Chemometrics and Intelligent Laboratory Systems 100 (2010) 127–135

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Residual bilinearization combined with kernel-unfolded partial least-squares: A new technique for processing non-linear second-order data achieving the second-order advantage

Alejandro García-Reiriz, Patricia C. Damiani, Alejandro C. Olivieri *

Departamento de Química Analítica, Facultad de Ciencias Bioquímicas y Farmacéuticas, Universidad Nacional de Rosario e Instituto de Química Rosario (IQUIR-CONICET), Suipacha 531, Rosario (2000), Argentina

ARTICLE INFO

Article history:
Received 9 June 2009
Received in revised form 20 November 2009
Accepted 23 November 2009
Available online 27 November 2009

Keywords:
Second-order calibration
Second-order advantage
Residual bilinearization
Kernel partial least-squares

ABSTRACT

A new second-order multivariate calibration model is presented which allows one to process matrix data showing a non-linear relationship between signal and concentration, and achieving the important second-order advantage. The latter property permits analyte quantitation even in the presence of unexpected sample components, i.e., those not present in the calibration set. The model is based on a combination of residual bilinearization, which provides the second-order advantage, and kernel partial least-squares of unfolded data, a flexible non-linear version of partial least-squares. The latter one involves projection of the measured data onto a non-linear space, which in the present case consists of a set of Gaussian radial basis functions. Simulations concerning two ideal systems are analyzed: one where the signal–concentration relation is quadratic with positive deviations from linearity, and another one where it is sigmoidal. The results are favorably compared with those provided by several artificial neural network approaches. Two experimental systems are also studied, involving the analysis of: 1) the lipid degradation product malondialdehyde in olive oil samples, where the background oil provides a strong interferent signal, and 2) the antibiotic amoxicillin in the presence of the anti-inflammatory salicylate as interferent. The results for these experimental cases are also encouraging.

1. Introduction

The processing of second- and higher-order data have attracted the attention of chemometricians in recent years for a variety of reasons: 1) they are now abundantly produced by modern analytical instruments, 2) they show peculiar mathematical characteristics which distinguish them from first-order data, and 3) they provide analytical chemists with the important second-order advantage, an intrinsic property which permits analyte quantitation in the presence of unexpected sample components (i.e., components not present in the calibration set of samples) [1].

Several algorithms are available for the convenient processing of second-order data, with specific characteristics which have been reviewed in recent years [1–3]. When the relationship between signal and concentration is linear, available algorithms achieving the second-order advantage are based on: 1) calibration with latent variables, such as unfolded partial least-squares (U-PLS) [4], where 'unfolded' refers to working with previously vectorized data matrices [5], and multi-way PLS (N-PLS) [6], both combined with residual bilinearization (RBL) [7–10], which is the procedure providing the second-order advantage to both of these PLS versions, 2) alternating least-squares (ALS), such as parallel factor analysis (PARAFAC) [11], some of its variants [12–14], and multivariate curve resolution-alternating least-squares (MCR-ALS) [15], 3) direct least-squares, such as bilinear least-squares (BLLS) in its several versions [16–18], also combined with RBL to obtain the second-order advantage, and 4) eigenvector–eigenvalue techniques, such as the generalized rank annihilation method (GRAM) [19].

Recently, the possibility of achieving the second-order advantage from non-linear second-order data has been advanced [20], and soon several hybrid algorithms combining unfolded-principal component analysis (U-PCA), residual bilinearization and a number of artificial neural networks (ANN) have been devised [21], and applied to both second- [22,23] and higher-order [24] experimental data. In all of these cases, neither U-PCA nor the different ANNs is able to achieve the second-order advantage, since they process data whose two-dimensional structure has been removed by the unfolding. The second-order advantage is provided by the RBL procedure, a genuinely matrix-based method which refolds the U-PCA residuals into the original two-dimensional structure in order to remove the contribution of the interferents from the test sample data.

The purpose of this work is to introduce a new technique, which combines residual bilinearization with a flexible non-linear PLS

* Corresponding author. Tel./fax: +54 341 4372704.
E-mail address: olivieri@iquir-conicet.gov.ar (A.C. Olivieri).

0169-7439/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2009.11.009

method (i.e., kernel-PLS) [25–27]. The latter one is being increasingly applied to process non-linear first-order instrumental data (mainly of spectroscopic origin) [28–30]. In the present report, we show its capability to handle non-linear second-order data, and, in combination with RBL, to quantitate the analytes in the presence of unexpected sample constituents. Both simulations and experimental examples show that the analytical figures of merit of the presently described algorithm compare well with those provided by previously discussed approaches, which were based on U-PCA/RBL processing followed by: 1) multiple perceptron back-propagation ANN, 2) radial basis functions and 3) support vector machines. Moreover, in the simulated cases, where the functional relationship between signal and concentration is known *a priori*, an analysis is made of the minimum root-mean-squared error which can be expected in the presence of non-linearities, concluding that the present approach provides a satisfactory approximation to this limit.

2. Simulations

Data were simulated for multi-component mixtures having two calibrated analytes, and a single potential interferent appearing in the test samples along with the analytes. Noiseless profiles for the analytes and the potential interferent are shown in Fig. 1A and B in both data dimensions, leading to data matrices of size 50×40 data points. They may be viewed as mimicking experimental systems such as fluorescence excitation–emission matrices, UV–visible-chromatographic retention time matrices, etc. Using the analyte profiles shown in Fig. 1, a calibration set of 25 samples was built having random concentrations

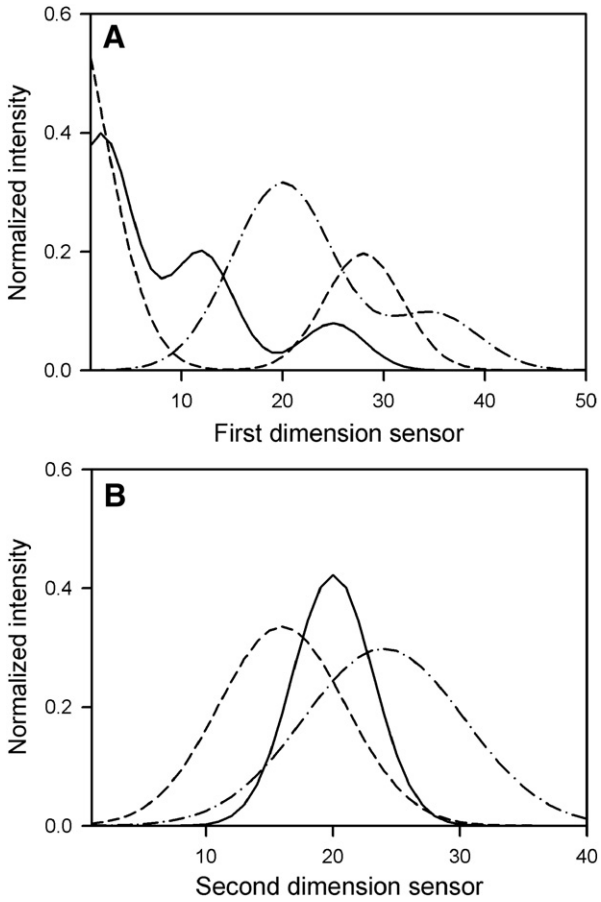

Fig. 1. Noiseless profiles employed for the simulations, in the first (A) and second (B) dimension. In both cases, the solid line corresponds to analyte 1, the dashed line to analyte 2, and the dash-dotted line to the potential interferent.

