JOURNAL OF CHEMOMETRICS
J. Chemometrics 2005; **19**: 615–624
Published online in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/cem.967

A combined artificial neural network/residual bilinearization approach for obtaining the second-order advantage from three-way non-linear data

Alejandro C. Olivieri*

Departamento de Química Analítica, Facultad de Ciencias Bioquímicas y Farmacéuticas, Universidad Nacional de Rosario, Suipacha 531, Rosario S2002LRK, Argentina

Received 8 March 2006; Revised 26 April 2006; Accepted 19 May 2006

Three-way instrumental data offer the second-order advantage to analysts, a property of great utility in the field of complex sample analysis in the presence of unsuspected components as potential interferents. The available multivariate methodologies for obtaining this advantage are all based on linear models, and hence they are not applicable to spectral information behaving in a non-linear manner with respect to target analyte concentrations. This work describes the combination of a back-propagation artificial neural network model with a technique known as residual bilinearization, applicable to second-order spectral information. The joint model allows one to efficiently extract analyte concentrations from intrinsically non-linear data, even in the presence of unsuspected constituents. Simulations have been performed by mimicking deviations from linearity brought about by: (1) exponential relationship between fluorescence and concentration, (2) kinetic evolution of responsive reaction products and (3) analytes acting as reaction catalysts. In all of these cases, successful prediction of the analyte concentrations was achieved on large test sample sets, which included the presence of overlapping components not included in the training step. The new method not only obtains the second-order advantage, but also correctly retrieves the contribution of the unsuspected components to the total test sample signals. The comparison with a multivariate methodology based on partial least-squares regression with second-order advantage shows that the presently described method displays better predictive ability. Copyright © 2006 John Wiley & Sons, Ltd.

KEYWORDS: second-order spectroscopic data; second-order advantage; artificial neural networks; residual bilinearization

1. INTRODUCTION

Instrumental data bearing a non-linear relationship with component concentration are ubiquitous in analytical chemistry. This phenomenon arises, for example, when significant interactions occur among sample components [1,2], or in fluorescence spectroscopy, due to the exponential relationship between emission and concentration [3]. They also appear in kinetic-spectroscopic systems, when a reaction product is followed which is the result of a pseudo first-order kinetics with respect to reagent [4], or when the analyte intervenes as a catalyst [5–7]. When the data structure is intrinsically non-linear, classical calibration methods with linear underlying models cannot be successfully applied. In these cases, an excellent alternative is the use of artificial neural networks (ANNs) [8–10]. These latter algorithms are based on concepts loosely related to the behavior of the human brain: the variables are assigned to mathematical objects called neurons, and a mathematical function is associated with the so-called intra-neural connections. A neural network model is composed of a number of neurons, organized into a sequence of layers [8]. Mathematically, an ANN transforms an input vector (a vector of variables assigned to a number of neurons) into an output vector, through the operation of a suitable transfer function. Neural networks show several advantages, namely: (1) they allow for better generalizations by modeling complex relationships without requiring prior knowledge of the model-related function, and (2) they display more flexibility in comparison

*Correspondence to: A. C. Olivieri, Departamento de Química Analítica, Facultad de Ciencias Bioquímicas y Farmacéuticas, Universidad Nacional de Rosario, Suipacha 531, Rosario S2002LRK, Argentina.
E-mail: aolivier@fbioyf.unr.edu.ar

Contract/grant sponsors: Consejo Nacional de Investigaciones Científicas y Técnicas; Agencia Nacional de Promoción Científica y Tecnológica; Universidad Nacional de Rosario.

with parametric techniques requiring the assumption of a specific model form. ANNs can be adequately trained to produce quantitative results, usually by employing the back propagation model, whose basic theory and application to chemical problems can be found in the literature [8,9].

A novel aspect of calibration of spectral data for making valid predictions concerning analyte concentrations in complex and/or strongly interfering systems is the processing of multi-way data [11]. The continuous progress in analytical instrumentation is yielding information of increasing complexity, which does also show distinct advantages in terms of analytical figures of merit. Second-order data, for example, provide analysts with the so-called second-order advantage [12], a property of immense utility in several fields of analysis, particularly when the samples carry unsuspected components which have not been included in the training step (notice that second-order data for a set of samples can be conveniently grouped into a three-way array, hence the name three-way data). There are basically two modes of obtaining the second-order advantage, which are pictorially illustrated in Figure 1: either data for the test sample has an influence on the regression coefficients leading to prediction (Figure 1A), or calibration is first performed using only training data, with the test sample leading to sample-specific regression coefficients in a subsequent step (Figure 1B). In either case, the underlying philosophy implies that the test sample becomes part of the whole calibration process.

Only a few applications of neural networks to non-linear spectroscopic second-order data are known: appropriate examples are the kinetic-spectrophotometric determination of three carbamate pesticides [13], the correlation between two-dimensional nuclear magnetic resonance data with the composition and properties of oil samples [14], and the monitoring of fermentation processes [15]. Little is known, however, as to whether the second-order advantage can be obtained from higher order information in the presence of unsuspected sample components, in any of the two modes depicted in Figure 1, or in additional, yet unimagined ways. The aim of this work is thus to provide with a starting algorithm which will close the gap between non-linear data and the second-order advantage.

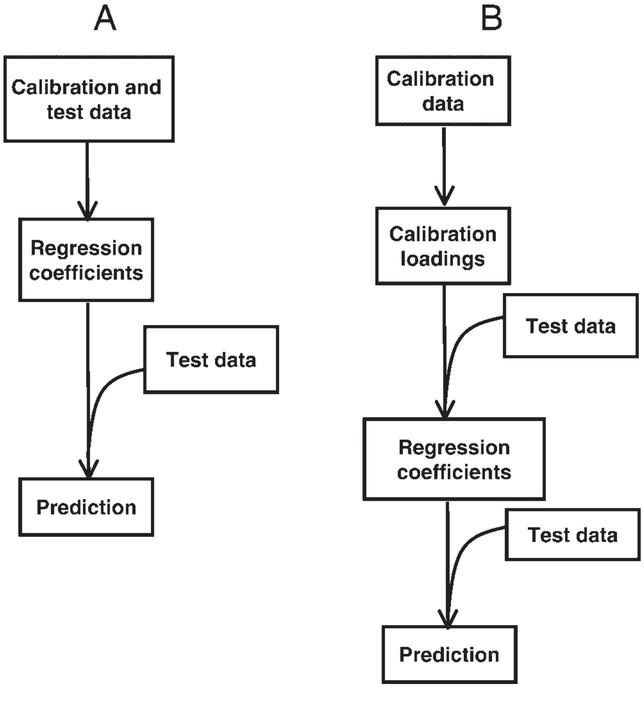

Figure 1. Two basic modes of obtaining the second-order advantage from higher-order information. (A) Combining data from calibration and test sample before computing the regression coefficients. (B) Estimating loadings from calibration data only, and then calculating regression coefficients after the test sample enters the scene.

