Chemometrics and Intelligent Laboratory Systems 170 (2017) 51–57

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Maximum likelihood unfolded principal component regression with residual bilinearization (MLU-PCR/RBL) for second-order multivariate calibration

Jez Willian Batista Braga [a,b], Franco Allegrini [a], Alejandro C. Olivieri [a,*]

[a] Departamento de Química Analítica, Facultad de Ciencias Bioquímicas y Farmacéuticas, Universidad Nacional de Rosario, Instituto de Química de Rosario (IQUIR-CONICET), Suipacha 531, Rosario, S2002LRK, Argentina
[b] Laboratório de Automação, Quimiometria e Química Ambiental, Instituto de Química, Universidade de Brasília, CEP 70904-970, Brasília-DF, Brazil

ARTICLE INFO

Keywords:
Second-order multivariate calibration
Maximum likelihood principal component regression
Residual bilinearization
Error covariance matrix

ABSTRACT

A maximum likelihood model is described for performing second-order multivariate calibration with unfolded principal component regression with residual bilinearization (MLU-PCR/RBL). It differs from the conventional RBL models based on U-PCR or U-PLS (unfolded partial least-squares) in the incorporation of the measurement error information into both the U-PCR calibration and the RBL model phases. The error information is represented by the instrumental error covariance matrix. Simulations were made by adding correlated and proportional noise to synthetic systems consisting of one analyte in the presence of a calibrated and unexpected interferent, under different conditions of overlapping profiles, noise levels and noise types (correlated and proportional). The results show that MLU-PCR/RBL outperforms conventional RBL methods in prediction ability, as confirmed by a detailed study on validation samples through the average prediction error as a convenient figure of merit. Results obtained in experimental data set based on flow injection analysis and UV detection for determination of acetylsalicylic and ascorbic acids in pharmaceutical products also support the theoretical conclusions.

1. Introduction

Second-order multivariate calibration is becoming popular for the quantitative analysis of complex mixtures, exploiting the second-order advantage which is inherent to matrix instrumental data [1]. This specific calibration methodology allows one to determine selected analytes in the presence of uncalibrated interferents using small calibration data sets, by mathematically modeling and separating the contribution of the interfering constituents to the test sample signals [2]. This has opened a rapidly expanding analytical field with a great potentiality towards samples of complex composition [3,4].

Among the various useful second-order calibration models which have already been discussed in the literature, unfolded partial least-squares with residual bilinearization (U-PLS/RBL) is a convenient alternative [5,6]. The model consists of two sequential phases: (1) a calibration phase where the second-order data for the training samples are unfolded and employed to build a classical PLS model, and (2) a residual bilinearization phase in which the contribution of the interferents is modeled using principal component analysis (PCA). The second phase isolates the interfering effect from the total test sample signal, producing analyte PLS scores which can be safely used for prediction using the calibration regression coefficients. The method has found a number of experimental applications, where its latent variable structure adequately modeled matrix data from different instrumental sources [3].

U-PLS/RBL is based on the assumption that the instrumental noise is independently and identically distributed (iid). Likewise, analytical figures of merit for U-PLS/RBL and other second-order calibration methodologies are also based on the iid noise assumption [7–9]. For example, the sensitivity has been defined as the ratio of iid instrumental noise to the prediction uncertainty propagated by the former [7]. Other noise structures are nevertheless possible for multivariate signals, including correlated and proportional noise [10]. It is important to note that for non-iid noise, the plain sensitivity is not a proper indicator of analytical performance, and thus a generalized analytical sensitivity has been proposed, which includes the noise structure in its definition [11]. Recently, figures of merit for second-order calibration with U-PLS/RBL were discussed for non-iid noise structures [12]. However, they were estimated for examples where the conventional U-PLS/RBL version was applied for analyte quantitation, which would in principle be sub-optimal for processing such data sets.

It is clear that alternative RBL calibration models are required when the noise structure is not iid. Maximum likelihood (ML) methods are able

* Corresponding author.
E-mail address: olivieri@iquir-conicet.gov.ar (A.C. Olivieri).

https://doi.org/10.1016/j.chemolab.2017.09.016
Received 31 March 2017; Received in revised form 21 September 2017; Accepted 26 September 2017
Available online 28 September 2017

to take into account the error structure and are therefore an appealing alternative [13]. In this report we describe the MLU-PCR/RBL model, which incorporates the ML philosophy in the two phases mentioned above. In the first one, MLU-PCR is applied to the unfolded calibration matrices [14]; in the second, ML-PCA is employed to model the contribution from the interfering agents within a weighted least-squares fitting procedure [15]. The newly proposed model complements the already developed ML versions of parallel factor analysis (ML-PARAFAC) [16] and of multivariate curve resolution-weighted alternating least-squares (MCR-WALS) [17]. In this way, the ML philosophy is extended to a second-order latent variable model, which may be able to handle analytical systems where trilinear PARAFAC or bilinear MCR-ALS cannot be applied [1].

A comparison with classical U-PCR/RBL and U-PLS/RBL is made using simulated second-order data for systems with controlled chemical compositions, degrees of overlapping among the constituent profiles in both data modes, noise levels and noise types (correlated and proportional). The results favor the new second-order MLU-PCR/RBL model, as can be judged from the average prediction error for a substantial number of independent test samples, all containing uncalibrated interferents. Experimental data sets obtained from a flow injection analysis (FIA) system with UV detection for the determination of acetylsalicylic and ascorbic acids in both synthetic mixtures and pharmaceutical formulations were also studied.

