analytical chemistry

Article

pubs.acs.org/ac

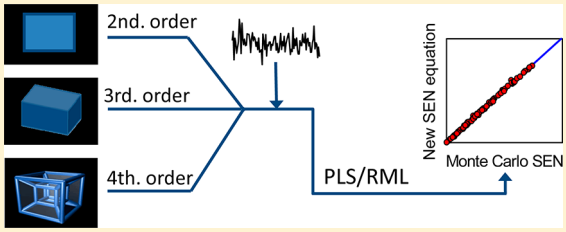
Analytical Figures of Merit for Partial Least-Squares Coupled to Residual Multilinearization

Franco Allegrini and Alejandro C. Olivieri*

Departamento de Química Analítica, Facultad de Ciencias Bioquímicas y Farmacéuticas, Universidad Nacional de Rosario and Instituto de Química Rosario (IQUIR-CONICET), Suipacha 531, Rosario (S2002LRK), Argentina

Supporting Information

ABSTRACT: A new expression is developed which allows estimating the sensitivity for the whole family of multivariate calibration algorithms based on partial least-squares regression combined with residual multilinearization. The sensitivity can be employed to compute other relevant figures of merit such as analytical sensitivity, limit of detection, limit of quantitation, and uncertainty in predicted concentration. The results are substantiated by extensive Monte Carlo noise addition simulations for a variety of systems with a different number of analytes and interfering agents, different degrees of overlapping in component profiles, and different numbers of instrumental data modes per sample, all requiring the achievement of the second-order advantage. The connection between the present approach and the intuitive concept of net analyte signal is discussed. An experimental example for which second-, third-, and fourth-order data are available is also studied, concerning the improvement in figures of merit on increasing the data order, which is consistent with the decrease in average prediction error.



Herman Wold and co-workers developed partial least-squares (PLS) for data modeling and regression.[1] Today PLS is a well-known chemometric resource in analytical chemistry, with numerous applications in areas as diverse as food, biomedical, and industrial analysis.[2,3] When PLS is applied to first-order multivariate calibration, i.e., when vector data are measured for each experimental sample (spectra, chromatograms, voltammograms, etc.), a sufficiently large and representative calibration sample set is required for model building, in order to account for the presence of potential interfering agents in new samples. Of considerable interest is the application of PLS in the area of higher-order multivariate calibration, i.e., when data arrays with two or more instrumental modes are measured for each sample, because analyte determination can be performed using calibration sets which are only built with the pure analyte, even when the new specimens contain uncalibrated interfering agents. This is possible thanks to the power of the so-called second-order advantage exhibited by higher-order data.[4−13]

The useful PLS strategy can in principle be applied to higher-order data in two different manners: (1) unfolding the original data into vectors followed by classical PLS regression analysis, giving rise to the U-PLS model,[14] or (2) processing the original data arrays employing a multidimensional PLS version called N-PLS.[15] None of these two strategies alone, however, is able to achieve the important second-order advantage. To reach the latter goal, residual bilinearization (RBL) was developed in 1990 and combined with PLS for second-order data analysis.[16] After an impasse of almost 15 years, RBL was rediscovered,[17] and subsequently applied to a variety of analytical systems, showing a great potentiality due to its intrinsically flexible structural model based on latent variables.[8] Extensions of RBL were then developed for third-order data analysis, i.e., residual trilinearization (RTL),[18] and very recently for fourth-order data studies, i.e., residual quadrilinearization (RQL).[19] They are all members of a family of tools collectively known as residual multilinearization (RML), which have been combined with U-PLS, N-PLS, and other multivariate methods. Applications exist to data of various orders, measured for samples of complex composition and diverse origins, as has been conveniently summarized and reviewed.[5,8−13]

Figures of merit can be reliably estimated in zeroth-order (univariate) and first-order calibration, as documented in IUPAC's Technical Reports.[20,21] Specifically for PLS, many works have focused on the estimation of concentration uncertainties, with emphasis on expression-based approaches.[22−29] In the latter ones, the most relevant figure of merit is the sensitivity, because it is the crucial element for estimating other important parameters such as analytical sensitivity, selectivity, limit of detection, limit of quantitation, and uncertainty in predicted concentrations.[30] Sensitivity may be defined as the change in (net) response for a given change in analyte concentration. While in univariate calibration, the sensitivity is numerically equal to the slope of the calibration curve,[20] in first-order multivariate calibration it is usually defined as the slope of a pseudounivariate calibration based on the so-called net analyte signal (NAS), which is the portion of the total signal uniquely ascribed to the analyte of interest.[31,32]