The present approach is partially based on a model described more than 15 years ago, in which the well-known partial least-squares (PLS) linear calibration method was combined with a procedure known as residual bilinearization (RBL), which was useful in providing PLS with the second-order advantage [16]. The combination PLS/RBL has only recently caught the attention of chemometricians, however, in order to analyze real systems of high complexity, and was shown to provide comparable analytical behavior as to the standard parallel factor (PARAFAC) model [17]. We show by means of simulated non-linear data for several systems mimicking real analytical applications, that the analogous combination ANN/RBL is adequate to train non-linear spectroscopic data, and to successfully predict analyte concentrations in the presence of unsuspected constituents, thus achieving the important second-order advantage. The model requires that the unsuspected contribution is bilinear, and can therefore be adequately modeled by RBL. The results can be considered as a ‘proof of principle’, suggesting that the model is indeed feasible, although subsequent phases are necessary for its demonstration.

2. THEORY

2.1. Terminology

It is important to define, in light of the forthcoming discussion, sample component categories, with particular focus on components generating a signal that overlaps with the signal of the analyte of interest, and can therefore be considered as potential interferents.

A distinction can be first made between components present in the training set of samples, and those which are only present in the unknown sample only. The former ones can be called ‘suspected’ components, because the analyst should include in the calibration set all components suspected to be present in unknown samples, in order to have a sufficiently representative training set. However, truly unknown samples may carry additional components: these are called ‘unsuspected’ ones. Note that the suspected constituents can be further divided into ‘calibrated’ and ‘uncalibrated’: calibrated refers to components for which calibration concentrations are available (including, as a specific case, the analyte of interest), whereas uncalibrated refers to components for which only a common subspace that contains them is accessible. Some multivariate calibration models require all suspected components to be properly calibrated, whereas the combinations PLS/RBL and

DOI: 10.1002/cem

ANN/RBL discussed in the present report do only require the analyte of interest to be calibrated.

Notice that potential interferents will not always produce an interference, in the sense of generating a systematic error in the analyte determination [18]. Whether the interference will be actual or will only remain as potential, depend on the type of measured instrumental signals and on the employed calibration methodology. For first-order instrumental data, unsuspected test sample components most likely constitute an interference. This may not be true, however, in the second-order multivariate domain involving the second-order advantage.

2.2. PLS/RBL

In the present report, the results provided by ANN/RBL will be compared with those obtained from the second-order multivariate model involving the combination of PLS and RBL. The latter methodology has already been described in the relevant literature [16]. A recent application discusses its use for the analysis of therapeutic drugs in biological fluids from second-order excitation-emission fluorescence data [19]. The present comparison could also be extended to other popular second-order multivariate models such as PARAFAC [20], generalized rank annihilation (GRAM) [21], or multivariate curve resolution coupled to alternating least-squares (MCR-ALS) [22]. They all exploit the second-order advantage in the mode shown in Figure 1A. However, these latter methodologies may not be appropriate, in general, to treat the complete set of data described in the present report. This is due to the fact that they are not suitable for data which show: (1) deviations from linearity, or (2) second-order signals having identical profiles in one of the dimensions (see below). Another possibility is to employ bilinear least-squares (BLLS) [23–25], which, once coupled to RBL, achieves the second-order advantage in the manner illustrated in Figure 1B. This method is similar to PLS/RBL, although admittedly less reliable with respect to non-linear systems. Therefore, the comparison was carried out with PLS/RBL, which is the most flexible second-order multivariate calibration technique, able to cope with mild non-linearities and also with linear dependencies in any of the second-order modes.

2.3. ANN training

In the present report, the back-propagation of errors method was selected for training the ANNs. Typically, an ANN suitable for the present analytical problems consists of three layers: (1) an input layer to accommodate for the input variables, which might be either the original variables or appropriate surrogate variables representing the spectral variability in the training set of samples, (2) a hidden layer of neurons, whose number is tuned on a trial an error basis and (3) an output layer with a single neuron, yielding the concentration of the analyte of interest in each sample (on an appropriate scale). Both the input and hidden layers do also include the so-called bias neurons, whose inputs are equal to 1. Since the number of original variables produced by modern instruments is large (data matrices have JK elements, where J and K are the number of channels in each dimension), a usual approach is to compress the raw information into a reduced number of latent variables, such as, for example, the first A principal components (PCs) or scores. In this approach the number of input neurons equals A, where in general $A \ll JK$. The value of A can be estimated, for example, by computing the % of variance explained by the PCs of the unfolded training data matrix (size $I \times JK$, I is the number of training samples), and selecting the first A PCs which explain more than a certain % (i.e. 99%) of the total variance.

Once the architecture of the net is established, the first A PCs are loaded into the A neurons of the input layer, and the outputs are calculated for each training sample using a set of randomly selected initial weights and a sigmoidal transfer function [8]. They are transferred from layer to layer through a suitable transfer function, which in the present case is the familiar sigmoidal function $f(x) = 1/[1 + \exp(-x)]$. The weights are then modified according to the back-propagation methodology, which compares the ANN outputs with the nominal values of the analyte concentration in the training and monitoring samples. The comparison yields the calibration root mean square error (RMSEC), which is computed every training cycle or 'epoch', and allows the correction of the network weights which leads to the decrease of the RMSEC. Simultaneously, the ANN performance is monitored by the results provided on an independent monitoring sample set, which helps to estimate the corresponding RMSEM (M for monitoring). Usually the net is trained during a number of epochs until a minimum in RMSEM (compatible with the noise level present in the system) is reached, in order to avoid overtraining. The set of weights so obtained is stored for future prediction on new samples. Two important parameters for network training are the learning rate and momentum, which pull the correction of weights in opposite directions: the learning rate tends towards a fast, steepest-descent convergence, while the momentum is a long-range function preventing the solution from being trapped into local minima. These parameters are usually tuned around a value of 0.5, also by trial and error.