2. Theory

The theory of U-PCR/RBL and U-PLS/RBL is well-known. In U-PCR/RBL, for example, the (unfolded) test sample signal $\mathbf{x}$ is first modeled as the sum of two contributions: (1) the portion of the test signal modeled by the calibration, and (2) the signal from the interferents modeled by RBL:

$\mathbf{x}$ = PCR calibration model of $\mathbf{x}$ + RBL model of interferents + $\mathbf{e}$ =

$$= \mathbf{Pt} + \sum_{n=1}^{N\text{int}} \mathbf{c}_{\text{int},n} \otimes \mathbf{b}_{\text{int},n} + \mathbf{e} \tag{1}$$

where $\mathbf{P}$ is the matrix of U-PCR calibration loadings (size = data points × number of latent variables), $\mathbf{t}$ is the (column) vector of test sample scores (number of latent variables × 1), $\mathbf{b}_{\text{int},n}$ and $\mathbf{c}_{\text{int},n}$ are the profiles in each data mode for the nth interferent, N_{int} is the number of interferents, ⊗ indicates the Kronecker product, and $\mathbf{e}$ is a vector of model errors. The number of latent variables for calibration is usually estimated by leave-one-out cross-validation using only the calibration data, as described by Haaland and Thomas [18].

In eq. (1), the product ($\mathbf{Pt}$) represents the part of $\mathbf{x}$ which can be modeled by the calibration parameters, while the summation of Kronecker products represents the contribution from the interferents in bilinear form. The latter contribution is decomposed into profiles in both data modes $\mathbf{b}_{\text{int},n}$ and $\mathbf{c}_{\text{int},n}$, which are found by PCA of the residual vector ($\mathbf{x} - \mathbf{Pt}$), after being reshaped into an appropriately dimensioned matrix.

The aim of the RBL procedure is to find the score vector $\mathbf{t}$ minimizing the sum of residuals SSR = $\mathbf{e}^{\mathrm{T}}\mathbf{e}$ in eq. (1), rendering at the same time the interferent profiles in each data mode. This can be done by Gauss-Newton minimization or using an iterative procedure (the latter will be described below in detail in connection with MLU-PCR/RBL).

Once $\mathbf{t}$ is found, prediction of the analyte concentration proceeds through the usual expression:

$$\hat{y} = \mathbf{t}^{\mathrm{T}}\mathbf{v} \tag{2}$$

where $\mathbf{v}$ is the vector of latent regression coefficients provided by the U-PCR calibration model.

In MLU-PCR/RBL, on the other hand, the error structure information is introduced in the form of the error covariance matrix (ECM) $\mathbf{\Sigma}_x^2$. In the calibration phase, this information is employed to build an MLU-PCR model with the unfolded signals, rendering a new set of calibration loadings $\mathbf{P}_{\text{ML}}$, and a new vector of latent regression coefficients $\mathbf{v}_{\text{ML}}$. The matrix $\mathbf{P}_{\text{ML}}$ and the vector $\mathbf{v}_{\text{ML}}$ play in MLU-PCR the analogous role of $\mathbf{P}$ and $\mathbf{v}$ in classical U-PCR respectively.

The RBL phase is similar to that for U-PCR/RBL, adapted to the maximum likelihood philosophy. In fact, the relevant RBL expression is analogous to eq. (1):

$$\mathbf{x} = \mathbf{P}_{\text{ML}}\mathbf{t}_{\text{ML}} + \sum_{n=1}^{N\text{int}} \mathbf{c}_{\text{MLint},n} \otimes \mathbf{b}_{\text{MLint},n} + \mathbf{e}_{\text{ML}} \tag{3}$$

where $\mathbf{c}_{\text{MLint},n}$ and $\mathbf{b}_{\text{MLint},n}$ are the profiles in both data modes for the interferent, as estimated by MLPCA of the residual matrix obtained by reshaping ($\mathbf{x} - \mathbf{P}_{\text{ML}}\,\mathbf{t}_{\text{ML}}$), and $\mathbf{t}_{\text{ML}}$ is the final score vector minimizing the weighted sum of squares:

$$\text{WSSR} = \mathbf{e}_{\text{ML}}^{\mathrm{T}}\left(\mathbf{\Sigma}_x^2\right)^{-1}\mathbf{e}_{\text{ML}} \tag{4}$$

In the present report, the RBL phase has been implemented as an iterative procedure following these steps:

(1) With the unfolded sample signal $\mathbf{x}$, the calibration matrix of loadings $\mathbf{P}_{\text{ML}}$ and the ECM matrix, the initial score vector $\mathbf{t}_{\text{ML}}$ is found by the maximum likelihood PCR expression [14]:

$$\mathbf{t}_{\text{ML}} = \left[\mathbf{P}_{\text{ML}}^{\mathrm{T}}\left(\mathbf{\Sigma}_x^2\right)^{-1}\mathbf{P}_{\text{ML}}\right]^{-1}\mathbf{P}_{\text{ML}}\left(\mathbf{\Sigma}_x^2\right)^{-1}\mathbf{x} \tag{5}$$

(2) The residuals of the sample model are obtained by subtracting the MLU-PCR modeled part of $\mathbf{x}$ from the overall sample signal $\mathbf{x}$:

$$\mathbf{r} = \mathbf{x} - \mathbf{P}_{\text{ML}}\,\mathbf{t}_{\text{ML}} \tag{6}$$

(3) The residual vector $\mathbf{r}$ is reshaped into a matrix, and MLPCA is applied to decompose the latter matrix into a sum of bilinear terms using N_{int} principal components. This leads to the estimation of the interferent profiles $\mathbf{b}_{\text{MLint},n}$ and $\mathbf{c}_{\text{MLint},n}$ (n indicates each of the N_{int} potential interferents).

