Analytica Chimica Acta 903 (2016) 51–60

Contents lists available at ScienceDirect

Analytica Chimica Acta

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

Generalized error-dependent prediction uncertainty in multivariate calibration

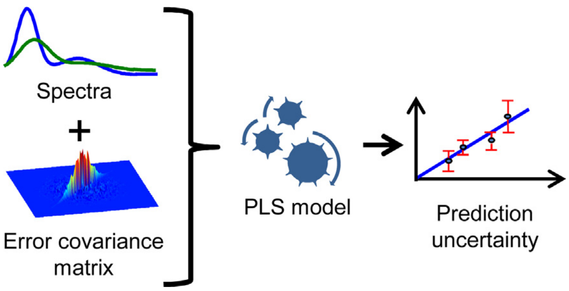
Franco Allegrini [a], Peter D. Wentzell [b, **], Alejandro C. Olivieri [a, *]

[a] Departamento de Química Analítica, Facultad de Ciencias Bioquímicas y Farmacéuticas, Universidad Nacional de Rosario, Instituto de Química de Rosario (IQUIR-CONICET), Suipacha 531, Rosario S2002LRK, Argentina
[b] Department of Chemistry, Dalhousie University, P.O. Box 15000, Halifax, Nova Scotia, B3H 4R2, Canada

HIGHLIGHTS

- Prediction uncertainty in multivariate calibration is addressed.
- Homo-, heteroscedastic and correlated error structures are studied.
- Closed-form expressions for prediction errors are derived.
- Different error sources can be discerned.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 31 July 2015
Received in revised form
19 November 2015
Accepted 23 November 2015
Available online 2 December 2015

Keywords:
Prediction errors
Measurement noise
Multivariate calibration
Error propagation
Heteroscedastic errors
Correlated errors

ABSTRACT

Most of the current expressions used to calculate figures of merit in multivariate calibration have been derived assuming independent and identically distributed (iid) measurement errors. However, it is well known that this condition is not always valid for real data sets, where the existence of many external factors can lead to correlated and/or heteroscedastic noise structures. In this report, the influence of the deviations from the classical iid paradigm is analyzed in the context of error propagation theory. New expressions have been derived to calculate sample dependent prediction standard errors under different scenarios. These expressions allow for a quantitative study of the influence of the different sources of instrumental error affecting the system under analysis. Significant differences are observed when the prediction error is estimated in each of the studied scenarios using the most popular first-order multivariate algorithms, under both simulated and experimental conditions.

1. Introduction

First-order multivariate calibration is today dominated by latent variable based models. Among them, the most popular ones are principal component regression (PCR) [1] and partial least-squares (PLS) regression [2–4]. The latter involves a modification of the former to model the data with fewer latent variables, but there is no clear advantage in terms of quantitative predictive ability [5]. Despite the widespread use of these calibration models in analytical chemistry, one important feature that has been somewhat neglected is the fact that they work optimally when the measurement errors are independently and identically distributed with a

[*] Corresponding author.
[**] Corresponding author.
E-mail addresses: peter.wentzell@dal.ca (P.D. Wentzell), olivieri@iquir-conicet.gov.ar (A.C. Olivieri).

http://dx.doi.org/10.1016/j.aca.2015.11.028

normal distribution (iid normal). The same situation stands for the estimation of important analytical figures of merit, most of which have been defined within the same iid context [6–12]. The subject has arisen considerable interest in recent years, particularly in the present Journal [7–12].

Traditional multivariate calibration methods can often obtain satisfactory results when modest deviations from the ideal iid conditions are present, although this leads to increasing prediction errors [13]. When the error structure significantly deviates from the ideal situation, specific actions may be required to improve calibration performance. There are two alternatives in this regard. One is to apply a suitable data preprocessing method prior to the classical calibration procedure, which modifies the error structure to approximate the iid case (i.e. match the error structure to the model). However, this approach is only possible for certain error structures and such preprocessing may yield suboptimal results when misapplied [13]. The second option is to use an algorithm based on maximum likelihood (ML) principles such as MLPCR (i.e. match the model to the error structure) [14,15]. This latter alternative requires the estimation of the error covariance matrix associated with the error structure by replication and/or modeling [16].

Irrespective of the applied model, the definition of the figures of merit necessary to evaluate and validate the performance of the models operating in a non-iid context is still unclear. Further research is needed to uncover how and to what extent the different error sources affect error propagation in the data under analysis [17]. An approximation based on the value of the mean square error of calibration (MSEC) has been proposed and tested for second-order data [18]. However, this approach is suitable if the measurement noise is the same during both calibration and prediction stages. Moreover, it takes into account the overall effect of the measurement noise, without insight into the specific properties of the individual error sources. This is a fundamental aspect in the development of analytical instrumentation. If one could separately identify the influence of each error source on the final prediction uncertainty, limiting sources of errors could be identified and possibly mitigated to improve the overall quality of the result.

A relevant figure of merit is the sensitivity, which lies at the core of the definition of most analytical quality metrics [19–21]. The sensitivity estimator is well-defined by a general expression covering different algorithms and data orders, i.e., from univariate to multiway calibration [6]. The general formula discussed in the latter report was derived by considering the sensitivity as the ratio of input to output noise, assuming that the input noise is iid. The latter is used as a small perturbing probe which allows one to investigate how it propagates to prediction. However, no assumptions are made regarding the properties of the real experimental noise affecting the system [6]. As a consequence, the interpretation of the sensitivity parameter remains invariable, even when the noise deviates from the iid structure.

However, the prediction uncertainty and other relevant figures of merit that depend on it are significantly affected by the noise structure, as will be clear below. Important reasons for conducting further studies in this field are: (1) all validation procedures require, as a good analytical practice, to report a result together with a reliable estimate of its uncertainty [22,23], and (2) uncertainty estimation is a key step in the calculation of other important figures of merit such as the limit of detection [24]. Even when replicate sample analysis may allow for the experimental estimation of the overall prediction uncertainty, studies such as the present one provide further insight into the different error sources affecting the latter. This is important regarding method optimization aimed at precision improvement, which can be achieved even in the absence of replicates [25].

The error covariance matrix is central to error propagation procedures estimating prediction uncertainty in first-order multivariate calibration. However, relevant expressions for prediction uncertainty have been derived under the iid assumption, without going deeper into the consequences of non-iid situations [23]. On the other hand, Wentzell et al. have highlighted the importance of estimating the noise structure of multivariate data, proposing and testing different methods to model the error covariance matrix [16]. Even in the absence of replicates, heteroscedastic noise can be characterized using a strategy based on a high-pass digital filter [25]. This is an important step to identify non-iid data sets, but does not cover the presence of correlated errors. These two lines of work, involving the estimation of the prediction uncertainty and of the error covariance matrix, are complementary, although no efforts have been undertaken to combine them.