(both analyte concentrations were taken as random numbers, distributed with equal probability in the range 0–1). The relationship between signal and concentration for analyte 1 was considered to take two alternative non-linear forms: a quadratic function with a positive deviation from linearity (system S1) and a sigmoidal function (system S2). These specific functional forms were chosen because they mimic the signal–concentration behavior in both of the experimental systems which will be described below.

In system S1, the signal–concentration relationship for analyte 1 is governed by the following quadratic equation:

$$\mathbf{X}_1 = \mathbf{S}_1(y_1 + ay_1^2) \tag{1}$$

where $\mathbf{X}_1$ is the matrix signal at concentration y_1, a is a parameter controlling the deviations from the ideal linearity (in our case $a = 0.8$) and $\mathbf{S}_1$ is a pure-analyte bilinear matrix given by the product of the corresponding spectral profiles in each dimension:

$$\mathbf{S}_1 = \mathbf{b}_1\mathbf{c}_1^{\mathrm{T}} \tag{2}$$

where $\mathbf{b}_1$ and $\mathbf{c}_1$ are the $(J\times 1)$ and $(K\times 1)$ profiles in dimension 1 and 2 respectively, (J and K are the number of channels in each dimension) and the superscript 'T' indicates matrix transposition. The profiles $\mathbf{b}_1$ and $\mathbf{c}_1$ are shown in Fig. 1A and B and were both normalized to unit length. Hence the Frobenius norm of the matrix $\mathbf{S}_1$ is unity, a fact which will be employed below in connection with the estimation of figures of merit for the present model.

In system S2, on the other hand, the non-linear relationship between analyte signal and concentration is sigmoidal:

$$\mathbf{X}_1 = \mathbf{S}_1\left[\frac{1}{1 + \exp(b - cy_1)} - \frac{1}{1 + \exp(b)}\right] \tag{3}$$

where $b = 1.5$ and $c = 3$, and $\mathbf{X}_1$ and $\mathbf{S}_1$ have the same meaning as in Eq. (1).

In both simulated systems, the signal for analyte 2 is considered to be linearly related to its concentration:

$$\mathbf{X}_2 = \mathbf{S}_2 y_2 \tag{4}$$

with $\mathbf{S}_2$ given as a bilinear product analogous to Eq. (2). The profiles for analyte 2 (Fig. 1A and B) were also normalized to unit length.

To produce the calibration data, the signal for a typical sample was given by the sum of the contributions of both analytes. For the 500 test samples, on the other hand, both analytes were considered to be present at concentrations which were also taken at random from the range 0–1, but different from the calibration values. These test samples did also contain the potential interferent, at concentrations taken at random from the range 0.5–1.5 (to ensure that a significant amount of interferent is always present), and having a signal $\mathbf{X}_3$ given by an equation analogous to Eq. (4). Hence these synthetic data require the second-order advantage to be achieved in order to adequately predict the analyte concentration in the test samples. The corresponding profiles (Fig. 1A and B) were normalized to unit length.

Once the noiseless calibration and test matrices were built, Gaussian noise was added to all signals. The standard deviation varied between 0.001 and 0.01 units (see below). These extremes represent 0.25% and 2.5% respectively with respect to the maximum calibration signal in the quadratic system, and 0.5% and 5% in the sigmoidal system. The data matrices were then processed using several non-linear second-order multivariate calibration methods.

3. Experimental

3.1. Apparatus

Fluorescence spectral measurements were performed on a fast-scanning Varian Cary Eclipse fluorescence spectrophotometer, equipped with two Czerny–Turner monochromators and a xenon flash lamp, and connected to a PC microcomputer via an IEEE 488 (GPIB) serial interface. Excitation–emission matrices were recorded in a 10 mm quartz cell. For the experimental system involving the analyte amoxicillin, the cell was thermostated at 80 °C for system E1 and at 15 °C for system E2.

3.2. Experimental system E1: analyte malondialdedyde

Aqueous malonaldehyde (1.00×10^{-3} M) was prepared from 1,1,3,3,-tetraethoxypropane (TEP, Sigma-Aldrich, Steinheim, Germany): to 20.0 μL of TEP (ca. 0.0200 g), 10 mL of HCl 0.01 M were added, and the mixture was heated at 50 °C during 60 min. The solution was then neutralized with NaOH, diluting with water to 100.00 mL. A methylamine solution (0.1 M) was prepared by dissolving a suitable amount of reagent (Fluka, Steinheim, Germany) in water.

A twenty one-sample calibration set with malonaldehyde in seven triplicate concentration levels (0.00, 0.30, 0.60, 0.90, 1.20, 1.79 and 2.39 mg L^{-1}, concentrations correspond to the measuring cell). Adequate volumes of malonaldehyde 1.00×10^{-3} M were placed in a 10.00 mL volumetric flask, adding 3 mL of isopropanol, 2.5 ml of sodium acetate/acetic acid buffer (1 M, pH = 3.8), 1 mL of methylamine 0.1 M, and completing to the mark with water. Spiked olive oils samples were prepared as follows: known volumes (0.500 mL) of different olive oils were artificially added with malonaldehyde, in order to contain the analyte in the range 5.00–20.0 μg. Blank, analyte-free oil samples were also studied. Each sample was mixed with 10.0 mL of hexane, placed in an extraction flask and extracted with 10.0 mL of a solution prepared by mixing 2.9 mL of sodium acetate/acetic acid buffer solution (1 M, pH = 3.8) and 3.5 mL of isopropanol. Then 8.50 mL of the aqueous phase were transferred to a 10.00 mL volumetric flask, 1.0 mL of methylamine solution was added, and the flask was completed to the mark with distilled water. In this manner, the final analyte concentrations in the spiked oils were within the calibration range (they were however different than those employed for model training).

All excitation–emission fluorescence matrices were recorded in the following ranges: excitation, 385–424 nm each 3 nm, emission 448–487 nm each 4 nm, after a reaction time of 30 min. The data arrays were thus of size 14×11, making a total of 154 data points. This analytical system has already been analyzed using third-order data, including the time dimension of the reaction of the analyte with methylamine [24]. In the present paper, the time reaction has been fixed at its optimum (30 min) in order to analyze the corresponding second-order data.

3.3. Experimental system E2: analyte amoxicillin

UV-induced fluorescence excitation–emission matrices were measured for the determination of the antibiotic amoxicillin in the presence of salicylate. Amoxicillin (LLCM Laboratory, Santa Fe, Argentina), potassium periodate (Fluka, Buchs, Switzerland) and potassium dihydrogen phosphate (Merck, Darmstadt, Germany) were employed. A stock solution of amoxicillin (0.130 g L^{-1}) was prepared by dissolving an appropriate amount of amoxicillin standard in water and sonicating during 15 min. Amoxicillin solutions were prepared every two weeks and stored in a refrigerator at 4 °C. The calibration set had seven duplicate concentration levels of amoxicillin in the range from 0.00 to 6.30 mg L^{-1}. They also contained a phosphate buffer (0.8 mol L^{-1}, pH = 5.84) and potassium periodate (0.007 mol L^{-1}). The test set contained five samples having amoxicillin at concentrations different than those employed in the training phase. They also contained salicylate as a fluorescent interferent, in concentrations ranging from 0.002 to 0.009 mg L^{-1}. All samples were subjected to photo-activated reaction during 30 min, by irradiating them with a 125 W high pressure mercury lamp.