Received: October 15, 2012
Accepted: November 21, 2012
Published: November 21, 2012

ACS Publications © 2012 American Chemical Society

Table 1. Sensitivity U-PLS/RML Expressions for Data of Increasing Order

Jacobian approach[a]	
general SEN_J expression	$SEN_J = \{\mathbf{v}^T [\mathbf{P}^T (\mathbf{I} - \mathbf{Z}_{int} \mathbf{Z}_{int}^{+}) \mathbf{P}]^{-1} \mathbf{v}\}^{-1/2}$ $\mathbf{Z}_{int} = [\mathbf{Z}_{int1} \mid \mathbf{Z}_{int2} \mid \ldots \mid \mathbf{Z}_{intN}]$
data order	generic interfering agent $\mathbf{Z}_{intn}$ block inside $\mathbf{Z}_{int}$
2	$[\mathbf{I}_c \otimes \mathbf{b}_{intn} \mid \mathbf{c}_{intn} \otimes \mathbf{I}_b]$
3	$[\mathbf{I}_d \otimes \mathbf{c}_{intn} \otimes \mathbf{b}_{intn} \mid \mathbf{d}_{intn} \otimes \mathbf{I}_c \otimes \mathbf{b}_{intn} \mid \mathbf{d}_{intn} \otimes \mathbf{c}_{intn} \otimes \mathbf{I}_b]$
4	$[\mathbf{I}_e \otimes \mathbf{d}_{intn} \otimes \mathbf{c}_{intn} \otimes \mathbf{b}_{intn} \mid \mathbf{e}_{intn} \otimes \mathbf{I}_d \otimes \mathbf{c}_{intn} \otimes \mathbf{b}_{intn} \mid \mathbf{e}_{intn} \otimes \mathbf{d}_{intn} \otimes \mathbf{I}_c \otimes \mathbf{b}_{intn} \mid \mathbf{e}_{intn} \otimes \mathbf{d}_{intn} \otimes \mathbf{c}_{intn} \otimes \mathbf{I}_b]$
net analyte signal approach[b]	
data order	specific expression
2	$SEN_{NAS2} = [\mathbf{v}^T (\mathbf{P}^T \mathbf{P}_C \otimes \mathbf{P}_B \mathbf{P})^{-1} \mathbf{v}]^{-1/2}$
3	$SEN_{NAS3} = [\mathbf{v}^T (\mathbf{P}^T \mathbf{P}_D \otimes \mathbf{P}_C \otimes \mathbf{P}_B \mathbf{P})^{-1} \mathbf{v}]^{-1/2}$
4	$SEN_{NAS4} = [\mathbf{v}^T (\mathbf{P}^T \mathbf{P}_E \otimes \mathbf{P}_D \otimes \mathbf{P}_C \otimes \mathbf{P}_B \mathbf{P})^{-1} \mathbf{v}]^{-1/2}$

[a]The subscripts "int1", "int2", "int*n*", "int*N*" are the numbering for the sources of interference, with profiles in the different data modes obtained during RML as **b**, **c**, **d**, **e** depending on the data order, and $\mathbf{I}_b$, $\mathbf{I}_c$, $\mathbf{I}_d$ and $\mathbf{I}_e$ are unit matrices of size $J \times J$, $K \times K$, $L \times L$, and $M \times M$, respectively. Notice that the **P** matrix in the general SEN_J expression depends on the data order and is of size $JK \times A$, $JKL \times A$, and $JKLM \times A$ for second-, third-, and fourth-order data, respectively (see text). [b]$\mathbf{P}_B = \mathbf{I} - \mathbf{B}_{int} \mathbf{B}_{int}^{+}$, with the loading matrix $\mathbf{B}_{int}$ containing the profiles for the interfering agent sources in one of the data modes. The definitions of $\mathbf{P}_C$, $\mathbf{P}_D$, and $\mathbf{P}_E$ are analogous to that for $\mathbf{P}_B$.

