Analytica Chimica Acta 811 (2014) 13–22

Contents lists available at ScienceDirect

Analytica Chimica Acta

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

The Successive Projections Algorithm for interval selection in trilinear partial least-squares with residual bilinearization

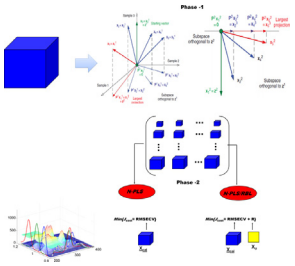
Adriano de Araújo Gomes [a], Mirta Raquel Alcaraz [b], Hector C. Goicoechea [a,b,*], Mario Cesar U. Araújo [a,**]

[a] Laboratório de Automação e Instrumentação em Química Analítica e Quimiometria (LAQA), Universidade Federal da Paraíba, CCEN, Departamento de Química, Caixa Postal 5093, CEP 58051-970, João Pessoa, PB, Brazil
[b] Laboratorio de Desarrollo Analítico y Quimiometría (LADAQ), Cátedra de Química Analítica I, Facultad de Bioquímica y Ciencias Biológicas, Universidad Nacional del Litoral-CONICET, Ciudad Universitaria, 3000 Santa Fe, Argentina

HIGHLIGHTS

- The Successive Projections Algorithm (SPA) is proposed to interval selection to multiway data.
- Discarding of non-informative variables using SPA was proposed for trilinear PLS.
- The feasibility of the proposed method was demonstrated using simulated study and two sets of real data.
- The iSPA-N-PLS models provide better predictions as compared to N-PLS full model and GA-N-PLS.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 21 September 2013
Received in revised form
12 December 2013
Accepted 18 December 2013
Available online 27 December 2013

Keywords:
Multiway data
Variable selection
Second order calibration
Second order advantage

ABSTRACT

In this work the Successive Projection Algorithm is presented for intervals selection in N-PLS for three-way data modeling. The proposed algorithm combines noise-reduction properties of PLS with the possibility of discarding uninformative variables in SPA. In addition, second-order advantage can be achieved by the residual bilinearization (RBL) procedure when an unexpected constituent is present in a test sample. For this purpose, SPA was modified in order to select intervals for use in trilinear PLS. The ability of the proposed algorithm, namely *i*SPA-N-PLS, was evaluated on one simulated and two experimental data sets, comparing the results to those obtained by N-PLS. In the simulated system, two analytes were quantitated in two test sets, with and without unexpected constituent. In the first experimental system, the determination of the four fluorophores (L-phenylalanine; L-3,4-dihydroxyphenylalanine; 1,4-dihydroxybenzene and L-tryptophan) was conducted with excitation-emission data matrices. In the second experimental system, quantitation of ofloxacin was performed in water samples containing two other uncalibrated quinolones (ciprofloxacin and danofloxacin) by high performance liquid chromatography with UV–vis diode array detector. For comparison purpose, a GA algorithm coupled with N-PLS/RBL was also used in this work. In most of the studied cases *i*SPA-N-PLS proved to be a promising tool for selection of variables in second-order calibration, generating models with smaller RMSEP, when compared to both the global model using all of the sensors in two dimensions and GA-NPLS/RBL.

* Corresponding author at: Laboratorio de Desarrollo Analítico y Quimiometría (LADAQ), Cátedra de Química Analítica I, Facultad de Bioquímica y Ciencias Biológicas, Universidad Nacional del Litoral-CONICET, Ciudad Universitaria, 3000 Santa Fe, Argentina. Tel.: +54 342 4575206x190; fax: +54 342 4575205.
** Corresponding author. Tel.: +55 83 3216 7438; fax: +55 83 3216 7437.
E-mail addresses: hgoico@fbcb.unl.edu.ar (H.C. Goicoechea), laqa@quimica.ufpb.br, mariougulino@gmail.com (M.C.U. Araújo).

0003-2670/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.aca.2013.12.022

1. Introduction

In recent years, the dramatic advances in analytical instrumentation allow obtaining a large amount of data per sample in a short time interval. Hyphenated analytical techniques like HPLC-DAD/MS, HPLC–MS-MS, GC–GC–TOFMS can be cited as good examples [1–3]. However, sometimes the large amount of acquired data would not be useful to build calibration models, and part of the recorded signal may be strongly correlated [4,5]. Moreover, some sensors may not provide relevant information for the system under consideration, and in the latter case their use may even compromise the model performance in terms of precision and accuracy [6].

In this context, variable selection as a previous stage in the construction of calibration models can be a tool to improve the predictive capability and robustness, eliminating non-informative variables [6–8]. In the context of first-order calibration, variable selection has been successfully employed building models for a variety of analytical techniques, for example, near-infrared spectroscopy [9,10], UV–vis [11] and plasma emission spectrometry [12]. Several strategies were reported in the literature, such as genetic algorithm (GA) [13], colony of ants (CA) [14], tabu search (TS) [15], simulated annealing (SA) [16] and many others [17–19]. These methods of selection of variables are usually coupled to regression models like multiple linear regressions (MLR) and partial least squares (PLS) [20,21].