All fluorescence excitation–emission matrices were read in the excitation range 300–360 each 4 nm, and emission range 370–470 each 4 nm, i.e., the size of each data matrix was $16\times26=416$ data points per sample. This analytical system has already been studied using artificial neural networks combined with residual bilinearization [21], and is included here for comparison of present and previous results.

4. Theory

The essentials of residual bilinearization have already been discussed in Refs. [7,9]. The important outcome of the RBL procedure, as applied to a given array of test sample data, is a vector of sample scores which have been freed from the effect of the potential interferents, permitting the reconstruction of the portion of the test sample data which can be explained by the calibration model. This important step provides the second-order advantage to the whole scheme.

For non-linear U-PLS calibration with second-order data, we propose to first unfold them, and then to apply the kernel-PLS method described in Ref. [25]. Briefly, it consists of building the kernel matrix **K**, whose individual elements are given by the projection of the measured data onto a non-linear Gaussian space:

$$\mathbf{K}(i,i') = \exp(-\| vec(\mathbf{X}\mathrm{cal},i) - vec(\mathbf{X}\mathrm{cal},i')\| / \sigma) \quad (5)$$

where vec() indicates the unfolding operation, **X**cal,i and **X**cal,i' are the second-order data matrices for two calibration samples, $\|\ \|$ is the norm of a vector, and σ is the width of the Gaussian transformation. As can be seen, the matrix **K** is of size $I_{cal}\times I_{cal}$, where I_{cal} is the number of calibration samples.

The next step is to build a PLS model between the matrix **K** and the analyte concentrations contained in the vector **y**, using a certain number of latent variables, a procedure which will provide a vector of regression coefficients **β**, of size $I_{cal}\times1$.

If there were no potential interferents in the unknown sample data matrix **X**unk, the latter would be projected as in Eq. (5):

$$\mathbf{k}_u(i) = \exp(-\| vec(\mathbf{X}\mathrm{cal},i) - vec(\mathbf{X}\mathrm{unk})\| / \sigma) \quad (6)$$

providing a kernel vector $\mathbf{k}_u$ for the unknown sample, which renders the analyte concentration y_u from:

$$y_u = \beta^T \mathbf{k}_u \quad (7)$$

When unexpected components are present in a test sample, however, Eq. (6) can no longer be employed to compute the test kernel vector. Nevertheless, successful analyte prediction is still possible, provided **X**unk is replaced in Eq. (6) by the reconstructed portion of the unknown data matrix which can be explained using the calibration model. This is provided by the above commented RBL procedure. The entire process is outlined in Fig. 2.

The RBL method is usually implemented using U-PCA to filter the test sample data from the contribution of unexpected components, as described in detail in Ref. [9]. In practice, it is convenient to employ both calibration and test sample scores instead of vectorized data matrices in the kernel-PLS procedure. This implies building the Kernels by replacing vec(**X**cal,i) in Eqs. (5) and (6) by the vector of calibration scores for the ith sample $\mathbf{t}_{cal}$ (size $A_{cal}\times1$, where A_{cal} is the number of principal components required to describe the variability

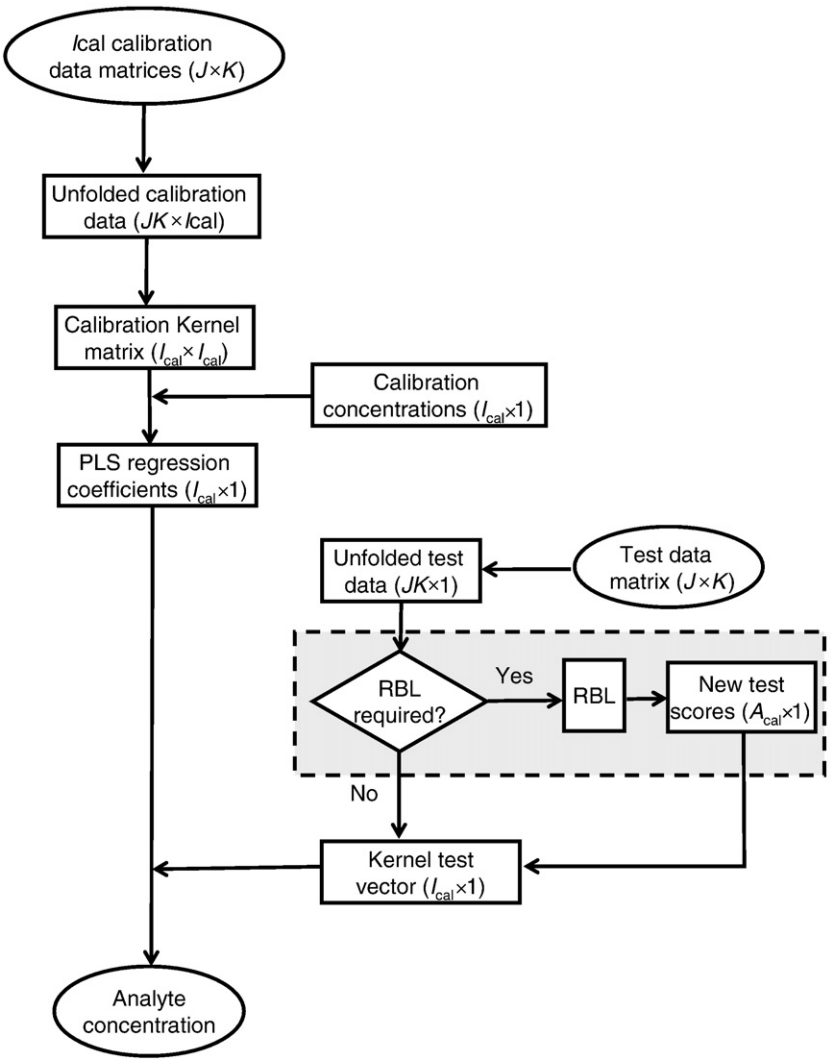

Fig. 2. Scheme outlining the entire kernel U-PLS/RBL procedure. The gray rectangle highlights the activities involved in the RBL procedure which provides the second-order advantage to the combined algorithm. The sizes of the different vector and matrices involved in the scheme are indicated.

in the unfolded calibration data). Likewise, vec(**X**unk) is replaced in Eq. (6) by the test sample scores $\mathbf{t}_{unk}$. The latter is directly provided by the RBL procedure.

From the above discussion it is clear that two parameters should be tuned before calibration and prediction: the width σ and the number of latent PLS variables N. We propose to estimate both values in a single procedure using leave-one-out cross-validation. In order to be conservative in the number of latent variables, we conduct cross-validation for a range of values of σ and N. Then a comparison is made between the concentrations predicted by the model for all pairs of (σ,N) values and those at the absolute minimum of the predicted error sum of squares (PRESS). We employed a suitable statistical test for this purpose, selecting N as the lowest value which provides a certain probability for the comparison of its PRESS with the minimum PRESS. This helps to avoid overfitting, and does not require and independent set of samples for tuning σ and N.

