Chemometrics and Intelligent Laboratory Systems 100 (2010) 99–109

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Evaluation of the number of factors needed for residual bilinearization in BLLS and UPLS models to achieve the second-order advantage

Jez Willian Batista Braga [1], Renato Lajarim Carneiro, Ronei Jesus Poppi *

Institute of Chemistry, University of Campinas—UNICAMP, P.O. Box 6154, 13083-970, Campinas, SP, Brazil

ARTICLE INFO

Article history:
Received 27 July 2009
Received in revised form 3 November 2009
Accepted 14 November 2009
Available online 8 December 2009

Keywords:
Second-order multivariate calibration
BLLS
UPLS
RBL
Fluorescence
Flow injection analysis

ABSTRACT

Bilinear least squares (BLLS) and unfold partial least squares (UPLS) are second-order multivariate calibration methods, which require the application of the residual bilinearization (RBL) algorithm to achieve the second-order advantage. The present work presents a study of the choice of the number of RBL factors, in BLLS and UPLS models, for two different datasets based on fluorescence and flow injection analysis (FIA) measurements. Confidence limits for the noise level and mean calibration residuals, based on a student-*t* distribution, are proposed as a criterion for determination of the number of RBL factors. Feasible results were obtained based on the proposed confidence limits, but divergences were observed in some situations in the FIA dataset due to either differences in the models or characteristics of the analyte signal. These results suggest, whenever possible, that the number of RBL factors should be checked with a dataset composed by samples where values of the property of interest are known from a reference method.

1. Introduction

Nowadays multivariate calibration can be considered one of the most active research areas in the analytical chemistry field. Advances in the development of additional tools for variable selection [1–3], detection of outliers [3–5], preprocessing methods [6–8], user-friendly software [9–11] and determination of figures of merit for validation of the models [12–15] have made possible its application in the most diverse areas and their acceptance by official regulatory organs, such as the American Society for Testing Materials (ASTM) [16] and the United States Pharmacopoeia [17]. Depending on the data complexity, different multivariate calibration methods are available, being classified as zero, first, second, third order calibrations and so on [18].

Second-order calibrations are particularly attractive for determinations in complex samples, essentially due to the ability to perform determinations even in the presence of interferences unmodeled in the calibration step, a property known as the second-order advantage. Parallel factor analysis (PARAFAC) [19], multivariate curve resolution (MCR) [20], bilinear least squares (BLLS) [21,22], unfold partial least squares (UPLS) [23], and multilinear partial least squares (NPLS) [24] are second-order methods currently used and described in the literature. However, from these methods only PARAFAC and MCR present an inherent second-order advantage. It is necessary to apply a separate process of bilinearization of residuals (RBL) [25,26] to enable BLLS [27–29], UPLS [30,31] and NPLS [32] handle uncalibrated interferences.

For PARAFAC and MCR the calibration samples and one or several test samples are put together and decomposed by the model, so the number of factors necessary to perform the regression model is determined at once and in theory is equal to the number of independent chemical species in the calibration samples plus the unknown interferences in the test samples. In BLLS, UPLS and NPLS this procedure is performed in two steps: (1) the calibration samples are decomposed separately and the number of factors is usually estimated by prior knowledge of the system or by cross-validation methods, except for BLLS, where this number is always equal to the number of calibrated species; (2) when a test sample is analyzed, first the RBL procedure is applied to capture the interference signals and then the prediction is accomplished by a model developed in step (1). In RBL, it is also necessary to estimate the number of factors and it is usually determined by the residuals of decomposition obtained by analyzing the residual matrices of the test sample [32].