The above scheme works properly provided the new test samples have a composition which is representative of the training set. When unsuspected constituents occur in the test sample, however, its scores will not be suitable for analyte prediction using the trained ANN. To cope with this problem, it is necessary to resort to a technique which is able to: (1) mark the new sample as an outlier, indicating that further actions are necessary before ANN prediction, and (2) isolate the contribution of the unsuspected component from that of the calibrated analytes, in order to recalculate appropriate surrogate variables for the test sample. PC analysis (PCA) is adequate in this regard. In this case, the sample will be considered as an outlier if the residuals of the PCA of $\mathbf{X}_u$ [s_p, see Equation (1)] are abnormally large in comparison with the typical instrumental noise (the latter is easily assessed by replicate measurements and the comparison can be carried out through an F test if the residuals can be assumed to be identically and independently distributed, otherwise other non-parametric techniques may be necessary):

$$\begin{aligned} s_p &= \|\mathbf{e}_p\|/(JK - A)^{1/2} = \|\text{vec}(\mathbf{X}_u) - \mathbf{P}\,\mathbf{P}^T\,\text{vec}(\mathbf{X}_u)\|/(JK - A)^{1/2} \\ &= \|\text{vec}(\mathbf{X}_u) - \mathbf{P}\,\mathbf{t}_u\|/(JK - A)^{1/2} \end{aligned} \tag{1}$$

DOI: 10.1002/cem

where $||\cdot||$ indicates the Euclidean norm, $\mathbf{P}$ is the matrix containing the first A loadings obtained by applying PCA to the unfolded training data, $\mathbf{t}_u$ is the vector of test sample scores, and vec() indicates the unfolding operation. The sizes of the relevant arrays in Equation (1) are as follows: $\mathbf{e}_p$ and vec($\mathbf{X}_u$), $JK \times 1$; $\mathbf{X}_u$, $J \times K$; $\mathbf{P}$, $JK \times A$ and $\mathbf{t}_u$, $A \times 1$. If s_p is indeed large, then no further analysis is possible, unless a procedure is devised which takes into account the presence of unsuspected sample components (see below).

2.4. Second-order advantage with RBL

RBL is a procedure which can handle the presence of unsuspected components in the test sample. It can be easily described by decomposing the signal for a test sample containing unsuspected components into two parts: one modeled using the calibration latent variables ($\mathbf{X}_{mod}$) and the remaining part which cannot be modeled ($\mathbf{X}_{unmod}$) with these variables, that is:

$$\mathbf{X}_u = \mathbf{X}_{mod} + \mathbf{X}_{unmod} \tag{2}$$

As is usual in PCA, the modeled part can be expressed as a function of the calibrated latent variables $\mathbf{P}$ and the unknown sample score $\mathbf{t}_u$, and hence:

$$\mathrm{vec}(\mathbf{X}_u) = \mathbf{P}\,\mathbf{t}_u + \mathbf{e}_{mod} + \mathrm{vec}(\mathbf{X}_{unmod}) \tag{3}$$

where $\mathbf{e}_{mod}$ is the vector of residuals not fitted to $\mathbf{X}_{mod}$ by the PCA model with A PCs. Typically, $\mathbf{e}_{mod}$ will contain elements of the order of the instrumental noise. If anything having a bilinear structure is present in $\mathbf{X}_{unmod}$ which rises above the noise level, it can be modeled using singular value decomposition (SVD). This allows one to estimate profiles for the unsuspected components ($\mathbf{b}_{uns}$ and $\mathbf{c}_{uns}$) by minimization of the norm of the residual vector $\mathbf{e}_u$, computed while fitting the sample data to the sum of the relevant contributions:

$$\mathrm{vec}(\mathbf{X}_u) = \mathbf{P}\,\mathbf{t}_u + \mathrm{vec}[g_{uns}\mathbf{b}_{uns}(\mathbf{c}_{uns})^T] + \mathbf{e}_u \tag{4}$$

During this procedure, $\mathbf{P}$ is kept constant at the calibration values, $\mathbf{t}_u$ is varied until $||\,\mathrm{e}_u\,||$ is minimized, and profiles for the unsuspected components are estimated by SVD of a residual matrix obtained after reshaping $\mathbf{e}_p$ [see Equation (1)] to a $J \times K$ matrix:

$$(g_{uns}, \mathbf{b}_{uns}, \mathbf{c}_{uns}) = \mathrm{SVD}_1(\mathbf{E}_p) \tag{5}$$

where $\mathbf{E}_p$ is the $J \times K$ matrix obtained after reshaping the $JK \times 1$ $\mathbf{e}_p$ vector, and SVD_1 indicates taking the first PC.

Minimization can be been carried out using either iterative or Gauss–Newton procedures, starting with $\mathbf{t}_u$ as given by the projection of the vector of responses for the test sample on the space spanned by the calibration A PCs:

$$\mathbf{t}_u = \mathbf{P}^T\,\mathrm{vec}(\mathbf{X}_u) \tag{6}$$

In all cases reported in the present paper, we have employed the Gauss–Newton procedure for achieving RBL.

Should it be necessary to consider a larger number of unsuspected components (N_{uns}) in the SVD analysis of $\mathbf{E}_p$ [Equation (5)], then N_{uns} can be assessed by comparing the final residuals s_u with the instrumental noise level, with s_u given by:

$$s_u = ||\mathbf{e}_u|| / [JK - (A + N_{uns})]^{1/2} \tag{7}$$

where $\mathbf{e}_u$ is from Equation (4). Typically, a plot of s_u computed for trial values of N_{uns} will show decreasing values, starting at s_p when $N_{uns} = 0$, until it stabilizes at a value compatible with the experimental noise, allowing to locate the correct number of unsuspected components. It should be noticed that for $N_{uns} > 1$, the profiles provided by the SVD analysis of $\mathbf{E}_p$ no longer resemble the true profiles for the unsuspected components, due to the rotational ambiguity which is intrinsic to SVD. Once $||\,\mathrm{e}_u\,||$ is minimized in Equation (4), and the correct test sample scores $\mathbf{t}_u$ have been found, the final $\mathbf{t}_u$ vector is introduced into the input neurons of the trained ANN, providing the analyte concentration as output. The entire process is schematized in the flow sheet shown in Figure 2, where it is apparent that the second-order advantage is achieved in the way shown in Figure 1B.

2.5. Software

All multivariate methods discussed in the present work were implemented in MATLAB 6.0 [26]. The specific scripts for applying PLS/RBL and ANN/RBL are available from the author on request, including a graphical user interface, useful for routine analytical chemistry studies, which provides access to a variety of second-order multivariate