5. Software

The presently discussed RBL/kernel U-PLS model was applied using an in-house MATLAB 7.0 routine [31]. It will be incorporated into a graphical interface for second-order multivariate calibration, and made available through the internet at http://www.chemometry.com/Index/Links%20and%20downloads/Programs.html.

Multiple perceptron networks were applied in the Bayesian regularization mode [32,33], which does not require an independent monitoring set for estimating the training parameters, as implemented in the MATLAB 7.0 Neural Network Toolbox. RBF networks were implemented using the forward selection method described by Orr in Ref. [34] and available at http://www.anc.ed.ac.uk/rbf/rbf.html. SVM were implemented using the LS-SVM lab toolbox (MATLAB/C Toolbox for Least-Squares Support Vector Machines) available at http://www.esat.kuleuven.ac.be/sista/lssvmlab/, and described in the accompanying manual [35]. All programs were run on an IBM-compatible microcomputer with an Intel core duo T7100, 1.80 GHz microprocessor and 2.00 GB of RAM.

6. Results and discussion

6.1. Simulated systems

We first concentrate in the quadratic non-linear system S1 with positive deviations from linearity, whose behavior with respect to the concentration of the analyte of interest is described by Eq. (1). The 25 calibration data matrices were subjected to estimation of the important parameters σ and N. Fig. 3A shows a three-dimensional plot of the obtained PRESS values when the signal noise level was equal to 0.003 units. It is customary to analyze the PRESS values for

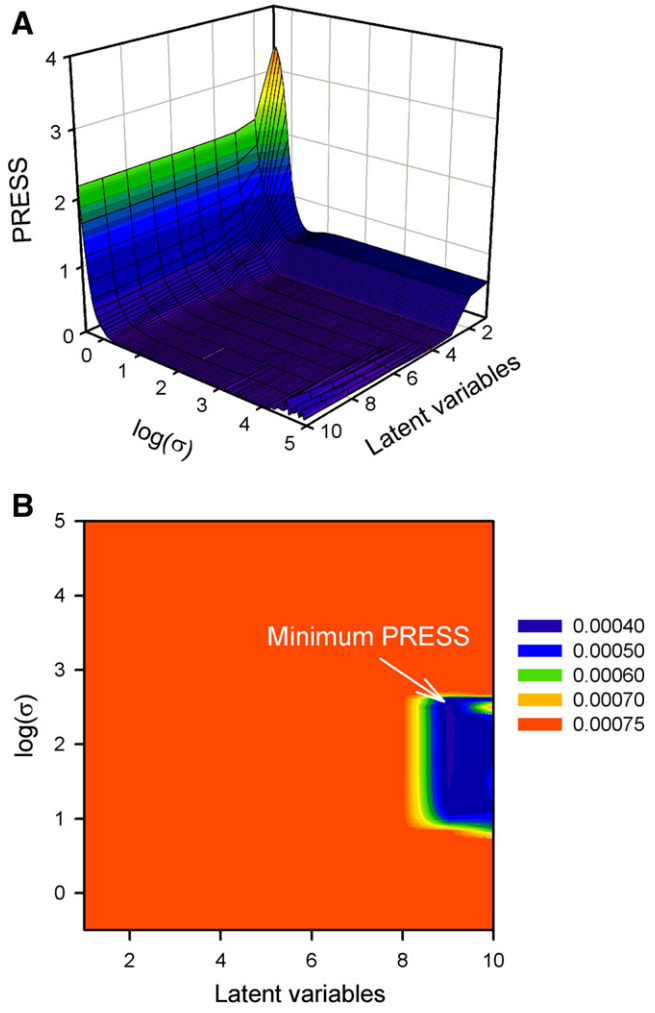

Fig. 3. A) Three-dimensional plot of leave-one-out cross-validation PRESS (predicted error sum of squares) as a function of the logarithm of σ and the number of latent PLS variables N. The values correspond to the simulated quadratic system S1 when the noise in signals was 0.003 units. B) Contour plot of the PRESS values of plot A) cut to a maximum equal to ($F_{0.75,25,25}\times$minimum PRESS). The position of the absolute minimum PRESS is indicated. The colored bar at the right shows the contour heights. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

selecting the number of latent variables using Haaland's criterion [36], i.e., choosing N as the lowest number for which the F ratio between the PRESS values and the absolute minimum PRESS does not exceed a probability of 0.75 (considering that the degrees of freedom are the number of calibration samples). When the values of PRESS are cut at the absolute minimum of Fig. 3A times $F_{0.75,25,25}$ (equal to 1.96), the corresponding contour plot is shown in Fig. 3B, with the position of the minimum indicated [$\sigma=316$, $\log(\sigma)=2.50$, $N=9$]. A rather shallow region is obtained around the absolute minimum (Fig. 3B), with many values of PRESS seemingly equivalent from a statistical point of view. A conservative criterion thus suggests that $N=8$ is a reasonable choice. Subsequent analysis of the PRESS values for $N=8$ and different values of σ gives $\sigma=316$ as the optimum for this specific N. Hence $\sigma=316$, $N=8$ were selected for further data processing.

Using these tuned parameters, predictions can be made for the analyte of interest (component no. 1) in the test samples. This requires to perform RBL for each of the 500 test samples, employing two principal components to model the calibration data and one principal component to model the contribution of the unexpected component to the signal. These numbers of PCs stem from the known composition of the calibration and test samples (two analytes and a single interferent), but in a general case further statistical considerations may be required, to be detailed below for the experimental systems. Once the sample signals were freed from the interferent effects, prediction proceeded as described in the theory section [Eq. (7)]. For the sake of comparison, results will be shown for processing the data under the following three conditions: 1) kernel U-PLS of raw test sample data, 2) kernel U-PLS of RBL-filtered test data, and 3) regular U-PLS (i.e., U-PLS not using the non-linear approach) of RBL-filtered data. This comparison is made in order to appreciate the need of combining both RBL and kernel U-PLS for obtaining reasonable predictions in this non-linear second-order system in the presence of interferents.

Fig. 4 plots the corresponding predictions in the above commented three different scenarios. It is apparent that the best approach is the presently discussed RBL/kernel U-PLS procedure. The plot of predicted values when no RBL is applied (Fig. 4A) shows the expected lack of accuracy (a systematic bias is observed) when unexpected sample components are not taken into account in the second-order model. On