In second-order calibration, different NAS definitions were developed, particularly in the framework of the parallel factor model (PARAFAC).[33,34] These initial alternative definitions were shown to be special cases of a general mathematical expression, revealing the difficulties of the NAS concept in the second-order scenario.[35] Extending the NAS approach to third-order data analysis and beyond has been even more troublesome,[36] although an improved closed-form PARAFAC sensitivity expression has been recently developed.[37] In the latter case, a new approach to concentration uncertainties was employed, which did not involve, at least explicitly, NAS-based arguments.

In regards to the useful PLS/RML methodologies, a provisional sensitivity expression is only known in the case of second-order PLS/RBL,[17] based on the rather conflicting second-order NAS concept (see below). No expressions for the remaining PLS/RBL figures of merit are known. Moreover, virtually no information is available in the case of the third- and fourth-order extensions PLS/RTL and PLS/RQL and neither for the unfolded or multidimensional PLS versions. In the present work, closed-form expressions are presented for estimating several important figures of merit for the complete family of PLS/RML calibration methods. The purpose is 2-fold: on one hand, to provide analytical chemists with the full battery of figures of merit for reporting meaningful results derived from higher-order PLS data analysis, and on the other, to derive a single mathematical expression for the sensitivity, applicable to all PLS/RML methodologies, with a link to the intuitively useful concept of net analyte signal.

An appropriate experimental example for which second-, third-, and fourth-order data are available is employed to illustrate a real application of the developed expressions and to demonstrate the improvement in figures of merit which is gained on increasing the data order.

THEORY

Sensitivity in U-PLS/RBL. Details on the specific implementation of U-PLS/RBL are provided in the Supporting Information and in the relevant literature.[17] In the present subsection, focus is directed toward the estimation of the sensitivity parameter, with the expressions for additional remaining figures of merit provided below.

We take as an example the specific case of second-order data when a single interfering agent occurs. The relevant U-PLS/RBL expression for modeling the test sample data can be written as

$$\mathbf{x} = \mathbf{P}\mathbf{t} + \mathbf{c}_{int\,1} \otimes \mathbf{b}_{int\,1} + \mathbf{e} \quad (1)$$

where **x** is the vector of unfolded test sample signals, **P** is the matrix of calibration U-PLS loadings, **t** is the vector of test sample scores, $\mathbf{b}_{int1}$ and $\mathbf{c}_{int1}$ are vectors containing the profiles representing the contribution of the single interfering agent in both data modes as obtained by principal component analysis (PCA) of residuals (see the Supporting Information), $\otimes$ is the Kronecker product operator, and **e** is an error term. To fit **x** to the model of eq 1, **P** is kept fixed at the calibration values, while RBL provides the sample score vector **t** and interfering agent profiles $\mathbf{b}_{int1}$ and $\mathbf{c}_{int1}$ by least-squares minimization of **e** (see the Supporting Information). If the data matrices for each sample are of size $J \times K$ (J and K are the number of instrumental channels in both data modes), **P** is of size $JK \times A$ (A is the number of latent variables employed to model the calibration data), and $\mathbf{b}_{int1}$ and $\mathbf{c}_{int1}$ are of size $J \times 1$ and $K \times 1$, respectively. Once accomplished the goal of RBL, analyte prediction proceeds through the usual expression:

$$y = \mathbf{v}^T\mathbf{t} \quad (2)$$

where **v** is the vector of PLS regression coefficients in latent variable space, as obtained during the calibration phase (if data are mean-centered, the mean calibration concentration should be added to the right-hand side of eq 2).