(4) The contribution of the interferents is subtracted from the sample signal $\mathbf{x}$, and a new $\mathbf{t}_{\text{ML}}$ score vector is found:

$$\mathbf{t}_{\text{ML}} = \left[\mathbf{P}_{\text{ML}}^{\mathrm{T}}\left(\mathbf{\Sigma}_x^2\right)^{-1}\mathbf{P}_{\text{ML}}\right]^{-1}\mathbf{P}_{\text{ML}}\left(\mathbf{\Sigma}_x^2\right)^{-1}\left(\mathbf{x} - \sum_{n=1}^{N\text{int}} \mathbf{c}_{ML\text{int},n} \otimes \mathbf{b}_{ML\text{int},n}\right) \tag{7}$$

(5) The RBL residual vector $\mathbf{e}_{\text{ML}}$ is found from eq. (3).

(6) The weighted sum of squared residuals WSSR is computed from eq. (4).

(7) The procedure returns to step (2) and is repeated until convergence, indicated by no significant changes in WSSR within a certain predefined tolerance, typically 0.001% for successive values of WSSR.

The number of potential interferents N_{int}, i.e., the dimensionality of the MLPCA model in step (3) is estimated by increasing N_{int} until stabilization of the final value of WSSR, as is usual in classical U-PCR/RBL or U-PLS/RBL [19].

(8) After finding $\mathbf{t}_{\text{ML}}$, prediction is made through an equation analogous to (2):

$$\hat{y}_{\text{ML}} = \mathbf{t}_{\text{ML}}^{\mathrm{T}}\mathbf{v}_{\text{ML}} \tag{8}$$

where $\mathbf{v}_{\text{ML}}$ is the calibration vector of latent regression vectors. If data are mean-centered, the mean calibration concentration should be added to

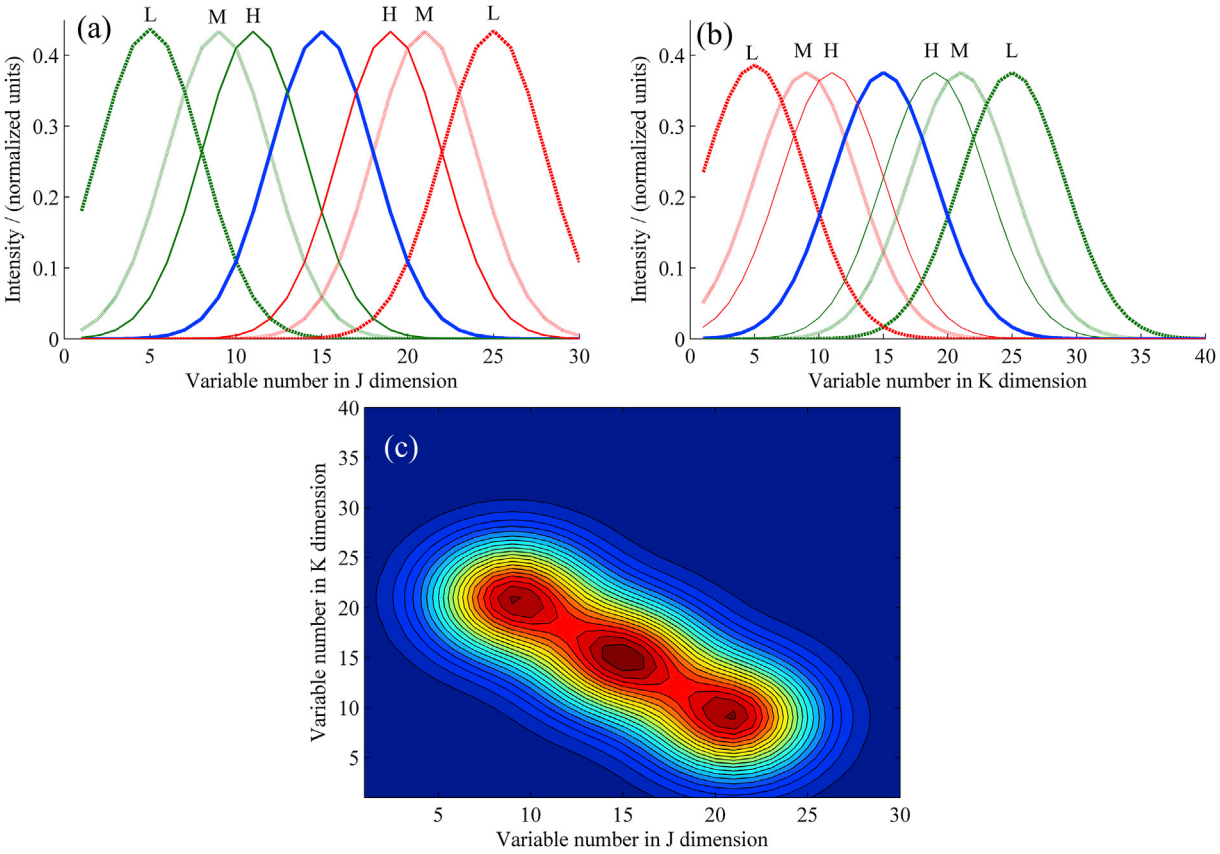

Fig. 1. Position of the simulated spectral profiles in the (a) *J* and (b) *K* modes for the (—) analyte, (—) calibrated interferent and (—) unexpected interferent in the conditions of (L) low, (M) moderate and (H) high overlap. (c) Contour plot of a sample with moderate overlapping condition of the analyte with both interferents.

the right-hand side of eq. (8).

3. Datasets

3.1. Simulated data sets

Simulated second-order data were generated according to the procedures already described in Ref. [12]. The calibration set was composed of 50 samples containing two components, the target analyte with concentrations linearly distributed between 0 and 1 and a single interferent with random concentrations, also in the range of 0–1. The test set was composed of 25 samples containing the two components present in the calibration phase and a single unexpected interferent, not included in the calibration samples, with concentrations in the range of 0.1–1.

Nine overlapping conditions were evaluated according to a 3^2 factorial design, wherein the levels were three degrees of overlap (low, moderate and high) and the factors were the overlap between: (1) the analyte and the calibrated interferent (factor 1) and (2) the analyte and the unexpected interferent (factor 2). The position of the spectral profiles in the different overlapping conditions used for the evaluation of the models are presented in Fig. 1, where the analyte is positioned between the calibrated and unexpected interferent.