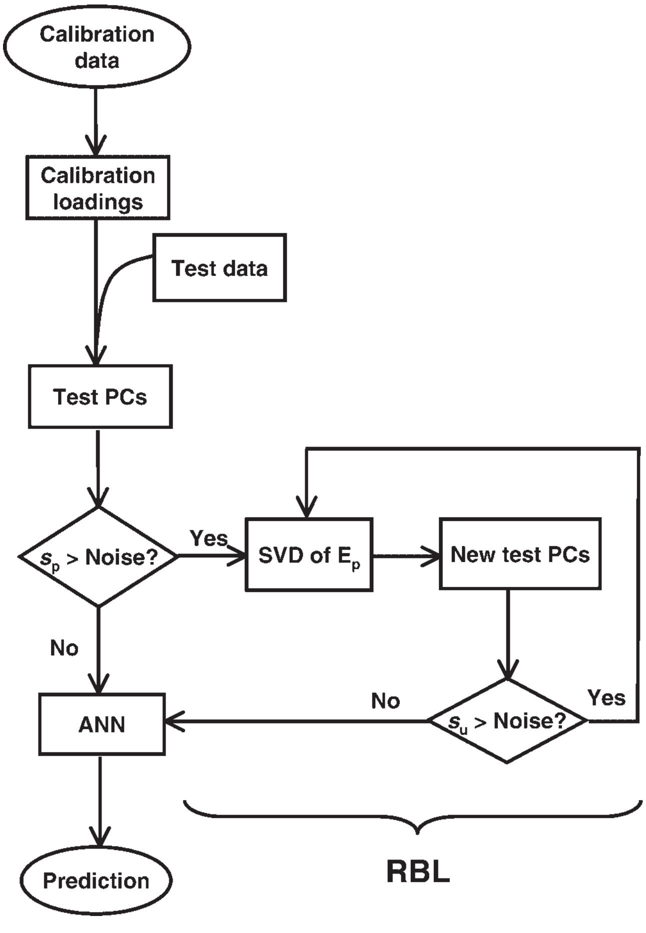

Figure 2. Flow sheet indicating how prediction is made by ANN/RBL on a new test sample having unsuspected component. The RBL procedure is indicated. For details on symbols see text.

DOI: 10.1002/cem

methodologies, of the type already described for first-order methods [27].

3. SIMULATIONS

3.1. General considerations

In all cases, second-order data for a calibration set of 20 samples were created starting from noiseless profiles for the analyte(s) (see below). All data matrices were of size 17 × 19 data points (17 points correspond to the first dimension and 19 to the second dimension, which are intended to mimic emission and excitation wavelengths, respectively, in data set 1, absorption wavelength and time in data set 2, and absorption wavelength and time in data set 3). The training concentrations of the analyte(s) were taken at random (equally distributed) from the range 0 to 1. A test set of 500 samples was also built, again with random concentrations of the analyte(s), all in the range 0.2–0.8. To each of these test samples, second-order signals corresponding to a single unsuspected component were added in random concentrations. Finally, random numbers taken from a Gaussian distribution were added to all signals. The standard deviation of the added Gaussian noise was taken as 2% of the mean calibration signal. The nominal concentrations of the calibrated component are assumed to carry a significantly lower error in comparison with the instrumental signals.

3.2. Data set 1

In the case of data set 1, the signals for the training samples were computed as the sum of the contributions of two analytes and noise:

$$\mathbf{X}_{c,i} = [1 - \exp(-k_{n1}y_{1,c,i})]\mathbf{S}_1 + y_{2,c,i}\mathbf{S}_2 + \mathbf{R}s_X \tag{8}$$

where $\mathbf{X}_{c,i}$ is the $J \times K$ matrix of second-order signals for the ith calibration sample, $y_{n,c,i}$ is the nominal concentration of each analyte, $\mathbf{S}_n = g_n\mathbf{b}_n\mathbf{c}_n^{\mathrm{T}}$ are the corresponding matrix signals at unit-concentration for analyte n ($\mathbf{b}_n$ and $\mathbf{c}_n$ are the profiles in the first and second dimension, both normalized to unit length, and g_n is a scaling factor, in all cases set at 1 unless stated otherwise), $\mathbf{R}$ is a matrix of appropriate size composed of Gaussian random numbers with unit standard deviation, and s_X is the standard deviation of the noise added to signals. Notice in Equation (8) the non-linear dependence of signals for analyte 1 with respect to concentration, with the parameter k_{nl} controlling the degree of departure from linearity. In the present case, $k_{nl} = 2$.

The test signals for data set 1, on the other hand, were built using the following expression, in which a signal due to a third partner was added:

$$\mathbf{X}_u = [1 - \exp(-k_{n1}y_{1,u})]\mathbf{S}_1 + y_{2,u}\,\mathbf{S}_2 + y_{\mathrm{uns,u}}\,\mathbf{S}_{\mathrm{uns}} + \mathbf{R}s_X \tag{9}$$

where $\mathbf{X}_u$ is the $J \times K$ matrix for the unknown sample, $y_{1,u}$ and $y_{2,u}$ are the nominal concentration of each analyte, $y_{\mathrm{uns,u}}$ is the concentration and $\mathbf{S}_{\mathrm{uns}}$ is the matrix signal for the unsuspected component ($\mathbf{S}_{\mathrm{uns}} = g_{\mathrm{uns}}\mathbf{b}_{\mathrm{uns}}\mathbf{c}_{\mathrm{uns}}^{\mathrm{T}}$). Notice that the presence of the latter in the test samples makes the use of the second-order mandatory to resolve the presently simulated mixtures.

The profiles in both dimensions for the above discussed data set are shown in Figure 3(A) and (B). They will be discussed in detail below.

3.3. Data set 2

Data set 2 is designed to mimic a kinetic experiment run under conditions where two non-responsive analytes (A and B) react with a third component (R), producing absorbing species (P_A and P_B) with spectral profiles shown in Figure 3A. The reaction scheme is thus:

$$\mathrm{A} + \mathrm{R} \xrightarrow{k_A} \mathrm{P_A} \tag{10}$$

$$\mathrm{B} + \mathrm{R} \xrightarrow{k_B} \mathrm{P_B} \tag{11}$$

where k_A and k_B are the corresponding kinetic constants. The non-absorbing reagent R is present in defect with respect to the analytes, making the kinetics pseudo-first order with respect to R. The time evolution of the product concentrations is therefore governed by the following equations:

$$y_1(t) = \frac{y_R[1 - \exp(-r_2t)]}{1 + r_1} \tag{12}$$

$$y_2(t) = \frac{y_Rr_1[1 - \exp(-r_2t)]}{1 + r_1} \tag{13}$$

where y_R is the initial concentration of R (taken as 0.01), $y_1(t)$ and $y_2(t)$ are the concentrations of P_A and P_B, respectively, at time t, $r_1 = (k_By_B/k_Ay_A)$, $r_2 = k_Ay_A + k_By_B$, y_A and y_B are the initial concentrations of both analytes and $k_A = 0.5$ and $k_B = 0.02$ are the kinetic constants. As can be seen, the time profiles for the reaction products are a non-linear function of the nominal analyte concentrations. Only combinations where $(y_A + y_B) > (10\ y_R)$ were selected, in order to fulfill the requirement that R is in defect with respect to A and B.