Since the past decade, the application of multiway methods [22] has become increasingly important, demonstrating its importance as chemometric tool. Rapid quantitation of several chemical species in complex matrices with minor chemical treatment of the sample has been reported in the literature [23–25]. However, in the context of multiway calibration, variable selection techniques have been scarcely explored. To the best of our knowledge only four papers have been published in this regard: (a) Wu and collaborators presented a method using GA for variable selection together with data modeling by parallel factor analysis (PARAFAC) [26]; (b) Lopes and Menezes applied GA in variable selection to model the performance of industrial fed-batch fermentation process by trilinear PLS [27]; (c) Gourvénec et al. proposed the use GA for variable selection in multivariate curve resolution (MCR) to improve the quality of the on-line monitoring of batch processes [28]; and (d) Carneiro et al. developed a method based on GA for variable selection for a bilinear least squares (BLLS) whit residual bilinearization (RBL) procedure to determination of pesticides and metabolites in wine [29]. Interestingly, the process of variable selection led to improved results in all the mentioned works.

In the field of variable selection, Araújo et al., proposed the Successive Projections Algorithm (SPA) [30,31]. This algorithm was initially implemented as a tool for variable selection in MLR. Afterward, it was modified to work in several areas of chemometrics such as variable selection combined with Uninformative Variables Elimination (UVE) [32], classification [33], calibration transfer [34], sample selection [35] and selection of intervals for PLS regression [36]. Several reports of successful applications of SPA can be found in the literature [37].

The present paper proposes a modification of SPA for the selection of intervals of variables to be used in a trilinear PLS model. The proposed algorithm, namely *i*SPA-N-PLS, combines the properties of trilinear PLS with the possibility of discarding non-informative variables in SPA and the second-order advantage, which is achieved by RBL procedure when unexpected constituents occur in a test sample. In order to evaluate the ability of the new method, one simulated and two experimental data sets were studied. In the simulated system, two analytes were quantitated in two test sets, with and without unexpected constituent. In the first experimental system, the determination of the four fluorophores (L-phenylalanine, L-3,4-dihydroxyphenylalanine, 1,4-dihydroxybenzene and L-tryptophan) was conducted with excitation-emission data. In the second experimental system, quantitation of ofloxacin was performed in water samples containing two other uncalibrated quinolones (ciprofloxacin and danofloxacin) by high performance liquid chromatography with UV–vis diode array detector.

2. Background and theory

2.1. Notation

In what follows, three-way arrays, matrices, vectors and scalars will be denoted by bold capital letters underlined, bold capital letters, bold lowercase letters and italic characters, respectively. The *T* superscript indicates the transpose of a vector or matrix.

2.2. Trilinear PLS

The trilinear PLS or more generally N-PLS, has been proposed by Bro [38] in 1996, although other studies have previously reported the use PLS for trilinear data [39,40]. N-PLS has been proposed as an alternative to the U-PLS (unfolded-PLS), wherein the three-dimensional data structure is unfolded in a two-dimensional matrix [38–40]. When considering the structure of the trilinear data, N-PLS models present more stability and less complexity when compared to U-PLS models [38]. In the other hand, from the point of view of computational effort, outperforms PARAFAC, since N-PLS is based on solving a problem of eigenvectors [38].

Interestingly, the algorithm proposed by Bro in essence is no different of PLS1, in both the decomposition of the independent variables are driven to maximize the covariance between **y** (dependent variable vector) and scores, but unlike the bilinear PLS1, the trilinear PLS1 decompose three-dimensional data arrays ($\underline{\mathbf{X}}_{i\times j\times k}$) in a set of triads. Each triad is comprised of a score vector **t** and two loading vectors, $\mathbf{w}^j$ and $\mathbf{w}^k$, which are the weights in the dimensions *j* and *k*, respectively. A triad can be expressed as shown in Eq. (1):

$$x_{ijk} = t_i w_j^j w_k^k \tag{1}$$

As in the bilinear PLS, in the trilinear model $\mathbf{w}^j$ and $\mathbf{w}^k$ are searched minimizing the squared residuals according to Eq. (2):

$$e^2 = (x_{ijk} - t_i w_j^j w_k^k) \tag{2}$$

The solution by the method of least squares is given by Eq. (3):

$$t_i = \sum_{j=1}^{j}\sum_{k=1}^{k}(z_{jk} w_j^j w_k^k) \tag{3}$$

where z_{jk} are the elements of a **Z** matrix of dimension $(j \times k)$ corresponding to the sum of each of the *i* matrix elements that comprise the three-dimensional data array ($\underline{\mathbf{X}}_{i\times j\times k}$), multiplied by the concentration of the analyte [41], as shown in Eq. (4).

$$\mathbf{Z} = \mathbf{X}_1 y_1 + \mathbf{X}_2 y_2 + \mathbf{X}_3 y_3 + \cdots + \mathbf{X}_I y_I \tag{4}$$

The next step is to determine $\mathbf{w}^j$ and $\mathbf{w}^k$, which is easily achieved by singular value decomposition (SVD) of the matrix **Z**, being **t** estimated using Eq. (3). In the next stage, the regression coefficient vector **v** is calculated employing Eq. (5).

$$\mathbf{v} = (\mathbf{t}^T\mathbf{t})^{-1}\mathbf{t}^T\mathbf{y} \tag{5}$$

The contribution of the factor *fth* is removed, and the next factor is computed in the remaining residue, where each sample $\mathbf{X}_i$ is replaced by $[\mathbf{X}_i - \mathrm{t}_i\mathbf{w}^j(\mathbf{w}^k)^T]$ and **y** by $(\mathbf{y} - \mathbf{Tv})$. The number of factors