In this work, a general scheme to estimate sample dependent uncertainties in first-order multivariate calibration is presented. It is based on a local linearization/error propagation approach, and requires an adequate estimation of the covariance matrix characterizing the error structure. Three possible situations are described, depending on the type of measurement noise structure for the samples under analysis. Comparison and validation of the results obtained by the proposed expressions is supported by noise addition simulations, and confirmed in some experimental data sets. The presently discussed strategy was developed and tested for both classical PCR and PLS calibration models, these representing the most widely applied inverse least squares methods (even when iid assumptions are not valid) and the most straightforward cases. The validity of the prediction error expressions is not contingent on the optimality of the model (providing it is unbiased). The obtained results are relevant to the estimation of further figures of merit which are a function of the prediction uncertainty, such as the limits of detection and quantitation.

2. Theory

2.1. Latent variable based regression methods

PCR and PLS are the most widespread regression techniques for first-order analytical calibration [3,4]. These models are similar in their basic philosophy: they project the original variables into a vectorial subspace defined to extract the maximum significant variance of the data [3]. This projection shows the main advantage of compressing the information contained in the original data, in such a way that only the relevant information concerning the quantitation of the analyte of interest is kept, while removing small and random noise variability [3]. This also allows one to deal with the usual problems of collinearity (similar spectral responses for the analyte and the interferences) and rank deficiency (number of instrumental sensors larger than number of calibration samples) [3]. These advantages readily explain the popularity of the PCR/PLS approaches, and their success compared to other less complex first-order algorithms such as classical least-squares (CLS) [26] and multiple linear regression (MLR) (sometimes also called inverse least-squares or ILS) [26].

The prediction step for PCR and PLS can be expressed as:

$$\widehat{y} = \mathbf{t}\,\mathbf{T}^{+}\mathbf{y}_{\mathrm{cal}} \tag{1}$$

where $\widehat{y}$ is the predicted analyte concentration (or other predicted quantity) in the test sample, the vector $\mathbf{t}$ contains the scores calculated for the test sample (size $1 \times a$), $\mathbf{y}_{\mathrm{cal}}$ is the vector of reference values used for calibration (size $m \times 1$, where m is the number of calibration samples) and $\mathbf{T}$ is the matrix of calibration scores (size $m \times a$, where a is the number of latent regression variables), resulting from the projection of the original calibration data matrix **X** containing as rows the calibration spectra (size $m \times n$, where n is the number of sensors) into the subspace defined during the compression step. The superscripts ‘T’ and ‘+’ indicate matrix transposition and generalized inverse, respectively. During the projection, the dimensionality is reduced from n sensors to a latent factors. The projection equation to obtain PCR and PLS scores can be written in a common way for both models as:

$$\mathbf{T} = \mathbf{X}\mathbf{V} \tag{2}$$

where **V** (size $n \times a$) is known as the loading matrix. In PCR, **V** is obtained by singular value decomposition (SVD) of the data matrix **X**, whereas in PLS the calculation of **V** is somewhat more complex. In fact, there are nine possible PLS algorithms [27], with the most popular and numerically stable being the non-linear iterative PLS (NIPALS) version, where:

$$\mathbf{V} = \mathbf{W}\left(\mathbf{P}^{\mathrm{T}}\mathbf{W}\right)^{-1} \tag{3}$$

In this expression, two additional sets of loadings **P** and weight loadings **W** are introduced to incorporate concentration information and to maintain orthogonality among scores [3,26].

2.2. The error covariance matrix

To understand the concept and utility of the covariance matrix, it is important to make clear the distinction between the terms “error” and “uncertainty” [17]. Although they are frequently used as synonymous, the uncertainty is a parameter, associated with the result of a measurement, that characterizes the dispersion of the values that could reasonably be attributed to the measurand, while the error is a specific value associated with one particular measurement, which can be expressed as a difference between a measured and a true value, and can thus be either positive or negative.

The error covariance matrix is a symmetric square matrix containing, as diagonal elements, the error variances associated with each measurement channel, and as off-diagonal elements, all of the covariances among measurement errors at different channels. It allows visualization of the statistical relationships among the errors of elements of a vector-form predictor in first-order calibration, as opposed to a single scalar in univariate calibration [17]. For the analysis of the error structure, the diagonal of the covariance matrix gives information about the noise heteroscedasticity, whereas the off-diagonal values describe the nature of the correlated noise.

2.3. Uncertainty propagation approach

The prediction uncertainty is a function of the input data. Every differentiable function can be approximated using a Taylor series expansion truncated after the linear term. This approximation is known as local linearization in statistics, and as error propagation in chemistry [18]. Considering a generalized scalar quantity z, as the predicted analyte concentration in a test sample, which is a function of several variables such that $z = f(x_1, x_2, \ldots)$, the general formula for error propagation is represented by eq. (4) [17]:

$$\sigma_z^2 = \sum_i \left(\frac{\partial z}{\partial x_i}\right)^2 \sigma_i^2 + 2\sum_i \sum_{j>i} \left(\frac{\partial z}{\partial x_i}\right)\left(\frac{\partial z}{\partial x_j}\right)\sigma_{ij} \tag{4}$$

where σ_i^2 is the variance of the errors in x_i and σ_{ij} is the covariance of the errors in x_i and x_j. To simplify the notation, it is convenient to represent eq. (4) in matrix form. This can be achieved by defining a Jacobian column vector **j** which contains the partial derivatives of z with respect to x:

$$\mathbf{j} = \begin{bmatrix} \dfrac{\partial z}{\partial x_1} \\ \dfrac{\partial z}{\partial x_2} \\ \cdots \\ \dfrac{\partial z}{\partial x_n} \end{bmatrix} \tag{5}$$

The variance in the errors of z is then given by:

$$\sigma_z^2 = \mathbf{j}^{\mathrm{T}}\, \Sigma_{\mathrm{x}}\, \mathbf{j} \tag{6}$$

where Σ_x (size $n \times n$) is the error covariance matrix for vector **x**. Eq. (6) is useful for describing the changes in the uncertainty that take place when a transformation is applied to the measurement vector **x**, producing a new scalar y. For example, if $y = \mathbf{c}^{\mathrm{T}}\,\mathbf{x}$, with **c** being a transformation vector, the variance in errors of y is:

$$\sigma_y^2 = \mathbf{c}^{\mathbf{T}}\boldsymbol{\Sigma}_{\mathbf{x}}\mathbf{c} \tag{7}$$

An interesting feature of the latter expression is that it can be applied to a variety of situations including smoothing, differentiation, subspace projection and wavelet transform. Moreover, it can be used to track how the errors in the original measurements are carried through different data analysis steps as is the case of PCR and PLS regression. From the prediction eq. (1), error propagation can be performed assuming that the main source of error is the instrumental signal of the unknown sample. In this case, eq. (1) can be conveniently expressed as:

$$\widehat{y} = \mathbf{t}\mathbf{v} \tag{8}$$

where **v** is the vector of regression coefficients in the latent variable space (size $a \times 1$). Since **t** is the projection of the unknown sample measurement vector **x** onto the subspace defined by the loadings **V**, eq. (8) can also be written as:

$$\widehat{y} = \mathbf{v}^{\mathrm{T}}\mathbf{V}^{+}\ \mathbf{x} = \mathbf{b}^{\mathrm{T}}\mathbf{x} \tag{9}$$

where **b** is the regression vector in the original variables space. If error propagation is performed on the latter equation, by analogy with eq. (7), the uncertainty in the predicted concentration, due to the uncertainty in the instrumental signal of the unknown sample, will be given by:

$$\sigma_{\widehat{y}}^2 = \mathbf{b}^{\mathrm{T}}\, \Sigma_{\mathbf{x}}\, \mathbf{b} \tag{10}$$