For the simulations, the concentrations of P_A and P_B were calculated at 19 different times, in the range 1–19, using Equations (12) and (13). Two column vectors $\mathbf{y}_{1i}$ and $\mathbf{y}_{2i}$ (size 19 × 1), were constructed for the ith calibration sample, and converted to $(g_{1i}\mathbf{c}_1)$ and $(g_{2i}\mathbf{c}_2)$, respectively, where $\mathbf{c}_1$ and $\mathbf{c}_2$ are unit-length normalized time profiles, and g_{1i} and g_{1i} are scaling factors. Calibration absorbance-time matrices were then built according to:

$$\mathbf{X}_{c,i} = g_{1,i}\,\mathbf{b}_1\,\mathbf{c}_1^{\mathrm{T}} + g_{2i}\,\mathbf{b}_2\,\mathbf{c}_2^{\mathrm{T}} + \mathbf{R}s_X \tag{14}$$

where $\mathbf{b}_1$ and $\mathbf{b}_2$ are the spectral profiles (each of size 17 × 1) shown in Figure 3A for each reaction product (notice that the scaling factors g_{1i} and g_{2i} will change from sample to sample and will differ from unity in this particular case, depending on the initial concentrations of the analytes). The normalized time profiles are depicted in Figure 3C, where it can be noticed that although the kinetic constants are different, the normalized time profiles are identical, due to the appearance of the common factor $[1 - \exp(-r_2t)]$ in the integrated rate law. The unsuspected component, only present in the test samples, adds to the contribution of the products in the following manner:

$$\mathbf{X}_u = g_{1u}\,\mathbf{b}_1\,\mathbf{c}_1^{\mathrm{T}} + g_{2u}\,\mathbf{b}_2\,\mathbf{c}_2^{\mathrm{T}} + g_{\mathrm{uns}}\,\mathbf{b}_{\mathrm{uns}}\,\mathbf{c}_{\mathrm{uns}}^{\mathrm{T}} + \mathbf{R}s_X \tag{15}$$

where g_{1u} and g_{1u} will again depend on the test concentrations and kinetic constants for each analyte, $\mathbf{b}_{\mathrm{uns}}$ is the absorption profile for the unsuspected component given in

DOI: 10.1002/cem

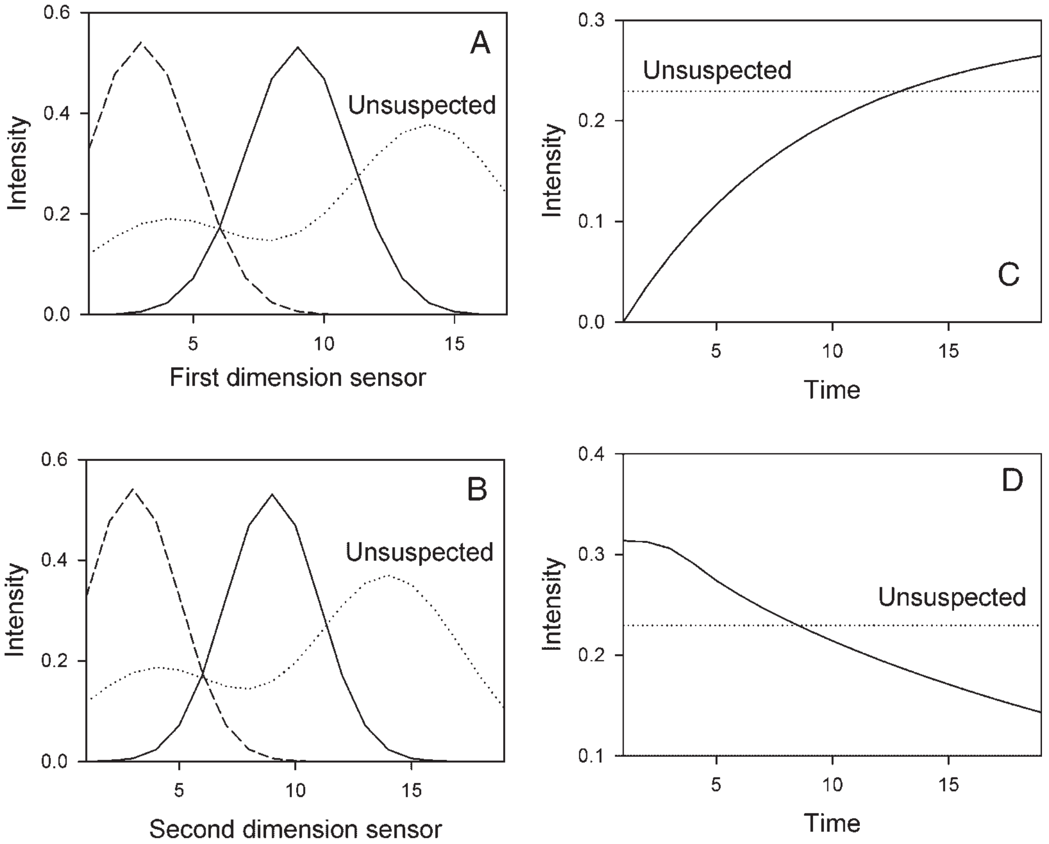

Figure 3. Simulated component profiles in both dimensions. (Plot A) shows the emission profiles for data set 1 (solid line, analyte 1, dashed line, analyte 2, dotted line, unsuspected component), the absorption profiles for data set 2 (solid line, reaction product P_A, dashed line, product P_B, dotted line, unsuspected component), and the absorption profiles for data set 3 (solid line, reagent C, dotted line, unsuspected component). (B) Excitation profiles for data set 1 (meaning of lines as in plot A). (C) Kinetic profiles for data set 2 (solid line, reaction products P_A and P_B, dotted line, unsuspected component). (D) Kinetic profiles for data set 3 (solid line, reagent C, dotted line, unsuspected component).

Figure 3A, $\mathbf{c}_{\text{uns}}$ is a normalized vector having identical elements (i.e. a constant time profile, see Figure 3C) and g_{uns} a unit scaling factor.

3.4. Data set 3

In the case of data set 3, a single analyte was considered, acting as the catalyst of an autocatalytic reaction scheme such as [28]:

$$\mathrm{C} + \mathrm{S} \xrightarrow{k_S} \mathrm{D} \tag{16}$$

$$\mathrm{C} + \mathrm{D} \xrightarrow{ka} 2\mathrm{D} \tag{17}$$

$$\mathrm{D} \underset{k_E}{\overset{k_D}{\rightleftarrows}} \mathrm{E} \tag{18}$$

where the first step is the one affected by the concentration of the catalyst, in such a way that $k_S = k_{Scat}$ [Catalyst]$^{n\text{cat}} = k_{Scat}$ $(y_1)^{n\text{cat}}$. In this latter equation, y_1 is the nominal analyte concentration, and $n$cat is the reaction order with respect to the catalyst, which in the present case was set to 2, with $k_{Scat} = 0.5$. The above mechanism can be solved by numerical calculations, for example, using a Runge–Kutta algorithm. This allowed to obtain the concentrations of all species as a function of time, using typical kinetic constants taken from the catalytic effect of Cu(II) on the reduction of the organic dye resazurin by sulfide ions [29]. The concentration of the responsive reagent C was calculated at 19 different reaction times, in the range 0–3. The remaining constants employed were: autocatalytic constant, $k_a = 60$, and direct and inverse constants for the equilibrium step, $k_D = 150$, $k_E = 0.2$. The absorption spectrum of C was taken as indicated in Figure 3A (solid line), with an unsuspected spectrum equal to the dashed line in this same Figure (S, D and E are assumed to be non-absorbing). Typical time profiles for data set 3 are shown in Figure 3D, including that for the unsuspected component. The matrix data for this system were built analogously to data set 2, that is, in the form $(g_1\mathbf{b}_1\mathbf{c}_1^{\mathrm{T}})$ for C (with g_1 varying depending on the analyte concentration in each sample). The analyte is present in both the calibration and test samples. For the unsuspected component, which does only affect the test samples, the signals were given as $(g_{\text{uns}}\mathbf{b}_{\text{uns}}\mathbf{c}_{\text{uns}}^{\mathrm{T}})$, with a unit scaling factor g_{uns}.