The dimensions of each simulated instrumental mode were $J = 30$ and $K = 40$. The code used to generate and execute the simulations were written in Matlab® version 7.14 (R2012a) using routines developed in our research group.

3.1.1. Noise addition

In order to evaluate the performance of the models, two noise structures were used: (1) correlated noise in two distinct levels of the inverse noise frequency, $1/f$ (pink noise) and $1/f^2$ (brown noise) and (2) proportional noise. Pink and brown noise vectors were generated by a MATLAB function [8] to match the size of the simulated instrumental mode with a larger number of sensors (*K*) and added to each row of this mode. Both for pink and brown noise, the noise sequence was scaled so that the standard deviation corresponds to a fixed percentage of the maximum signal of the sample. The following levels of correlated noise were tested: 0.1, 0.25, 0.5, 0.75 and 1.0%.

Proportional noise is characterized by a heteroscedastic behavior, varying as a function of the magnitude of the instrumental signal, according to a scaling factor or noise level. The proportional noise vector to be added to each sample was determined by multiplying the noise level by the unfolded data and by a vector containing normally distributed random numbers. According to this procedure, the noise in each data sensor will be the specific noise level (i.e. 1%) of its signal intensity. It should be noticed that data simulations containing only proportional noise can lead to mathematical issues that prevent the inversion of the error covariance matrices. To avoid this potential problem, an iid noise vector was added to each sample containing proportional noise, scaled to represent 1% of the proportional noise level present in the sensor with the highest signal intensity. The following levels of proportional noise were evaluated: 0.5, 1.0, 2.5, 5.0, 7.5 and 10.0%.

Considering all overlapping conditions, type of noises and noise levels, 144 different calibration and test set samples were created. Additionally, for the estimation of the root mean square error of prediction (RMSEP) for MLU-PCR/RBL, U-PCR/RBL and U-PLS/RBL, all the procedures were repeated 50 times. Further details of the noise addition and Monte Carlo simulations can be found in Refs. [8,12].

3.1.2. Estimation of the error covariance matrices

In the simulated data, the error sources and their structures are known and well characterized, so that the theoretical prediction of the

Table 1
RMSEP values obtained in the different types and levels of noise for the three overlapping conditions between the analyte and the unexpected interferent, and for the highest overlap with the calibrated interferent.[a]

Model		RMSEP × 10^2								
		U-PLS/RBL			U-PCR/RBL			MLU-PCR/RBL		
	Overlap level	L	M	H	L	M	H	L	M	H
Pink Noise level/(%)	0.1	0.09	0.09	0.15	0.08	0.09	0.14	0.07	0.07	0.11
	0.25	0.21	0.23	0.36	0.20	0.22	0.33	0.18	0.19	0.28
	0.5	0.43	0.45	0.70	0.40	0.42	0.64	0.36	0.37	0.55
	0.75	0.64	0.68	1.04	0.61	0.65	0.96	0.54	0.55	0.87
	1	0.85	0.89	1.39	0.81	0.84	1.30	0.71	0.72	1.19
Brown Noise level/(%)	0.1	0.08	0.09	0.13	0.08	0.08	0.12	0.05	0.05	0.07
	0.25	0.20	0.22	0.31	0.19	0.21	0.28	0.12	0.13	0.17
	0.5	0.42	0.42	0.59	0.39	0.40	0.54	0.26	0.26	0.35
	0.75	0.59	0.60	0.87	0.56	0.57	0.81	0.37	0.38	0.53
	1	0.76	0.81	1.13	0.72	0.77	1.06	0.50	0.53	0.73
Proportional Noise level/(%)	0.5	0.09	0.10	0.15	0.08	0.09	0.14	0.08	0.10	0.13
	1.0	0.16	0.21	0.29	0.15	0.19	0.27	0.14	0.20	0.25
	2.5	0.41	0.50	0.69	0.38	0.47	0.65	0.38	0.48	0.62
	5.0	0.83	0.97	1.28	0.78	0.93	1.22	0.75	0.94	1.12
	7.5	1.22	1.41	1.98	1.22	1.41	2.00	1.18	1.41	1.84
	10	1.61	1.96	2.71	1.55	1.90	2.68	1.48	1.95	2.48

[a] L: low overlap; M: moderate overlap; H: high overlap. The values of RMSEP × 10^2 represent the relative prediction error (in %).

ECMs can be obtained. For the simulations with pink and brown noise, all calibration and test samples present the same ECM, which were obtained directly by the correlated noise function presented in Refs. [8,12]. The calibration phase of MLU-PCR/RBL will require, in general, an ECM for the training data matrix of size $JKI \times JKI$. However, the fact that no correlations are assumed among different samples considerably simplifies the calculations to a single ECM of size $JK \times JK$ [8]. We assume that second-order data collection involves the measurement of J vectors, each with K channels, with no correlation among them (for example, among emission spectra at different excitation wavelengths in excitation-emission fluorescence spectroscopy, or among spectra at different elution times in chromatography or FIA with multivariate detection). Therefore, this ECM is a block-diagonal matrix, each of the identical J blocks of size $K \times K$, corresponding to the ECM along the K channels. In the subsequent RBL phase, i.e., eq. (4), the ECM corresponds to the residuals, after refolding them into a matrix. Here the MLPCA modeling of the interferents employs an ECM composed of just one of the blocks, of size $K \times K$. Finally, in eq. (5), an ECM with the same general characteristics as the one for calibration was employed.

