a small number of informative and non-redundant variables, in order to yield an MLR model that is parsimonious and easy to interpret. However, in comparison with a full-spectrum PCR or PLS model, such an MLR model tends to be less robust to the presence of unknown interferent species not included in the calibration data set. In fact, if the interferent exhibits spectral occurrences in the same region of the selected variables, the model predictions can be severely compromised. In a full spectrum model, such an effect is less pronounced because the interference is averaged over a larger number of analytical channels. On the other hand, if the spectral profile of the interferent does not overlap the spectra of the analytes over the entire working range, it may be possible to select informative variables that are not significantly affected by the interferent. In that case, the resulting MLR model could even outperform full-spectrum models, provided that a suitable criterion can be devised to guide the selection of variables in the presence of interferents.

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

0003-2670/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.aca.2011.01.022

If the spectral profile of the interferent species is known beforehand, the spectral regions in which the interference is more intense could be eliminated *a priori* from the variable selection process. However, such an assumption may be unrealistic, as a precise knowledge of all possible interferents involved in an analytical problem is seldom available. Alternatively, samples containing interferents could be included in the calibration process in order to obtain more robust models. In this case, variable selection algorithms could be employed, for instance, to minimize the error in a set of samples with interferents. Such an idea has been employed in the context of calibration transfer to select variables that are robust with respect to response differences between instruments [15,16]. However, this approach requires the analyst to obtain a set of samples that are representative of the possible interferents involved in the problem, which may not be a straightforward task.

In this context, the present work proposes an approach that does not require *a priori* knowledge of the interferent species, nor the availability of modelling samples containing interferents. The proposed scheme entails a modification to the successive projections algorithm, which was originally developed to minimize collinearity problems in MLR modelling. The modified algorithm retains the collinearity avoidance mechanism of the original SPA formulation, which is useful to prevent the selection of redundant variables. The main modification consists of a change in the cost function, in order to favour the selection of variables in which the effect of interferents is less pronounced.

For this purpose, given a set of unknown samples to be analyzed, it is assumed that: (1) a full-spectrum outlier detection algorithm will be initially employed to separate the samples that display abnormal spectral features as compared to the calibration samples; (2) such an abnormality is caused by the presence of interferents that overlap part of the spectrum of the analytes; and (3) higher-order measurements are not available to remove the interference effects [17]. Under these assumptions, the proposed SPA modification is aimed at selecting a subset of variables specifically tailored to the analysis of the unknown samples under consideration. Such a procedure can be regarded as an adaptive modelling technique, which takes into account the spectral features of the samples to be analyzed.

The advantages of the proposed technique are demonstrated in two analytical problems. The first problem concerns ultraviolet–visible (UV–vis) spectrometric determination of three colorants (tartrazine, allure red and sunset yellow) in aqueous solutions under the interference of erythrosine. The second problem involves near-infrared (NIR) spectrometric determination of ethanol in gasoline under the interference of toluene. In both cases, the detection of samples with the presence of interferents is carried out by using a full-spectrum SIMCA (soft independent modelling of class analogies) classifier. The proposed technique is compared with the standard MLR-SPA, as well as MLR with variables selected by stepwise regression (MLR-SW) [4], full-spectrum PLS and PLS with variables selected by a genetic algorithm (PLS-GA) [18].

2. Background and theory

2.1. Notation

Matrices are represented by bold capital letters, column vectors by bold lowercase letters, and scalars by italic characters. The transpose of a matrix is denoted by superscript *T*. The matrix of instrumental responses for the calibration data set is denoted by $\mathbf{X_{cal}}$ ($N_{cal} \times K$), where N_{cal} and K indicate the number of samples and variables, respectively. The kth variable x_k is associated to the kth column of matrix $\mathbf{X_{cal}}$, which is denoted by $\mathbf{x}_k$ ($N_{cal} \times 1$). The maximum number of variables that can be included in an MLR model with intercept term is given by $M = \min(N_{cal} - 1, K)$.

2.2. The successive projections algorithm

SPA comprises three phases [19,20]. The first phase consists of projection operations involving the columns of matrix $\mathbf{X_{cal}}$, which generate K chains of M variables each (with K, M defined as in Section 2.1). Each element in a chain is selected in order to display the least collinearity with the previous ones [10,11]. 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
$\mathbf{z}^1 = \mathbf{x}_k$ (vector that defines the initial projection operations)
$\mathbf{x}_j{}^1 = \mathbf{x}_j$, $j = 1, \ldots, K$
$SEL(1, k) = k$
$i = 1$ (iteration counter)

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

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

where $\mathbf{I}$ is a ($N_{cal} \times N_{cal}$) identity matrix.

Step 3: Calculate the projected vectors $\mathbf{x}_j{}^{i+1}$ as

$$\mathbf{x}_j^{i+1} = \mathbf{P}^i \mathbf{x}_j^i \quad (2)$$

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 **SEL** matrix:

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

$$SEL(i + 1, k) = j^* \quad (4)$$

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 $\{SEL(1, k), SEL(2, k), \ldots, SEL(m, k)\}$. Since m ranges from one to M and k ranges from one to K and, a total of $M \times K$ subsets of variables are tested. The best subset of variables is selected on the basis of a suitable cost function related to the performance of the resulting MLR model. Usually, the cost function is defined as the *RMSE* (root mean square error) value obtained by cross-validation or by applying the MLR model to a separate validation set [21]. It is worth noting that the cost function evaluation for different subsets of variables is a repetitive task that can be easily implemented in parallel computation platforms, as described in [22].

The third phase consists of a variable elimination procedure aimed at discarding uninformative variables and thus improving the parsimony of the model. For this purpose, the variables selected in phase 2 are sorted according to a relevance index and a scree plot of *RMSE* against the number of variables included in the model is generated. Let $RMSE_{min}$ be the smallest *RMSE* value observed in the scree plot. The optimal solution is then taken as the smallest number of variables such that *RMSE* is not significantly larger than $RMSE_{min}$, according to a statistical hypothesis test. More details concerning phase 3 are given elsewhere [20].

2.3. Proposed modification for variable selection in the presence of unknown interferents

The proposed approach consists of modifying the cost function employed in phase 2 of SPA to take into account spectral abnormalities caused by the presence of interferents. It is assumed that the MLR-SPA model will be applied to the spectra of a set of unknown