3.5. Software

All simulations, including the Runge–Kutta algorithm for building data set 3, were run using suitable MATLAB 6.0 scripts.

DOI: 10.1002/cem

4. RESULTS AND DISCUSSION

4.1. Data set 1

In this data set 2 analytes are calibrated from excitation-emission fluorescence matrix data, and a single unsuspected component occurs in the test samples. One of the analytes (number 1 in this case) significantly deviates from the linearity between signal and concentration. Profiles for all components in data set 1 are shown in Figure 3, where it can be noticed that the unsuspected signal overlaps the analyte signals across the whole spectral ranges in both dimensions.

The first issue to be assessed in connection with the achievement of the second-order advantage is the ability of the multivariate methods to isolate the non-linear contribution of the analyte of interest (component 1 in the present case) from the combined contribution of the second analyte and the unsuspected component in the test samples. The reference PLS model was calibrated using a number of latent variables obtained from the well-known leave-one-out cross-validation procedure, employing mean-centered calibration data, and selecting the number of latent variables as suggested by Haaland [30], that is, the least number leading to a variance which is statistically undistinguishable from the number leading to the minimum variance. Alternative procedures for establishing the correct number of PLS latent variables, such as Monte Carlo cross-validation [31], were also applied with identical results to those reported in the present paper. This allowed to establish the parameter A required for calibration. The quality of the cross-validation results can be judged from the root mean square error RMSECV (Table I), which in this case is rather high due to the presence of strong non-linearities in the behavior of analyte 1. For applying the PLS/RBL procedure to each of the analyzed test samples, the number of unsuspected components (N_{uns}) was set to 1. In general, this number can be estimated by inspection of the variation of s_{u} [see Equation (4)] with a trial number of unsuspected components, as commented above.

Application of the PLS/RBL method rendered profiles for the unsuspected component which are shown in Figure 4(A) and (B) in both dimensions. They are similar to those employed for simulation, confirming the ability of this multivariate method in retrieving correct information from the test samples in order to achieve the second-order advantage. Prediction results for a 500 sample test set including the unsuspected constituent shows, however, a poor analytical performance (Table I), with a high RMSEP (the RMSE for prediction on new samples). Furthermore, the plot of prediction errors (i.e. predicted minus nominal concentration) versus nominal concentration values for analyte 1 (Figure 5A) clearly shows the U-shaped, non-linear behavior for this analyte, which is not adequately covered by the PLS model.

Table I. PLS/RBL and ANN/RBL results on the different simulated data sets

Method	Parameter	Value
Data set 1		
PLS/RBL	Number of latent variables A	2
	RMSECV	0.08 (16%)
	RMSEP	0.10 (20%)
ANN/RBL	Architecture (input-hidden-output neurons)	3-5-1
	Number of training epochs	470
	Learning rate	0.5
	Momentum	0.5
	RMSEC	0.017 (3.4%)
	RMSEM	0.024 (4.8%)
	RMSEP	0.025 (5.0%)
Data set 2		
PLS/RBL	Number of latent variables A	3
	RMSECV	0.021 (4.2%)
	RMSEP	0.027 (5.4%)
ANN/RBL	Architecture (input-hidden-output neurons)	3-4-1
	Number of training epochs	20 000
	Learning rate	0.5
	Momentum	0.5
	RMSEC	0.010 (2.0%)
	RMSEM	0.012 (2.4%)
	RMSEP	0.014 (2.8%)
Data set 3		
PLS/RBL	Number of latent variables A	2
	RMSECV	0.040 (8.0%)
	RMSEP	0.029 (5.8%)
ANN/RBL	Architecture (input-hidden-output neurons)	3-4-1
	Number of training epochs	362
	Learning rate	0.5
	Momentum	0.5
	RMSEC	0.011 (2.2%)
	RMSEM	0.014 (2.8%)
	RMSEP	0.015 (3.0%)

RMSE, root mean square error; CV, cross validation; P, prediction; T, training; M, monitoring. Relative % errors in parenthesis (calculated with respect to the mean calibration concentration).

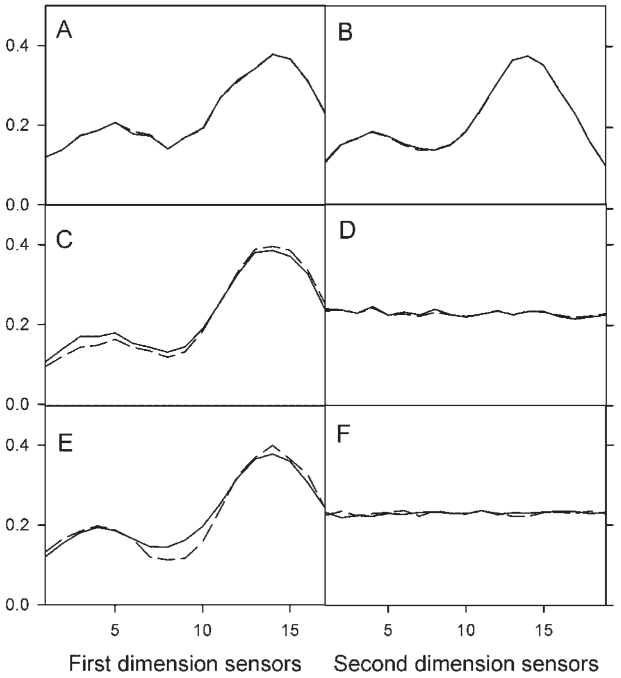

Figure 4. Profiles for the unsuspected component in both dimensions, as retrieved by ANN/RBL (solid lines) and PLS/RBL (dashed lines). (A, C and E) correspond to the first dimension in data sets 1, 2 and 3, respectively. (B, D and F) correspond to the second dimension. All profiles have been normalized to unit length, and are plotted on a common vertical scale.