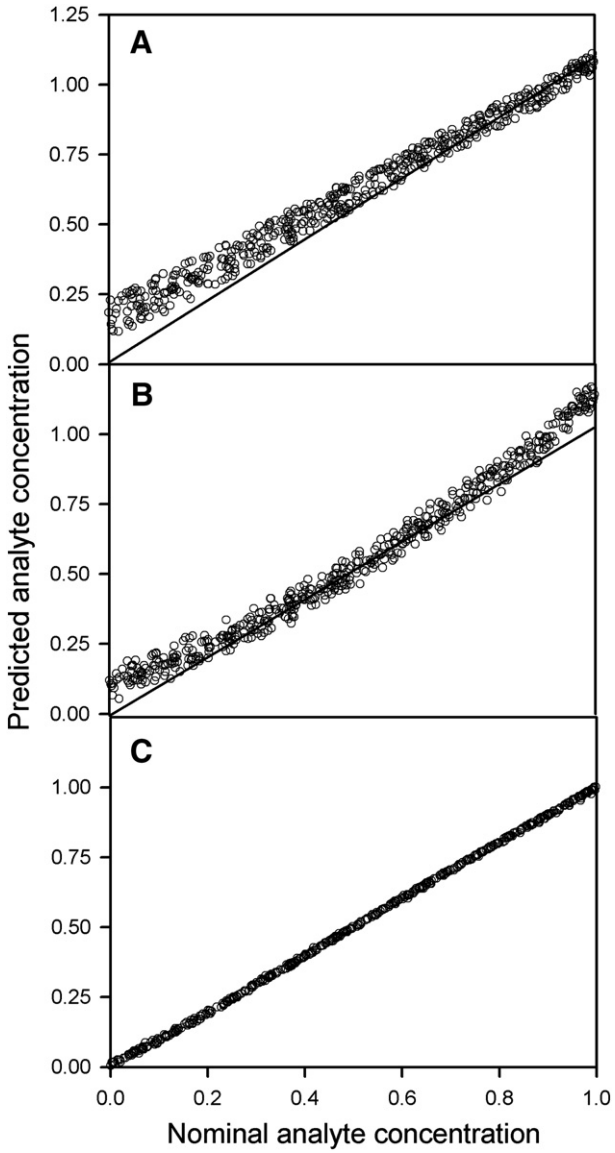

Fig. 4. Predicted vs. nominal concentrations for the simulated quadratic system S1 when the noise in signals was 0.003 units. A) kernel U-PLS analysis of the raw data, i.e., without RBL to model the interferents, B) RBL followed by normal PLS regression of unfolded data, C) RBL combined with kernel U-PLS regression. In all cases the open circles indicate the individual predictions and the solid line the ideal fit.

the other hand, even when RBL is applied (Fig. 4B), the use of a linear model predicts concentration values with a large positive deviation, as expected from the known behavior of this non-linear system [Eq. (1)]. Although visual inspection is evidently favoring the present approach (Fig. 4C), root-mean-square errors of prediction (RMSEP) speak by themselves: 0.12 concentration units for the raw data, 0.068 for RBL with normal U-PLS, and 0.0053 for RBL with kernel U-PLS.

The sigmoidal system S2 leads to similar results to those commented above for the quadratic system S1, except that the figures of merit are poorer because of the lower sensitivity provided by Eq. (2) in comparison with Eq. (1). In fact, predictions lead to the following RMSEP values: 1) 0.30 concentration units for processing the raw data, 2) 0.28 for RBL followed by normal U-PLS, and 3) 0.019 for RBL with kernel U-PLS. Comparing the best RMSEP for both systems, the sigmoidal best RMSEP is more than 3 times larger than the best quadratic RMSEP, in agreement with the higher sensitivity for system S1 with respect to system S2.

It is interesting to wonder if the achieved figures of merit with RBL/kernel U-PLS are indeed satisfactory. This question can be answered by comparing the results from the presently discussed procedure with those already described in the literature, which are known to produce acceptable results, i.e., U-PCA/RBL followed by different artificial neural network approaches, as discussed in Ref. [21]. Table 1 shows such results, which point to a good agreement among the different techniques. Details for the implementation of the ANN approaches can be found in Ref. [21].

Further insight into the comparison of the above non-linear approaches involves the estimation of the expected concentration uncertainties. Using an approach which has already been found to be useful in linear systems [37], one can estimate the lowest uncertainty which could be expected for the predicted concentrations, knowing the level of instrumental noise introduced in the simulated systems, the degree of spectral overlapping among the component profiles, and the functional relationship between signal and concentration for analyte 1. In cases of linear signal–concentration relationship, the RMSEP has been found to be similar to the estimated concentration uncertainty [37].

It has already been shown that when two analytes and one interferent occur, such as in the present study, the sensitivity is given (in the linear case) by the following expression [38]:

$$S_n = s_n\{[(\mathbf{B}_{exp}^{T}\mathbf{P}_{b,unx}\mathbf{B}_{exp})^{*}(\mathbf{C}_{exp}^{T}\mathbf{P}_{c,unx}\mathbf{C}_{exp})]^{-1}\}_{nn}^{-1/2} = s_n O_f \quad (8)$$

where s_n is the integrated total signal for component n at unit concentration, and the overlapping factor O_f is computed from the matrices $\mathbf{B}_{exp}$ and $\mathbf{C}_{exp}$ containing the profiles for all expected components in each dimension [i.e., those present in the calibration sample set, with '*nn*' implying the selection of the (*n*,*n*) element corresponding to the *n*th analyte of interest], and from two projection matrices which are orthogonal to the space spanned by the unexpected components in each of the data modes:

$$\mathbf{P}_{b,unx} = \mathbf{I} - \mathbf{B}_{unx}(\mathbf{B}_{unx})^{+} \quad (9)$$

$$\mathbf{P}_{c,unx} = \mathbf{I} - \mathbf{C}_{unx}(\mathbf{C}_{unx})^{+} \quad (10)$$

where $\mathbf{B}_{unx}$ and $\mathbf{C}_{unx}$ contain the profiles for the unexpected components as columns, $\mathbf{I}$ is an appropriately dimensioned unit matrix, and '+' stands for the pseudo-inverse.

In a non-linear system, an appropriate correction to Eq. (8) is to replace s_n by the derivative of pure-analyte signal vs. concentration, i.e., by:

$$(d\|\mathbf{X}_1\|/dy) = \|\mathbf{S}_1\|(1-2ay_1) \quad (11)$$

in the quadratic system, and by

$$(d\|\mathbf{X}_1\|/dy) = \|\mathbf{S}_1\|\frac{c\exp(b-cy)}{[1+\exp(b-cy)]^2} \quad (12)$$

in the sigmoidal system. Recall that $\|\mathbf{S}_1\| = 1$ in both cases ($\|\ \|$ represents the Frobenius norm of a matrix), as discussed above when describing the data simulations. Both Eqs. (11) and (12) predict a sensitivity parameter which will vary with the analyte concentration, and hence one should strictly speak of average figures of merit for a non-linear analytical system.

The average uncertainty in predicted concentrations is expected to be given by the ratio of signal noise to average sensitivity $\langle S_n\rangle$ [37]:

$$\langle SD(y)\rangle = \frac{SD(X)}{\langle S_n\rangle} = \frac{SD(X)}{\langle(d\|\mathbf{X}_1\|/dy)\rangle O_f} \quad (13)$$

where $(d\|\mathbf{X}_1\|/dy)$ is computed as the average of either Eqs. (11) or (12) over the range of analyte concentrations. The average derivative is easily computed as:

$$(d\|\mathbf{X}_1\|/dy) = \frac{\int_0^1 (d\|\mathbf{X}_1\|/dy)dy}{\int_0^1 dy} = \|\mathbf{X}_1(y)\|\|_0^1 = \|\mathbf{X}_1(1)\| \quad (14)$$

In the present case, $\|\mathbf{X}_1(1)\|$ are 1.8 and 0.635 for the quadratic and sigmoidal system respectively. The overlapping factor O_f, on the other hand, is equal to 0.55 for both systems [as computed from Eq. (8) and the noiseless component profiles shown in Fig. 1]. These figures justify the higher average sensitivity of system S1 with respect to S2 by a factor of $(1.8/0.635) = 2.8$.