If calibration is precise, as usually assumed when estimating the sensitivity for U-PLS/RBL,[35] it is apparent from eq 2 that the variance in concentration is given by

$$\mathrm{var}(y) = \mathbf{v}^T\mathbf{V}_t\mathbf{v} \quad (3)$$

where $\mathbf{V}_t$ is the variance-covariance matrix for the elements of the **t** vector. Using a previously discussed approach based on the computation of the Jacobian matrix associated to the fitted parameters in eq 1, var(y) can be expressed as a function of the instrumental uncertainty var(x), the calibration parameters and the interfering agent profiles as follows (see the Supporting Information):

dx.doi.org/10.1021/ac302996d | Anal. Chem. 2012, 84, 10823−10830

$$\mathrm{var}(y) = \mathrm{var}(x)\mathbf{v}^{\mathrm{T}}[\mathbf{P}^{\mathrm{T}}(\mathbf{I} - \mathbf{Z}_{\mathrm{int}}\mathbf{Z}_{\mathrm{int}}{}^{+})\mathbf{P}]^{-1}\mathbf{v} \quad (4)$$

where $\mathbf{I}$ is a $JK \times JK$ unit matrix, and $\mathbf{Z}_{\mathrm{int}}$ contains information regarding the interfering agent in the following form:

$$\mathbf{Z}_{\mathrm{int}} = [\mathbf{I}_{\mathrm{c}} \otimes \mathbf{b}_{\mathrm{int}\,1} | \mathbf{c}_{\mathrm{int}\,1} \otimes \mathbf{I}_{\mathrm{b}}] \quad (5)$$

where $\mathbf{I}_{\mathrm{c}}$ and $\mathbf{I}_{\mathrm{b}}$ are $J \times J$ and $K \times K$ unit matrices, respectively. From eq 4, the sensitivity can be deduced as the ratio of uncertainties in signal and concentration:[37]

$$\mathrm{SEN}_{\mathrm{J}} = [\mathrm{var}(x)/\mathrm{var}(y)]^{1/2} = \{\mathbf{v}^{\mathrm{T}}[\mathbf{P}^{\mathrm{T}}(\mathbf{I} - \mathbf{Z}_{\mathrm{int}}\mathbf{Z}_{\mathrm{int}}{}^{+})\mathbf{P}]^{-1}\mathbf{v}\}^{-1/2} \quad (6)$$

where the subscript "J" stands for the Jacobian approach. Generalization to more interfering agents needs only the appropriate expansion of $\mathbf{Z}_{\mathrm{int}}$, as shown in Table 1.

An alternative approach to U-PLS/RBL sensitivity has been previously developed based on the concept of net analyte signal.[17] This involves two projection matrices, orthogonal to the spaces spanned by the interfering agents in each of the data modes, i.e., $\mathbf{P}_{\mathrm{B}} = \mathbf{I} - \mathbf{B}_{\mathrm{int}}\,\mathbf{B}_{\mathrm{int}}{}^{+}$ and $\mathbf{P}_{\mathrm{C}} = \mathbf{I} - \mathbf{C}_{\mathrm{int}}\,\mathbf{C}_{\mathrm{int}}{}^{+}$, with the columns of $\mathbf{B}_{\mathrm{int}}$ and $\mathbf{C}_{\mathrm{int}}$ collecting the profiles of the various sources of interference detected by RBL in each of the data modes. The matrices $\mathbf{P}_{\mathrm{B}}$ and $\mathbf{P}_{\mathrm{C}}$ served to remove the matrix signal contributed by the interfering agents, leading to the following expression:[17]

$$\mathrm{SEN}_{\mathrm{NAS2}} = [\mathbf{v}^{\mathrm{T}}(\mathbf{P}^{\mathrm{T}}\mathbf{P}_{\mathrm{C}} \otimes \mathbf{P}_{\mathrm{B}}\mathbf{P})^{-1}\mathbf{v}]^{-1/2} \quad (7)$$

where "NAS2" indicates a NAS-based expression for second-order data. Although eq 7 has been supported by Monte Carlo noise addition simulations,[17] and the numerical $\mathrm{SEN}_{\mathrm{NAS2}}$ values are identical to those of $\mathrm{SEN}_{\mathrm{J}}$, when it comes to extending the idea to third- and fourth-order data, this intuitive NAS approach severely fails (see below).