On the other hand, for proportional noise, each calibration and test sample will present a distinct ECM, which formally requires the application of a general version of MLPCR, as demonstrated by Wentzell et al. [13,15]. However, from an experimental point of view, it might be problematic to access individual ECMs for every sample, since this requires many replicates for each sample. In most experimental situations, the usually viable solution for estimating ECMs is the simplifying assumption that all samples present approximately the same (average) ECM. Although this is not formally true, it may represent a better approximation than the naive condition of a constant diagonal ECM (only consistent with the presence of iid noise). This approximation was already reported in first-order calibration, leading to significant improvements in prediction ability in the presence of proportional noise [14]. Therefore, in the present simulations involving proportional noise, a single ECM was considered for both calibration and test sets. For the estimation of this ECM, a three-dimensional array $\underline{\mathbf{X}}$ (size $I \times J \times K$) containing the I calibration samples was averaged along the modes I and J, resulting in an average vector containing the spectral profile with K data points, assumed to be the simulated instrumental detection mode. This average spectrum was multiplied by the noise level and placed at the diagonal of the ECM. Finally, a constant value equal to the variance of the iid noise introduced in the data was added to the diagonal of the ECMs (see above).

3.2. Experimental datasets

The experimental data were previously presented in Refs. [20,21], and comprise the determination of both ascorbic acid (AA) and acetylsalicylic acid (ASA) in two different systems of synthetic mixtures and four different pharmaceutical products. A flow injection analysis system with pH gradient and UV detection was employed for collecting the second-order pH-spectral data. The dimensions of each data matrix are as follows: 321 data points in the spectral mode (range of wavelengths: 180–820 nm with a resolution of 2 nm) and 291 data points in the time (pH) mode.

The calibration set was composed of eleven samples, formed by nine synthetic mixtures of ASA and AA following a central composite design [20], and by two additional samples containing either of the two analytes. Three independent replicates of these eleven samples were prepared and analyzed, leading to a total of 33 calibration matrices. The concentrations of the analytes ranged from 0 to 136.4 mg L^{-1} for ASA and 0–82.0 mg L^{-1} for AA.

Two validation sets composed of synthetic mixtures were prepared to evaluate the models in samples containing a single analyte (validation set 1, VAL 1) and in the presence of an uncalibrated interferent (validation set 2, VAL 2). Validation set 1 was composed by twelve mixtures containing only ASA and AA in concentrations ranging from 31.7 to 128.3 mg L^{-1} and 19.0–77.0 mg L^{-1}, respectively. On the other hand, validation set 2 was composed by five synthetic mixtures containing ASA in concentrations varying between 40.0 and 120.0 mg L^{-1}, AA between 24 and 72 mg L^{-1} and caffeine at a constant concentration of 5.00 mg L^{-1}. This concentration level of caffeine was chosen based on its usual concentration present in the commercial pharmaceutical drug named Doril®. Three replicates of these samples were also analyzed, giving a total number of 36 and 15 matrices for validation sets 1 and 2, respectively.

The analyzed pharmaceutical products (PP) were: PP1, Aspirina® +C (Bayer S.A.), PP2, Melhoral® C (DM Indústria Farmacêutica LTDA), PP3, Doril® (DM Indústria Farmacêutica LTDA) and PP4, Sandoz® (Novartis Biociências S.A.). The first two products contained both analytes as active principles and possibly additional excipient interferences. In contrast, Doril® contains only ASA and caffeine as active principle, while Sandoz® just AA as active and potential excipient interferences.

For the preparation of the pharmaceutical samples, a specific mass of each of the four PPs was weighed, dissolved and diluted in volumetric flasks to obtain concentrations approximately equal to the center of the