The expected value of $\langle SD(y)\rangle$ for the quadratic system is thus computed as 0.003 concentration units, using $SD(X) = 0.003$ in Eq. (13). Although the estimated uncertainty is lower than the RMSEP values quoted in Table 1, it should be remembered that the latter values contain a mixture of random uncertainty and systematic bias [39]. The latter may arise from a slight misfit of the kernel U-PLS method of the non-linearity introduced to the system. The bias is difficult to be precisely determined, although an approximation can be obtained by the average value of the absolute bias $|(y_{pred} - y_{nom})|$, where 'pred' stands for predicted and 'nom' for nominal. This allows one to obtain an approximation to the value of $\langle SD(y)\rangle$ as [40]:

$$\langle SD(y)\rangle = (RMSEP^2 - Bias^2)^{1/2} \quad (15)$$

Table 1 shows that all biases are similar for the different non-linear approaches. Moreover, the expected concentration uncertainty

Table 1
Prediction results on the 500 test sample set using the present approach and several neural network models.[a]

System	Figure of merit	RBL/MP[b]	RBL/RBF[c]	RBL/SVM[d]	RBL/kernel U-PLS[e]
Quadratic	RMSEP	0.0050	0.0048	0.0050	0.0053
	Bias	0.0040	0.0040	0.0041	0.0043
	$\langle SD(y)\rangle$	0.0030	0.0027	0.0030	0.0031
Sigmoidal	RMSEP	0.020	0.018	0.017	0.019
	Bias	0.014	0.013	0.013	0.014
	$\langle SD(y)\rangle$	0.014	0.012	0.011	0.012

[a] The signal noise employed in the simulations was 0.003 units. In all cases, the number of unexpected RBL components was 1, and the number of input nodes was 2 (corresponding to two analytes in the calibration set).
[b] MP = multiple perceptron, hidden neurons, 5 (both systems), Bayesian regularization, training epochs = 413 (quadratic) and 234 (sigmoidal), effective number of parameters/total parameters = 10/21 (quadratic) and 7/21 (sigmoidal).
[c] RBF = radial basis functions, hidden neurons = 22 (quadratic) and 7 (sigmoidal), Gaussian width = 1 (quadratic) and 5 (sigmoidal).
[d] SVM = support vector machines, $\gamma = 3.93\times10^6$, $\sigma = 60.7$ (quadratic) and $\gamma = 1.19\times10^7$, $\sigma = 93.0$ (sigmoidal).
[e] $N = 8$, $\sigma = 360$ (quadratic) and $N = 4$, $\sigma = 3.2\times10^5$ (sigmoidal).

[0.003, Eq. (13)] appears to be comparable to the average <SD(y)> value (0.0033) provided by Eq. (15) from the simulation study.

If the above calculations are repeated for different values of the signal noise SD(X), different uncertainties in concentration are obtained. Fig. 5 shows the expected values of SD(y) as a function of signal noise SD(X) as provided by Eq. (13) (straight lines), with circles indicating the values obtained on simulation using Eq. (15). Also included in Fig. 5 are the results for the sigmoidal system, which shows correspondingly larger uncertainties because of its lower intrinsic sensitivity. All results are stimulating and provide support for the accuracy and precision of the presently discussed approach based on kernel U-PLS processing of RBL-filtered second-order data (the latter allowing to achieve the important second-order advantage).

6.2. Experimental systems

Fig. 6 shows the variation of fluorescent signal at fixed wavelengths (excitation and emission maxima) for both experimental systems in their corresponding analytical dynamic ranges. As can be appreciated, the malondialdehyde system E1 (Fig. 6A) shows positive deviations from linearity which were mimicked by the quadratic simulated system S1 described in the simulations. On the other hand, the changes in signal vs. concentration which are shown in Fig. 6B for the amoxicillin system E2 are visually described as sigmoidal. The behavior is similar to that described by the sigmoidal expression (2) employed during simulations of system S2.

First system E1 was analyzed using regular U-PLS for calibrating the model (using three latent variables as suggested by leave-one-out cross-validation), and RBL for modeling the interferent contribution (using a single RBL component). The results provided poor figures of merit (Table 2). Moreover, a plot of U-PLS calibration scores vs. analyte concentration shows a clear non-linear behavior for the first score (Fig. 7A) and hints on non-linearity for the remaining two scores (Fig. 7B and C). All these results clearly point to the need of a model which adequately covers the non-linearity of the system.

When system E1 was studied using the presently proposed model, the first issue was to assess optimal calibration values for N and σ. Using the same cross-validation methodology described for the simulated data, they were estimated as 4 and 42 respectively. The next activity was the estimation of the number principal components to be employed in the modeling of the calibration data and in the RBL modeling of the interferent signal. The first number was estimated by leave-one-out cross-validation, as described in detail in Ref. [41]. Briefly, for each left-out sample, U-PCA is performed on the remaining training samples, and the unfolded signal for the left sample is predicted using its scores and the U-PCA loadings. The squared prediction error is then saved and summed to those corresponding to each of the remaining samples, leading to the LORSS (left-out residual sum of squares). These calculations are repeated for a number of PCs from 1 to a certain maximum (10 in our case). The values of LORSS using ($A_{cal}+1$) PCs

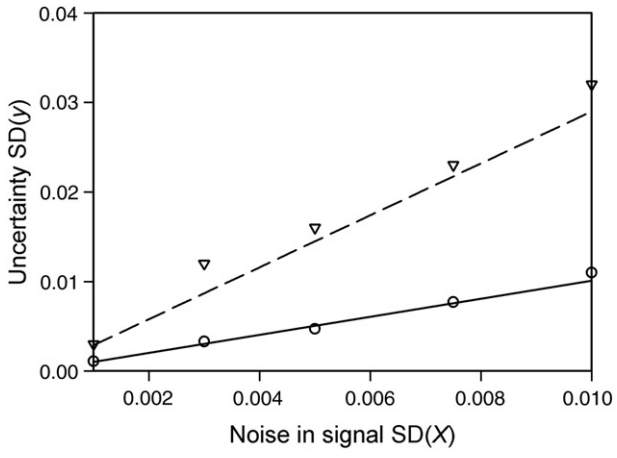

Fig. 5. Open circles (quadratic system S1) and triangles (sigmoidal system S2): average concentration uncertainty <SD(y)> [Eq. (15)] as a function of noise in signal for the simulated systems. The lines correspond to the theoretical expectations of concentration uncertainty as a function of signal noise [Eq. (13)]: solid line, quadratic system S1, dashed line, sigmoidal system S2.