Sensitivity in U-PLS/RTL and U-PLS/RQL. The extension of both the Jacobian and NAS sensitivity approaches to third-order data and beyond is straightforward. In the case of the former methodology, the basic eq 6 can be applied to higher-order cases after suitably adapting the matrix $\mathbf{Z}_{\mathrm{int}}$ (see the Supporting Information). The complete family of $\mathbf{Z}_{\mathrm{int}}$ matrices is shown in Table 1 for various data orders and interfering agent sources. Notice however that the $A \times 1$ $\mathbf{v}$ vector in eq 6 stems from a U-PLS calibration model built for a specific calibration data set and data order as well as the U-PLS calibration matrix $\mathbf{P}$, which is of size $JK \times A$, $JKL \times A$, and $JKLM \times A$ for second-, third-, and fourth-order data, respectively.

On the other hand, the NAS-based eq 7 admits corresponding extensions to third- and fourth-order data ($\mathrm{SEN}_{\mathrm{NAS3}}$ and $\mathrm{SEN}_{\mathrm{NAS4}}$, respectively), which are also provided in Table 1. They are based on the idea of removing the contributions of the interfering agents by orthogonal projection matrices, in the same manner as for the derivation of eq 7. However, the $\mathrm{SEN}_{\mathrm{NAS3}}$ and $\mathrm{SEN}_{\mathrm{NAS4}}$ expressions seriously underestimate the U-PLS/RTL and U-PLS/RQL sensitivities, as shown in the Supporting Information. This should not be taken as indicative that the concept of NAS is not useful in the field of higher-order multivariate calibration but rather that it should be modified to make it compatible with the Jacobian approach, because the latter is consistent with the Monte Carlo simulations. This interesting issue calls for additional theoretical and experimental research.

Sensitivity in N-PLS/RML. The expressions discussed above for U-PLS/RML can equally be applied to N-PLS/RML, noting that in N-PLS a $\mathbf{v}$ vector exists as in eqs 2 and 6. Likewise, an analogous $\mathbf{P}$ matrix of eqs 4 and 6 can be defined from the different weight matrices provided by N-PLS in each data mode (see the Supporting Information).[15]

Other Figures of Merit. The basic assumption throughout this work is that the standard error in the predicted analyte concentration by a PLS model [$\mathrm{SD}(y)$] is given by the well-known expression:[21,22,26,27,38,39]

$$\mathrm{SD}(y) = [\mathrm{var}(y)]^{1/2} = [\mathrm{SEN}^{-2}\,\mathrm{var}(x) + h\,\mathrm{SEN}^{-2}\,\mathrm{var}(x) + h\,\mathrm{var}(y_{\mathrm{cal}})]^{1/2} \quad (8)$$

where SEN is the sensitivity, $\mathrm{var}(x)$ the variance in instrumental signals, h the sample leverage, and $\mathrm{var}(y_{\mathrm{cal}})$ the variance in calibration concentrations. Details on the different parameters appearing in eq 8 are given below.

The three terms in the right-hand side of eq 8 account for the propagation of uncertainties derived from (in the order in which they appear): (1) instrumental signals in the test sample data, (2) instrumental signals in the calibration data, and (3) calibration concentrations. The first and probably the most relevant of these contributions is transmitted directly *via* the inverse squared sensitivity, which is the key ingredient in eq 8 and whose computation is rather involved in the field of higher-order data, as explained in the next subsection. The second and third terms arise from calibration uncertainties and are both scaled by the sample leverage, a dimensionless parameter measuring the position of the sample in the calibration space. The latter can be expressed in terms of concentrations (h_{c}), instrumental variables (h_{X}), or latent variables (h_{L}). An appropriate expression for h_{L} is[22]

$$h_{\mathrm{L}} = \mathbf{t}^{\mathrm{T}}(\mathbf{T}^{\mathrm{T}}\mathbf{T})^{-1}\mathbf{t} \quad (9)$$

where $\mathbf{T}$ is the matrix of PLS calibration scores and $\mathbf{t}$ the vector of test sample scores. Equation 9 is remindful of the classical least-squares counterpart $h_{\mathrm{c}} = \mathbf{y}^{\mathrm{T}}\,(\mathbf{Y}^{\mathrm{T}}\,\mathbf{Y})^{-1}\,\mathbf{y}$, where $\mathbf{Y}$ is the matrix of calibration concentrations for all analytes and $\mathbf{y}$ the vector of test sample concentrations.[40] Recall that if data are mean-centered, then $(1/I_{\mathrm{cal}})$ should be added to the leverage in eq 9, where I_{cal} is the number calibration samples. Alternatively, eq 8 can be modified by replacing h by $(h + 1/I_{\mathrm{cal}})$ when mean-centering is applied.