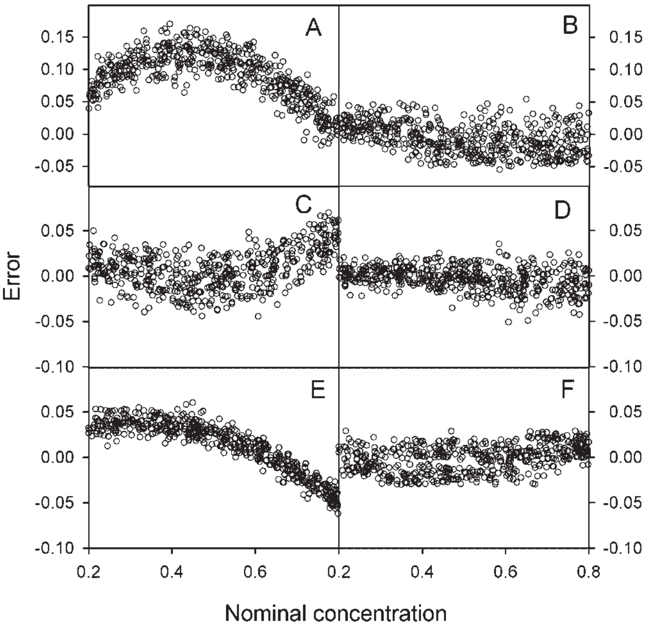

Figure 5. Prediction errors versus nominal concentration values found in a 500 sample test set containing an unsuspected component. (A) PLS/RBL in data set 1. (B) ANN/RBL in data set 1. (C) PLS/RBL in data set 2. (D) ANN/RBL in data set 2. (E) PLS/RBL in data set 3. (F) ANN/RBL in data set 3.

In the case of the application of ANN/RBL to this data set, the first step was the training of an appropriate ANN using only calibration data. The whole calibration set of 20 samples was first randomly divided into a training set (60% of samples) and a monitoring set (the remaining 40%). PCA analysis of the training set suggested the use of three PCs to employ as input variables, and hence the number of input neurons was set to 3. Training of several different nets, all starting from random weights, suggested that an appropriate number of hidden neurons was 5. Finally, a single output neuron was reserved for predicting the concentration of analyte 1. Table I shows the final architecture along with other ANN training parameters. The quality of this training step, as judged from the RMSEC and RMSEM (the RMSEs corresponding to the training and monitoring sets of samples, respectively) is significantly better than for PLS (Table I). Recall that the noise added to the system was 2% of the mean calibration signal, and hence that a minimum of 2% error is expected in predicting a set of new samples. After the training, application of the ANN/RBL procedure (Figure 2) to the test samples produced typical unsuspected profiles which are shown in Figure 4(A) and (B) as dashed lines. Comparison with those retrieved by PLS/RBL shows good agreement. This confirms the ability of ANN/RBL in recovering unsuspected profiles, the first requirement in the obtainment of the second-order advantage. Predictions on the same 500 sample test set studied by PLS/RBL led to the RMSEP quoted in Table I, which is significantly better than for PLS/RBL. Furthermore, the plot of prediction residuals versus nominal concentrations (Figure 5B) is acceptable, and demonstrates the ability of the presently described methodology in producing valid predictions for an overlapping, severely interfering and strongly non-linear spectroscopic system.

Although on a different spectroscopic area, this behavior resembles that of the absorption spectra of pharmaceutical dexamethasone in mixtures with chlorpheniramine and naphazoline, which has been previously studied using an ANN approach [2]. However, in these latter pharmaceutical samples no unsuspected components occurred, and hence the second-order advantage issue was unimportant.

4.2. Data set 2

This data set mimics a spectroscopic-kinetic experiment, where two analytes react with a single reagent (present in defect with respect to the former ones), and producing distinctly responsive reaction products. The profiles for this system are shown in Figure 3(A) and (C) and were already commented. Notice, however, that the spectral profiles for the reaction products are different (Figure 3A), but the kinetic profiles (normalized to unit length) are identical (Figure 3B). This produces three-way data in which not only non-linearities are present, but also where component profiles in the second dimension are equal, which constitutes a specific case of linear dependency. This latter type of problems becomes a difficult task for methods such as PARAFAC or GRAM, although it can be efficiently tackled by PLS/RBL [17]. Therefore, the latter was again the standard multivariate methodology used for comparison.

The application of both PLS/RBL and ANN/RBL to this data set was analogous to that described above in connection with data set 1. The relevant parameters for calibration and prediction with PLS/RBL, and for training, monitoring and prediction with ANN/RBL, are collected in Table I. Notice that PLS employs three latent variables for calibration in this case, indicating that the non-linearities present in the calibration set require one additional factor than the number of chemical constituents. As can be seen in Table I, although PLS/RBL produces results which are better than for data set 1, ANN/RBL outperformed the former for data set 2, yielding significantly lower RMSE values.

As regards the retrieval of unsuspected profiles, both PLS/RBL and ANN/RBL were able to distinguish the contribution of the calibrated analytes from that of the unsuspected component, yielding correct profiles for the latter (both in the spectral and time dimensions), as is evident in Figure 4(C) and (D).

The non-linear behavior of this system is apparent in Figure 5C on inspecting the plot of PLS/RBL errors versus nominal concentration values. In comparison, the corresponding ANN/RBL plot (Figure 5D) is indicative of a better predictive ability towards this non-linear data set.

The results are similar to those obtained when mixtures of amines react with salycilaldehyde to yield colored products, which has been previously explored using ANN, except that in this latter case unsuspected components were absent in the test samples [4].

4.3. Data set 3

In this case a single analyte occurs in all samples, and a single unsuspected component affects the test samples. The analyte is the catalyst of an autocatalytic reaction, and hence the

DOI: 10.1002/cem

relationship between signal (in this case the absorption of a given reagent) and analyte concentration is non-linear, as in the recently described reaction of the dye resazurin with sulfide ions catalyzed by copper(II). The latter reaction allowed for the determination of copper in real samples using ANNs [7], although in the absence of unsuspected constituents.

The profiles for data set 3 are shown in Figure 3(A) and (B) for the responsive reagent C and for the unsuspected component, in the form of spectral and time profiles, respectively. The calibration process with the PLS/RBL methodology, and the corresponding training, monitoring and prediction steps with ANN/RBL, were as described above for the remaining two data sets. Numerical results and parameter values are quoted in Table I, showing similar overall results to those discussed above with regard to data set 2. In this specific case, however, the improvement in analytical figures of merit is more apparent in going from PLS/RBL to the ANN/RBL approach.

As with the other studied systems, the unsuspected profiles (both in the spectral and time dimensions) were adequately retrieved by PLS/RBL and ANN/RBL, a necessary step before proceeding to achieve the second-order advantage. Figure 4(E) and (F) confirm this assertion by showing the corresponding plots.

Finally, the relevant plot of prediction errors versus nominal concentration values for the analyte shows the expected deviation of linearity in the case of PLS/RBL processing (Figure 5E), and the apparently better performance of ANN/RBL (Figure 5F).