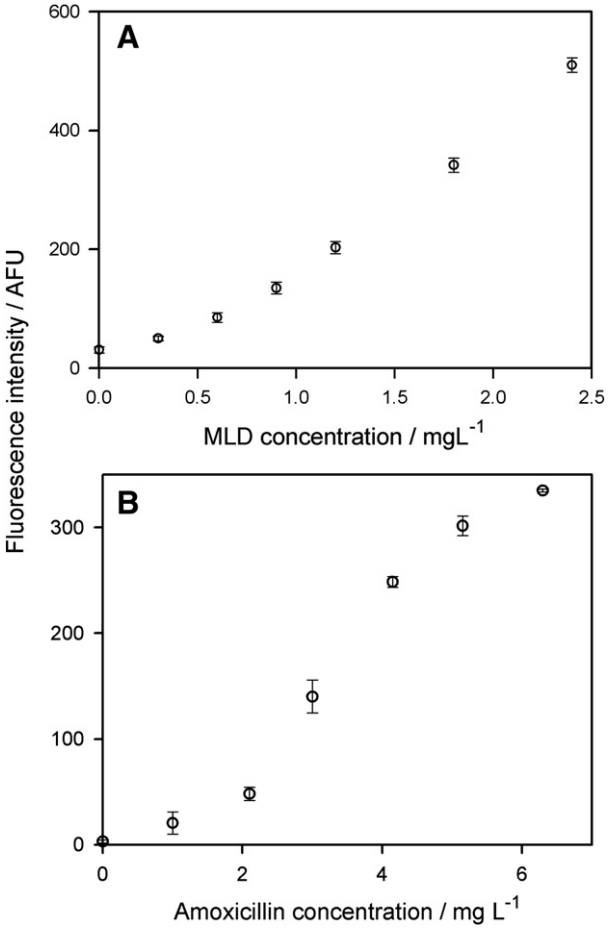

Fig. 6. Fluorescence intensity (AFU = arbitrary fluorescence units) as a function of analyte concentration for both experimental systems: A) system E1, B) system E2. In both cases, open circles indicate the averages of duplicate measurements, while the bars represent the corresponding standard deviation.

Table 2
Predictions and statistical analysis for malondialdehyde.[a]

Sample	Malondialdehyde/mg L^{-1}			
	Nominal	RBL/U-PLS	RBL/RBF	RBL/kernel U-PLS
1	0.00	0.28	0.19	0.10
2	0.00	0.26	0.17	0.07
3	0.48	0.44	0.56	0.63
4	0.48	0.46	0.54	0.55
5	0.96	1.03	1.09	0.88
6	0.96	0.83	0.81	0.76
7	1.44	1.06	1.31	1.20
8	1.91	0.95	1.74	1.72
RMSEP/mg L^{-1}		0.42	0.14	0.15
REP/%		41	12	12

[a] All samples were prepared from different olive oils, spiked with the analyte. Subsequent dilution led to the quoted concentrations in the measuring cell. For both chemometric approaches, number of calibration components = 3, number of unexpected components = 1. For RBF, input neurons = 3, hidden neurons = 19, Gaussian width = 1. For RBL/kernel U-PLS, N = 4, σ = 42.

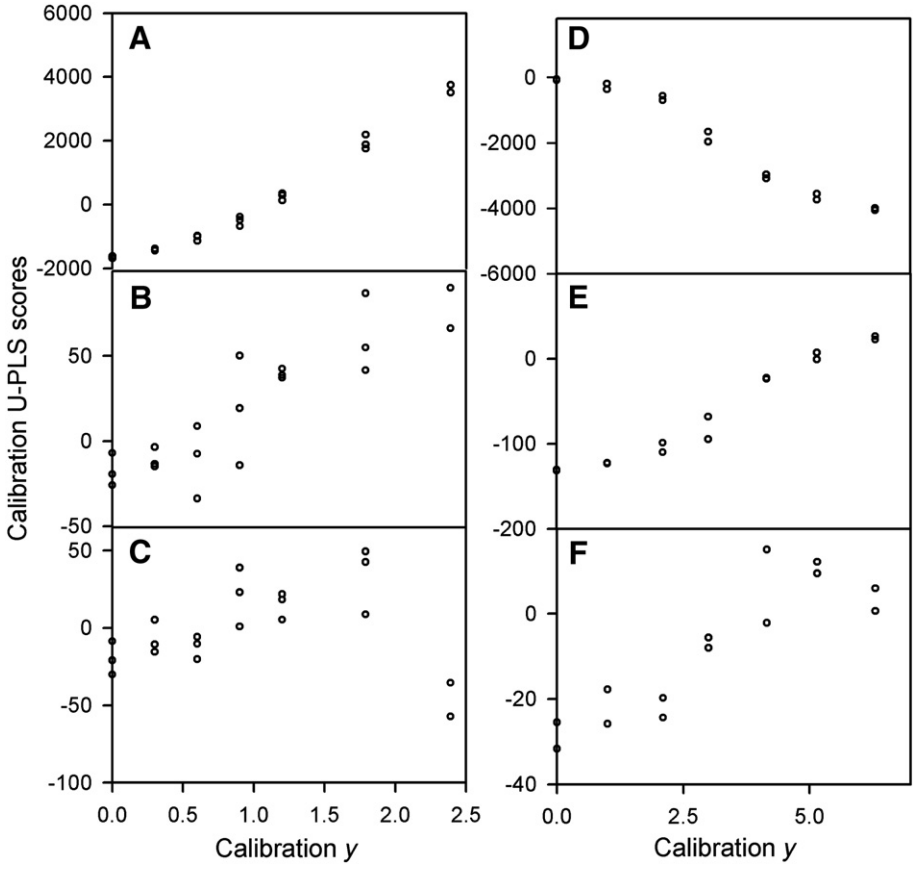

Fig. 7. Plots of U-PLS calibration scores as a function of analyte concentration for both experimental systems. Plots A), B) and C) show the first, second and third scores for the experimental system E1, while plots D), E) and F) show the analogous plots for system E2.

were compared to the APRSS values using A_{cal} PCs (APRSS, where APRSS indicates the autopredicted residual sum of squares). If the latter error is larger, then the extra PCs are modeling noise and are not significant. Usually, the ratio LORSS(A_{cal})/APRSS($A_{cal}-1$) is computed, and if this exceeds 1, ($A_{cal}-1$) PCs are employed to model the data [41]. This analysis led to the conclusion that three principal components were enough to account for the variability in the training data. Finally, the number of unexpected components in the RBL analysis can be assessed by comparing the final residuals of the RBL model with the instrumental noise level, as already described [42].

Using three principal components for modeling the calibration data and a single unexpected component for RBL, the RBL/kernel U-PLS methodology can be applied to the test samples, all consisting of olive oils spiked with the analyte in controlled concentrations. The prediction results are compared in Table 2 with those provided by a combination of RBL (which provides the second-order advantage) and the artificial neural network approach which provided the best analytical results, in this case the RBF approach. As can be seen, the results point to comparable predicting abilities of both employed combination of techniques.

In the case of the experimental system E2, normal U-PLS using three latent variables, followed by RBL for interferent modeling with a single RBL component yielded the results shown in Table 3. They are disappointing regarding analyte prediction. As with system E1, all plots of U-PLS calibration scores vs. analyte concentration are non-linear (Fig. 7D, E and F). This implies that a suitable non-linear model is required to account for this behavior.

For applying the presently discussed non-linear model to system E2, a similar approach was employed in what concerns the estimation of the parameters N and σ, and the numbers of principal components used to model the calibration data and the contribution of the interferent to the test sample signals. They are reported in Table 3, along with the specific prediction results. The comparison is now made with the best of the previously applied approaches to this same system, i.e., RBL (giving the second-order advantage) and support vector machines (Table 3). The results indicate that the present combined approach presents a comparable predictive ability.