2.4. General scheme for uncertainty estimation

It has been previously shown that the global uncertainty in prediction in inverse multivariate models can be estimated as a sum of three independent terms. They account for the propagation of uncertainties derived from: (1) instrumental signals which are measured for the test sample (σ_1), (2) instrumental signals measured to build the calibration data set (σ_2), and (3) calibration concentrations for the analyte or property of interest (σ_3) [22]. This can be expressed in a general way as a variance propagation expression:

$$\sigma^2_{\hat{y}} = \sigma^2_1 + \sigma^2_2 + \sigma^2_3 \tag{11}$$

From this starting point, the error propagation can be performed based on the initial assumptions about the error structure of the data set under analysis. There are three possible scenarios with respect to instrument noise, as summarized in Table 1: (1) iid noise (Case 1), (2) non-iid noise with all calibration and test samples having the same error structure (Case 2), (3) non-iid noise with calibration and test samples having different error structures (Case 3). Table 1 also shows the final expressions obtained for each of the previously described scenarios. The principles to derive these expressions have been briefly presented in the previous section, and a more detailed explanation is presented in the Supplementary Material. To the best of our knowledge, the obtained expressions for the second term of eq. (11) in Cases 2 and 3 are novel, as well as their specific application to the field of analytical chemistry. The expression for the first term has already been discussed, although in the context of classical and inverse least-squares multivariate calibration [28].

As expected for Case 1, where the error structure is iid (Table 1), the error covariance matrix is an identity matrix multiplied by the measurement error variance, and the equation for $\sigma^2_{\hat{y}}$ agrees with the one proposed by Faber and Kowalski [23]. The iid assumption is, however, not necessary for the derivation of a more general expression based on the estimation of the covariance matrix, as proposed in the present report for cases 2 and 3 (Table 1 and Supplementary Material).

The expressions for terms 2 and 3 in Case 2 (Table 1) are natural extensions of Case 1. It may be noticed that the error covariance matrices are represented with different symbols: the subscript ‘x’ in Σ_x indicates the error covariance matrix for the test sample, whereas ‘X’ in Σ_X is reserved for the calibration data. However, in Case 2 all samples are assumed to have the same error structure, and hence the error covariance matrices for test and calibration samples will normally match. The nomenclature in Table 1 aims at distinguishing the independence that in principle exists between terms 1 and 2.

Case 3 of Table 1 is the most general of the presently analyzed situations. Of particular interest, because of its complexity, is the expression corresponding to term 2 in Case 3. Here the error structure varies from sample to sample in the calibration set, making it necessary to include individual error covariance matrices belonging to the calibration samples (see Supplementary Material). The result is an expression closely related to the analogous one for Case 2, replacing the common calibration error covariance matrix by an effective one. The latter is the weighted average of all calibration error covariance matrices, i.e.:

$$\mathbf{\Sigma}_{\mathrm{X,eff}} = \frac{1}{h}\left(\mathbf{\Sigma}_{\mathrm{X},1}h_1^2 + \mathbf{\Sigma}_{\mathrm{X},2}h_2^2 + \ldots + \mathbf{\Sigma}_{\mathrm{X},m}h_m^2\right) \tag{12}$$

where $\mathbf{\Sigma}_{\mathrm{X},1}$, $\mathbf{\Sigma}_{\mathrm{X},2}$, … are the error covariance matrices for each calibration sample, h_1, h_2, … are the elements of the $(1 \times m)$ vector $\mathbf{h} = \mathbf{tT}^{+}$, and h is the test sample leverage, as detailed in the Supplementary Material.

Table 1
Expressions derived by error propagation for each of three possible error structure cases.[a]

$\sigma^2_{\hat{y}}$	σ^2_1	σ^2_2	σ^2_3
Case 1	$\mathbf{b}^{\mathrm{T}}\mathbf{b}\sigma^2_x = \frac{\sigma^2_x}{SEN^2}$	$h\mathbf{b}^{\mathrm{T}}\mathbf{b}\sigma^2_X = h\frac{\sigma^2_X}{SEN^2}$	$h\ \sigma^2_{y_{cal}}$
Case 2	$\mathbf{b}^{\mathrm{T}}\ \Sigma_x\ \mathbf{b}$	$h\ \mathbf{b}^{\mathrm{T}}\ \Sigma_X\ \mathbf{b}$	$h\ \sigma^2_{y_{cal}}$
Case 3	$\mathbf{b}^{\mathrm{T}}\ \Sigma_x\ \mathbf{b}$	$h\ \mathbf{b}^{\mathrm{T}}\ \Sigma_{X,\mathrm{eff}}\ \mathbf{b}$	$h\ \sigma^2_{y_{cal}}$

[a] SEN = analyte sensitivity (see Supplementary Material), h = sample leverage, σ^2_x = error variance in test sample signals, σ^2_X = error variance in calibration signals, $\sigma^2_{y_{\mathrm{cal}}}$ = error variance in calibration concentrations, $\mathbf{b}$ = vector of regression coefficients, $\mathbf{y}_{\mathrm{cal}}$ = vector of calibration concentrations, Σ_x, Σ_X = error covariance matrices for test sample signals and calibration signals respectively, and $\Sigma_{X,\mathrm{eff}}$, effective error covariance matrix for the calibration set. For further details on the latter parameter, see Supplementary Material.

3. Data sets

3.1. Simulated data

Synthetic data sets were created by mimicking a three-component analytical system, with component 1 being the analyte of interest. Individual calibration and test spectra ($\mathbf{x}$) were built using the following expression:

$$\mathbf{x} = y_1\ \mathbf{s}_1 + y_2\ \mathbf{s}_2 +\ y_3\ \mathbf{s}_3 \tag{13}$$

where $\mathbf{s}_1$, $\mathbf{s}_2$ and $\mathbf{s}_3$ are the pure component spectra at unit concentration defined in a range of 200 data points and y_1, y_2 and y_3 are the sample component concentrations. The pure component signals $\mathbf{s}_1$, $\mathbf{s}_2$ and $\mathbf{s}_3$ are Gaussian shaped functions, with the analyte to be quantified centered at sensor 100 and the other two components varying their positions to generate different sensitivity values depending on the degree of overlap. For more information about the way these data are generated, the reader is referred to references 24 and 35. In all simulated cases, calibration sets were created with 50 samples, and test sets with 4 samples, in all cases having component concentrations taken randomly from the range 0–1.

3.2. Noise addition

For the present noise addition simulations, uncertainty was added in four different manners: (1) in calibration concentrations only [ERRCON, leading to σ_3 in eq. (11) and Table 1], (2) in calibration signals only (ERRCAL, leading to σ_2), (3) in test sample signals only (ERRTST, leading to σ_1), (4) in all concentrations and signals (ERRALL). This strategy makes it possible to analyze the error propagation results for each of the terms presented in eq. (11) separately and in combination.