5. CONCLUSIONS AND OUTLOOK

The approach based on combining ANNs with RBL shows up as a new and efficient chemometric tool for processing second-order spectroscopic or spectral-kinetic information in several non-linear systems, with particular focus on the achievement of the second-order advantage. Its behavior towards simulated data of various kinds indicates good analytical performance, in all cases significantly better than the closest competitor, which involves a flexible PLS model and RBL to provide the latter with the second-order advantage. Although different approaches for achieving the second-order advantage from non-linear data may be developed in the future, the present report provides analysts with a viable alternative based on the combination of two well-tested procedures.

Acknowledgements

Universidad Nacional de Rosario, CONICET (Consejo Nacional de Investigaciones Científicas y Técnicas, Project No. PIP 5303) and ANPCyT (Agencia Nacional de Promoción Científica y Tecnológica, Project No. PICT05-25825) are gratefully acknowledged for financial support.

REFERENCES

1. Blanco M, Coello J, Iturriaga H, Maspoch S, Redon M. Partial least-squares regression for multicomponent kinetic determinations in linear and non-linear systems. *Anal. Chim. Acta* 1995; **303**: 309–320.
2. Goicoechea HC, Collado MS, Satuf ML, Olivieri AC. Complementary use of partial least-squares and artificial neural networks for the non-linear spectrophotometric analysis of pharmaceutical samples. *Anal. Bioanal. Chem.* 2002; **374**: 460–465.
3. Guilbault GG. *Practical Fluorescence* (2nd edn). Marcel Dekker: New York, 1990, Chapter 1.
4. Blanco M, Coello J, Iturriaga H, Maspoch S, Redon M, Villegas N. Artificial neural networks and partial least squares regression for pseudo-first-order with respect to the reagent multicomponent kinetic-spectrophotometric determinations. *Analyst* 1996; **121**: 395–400.
5. Absalan G, Safavi A, Maesum S. Application of artificial neural networks as a technique for interference removal: kinetic—spectrophotometric determination of trace amounts of Se(IV) in the presence of Te(IV). *Talanta* 2001; **55**: 1227–1233.
6. Safavi A, Absalan G, Maesum S. Simultaneous determination of V(IV) and Fe(II) as catalyst using ''neural networks'' through a single catalytic kinetic run. *Anal. Chim. Acta* 2001; **432**: 229–233.
7. Magni DM, Olivieri AC, Bonivardi AL. Artificial neural networks study of the catalytic reduction of resazurin. Stopped flow injection kinetic spectrophotometric determination of Cu(II) and Ni(II). *Anal. Chim. Acta* 2004; **528**: 275–284.
8. Zupan J, Gasteiger J. *Neural Networks in Chemistry and Drug Design*. Wiley: New York, 1999.
9. Despagne F, Massart DL. Neural networks in multivariate calibration. *Analyst* 1998; **123**: 157R–178R.
10. Zupan J, Novic M, Ruisánchez I. Kohonen and counterpropagation artificial neural networks in analytical chemistry. *Chemom. Intell. Lab. Syst*. 1997; **38**: 1–23.
11. Smilde A, Bro R, Geladi P. *Multi-way Analysis: Applications in the Chemical Sciences*. Wiley: New York, 2004.
12. Booksh KS, Kowalski BR. Theory of analytical chemistry. *Anal. Chem*. 1994; **66**: 782A–791A.
13. Ni Y, Huang C, Kokot S. Application of multivariate calibration and artificial neural networks to simultaneous kinetic-spectrophotometric determination of carbamate pesticides. *Chemom. Intell. Lab. Syst*. 2004; **71**: 177–193.
14. Vaananen T, Koskela H, Hiltunen Y, Ala-Korpela M. Application of quantitative artificial neural network analysis to 2D NMR spectra of hydrocarbon mixtures. *J. Chem. Inf. Comput. Sci*. 2002; **42**: 1343–1346.
15. Lopes JA, Menezes JC. Multivariate monitoring of fermentation processes with non-linear modelling methods. *Anal. Chim. Acta* 2004; **515**: 101–108.
16. Öhman J, Geladi P, Wold S. Residual bilinearization. Part I. Theory and algorithms. *J. Chemometrics* 1990; **4**: 79–90.
17. Olivieri AC. On a versatile second-order multivariate calibration method based on partial least-squares and residual bilinearization. Second-order advantage and precision properties. *J. Chemometrics* 2005; **19**: 253–265.
18. Van der Linden WE. Definition and classification of interferences in analytical procedures. *Pure Appl. Chem.* 1989; **61**: 91–95.
19. Culzoni MJ, Goicoechea HC, Pagani AP, Cabezón MA, Olivieri AC. Evaluation of partial least-squares with second-order advantage for the multi-way spectroscopic analysis of complex biological samples in the presence of analyte-background interactions. *Analyst* **131**: 718–723.
20. Bro R. PARAFAC. Tutorial and applications. *Chemom. Intell. Lab. Syst*. 1997; **38**: 149–171.
21. Sanchez E, Kowalski BR. Generalized rank annihilation factor analysis. *Anal. Chem*. 1986; **58**: 496–499.

DOI: 10.1002/cem

22. Tauler R. Multivariate curve resolution applied to second order data. *Chemom. Intell. Lab. Syst*. 1995; **30**: 133–146.
23. Linder M, Sundberg R. Second-order calibration: bilinear least squares regression and a simple alternative. *Chemom. Intell. Lab. Syst*. 1998; **42**: 159–178.
24. Linder M, Sundberg R. Precision of prediction in second-order calibration, with focus on bilinear regression methods. *J. Chemometrics* 2002; **16**: 12–27.
25. Goicoechea HC, Olivieri AC. A new robust bilinear least-squares method for the analysis of spectral-pH matrix data. *Appl. Spectrosc*. 2005; **59**: 926–933.
26. MATLAB 6.0, The MathWorks Inc., Natick, Massachusetts, USA, 2000.
27. Olivieri AC, Goicoechea HC, Iñón FA. MVC1: an integrated Matlab toolbox for first-order multivariate calibration. *Chemom. Intell. Lab. Syst*. 2004; **73**: 189–197.
28. Afkhami A, Safavi A, Massoumi A. Catalytic determination of Pb(II) in the presence of Cu(II). *Anal. Lett*. 1991; **24**: 1643–1655.
29. Magni DM, Olivieri AC, Bonivardi AL. (in preparation).
30. Haaland DM, Thomas EV. Partial least-squares methods for spectral analyses. 1. Relation to other quantitative calibration methods and the extraction of qualitative information. *Anal. Chem*. 1988; **60**: 1193–1202.
31. Xu Q-S, Liang Z. Monte Carlo cross validation. *Chemom. Intell. Lab. Syst*. 2001; **56**: 1–11.

DOI: 10.1002/cem