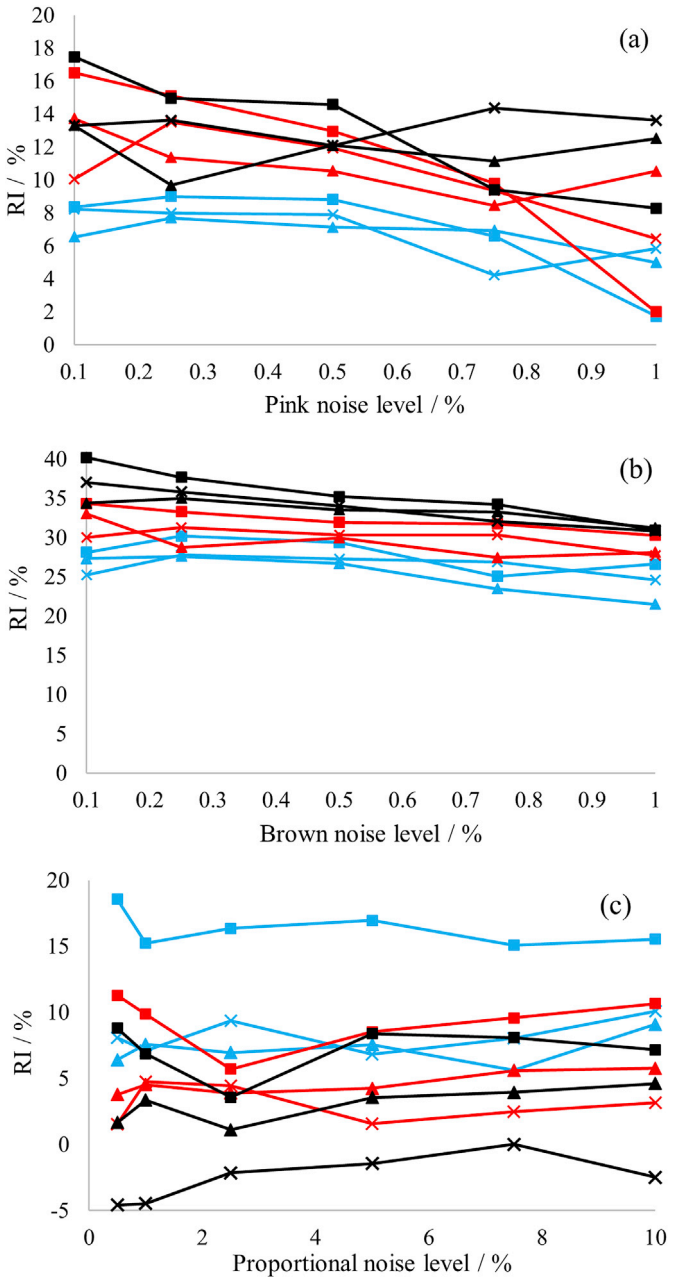

Fig. 2. Relative improvement of the RMSEP values comparing U-PCR/RBL and MLU-PCR/RBL for (a) pink noise, (b) brown noise and (c) proportional noise in the nine overlap conditions: (▲) low, (x) moderate and (■) high overlap with the calibrated interferent; (▬) low, (▬) moderate and (▬) high overlap with the unexpected interferent. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

analytical range defined by the calibration samples. Additionally, ASA and/or AA were added at three distinct levels to each one of the previously prepared pharmaceutical samples that contained ASA and/or AA. Overall, six different samples were prepared for each pharmaceutical product (three without and three with the addition of one or both analytes depending on the pharmaceutical product) and all these samples were analyzed in triplicate in the FIA system, producing 18 matrices for each one of the four pharmaceutical products investigated in this work and a total number of 72 (18 × 4) pharmaceutical samples were analyzed by FIA. The reference values for ASA e AA in the PP were obtained by HPLC and iodimetric titration [22], respectively.

4. Results

4.1. Simulated data

The results for all models were obtained using two latent variables in the calibration phase and one RBL component, due to the known composition of the calibration and test samples. Table 1 presents the RMSEP values for the three overlapping conditions between the analyte and the unexpected interferent, when the overlap with the calibrated interferent is set at the highest level. The detailed results for all nine overlapping conditions are presented in Tables S1, S2 and S3 of the Supplementary Material. Overall, the RMSEP follows, for all situations, an increasing linear relation with the noise level. At the same time, the RMSEP values also increase as the overlapping increases, with a larger effect of the unexpected interference compared to the calibrated interferent.

Comparing the performance of the three models, U-PCR/RBL and U-PLS/RBL present very similar results, while for almost all situations, MLU-PCR/RBL shows lower RMSEPs. In terms of average prediction error, the relative improvement (RI) in the results related to the maximum likelihood approach can be expressed as:

$$RI = 100 \times (RMSEP_{U\text{-}PCR/RBL} - RMSEP_{MLU\text{-}PCR/RBL})/ RMSEP_{U\text{-}PCR/RBL} \quad (9)$$

Fig. 2 shows more clearly the behavior observed for the RI values in all conditions. For correlated noise, the best results were obtained with brown noise, where RI ranged from 21 to 40%, whereas for pink noise it ranged from 2 to 17%. This behavior agrees with the fact that brown noise presents long-range correlations in the data, which can only be successfully modeled by the ML approach. Additionally, depending on the overlapping condition, the value of RI appears to be approximately constant or following a small decrease as a function of the noise level. On the other hand, as the overlapping increases, the improvement is larger, which is more clearly noticed for brown noise.

The results with proportional noise follow a similar trend, in that ML presents the best analytical results. However, as the pooled calibration ECM was employed in the modeling, the complete structure of the noise was not considered. As a result, the simplification provides RI values ranging only from −5 to approximately 20%, with no improvement observed in the conditions of moderate and high overlap with the calibrated and unexpected interferent. In addition, in contrast to correlated noise, the conditions presenting low overlap were the ones leading to the best results. To explain this behavior, further simulations were conducted changing the position of the spectral maxima for the calibrated and unexpected interferent in both modes (resulted not shown), which revealed that the overlapping effect in the ML improvement vary in a complex way with the interferent positions. The elucidation of this behavior will be the focus of future simulation studies. However, in all simulated cases, the ML approach showed a clear improvement in prediction ability, showing that the presently adopted simplification can handle, at least in part, the consequences of the presence of proportional noise in the data.

4.2. Experimental data

As in the simulation studies, experimental data sets described in Section 3.2 were analyzed using three models: classical U-PCR/RBL, classical U-PLS/RBL, and MLU-PCR/RBL. To estimate the ECM for the latter ML model, a similar strategy to the one described in Ref. [12] was applied. This analogy was possible due to the following similarities between the system presented in Ref. [12] and the presently described one: (1) both have a low number of replicates and (2) it is possible to find a "stable region" for each data matrix, which in our experimental data corresponds to the time range between 180 and 291 s, where the spectra remain almost constant, as can be observed in Fig. S1 of the Supplementary Material. In addition, as presented in Fig. S1, the spectral range between 208 and 298 nm was selected. Assuming that the main source of

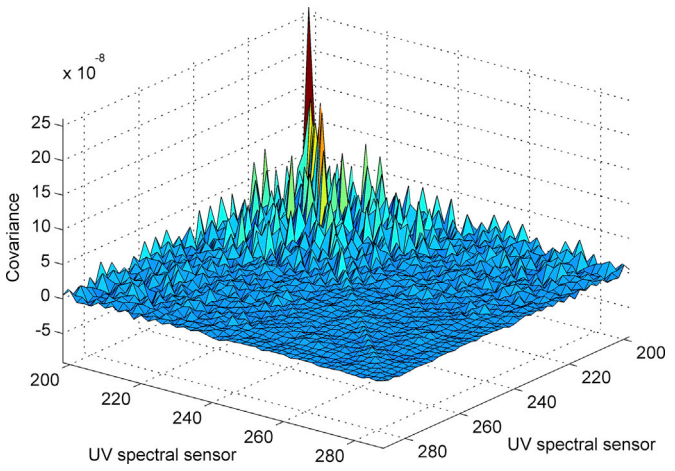

Fig. 3. Pooled error covariance matrix calculated from the stable region of FIA calibration matrices of the samples belonging to the experimental data set.