To test each of the cases presented in Table 1, three typical measurement noise structures were selected. For each of these, the standard deviation in the calibration concentrations (when required for ERRCON and ERRALL) was set to 0.2% of the mean calibration concentration. Otherwise, the structures differed only in the characteristics of the noise in the calibration and test signals when this was included (ERRCAL, ERRTST, ERRALL). For the first structure (designated as IID, consistent with the Case 1 equations), Gaussian iid noise was added in such a way that the signal standard deviation was 4% of the mean spectral intensity value for the calibration samples. This value was selected to balance the relative impacts of calibration instrumental uncertainty (term 2) and the concentration uncertainty (term 3). For the second structure (designated as COR, consistent with Case 2), we chose partially correlated 1/*f* noise, also known as pink noise [17,29,30]. This low frequency noise corresponds to systems where the errors in adjacent measurements are more correlated than those which are farther apart. In the time domain, it is often associated with low frequency drift in a source or detector [17]. Given the nature of this noise, it can be assumed that its correlation structure will not change significantly from sample to sample. The magnitude of the 1/*f* noise standard deviation was 2% of the mean spectral intensity and the noise sequences were generated using a MATLAB [31] function written by the authors (see

Supplementary Material). Finally, for the third error structure (designated as HET and consistent with Case 3), heteroscedastic proportional noise was added with a proportionality factor of 0.001 with respect to the signal at each sensor. This error shows a non-uniform standard deviation which is proportional to the signal magnitude, and therefore each specific sample will have a unique error covariance matrix.

For each of the above noise types, 100 different systems were created, with different calibration and test concentrations (randomly selected in the range 0–1), and different degrees of overlap among component spectra (see above). This generated 100 different sensitivity values. In each of these 100 systems, the calibration/prediction process was repeated 1000 times for each sensitivity value, using different random seeds for the signal and/or concentration uncertainty but keeping the same component profiles, calibration concentrations and test concentrations. PCR and PLS gave very similar results; we report here those obtained using PCR. Once the concentrations of the analyte of interest (constituent 1 in all cases) were estimated, the corresponding variance was computed across the 1000 test samples of each of the 100 systems. A flow chart summarizing the applied noise addition procedure is presented in Fig. 1. All calculations were done using MATLAB version 7.0 [31], using in-house routines.

3.3. Experimental data

This data set was generated by Shreyer et al. [32] to test the application of MLPCR to fluorescence emission spectra. The spectroscopic data were obtained for 27 mixtures of three polycyclic aromatic hydrocarbons, acenapththylene, naphthalene and phenanthrene, using a three level, three-factor factorial design with five replicates per mixture. The replicates were scanned in a randomized order of five blocks. The emission spectra were obtained through a 1 cm quartz cuvette on a Shimadzu RF-5301PC spectrofluorometer with a xenon lamp excitation source. The excitation wavelength was 278 nm, and the emission wavelength was scanned at 0.2 nm intervals between 310 and 460 nm. A medium scan speed was used, and the excitation and emission slit widths were both set to 3.0 nm. This data set is included in MATLAB format with the Supplementary Material.

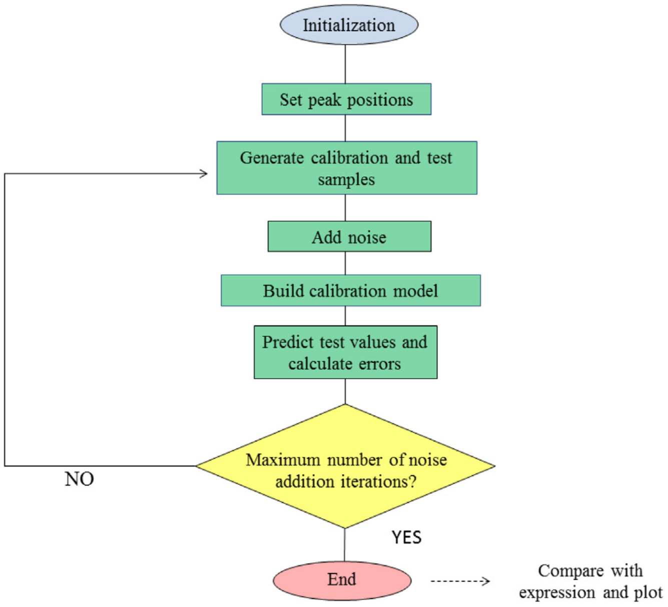

Initialization
Set peak positions
Generate calibration and test samples
Add noise
Build calibration model
Predict test values and calculate errors
Maximum number of noise addition iterations?
NO
YES
End
Compare with expression and plot

Fig. 1. Flow chart summarizing the noise addition simulations.

4. Results and discussion

4.1. Estimation of the error covariance matrix

As discussed in the theoretical section, it is apparent that the error covariance matrix plays a fundamental role for estimating the prediction uncertainty in the presence of non-iid noise. For this reason, it is important to discuss the way in which this matrix can be estimated. There are three main alternatives: (1) experimental replication, (2) theoretical prediction and (3) empirical modeling [17].

Experimental replication consists in taking N replicate measurement vectors $\mathbf{x}_j$ and then calculating the error covariance matrix as:

$$\Sigma_{\mathrm{X}} = \frac{1}{N-1}\sum_{j=1}^{N}\left(\mathbf{x}_j-\overline{\mathbf{x}}\right)\left(\mathbf{x}_j-\overline{\mathbf{x}}\right)^{\mathrm{T}} \tag{14}$$

There are two crucial issues in this procedure. One is the level of replication used, which refers to the way in which replicates are defined. For example, a replicate could refer to repeated scans without removing the sample from the instrument, or to replicate subsamples prepared and measured separately. Another important factor to consider is the number of replicates that is required to avoid a high uncertainty in the estimated values of variance, i.e. to minimize the "noise in the noise". It is well known that a high number of replicates (more than 100) is needed to reach reasonable standard deviations in the variance [17]. However, as this number is not realistic in practical terms, the best solution involves pooling the error covariance estimates by averaging the calculated covariance matrix from different subsets of samples, each of which has a relatively small number of replicates [16]. This is only possible as long as measurement vectors do not significantly change between samples, a typical situation for many analytical instruments. This strategy was employed for the experimental data set used in this work.

Theoretical prediction of the error covariance matrix is also possible for simulated data and for certain experimental data sets in which the error sources are well characterized. This strategy was employed for the simulated data sets used in this work. Fig. 2 shows typical error sequences and theoretical error covariance matrices (presented as image plots) for each of the signal noise structures discussed earlier (IID, COR, HET), as well as their combination.

Finally, empirical modeling represents an intermediate alternative between the previous approaches. It aims to find a model capable of providing a reliable estimate of the error covariance [16,17]. Although this option is not as simple and direct as experimental replication and requires some expertise from the analyst, it has several advantages such as a better understanding of the limiting measurement error in the system under study, a reduction in the need for replicates, and a smoothing of the stochastic variations inherent in the estimation of experimental error covariance matrices. While this approach was not used here, it could prove to be the most practical method for future applications, highlighting the importance and potential of the proposed expressions for real world applications. This is due to the fact that once a reliable covariance matrix model has been achieved through calibration samples, it can be further applied to estimate uncertainty in test samples without the requirement of new replicates.