Following the procedure used by both BLLS and UPLS when RBL is applied, the choice of the number of factors that is necessary in RBL, in principle, may be different for each sample, since the concentration of the analyte in the prediction samples is calculated in a sequential way. Based on papers already published [3,27–53], the estimate of the number of factors is performed by a plot of the RBL residuals of decomposition, computed for a given trial number of factors in RBL, and the number of factors that provides residuals at a value compatible or not statistically different to the instrumental noise are chosen. However, no statistical test to compare these residuals with the noise level is suggested, which makes this comparison subject to

* Corresponding author. Tel.: +55 19 35213126; fax: +55 19 35213023.
E-mail address: ronei@iqm.unicamp.br (R.J. Poppi).
[1] Universidade de Brasília, Instituto de Química, C.P. 04478, 70904-970 Brasília, DF, Brazil.

0169-7439/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2009.11.006

possible wrong decisions. In addition, there is a lack of papers presenting the residual values of a significant number of samples to verify the agreement of the residuals of decomposition in the plots with the prediction errors observed in the test samples.

In the present work, a study of the choice of the number of RBL factors in BLLS and UPLS models is presented for two different datasets with a significant number of samples and replicates. Confidence limits based on a student-*t* distribution are proposed and evaluated to compare the decomposition residuals of the test samples with the instrumental noise or the residual mean observed in the calibration samples. The applicability of the comparison of the decomposition residuals with these confidence limits or just using the noise level, as suggested in the literature, is discussed in systems where different numbers of interferences are expected. The first dataset studied consisted of fluorescence measurements of ternary mixtures of the polycyclic aromatic hydrocarbons anthracene (AN), benzo[a]pyrene (BAP) and phenanthrene (PHEN), which were chosen due to the overlap observed in their excitation/emission fluorescence signals. In this study four different calibrations were performed to determine BAP, in situations where no interference, a single and two interferences are present in the prediction samples. The second study is a simultaneous determination of ascorbic (AA) and acetylsalicylic acids (ASA) in pharmaceutical preparations by a flow injection analysis system (FIA) with pH gradient and ultraviolet detection, which constitutes a more complex system, presenting two equilibrium species for each analyte. For this dataset both synthetic and commercial pharmaceutical samples were analyzed.

2. Material and methods

2.1. Spectrofluorimetric determination of Benzo[a]pyrene in polycyclic aromatic hydrocarbons mixtures

2.1.1. Reagents and standards

Calibrated volumetric flasks and ultrapure water (Milli-Q, Millipore) were used for all solutions. Calibration and validation mixtures were prepared from 96% (w/w) anthracene (AN) (Fluka), 97% (w/w) benzo[a]pyrene (BAP) (Aldrich) and 97% (w/w) phenanthrene (PHE) (Fluka). Stock solutions of the compounds were prepared in 100 mL volumetric flasks by dissolving 20 mg with acetonitrile (HPLC-grade 99.99%, Tedia). Working solutions with concentrations of 0.100 mg/L, 0.150 mg/L and 0.500 mg/L for AN, BAP and PHE, respectively, were obtained by two appropriate dilutions of the stock solutions using 50:50(v/v) and 25:75 (v/v) acetonitrile:water as solvent, respectively.

The four different calibration conditions consisted of the tertiary mixture BAP|AN|PHE, the binary mixtures containing BAP|AN and BAP|PHE and finally only BAP. The validation dataset consisted of four tertiary mixtures prepared in 21 replicates (84 test samples). Thus, there are three distinct situations that allow the study of the number of RBL factors in the second-order calibration models: (a) three species calibrated, BAP|AN|PHE, and no interference, (b) two calibrated species, BAP|AN and BAP|PHE, with one interference present, PHE or AN, respectively, and (c) just BAP calibrated and the remaining two compounds as interferences. All four calibration sets were formed by five different standard solutions, with concentrations ranges of: 2.5 to 24.5 for AN, 9.5 to 28.0 for BAP and 13.0 to 129.0 for PHE, in μg/L. The concentrations for the four validation samples were 19.5, 4.87, 9.8, and 14.5; 11.5, 16.2, 21.0 μg/L and 25.5; 103.0, 51.7, 25.9 and 77.8 μg/L for AN, BAP and FEN, respectively.