Table 2
Comparison of RMSEP values obtained for the experimental data sets processed using different models and relative improvement (RI) values.[a]

Data set	UPCR/RBL		UPLS/RBL		MLUPCR/RBL		
	N_{int}	RMSEP	N_{int}	RMSEP	N_{int}	RMSEP	RI[b]
AAS							
VAL1	0	1.0	0	0.98	0	1.3	−30
VAL2	1	9.8	1	9.5	1	2.4	76
PP 1	2	6.2	2	13	2	1.5	76
PP 2	2	14	2	27	2	4.5	68
PP 3	1	8.8	1	9.9	1	2.9	67
PP 4	1	6.1	1	4.5	1	4.9	20
AA							
VAL1	0	0.81	0	0.49	0	0.57	30
VAL2	1	1.5	1	1.4	1	1.3	13
PP 1	2	2.8	2	2.9	2	1.2	57
PP 2	2	8.4	2	7.6	2	2.5	70
PP 3	1	2.0	1	2.0	1	0.87	57
PP 4	1	2.2	1	2.1	1	0.95	57

[a] Concentration values expressed as mg L^{-1}. N_{int}: number of RBL latent variables used to model unexpected interferents. VAL1: validation set 1. VAL2: validation set 2. PP: Pharmaceutical Products. PP1: Aspirina® +C (Bayer S.A.). PP2: Melhoral® C (DM Indústria Farmacêutica LTDA). PP3: Doril® (DM Indústria Farmacêutica LTDA). PP4: Sandoz® (Novartis Biociências S.A.).
[b] Computed with respect to UPCR/RBL.

error comes from the UV spectral measurements at each time, it is possible to find an empirical estimation for the instrumental error, by pooling along the stable region the ECMs built from the spectra extracted at different times and using both replicates, as described in Ref. [12]. The procedure consists of two steps: an ECM is first built from replicate spectra within a certain time region. This step is repeated for each spectrum in the pooling region, and can be summarized by:

$$\mathbf{\Sigma}_j^2 = \sum_{n=1}^{N} \left(\mathbf{x}_{nj} - \overline{\mathbf{x}}_j\right)\left(\mathbf{x}_{nj} - \overline{\mathbf{x}}_j\right)^{\mathrm{T}} \tag{10}$$

where $\mathbf{\Sigma}_j^2$ is an ECM from spectral replicates at time j, $\mathbf{x}_{nj}$ is the spectrum for replicate n and time j, and $\overline{\mathbf{x}}_j$the mean replicate spectrum calculated at time j. These matrices may be extremely noisy due to the small number of replicates. Thus, in a second step an ECM is built by pooling the matrices from the previous step:

$$\mathbf{\Sigma}_{pools}^2 = \left(\mathbf{\Sigma}_1^2 + \mathbf{\Sigma}_2^2 + \ldots + \mathbf{\Sigma}_j^2\right) \Big/ J_r \tag{11}$$

where $\mathbf{\Sigma}_{pools}^2$ is the pooled ECM for a given sample and J_r is the number of time points included in the pooling region. eq. (11) provides an ECM with considerably lower noise, and is a valid resource to infer about the error structure along the spectral mode.

In a recently published article by Wentzell et al. [23], an alternative strategy to model multivariate measurement errors using an objective function based on the Wishart distribution has been presented. This work showed that it is possible to obtain simple models that adequately describe the error structure. However, no further studies were made regarding the use of these models for calibration and prediction. Moreover, as recognized by the authors, a significant amount of work still remains to characterize other first-order analytical systems. The application of the presently discussed modeling strategies to second-order systems remains as a promising perspective.

Fig. 3 illustrates the resulting ECM, estimated from the calibration samples of the experimental data set, which was subsequently used to build the MLU-PCR/RBL model. Its visual inspection suggests that spectral proportional noise along the spectral mode contributes as a significant error source affecting the system. To confirm this suspicion, the mean signal intensity is plotted against its standard deviation at each spectral sensor, as shown for all calibration matrices in Fig. S2 (Supplementary Material).

In the calibration phase of the models compared in this work, the optimal number of calibration latent variables was 5. This value can be explained on the basis of the properties of the analytes used to build the samples, because of the two different acid-base forms adopted by AA and ASA according to the pH value in the gradient, and an additional physical effect generated by the change of the refraction index in the gradient zone, known as Schlieren effect. However, unexpected interferents may arise in test samples, making it essential to determine the number of latent variables that optimally model the contribution of these interferents in the RBL phase. The criterion here adopted to estimate this latter number was the obtainment of root mean squared values of the RBL decomposition residuals at the same level to those observed in the calibration.

There are some important observations to make regarding the model development. First, the stable region mentioned above only contains useful information regarding error structure elucidation, but does not significantly contribute in predictive terms. The initial half of the gradient region contains the same information of the second half, but shows a larger Schlieren effect. Moreover, its inclusion would lead to a considerable increase in the calculation time of the algorithm, given the large size of the data matrices in both modes. Therefore, the useful time interval of 57 s in the range from 94 to 150 s was used for model development. Second, the same simplification applied for the case of proportional noise in the simulated data sets was also employed for the experimental data sets. This means that both in the calibration and RBL phases, the same ECM was employed, calculated by pooling individual ECMs obtained by the procedure described in Ref. [12] for each calibration sample (Fig. 3).

Table 2 shows the prediction results for the concentration of AA and ASA in the different test sets. In general, a significant decrease in RMSEP is observed when the ML strategy is used, compared to the classical models, with RI values reaching a maximum of 76%. Although the improvement appears to be significantly larger than for the simulated datasets, the trend is in agreement with the theoretical results. Furthermore, even though the pooled ECM indicates that proportional noise contributes significantly to the overall noise, the off-diagonal elements also indicate the presence of correlated noise in the data. The contribution of both noise types is consistent with the significantly better results obtained in the experimental data with the ML model. Additionally, the most notorious changes were obtained when the second-order advantage is employed, which highlights the benefits of implementing the ML strategy in the RBL procedure.