A complementary concept, useful to analyze the error structure, is the correlation matrix, obtained by dividing each element of the error covariance matrix by the two contributing standard deviations [16,17]. The diagonal elements of the correlation matrix are all ones, while the off-diagonal elements indicate the degree of measurement error correlation.

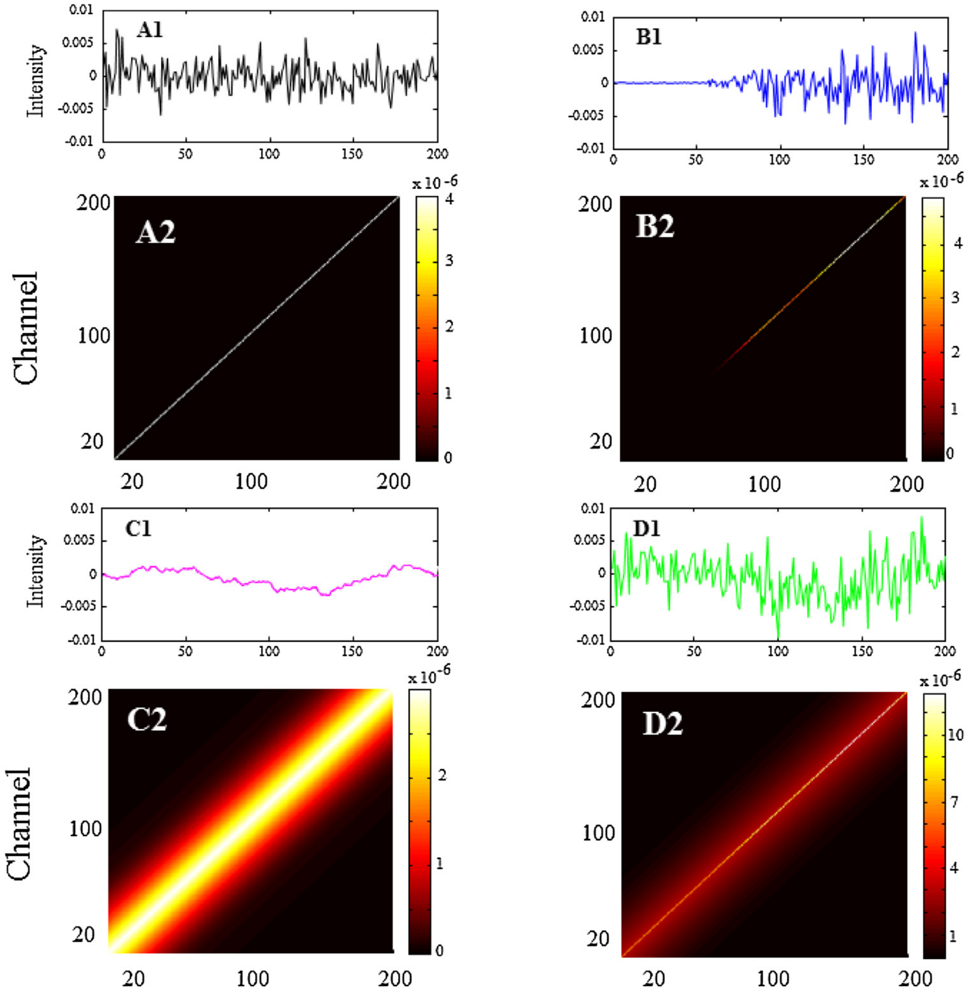


Fig. 2. Typical noise sequences (A1, B1, C1 and D1) and theoretical error covariance matrices (A2, B2, C2 and D2) used in the simulations. The matrices are illustrated with images resulting from scaling the data to the full range of the current color map. The different sub-panels illustrate iid noise (A1 and A2), proportional noise (B1 and B2) pink noise (C1 and C2) and the sum of the three noise sequences (D1 and D2).

4.2. Simulated data

To check the performance of the equations shown in Table 1, a noise addition calibration/prediction procedure was applied (see Fig. 1). For each signal error structure (IID, COR, HET), this was done, as discussed above, in four different situations, including noise in calibration concentrations only (ERRCON), in calibration signals only (ERRCAL), in test sample signals only (ERRTST), and in all of them (ERRALL). To test the adequacy of the equations in a reliable way, the theoretical error covariance matrices were used (see Fig. 2).

Because of the large number of systems studied, a convenient way to summarize the results is by plotting the concentration uncertainties estimated by the corresponding expression in Table 1 vs. those estimated by computing the variance of the predicted concentrations in the 1000 samples of each of the 100 simulated systems. The different uncertainty sources are identified using specific symbols. This strategy has already been used to test sensitivity expressions in multiway systems [33–35]. Fig. 3 corresponds to the presently simulated data sets under the effect of the above described noise types.

The overall results presented in Fig. 3 suggest that the error propagation approach based on the inclusion of the proper error covariance matrix to estimate prediction uncertainty is appropriate, since noise addition uncertainties reasonably match those computed from the corresponding derived equation (Fig. 3A, C and 3F). As expected, when iid noise is assumed in systems with a different noise structure, the classical Case 1 expressions notoriously fail in estimating the uncertainty except when errors are present only in the calibration concentrations (easily appreciated in Fig. 3B and D). The differences with respect to expressions taking into account correlation and heteroscedasticity are of at least an order of magnitude.

It is interesting to note that, for the simulation conditions employed here, the prediction uncertainty calculation assuming iid errors (Case 1) underestimates the true prediction uncertainty when correlated errors are present (Fig. 3B), whereas for proportional noise this relationship is not straightforward (Fig. 3D). The former observation can be qualitatively understood with reference to eq. (4), where the iid equations exclude the contribution of the error covariance, which is positive in this case. In other words, the correlation means that the random errors do not cancel in the way that is assumed. The prediction uncertainty results for proportional noise can be appreciated by recognizing that the regression vector gives differential weights to different spectral regions in the prediction uncertainty eq. (10). Consequently, if spectral regions where the regression vector is significant show higher intensities than other regions, smaller uncertainties will be calculated using an averaged measurement uncertainty assuming iid errors. The opposite will be true for high spectral intensities in regions where