2.1.2. Spectrofluorimetric measurements

The excitation/emission fluorescence data were acquired in the range of wavelengths of 230–400 nm and 240–600 with steps of 5 and 2 nm, for excitation and emission, respectively. A fluorescence spectrometer, Varian Cary Eclipse, with both excitation and emission slits at 5 nm, speed scan in the fastest mode, photomultiplier tube in the medium level and a cell with 1 cm of optical path, was employed. For model building a range of 240–280 and 368–468 nm for excitation and emission, respectively, was selected, resulting in a data matrix of size 9×51. The instrumental noise level was estimated based on 15 measurements of a blank solution.

2.2. Simultaneous determination of ascorbic (AA) and acetylsalicylic acids (ASA) by flow injection analysis

2.2.1. Reagents and standards

Acid carrier solution (H_3PO_4, 0.01 mol L^{-1}) was prepared from 85 % (w/w) phosphoric acid solution (Synth) and the basic solution (Na_2HPO_4, 0.05 mol L^{-1}) was made from a 99.0% (w/w) sodium hydrogen phosphate (Fluka). Standard solutions for calibration and validation were prepared from analytical grade 99.75% (w/w) ascorbic acid (AA) (Fluka), 99.9% (w/w) acetylsalicylic acid (ASA) (Synth) and 99.9 % (w/w) caffeine (Synth).

For the calibration and validation, synthetic mixtures were prepared daily from solutions of 400.0 mg/L of ASA, 240.0 mg/L of AA and 50.0 mg/L of caffeine (used as known interference only in the synthetic validation dataset 2). Due to the high decomposition presented by AA, the solutions (of standards and samples) were prepared at approximately 10 °C in ultrapure water saturated with nitrogen gas and measured immediately after preparing.

The calibration samples (CAL) were composed of 11 sample solutions, formed by 9 synthetic mixtures of ASA and AA, following a composite central design, and by 2 additional samples containing just one of the two analytes (although each of these two analytes gives two acid–base species, both of them spectrophotometrically active). Three independent replicates of these 11 samples were analyzed, giving therefore a total number of 33 samples for the calibration. The total concentration ranges of the two analytes were established based on their absorption in the ultraviolet region and varied between 0 and 136.4 mg/L for ASA and between 0 and 82.0 mg/L for AA.

The validation samples were composed of two independent datasets, the first (VAL1) was composed of 12 synthetic mixtures contain only ASA and AA with concentrations in the ranges defined by the calibration samples. The second one (VAL2) were composed of 5 synthetic mixtures of ASA and AA, to which caffeine was added as interference at a constant concentration of 5.00 mg/L. This concentration level of caffeine was chosen based on its usual concentration present in a commercial pharmaceutical drug (MED3), see below. Three replicates of each validation sample were also analyzed, giving a total number of 36 and 15 data matrices for VAL1 and VAL2, respectively. Fig. 1 shows the experimental design for CAL, VAL1 and VAL2 datasets, where the concentration values of each sample can be observed.

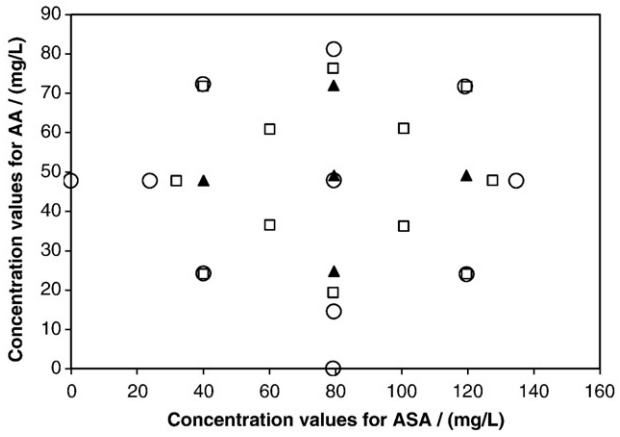

Fig. 1. Experimental design for the datasets (o) CAL, (□) VAL1 and (▲) for VAL2.