7. Conclusions and outlook

A new non-linear second-order model achieving the second-order advantage has been described, combining residual bilinearization (which provides the second-order advantage) and kernel partial least-squares regression of unfolded data (which adequately models non-linear data). Its analytical ability is comparable to already discussed non-linear models combining residual bilinearization and

Table 3
Predictions and statistical analysis for amoxicillin.[a]

Sample	Amoxicillin/mg L^{-1}			
	Nominal	RBL/U-PLS	RBL/SVM	RBL/kernel U-PLS
1	1.98	1.80	1.88	1.70
2	3.24	3.05	3.23	3.16
3	4.49	4.24	4.39	4.31
4	5.74	5.30	5.88	5.79
5	6.47	6.10	6.14	6.13
RMSEP/mg L^{-1}		0.30	0.17	0.20
REP/%		8.0	4.0	5.3

[a] In both cases, number of calibration principal components = 3, and number of unexpected components = 1. For SVM, input nodes, 3, $\gamma = 6\times10^4$, $\sigma^2 = 5\times10^4$. For RBL/kernel U-PLS, $N = 6$, $\sigma = 8$.

artificial neural networks, based on the study of two simulated and two experimental systems. Its implementation, however, is significantly simpler. Since the latent variable approach employed to model the calibration data is able to handle non-bilinear data, the present combination of algorithms represents a highly flexible technique for the processing of second-order instrumental data. The only important restriction is that the matrix signal from the interferent should have a bilinear form and could be adequately modeled using a few principal components during the residual bilinearization procedure.

Acknowledgments

The following institutions are gratefully acknowledged for financial support: Universidad Nacional de Rosario, CONICET (Consejo Nacional de Investigaciones Científicas y Técnicas, Project PIP 5303) and ANPCyT (Agencia Nacional de Promoción Científica y Tecnológica, Project PICT 25825). A.G.R. thanks CONICET for a fellowship.

References

[1] A.C. Olivieri, Anal. Chem. 80 (2008) 5713–5720.
[2] G.M. Escandar, N.M. Faber, H.C. Goicoechea, A. Muñoz de la Peña, A.C. Olivieri, R.J. Poppi, Trends Anal. Chem. 26 (2007) 752–765.
[3] R. Bro, Crit. Rev. Anal. Chem. 36 (2006) 279–293.
[4] S. Wold, P. Geladi, K. Esbensen, J. Øhman, J. Chemom. 1 (1987) 41–56.
[5] H.A.L. Kiers, J. Chemom. 14 (2000) 105–122.
[6] R. Bro, J. Chemom. 10 (1996) 47–61.
[7] J. Öhman, P. Geladi, S. Wold, J. Chemom. 4 (1990) 79–90.
[8] V.A. Lozano, G.A. Ibañez, A.C. Olivieri, Anal. Chim. Acta 610 (2008) 186–195.
[9] A.C. Olivieri, J. Chemom. 19 (2005) 253–265.
[10] M.J. Culzoni, H.C. Goicoechea, A.P. Pagani, M.A. Cabezón, A.C. Olivieri, Analyst 131 (2006) 718–723.
[11] R. Bro, Chemom. Intell. Lab. Syst. 38 (1997) 149–171.
[12] Z.-P. Chen, H.-L. Wu, J.-H. Jiang, Y. Li, R.-Q. Yu, Chemom. Intell. Lab. Syst. 52 (2000) 75–86.
[13] H.L. Wu, M. Shibukawa, K. Oguma, J. Chemom. 12 (1998) 1–26.
[14] A.L. Xia, H.L. Wu, D.M. Fang, Y.J. Ding, L.Q. Hu, R.-Q. Yu, J. Chemom. 19 (2005) 65–76.
[15] R. Tauler, Chemom. Intell. Lab. Syst. 30 (1995) 133–146.
[16] M. Linder, R. Sundberg, Chemom. Intell. Lab. Syst. 42 (1998) 159–178.
[17] M. Linder, R. Sundberg, J. Chemom. 16 (2002) 12–27.
[18] H.C. Goicoechea, A.C. Olivieri, Appl. Spectrosc. 59 (2002) 926–933.
[19] E. Sanchez, B.R. Kowalski, Anal. Chem. 58 (1986) 496–499.
[20] A.C. Olivieri, J. Chemom. 19 (2005) 615–624.
[21] A. García-Reiriz, P.C. Damiani, M.J. Culzoni, H.C. Goicoechea, A.C. Olivieri, Chemom. Intell. Lab. Syst. 92 (2008) 61–70.
[22] A. García-Reiriz, P.C. Damiani, A.C. Olivieri, Anal. Chim. Acta 588 (2007) 192–199.
[23] M.J. Culzoni, P.C. Damiani, A. García-Reiriz, H.C. Goicoechea, A.C. Olivieri, Analyst 132 (2007) 654–663.
[24] A. García-Reiriz, P.C. Damiani, A.C. Olivieri, F. Cañada-Cañada, A. Muñoz de la Peña, Anal. Chem. 80 (2008) 7248–7256.
[25] B. Walczak, D.L. Massart, Anal. Chim. Acta 331 (1996) 177–185.
[26] B. Walczak, D.L. Massart, Chemom. Intell. Lab. Syst. 50 (2000) 179–198.
[27] T. Czekaj, W. Wu, B. Walczak, J. Chemom. 19 (2005) 341–354.
[28] B.M. Nicolaï, K.I. Theron, J. Lammertyn, Chemom. Intell. Lab. Syst. 85 (2007) 243–252.
[29] K. Kim, J.-M. Lee, I.-B. Lee, Chemom. Intell. Lab. Syst. 79 (2005) 22–30.
[30] L. Gao, S. Ren, Chemom. Intell. Lab. Syst. 45 (1999) 87–93.
[31] MATLAB 7.0, The Mathworks, Natick, Massachusetts, USA, 2007.
[32] D.J.C. MacKay, Neural Comput. 4 (1992) 448–472.
[33] M.W. Pedersen, L.K. Hansen, Proc. Neural Inform. Process. Syst. 7 (1994) 673–680.
[34] M.J.L. Orr, Matlab functions for radial basis function networks. Technical report, Institute for Adaptive and Neural Computation, Division of Informatics, Edinburgh University, 1999.
[35] K. Pelckmans, J.A.K. Suykens, T. Van Gestel, J. De Brabanter, L. Lukas, B. Hamers, B. De Moor, J. Vandewalle, LS-SVMlab Toolbox User's Guide, Version 1.5, Department of Electrical Engineering, ESAT-SCD-SISTA, Katholieke Universiteit Leuven, Belgium, 2003.
[36] D.M. Haaland, E.V. Thomas, Anal. Chem. 60 (1988) 1193–1202.
[37] A.C. Olivieri, Anal. Chem. 77 (2005) 4936–4946.
[38] A.C. Olivieri, N.M. Faber, J. Chemom. 19 (2005) 583–592.
[39] D.L. Massart, B.G.M. Vandeginste, L.M.C. Bydens, D. De Jong, P.J. Lewi, J. Smeyers-Verbeke, Handbook of Chemometrics and Qualimetrics, Part A, Elsevier, Amsterdam, 1997 Chapter 13.
[40] J. Tellinghuisen, Fresenius J. Anal. Chem. 368 (2000) 585–588.
[41] R.G. Brereton, Chemometrics, Data Analysis for the Laboratory and Chemical Plant, Wiley, Chichester, UK, 2003, p. 199.
[42] P.C. Damiani, I. Durán-Merás, A. García Reiriz, A. Jiménez Girón, A. Muñoz de la Peña, A.C. Olivieri, Anal. Chem. 79 (2007) 6949–6958.