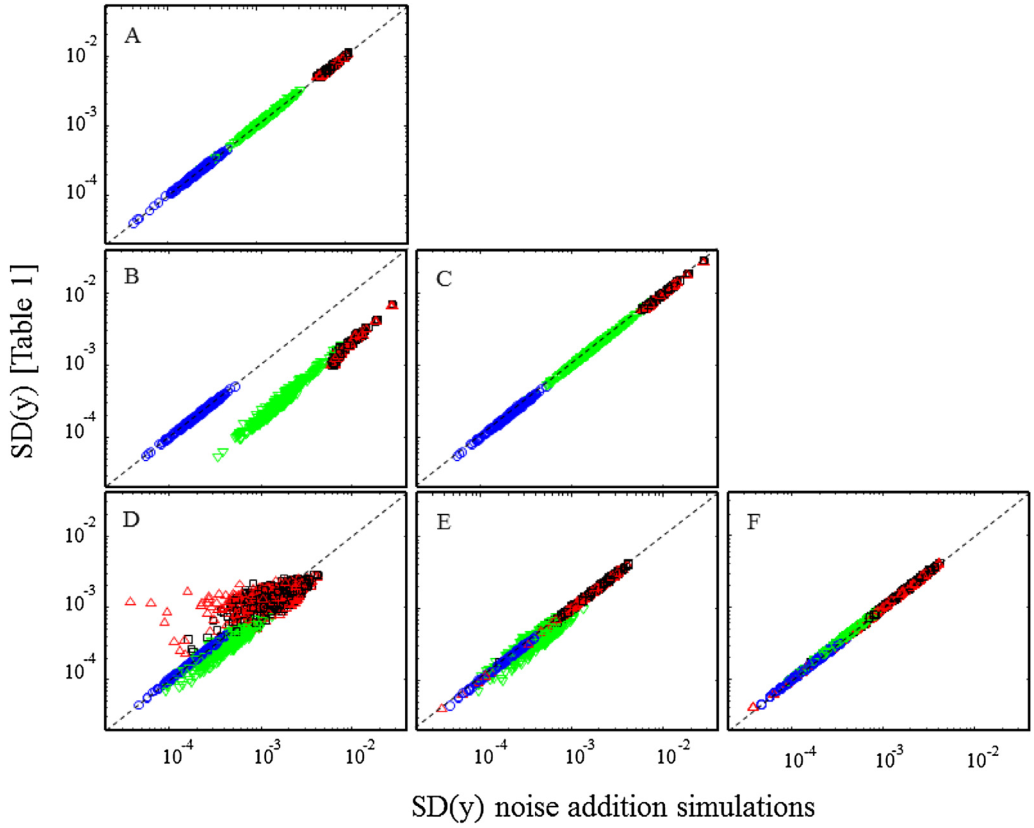


Fig. 3. Plots of calculated uncertainties in predicted concentrations, as a function of noise addition results. The different panels show the results for the addition of: iid noise (A), pink noise (B and C), and proportional noise (D, E and F). The theoretical values were estimated using the following information from Table 1: (A), (B) and (D), expressions for Case 1, (C) and (E), expressions for Case 2 and (F), expressions for Case 3. In plot (E), corresponding to uncertainties in calibration signals, $\boldsymbol{\Sigma}_{\mathrm{X}}$ was given the value of the average error covariance matrix over the calibration set. In all plots, the symbols identify the following cases: blue circles, noise only in calibration concentration, green down triangles, noise only in calibration signals, red up triangles, noise only in test sample signals, and black squares, noise in all concentrations and signals. The thin solid line indicates perfect correlation. All axes are in logarithmic scale.

the regression vector is small, due to sample components other than the analyte.

Proportional noise deserves special attention. The corresponding expression is shown in Table 1 for Case 3, and includes a second term accounting for the variation in error covariance matrices within the calibration set of samples. A simpler approach involves the use term 2 of Case 2 instead, which depends on the mean error covariance for all calibration samples. This represents an approximation to Case 2 when the calibration spectra are similar and results in a much more straightforward calculation. However, as the individual error covariance matrices deviate further from the mean, the accuracy of the predicted uncertainties diminishes, as shown in Fig. 3E, and the Case 3 expressions should be used (Fig. 3F).

A final integrated system was studied by adding all of the previously mentioned error sources together to analyze the combined effect of different noise sources. For this purpose, the relative proportion of each kind of noise was varied, leading to three different cases: (1) iid noise dominant (Fig. 4A), (2) partially correlated noise (pink noise) dominant (Fig. 4B) and (3) proportional noise dominant (Fig. 4C). In each of these cases, the prediction uncertainty was estimated using four different covariance matrices: one including all the error sources (green circles), and the remaining three separately considering each error source (blue, black and pink circles respectively). From Fig. 4, it is apparent that the latter three equations underestimate the prediction uncertainty, as expected, and the largest deviation from the ideal line corresponds to estimations based on the covariance matrix for the error source which is in the lowest relative proportion. Conversely, if only the most important error source is taken into account, the estimation is still good, despite the fact that other error sources are present in relatively smaller proportions.

4.3. Experimental data

The impact of the specific noise structure in uncertainty estimation should be apparent in experimental data sets, where the noise is rarely homoscedastic. In these systems, the noise addition calibration/prediction strategy is not viable, because error-free reference values are not available. Therefore, it is necessary to find a reliable way to assess which expression is more realistic.

An interesting alternative is the approximation proposed by Faber and Bro, based on the consideration of the following random variable [18]:

$$t = \frac{\widehat{y} - y_{\mathrm{ref}}}{s_{\widehat{y}}} \tag{15}$$

where y_{ref} and $\widehat{y}$ are the reference and predicted concentration values and $s_{\widehat{y}}$ is the standard deviation estimated from the equation for prediction uncertainty. When valid estimates of prediction uncertainty are obtained, the values of t for a group of prediction samples should be approximately distributed as a Student's t coefficient, which, for a large number of degrees of freedom, has a standard deviation close to 1 [18]. Consequently, demanding that the true prediction error to be correctly estimated on average amounts to demanding the expected standard deviation of t,

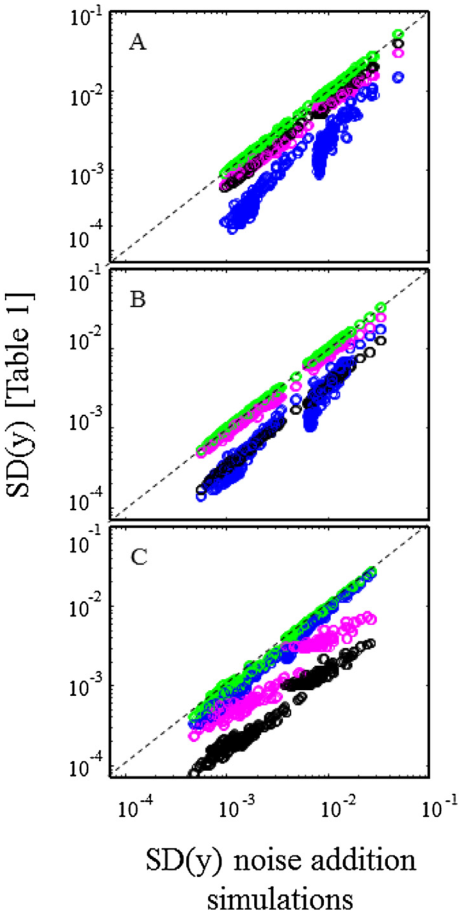


Fig. 4. Plot of uncertainties in predicted concentrations as a function of noise addition results when a combination of iid, pink and proportional noise is added in different relative proportions. The main error sources are: (A) iid noise, (B) pink noise, and (C) proportional noise. The green circles identify the standard deviation calculated using the expressions in Table 1, with an error covariance matrix which is the sum of the three individual sources of error. Black circles correspond to uncertainties calculated when only iid noise is considered in the covariance matrix, pink circles when only pink noise is included, and blue circles when only proportional noise is taken into account.