5. Conclusions

The results obtained in both simulated and experimental data sets

strongly indicate that the introduction of the maximum likelihood approach in the conventional second-order U-PCR/RBL model leads to better analytical performance in terms of lower prediction errors. The use of the pooling strategy for estimating the error covariance matrix showed significant analytical improvement, even with a relevant contribution of proportional noise is present in the second-order data. The newly introduced model complements the already known maximum likelihood versions of trilinear parallel factor analysis and bilinear multivariate curve resolution, and can be applied when the data cannot be adequately modeled by the latter two methods.

Acknowledgements

AO and FA thank Universidad Nacional de Rosario (Project No. 19B/487), CONICET (Consejo Nacional de Investigaciones Científicas y Técnicas, Project No. PIP 0163) and ANPCyT (Agencia Nacional de Promoción Científica y Tecnológica, Project No. PICT-2016-1122) are gratefully acknowledged for financial support. JWBB thanks CAPES/PROFORENCE (process AUXPE nº 3509/2014), CNPq (process nº 308748/2015-8) for the financial aid, and the Institute of Chemistry of the UNICAMP for all the infrastructure used for the experimental measurements of the FIA data set. FA thanks CONICET for a post-doctoral fellowship.

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.chemolab.2017.09.016.

References

[1] A.C. Olivieri, G.M. Escandar, Practical Three-way Calibration, Elsevier, Waltham (USA), 2014.
[2] K.S. Booksh, B.R. Kowalski, Theory of *analytical chemistry*, Anal. Chem. 66 (1994) 782A–791A.
[3] G.M. Escandar, H.C. Goicoechea, A. Muñoz de la Peña, A.C. Olivieri, Second- and higher-order data generation and calibration: a tutorial, Anal. Chim. Acta 806 (2014) 8–26.
[4] A.C. Olivieri, Recent advances in analytical calibration with multi-way data, Anal. Meth. 4 (2012) 1876–1886.
[5] J. Öhman, P. Geladi, S. Wold, Residual bilinearization. Part I. Theory and algorithms, J. Chemom. 4 (1990) 79–90.
[6] A.C. Olivieri, On a versatile second-order multivariate calibration method based on partial least-squares and residual bilinearization. Second-order advantage and precision properties, J. Chemom. 19 (2005) 253–265.
[7] A.C. Olivieri, Analytical figures of merit: from univariate to multiway calibration, Chem. Rev. 114 (2014) 5358–5378.
[8] F. Allegrini, P.D. Wentzell, A.C. Olivieri, Generalized error-dependent prediction uncertainty in multivariate calibration, Anal. Chim. Acta 903 (2016) 51–60.
[9] A.C. Olivieri, F. Allegrini, Recent advances in analytical figures of merit: heteroscedasticity strikes back, Anal. Meth. 9 (2017) 739–743.
[10] P.D. Wentzell, Measurement errors in multivariate chemical data, J. Braz. Chem. Soc. 25 (2014) 183–196.
[11] W. Fragoso, F. Allegrini, A.C. Olivieri, A new and consistent parameter for measuring the quality of multivariate analytical methods: generalized analytical sensitivity, Anal. Chim. Acta 933 (2016) 43–49.
[12] F. Allegrini, A.C. Olivieri, Multi-way figures of merit in the presence of heteroscedastic and correlated instrumental noise: unfolded partial least-squares with residual multi-linearization, Chemom. Intell. Lab. Syst. 158 (2016) 200–209.
[13] P.D. Wentzell, D.T. Andrews, B.R. Kowalski, Maximum likelihood multivariate calibration, Anal. Chem. 69 (1997) 2299–2311.
[14] S.K. Schreyer, M. Bidinosti, P.D. Wentzell, Application of maximum likelihood principal components regression to fluorescence emission spectra, Appl. Spectrosc. 56 (2002) 789–796.
[15] P.D. Wentzell, D.T. Andrews, D.C. Hamilton, K. Faber, B.R. Kowalski, Maximum likelihood principal component analysis, J. Chemom. 11 (1997) 339–366.
[16] L. Vega-Montoto, P.D. Wentzell, Maximum likelihood parallel factor analysis (MLPARAFAC), J. Chemom. 17 (2003) 237–253.
[17] M. Dadashia, H. Abdollahi, R. Tauler, Application of maximum likelihood multivariate curve resolution to noisy data sets, J. Chemom. 27 (2013) 34–41.
[18] D.M. Haaland, E.V. Thomas, Partial least-squares methods for spectral analyses. 1. Relation to other quantitative calibration methods and the extraction of qualitative information, Anal. Chem. 60 (1988) 1193–1202.
[19] Ref. [1], Partial Least-squares with Residual Bilinearization, Chapter 9.
[20] R.L. Carneiro, J.W.B. Braga, R.J. Poppi, R. Tauler, Multivariate curve resolution of pH gradient flow injection mixture analysis with correction of the Schlieren effect, Analyst 133 (2008) 774–783.
[21] J.W.B. Braga, R.L. Carneiro, R.J. Poppi, Evaluation of the number of factors needed for residual bilinearization in BLLS and UPLS models to achieve the second-order advantage, Chemom. Intell. Lab. Syst. 100 (2010) 99–109.
[22] Official Monographs, United States Pharmacopoeia USP28NF23, 2005, p. 178.
[23] P.D. Wentzell, C.D. Cleary, M. Kompany-Zareh, Improved modeling of multivariate measurement errors based on the Wishart distribution, Anal. Chim. Acta 959 (2017) 1–14.