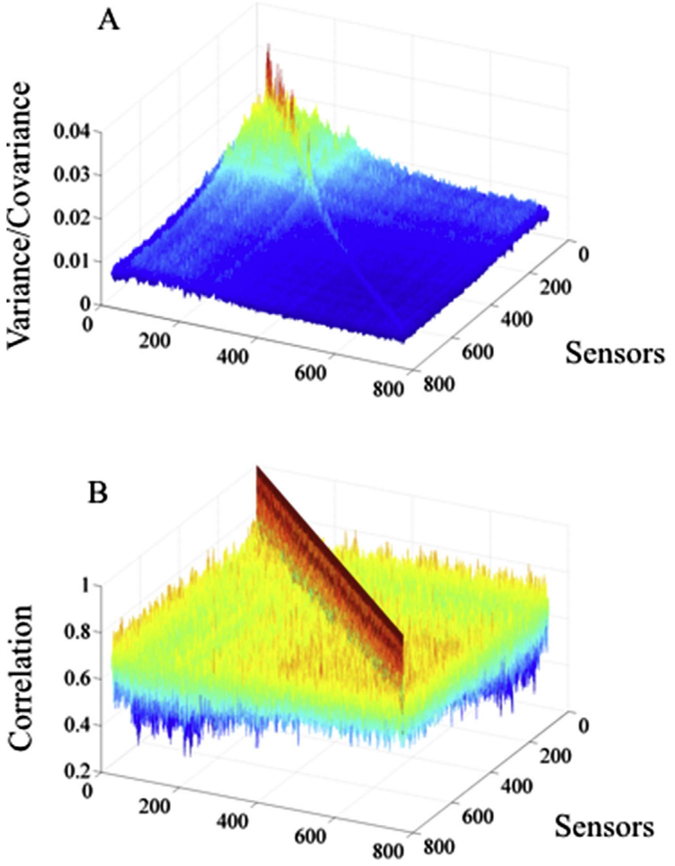


Fig. 5. Error covariance matrix (A) and correlation matrix (B) for experimental fluorescence emission data.

calculated from the pool of predicted samples, to approach unity.

The error covariance matrix of the fluorescence emission data set analyzed in this work is shown in Fig. 5 together with the correlation matrix. These matrices have already been characterized and modeled by Wentzell et al. using a target testing approach to determine the most significant error sources, together with principal component analysis to calculate the corresponding contributions [16]. The most important conclusions were that the main contributions to the noise structure are: (1) the proportional noise related to the well-known shot noise in fluorescence spectroscopy, (2) a constant offset noise which arises from cell positioning or blank measurement, and (3) an offset noise proportional to the square root of the spectrum.

For each of the calibration samples, the five replicates were left out from the data set, and calibration together with the covariance matrix estimation was carried out using the remaining samples with their corresponding replicates. Prediction was then performed using PCR and PLS with four latent factors, with very similar results. Those for PCR are reported below. Although one might expect an optimum of three latent variables on the basis of the sample composition, other factors such as baseline offset may increase this number [32]. Table 2 shows the root mean square errors of prediction (RMSECV) (estimated using cross-validation) and mean standard deviations (MSD) (calculated using equations in Table 1 for Case 2) for the experimental samples under different assumptions of noise structure. As can be seen, the standard prediction error is notoriously underestimated under the iid assumption (consistent with Fig. 3B in the simulations). However, if uncorrelated heteroscedastic errors are assumed, i.e., considering only the diagonal elements of the error covariance matrix calculated by pooling replicates, the MSD value increases and approaches the experimental RMSECV. The relative improvement is greatest for acenaphthylene, suggesting that the heteroscedasticity is relatively more important than correlated noise in this case. The uncertainty estimation improves even more if the complete error covariance matrix is considered, and achieves remarkably good agreement with the experimentally observed value.

Table 2 also shows the standard deviation of the *t* values (σ_t) given by eq. (14). A value of σ_t less than unity implies overestimation of the standard prediction error and vice-versa. The results agree with those discussed above when comparing RMSECV and MSD values. The value of σ_t calculated for the three analytes under the iid assumption is ca. 2, implying a significant underestimation of the standard prediction error. However, σ_t moves closer to unity when heteroscedasticity is included, and close to the ideal value for all three analytes when the correlated noise is accounted for as well. As was the case for the MSD values, the presence of correlated errors seems to have the greatest effect on for naphthalene and phenanthrene. For all three analytes, the prediction uncertainty appears to be slightly underestimated, and this may be due to the use of an average covariance matrix instead of individual values, as implied in Fig. 3E.

The presently discussed trends clearly show the importance of

Table 2
Uncertainty estimation assessment for the models built for each of the compounds used of the fluorescence emission data set.[a]

	a	RMSECV (μg/g)	MSD (μg/g)			σ_t (μg/g)		
			iid	het	niid	iid	het	niid
Acenaphthylene	4	0.0081	0.0036	0.0072	0.0075	2.27	1.14	1.10
Naphthalene	4	0.0011	0.0004	0.0008	0.0011	2.63	1.46	1.07
Phenanthrene	4	0.0003	0.0001	0.0002	0.0003	2.46	1.83	1.02

[a] a = number of latent variables, RMSECV = root mean square error of prediction, MSD = mean standard deviation, σ_t, standard deviation of t values, het = heteroscedastic and uncorrelated, niid = non-iid, i.e., heteroscedastic and correlated.

drawing attention to the different contributions affecting the error structure of the data under analysis to estimate reliable uncertainty values. It is important to note that proposed approach allows one to estimate prediction uncertainties for different error sources, without depending on the RMSECV value, which is not sample specific. The estimator is also useful to analyze the influence of specific error sources. Since in this experimental case a replication strategy was used to build the error covariance matrix, the error covariance effects were analyzed only in terms of heteroscedastic errors (diagonal elements) or correlated ones (non-diagonal elements). However, if an empirical noise modeling strategy is used, a comparison can be performed on a source by source basis, as previously done for the simulated data sets.

5. Conclusions

Since the beginning of chemometrics as a discipline assisting analytical chemistry, pioneers have remarked that the developments made in this field should help as a decision making guide to instrument manufacturers in order to improve equipment performance. In this sense, a better understanding of the origins of error sources for a particular instrument can be used to improve the quality of measurements. With this aim, we have investigated an estimator for prediction uncertainty based on the determination of the error covariance matrix. To quantify the impact of different error sources, the multivariate error structure has been propagated to the final result. Comparison of the proposed expressions with the classical iid formula shows significant differences in the prediction uncertainty, depending on the variation of the error structure for test and calibration samples. Finally, an important conclusion regarding experimental data, is that the use of a reasonable number of replicate measurements during the calibration stage, combined either with a pooling strategy (if possible) or an empirical model, allows the estimation of the prediction uncertainty of future test samples even in the absence of replicates. Future research work would involve extending the present analysis to maximum likelihood methods such as MLPCR, which take into account the error structure for building the multivariate models.

Acknowledgments

Universidad Nacional de Rosario, CONICET (Consejo Nacional de Investigaciones Científicas y Técnicas) (Project No. PIP 163), ANPCyT (Agencia Nacional de Promoción Científica y Tecnológica, Project No. PICT-2013-0136) and NSERC (Natural Sciences and Engineering Research Council of Canada) (Grant No. 46316) are gratefully acknowledged for financial support. F. A. thanks CONICET for a doctoral fellowship and Fundación Josefina Prats for financial support.

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.aca.2015.11.028.

References

[1] S. Wold, K. Esbensen, P. Geladi, Principal component analysis, Chemom. Intell. Lab. Syst. 2 (1987) 37–52.

[2] S. Wold, M. Sjöström, L. Eriksson, PLS-regression: a basic tool of chemometrics, Chemom. Intell. Lab. Syst. 58 (2001) 109–130.

[3] H. Martens, T. Næs, Multivariate Calibration, John Wiley, Chichester, 1989.

[4] D.L. Massart, B.G.M. Vandeginste, L.M.C. Buydens, S. De Jong, P.J. Lewi, J. Smeyers-Verbeke, Handbook of Chemometrics and Qualimetrics, Elsevier, Amsterdam, 1997.

[5] P.D. Wentzell, L.V. Montoto, Comparison of principal components regression and partial least squares regression through generic simulations of complex mixtures, Chemom. Intell. Lab. Syst. 65 (2003) 257–279.

[6] A.C. Olivieri, Analytical figures of merit: from univariate to multiway calibration, Chem. Rev. 114 (2014) 5358–5378.

[7] A.C. Olivieri, Practical guidelines for reporting results in single- and multi-component analytical calibration: a tutorial, Anal. Chim. Acta 868 (2015) 10–22.

[8] R. Boqué, N.M. Faber, F.X. Rius, Detection limits in classical multivariate calibration models, Anal. Chim. Acta 423 (2000) 41–49.

[9] A. Singh, Multivariate decision and detection limits, Anal. Chim. Acta 277 (1993) 205–214.

[10] J. Ferré, R. Boqué, B. Fernandez Band, M.S. Larrechi, F.X. Rius, Figures of merit in multivariate calibration. Determination of four pesticides in water by flow injection analysis and spectrophotometric detection, Anal. Chim. Acta 348 (1997) 167–175.

[11] M.B. Sanz, L.A. Sarabia, A. Herrero, M.C. Ortiz, Capability of discrimination: application to soft calibration methods, Anal. Chim. Acta 446 (2001) 297–311.

[12] M.B. Sanz, L.A. Sarabia, A. Herrero, M.C. Ortiz, Multivariate analytical sensitivity in the determination of selenium, copper, lead and cadmium by stripping voltammetry when using soft calibration, Anal. Chim. Acta 489 (2003) 85–94.

[13] C.D. Brown, L. Vega-Montoto, P.D. Wentzell, Derivative preprocessing and optimal corrections for baseline drift in multivariate calibration, Appl. Spectrosc. 54 (2000) 1055–1068.

[14] P.D. Wentzell, D.T. Andrews, B.R. Kowalski, Maximum likelihood multivariate calibration, Anal. Chem. 69 (1997) 2299–2311.

[15] P.D. Wentzell, D.T. Andrews, D.C. Hamilton, K. Faber, B.R. Kowalski, Maximum likelihood principal component analysis, J. Chemom. 11 (1997) 339–366.

[16] M.N. Leger, L. Vega-Montoto, P.D. Wentzell, Methods for systematic investigation of measurement error covariance matrices, Chemom. Intell. Lab. Syst. 77 (2005) 181–205.

[17] P.D. Wentzell, Measurement errors in multivariate chemical data, J. Braz. Chem. Soc. 25 (2014) 183–196.

[18] N.M. Faber, R. Bro, Standard error of prediction for multiway PLS: 1. Background and a simulation study, Chemom. Intell. Lab. Syst. 61 (2002) 133–149.

[19] K. Danzer, L.A. Currie, Guidelines for calibration in analytical chemistry. Part I. Fundamentals and single component calibration, Pure Appl. Chem. 70 (1998) 993–1014.

[20] K. Danzer, M. Otto, L.A. Currie, Guidelines for calibration in analytical chemistry. Part 2: Multicomponent calibration, Pure Appl. Chem. 76 (2004) 1215–1225.

[21] A.C. Olivieri, N.M. Faber, J. Ferre, R. Boque, J.H. Kalivas, H. Mark, Uncertainty estimation and figures of merit for multivariate calibration, Pure Appl. Chem. 78 (2006) 633–661.

[22] K. Faber, A. Lorber, B.R. Kowalski, Analytical figures of merit for tensorial calibration, J. Chemom. 11 (1997) 419–461.

[23] K. Faber, B.R. Kowalski, Propagation of measurement errors for the validation of predictions obtained by principal component regression and partial least squares, J. Chemom. 11 (1997) 181–238.

[24] F. Allegrini, A.C. Olivieri, IUPAC-consistent approach to the limit of detection in partial least-squares calibration, Anal. Chem. 86 (2014) 7858–7866.

[25] P.D. Wentzell, A.C. Tarasuk, Characterization of heteroscedastic measurement noise in the absence of replicates, Anal. Chim. Acta 847 (2014) 16–28.

[26] D.M. Haaland, E.V. Thomas, Partial least-squares methods for spectral analyses. 1. Relation to other quantitative calibration methods and the extraction of qualitative information, Anal. Chem. 60 (1988) 1193–1202.

[27] M. Andersson, A comparison of nine PLS1 algorithms, J. Chemom. 23 (2009) 518–529.

[28] C.D. Brown, Discordance between net analyte signal theory and practical

multivariate calibration, Anal. Chem. 76 (2004) 4364–4373.

[29] W.H. Press, Flicker noises in astronomy and elsewhere, Comments Astrophys. 7 (1978) 103–119.

[30] P.H. Handel, A.L. Chung, International Conference on Noise in Physical Systems and 1/f Fluctuations, American Institute of Physics, New York, 1993.

[31] MATLAB, Version 7.0, The Mathworks Inc., Natick, Massachusetts, 2010.

[32] S.K. Schreyer, M. Bidinosti, P.D. Wentzell, Application of maximum likelihood principal components regression to fluorescence emission spectra, Appl. Spectrosc. 56 (2002) 789–796.

[33] A.C. Olivieri, K. Faber, New developments for the sensitivity estimation in four-way calibration with the quadri-linear parallel factor model, Anal. Chem. 84 (2012) 186–193.

[34] M.C. Bauza, G.A. Ibanez, R. Tauler, A.C. Olivieri, A sensitivity equation for quantitative analysis with multivariate curve resolution – alternating least-squares. Theoretical and experimental approach, Anal. Chem. 84 (2012) 8697–8706.

[35] F. Allegrini, A.C. Olivieri, Analytical figures of merit for partial least-squares coupled to residual multi-linearization, Anal. Chem. 84 (2012) 10823–10830.