Other important figures of merit are the analytical sensitivity γ, the limit of detection (LOD), and the limit of quantitation (LOQ):[21]

$$\gamma = \mathrm{SEN}/[\mathrm{var}(x)]^{1/2} \quad (10)$$

$$\mathrm{LOD} = 3.3\,\mathrm{SD}(y_0) \quad (11)$$

$$\mathrm{LOQ} = 10\,\mathrm{SD}(y_0) \quad (12)$$

where the factor 3.3 corresponds to 5% for the so-called errors of types I and II, and y_0 is the concentration for a blank sample or a sample containing a low analyte concentration.

Software. All calculations were implemented with MATLAB[41] routines, available from the authors by request.

SIMULATED DATA

Data Sets. The first step in the simulations consists in creating various synthetic data sets. For each synthetic sample, a data array was built with the following dimensions: $J \times K$ for second-order data, $J \times K \times L$ for third-order data, and $J \times K \times L \times M$ for fourth-order data, where J, K, L, and M are the number of data points (or sensors) in each data mode. The number of samples to be submitted to PLS/RML analysis is in all cases $I = I_{cal} + 1$, i.e., the number of calibration samples plus the test sample.

To represent the instrumental measurements, noiseless Gaussian-shaped profiles for four different constituents (1, 2, 3, and 4) were defined in each data mode. The profiles correspond to the components at unit concentration and were normalized so that the area under the profile (the total signal for each pure constituent) is one (Figure 1). These simulated profiles spanned the following numbers of data points in the different modes: $J = 20$, $K = 15$, $L = 10$, and $M = 10$. In all cases, the peak maxima for the Gaussian profiles of constituent **1** (the analyte of interest) were fixed at the center of each of the data ranges. The peak maxima for the remaining constituents were placed in all data modes at 10 different random positions, giving rise to 10 different degrees of spectral overlapping. Figure 1 shows a particular situation for the four possible constituents.

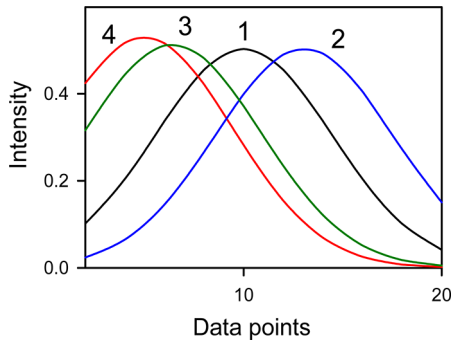


Figure 1. Representative noiseless component profiles employed to build the simulated data sets in the first mode of second-order systems. The black line identifies the analyte of interest; the remaining colors correspond to other sample components. Profiles and relative overlapping in the remaining data modes and orders are similar to those presently shown.

In all the simulated data sets, calibration sets of samples were created with analyte concentrations taken at random and uniformly distributed in the range 0−1, having the following number of samples: 10 for a single calibrated analyte, 20 for two analytes, and 30 for three analytes. Four test samples were also produced, having component concentrations taken randomly from the range 0−1 and analyzed by joining them, each at a time, with the set of calibration samples.

The data sets were identified according to the total number of components (B for binary, T for ternary, and Q for quaternary) and with three numbers identifying the data order, the number of analytes, and the number of interfering agents. The complete list of studied simulated systems is shown in Table 2. Overall, 960 different systems were analyzed, corresponding to 3 different data orders, 6 systems with varying number of analytes and interfering agents, 4 different component concentrations, and 10 different sensitivities according to spectral overlapping.

Table 2. Simulated Systems, Nomenclature, Data Order, and Component Numbering

system[a]	data order	component(s) in calibration set	interfering agent(s)
B2_11	2	1	2
T2_12	2	1	2 and 3
T2_21	2	1 and 2	3
Q2_13	2	1	2, 3, and 4
Q2_22	2	1 and 2	3 and 4
Q2_31	2	1, 2, and 3	4
B3_11	3	1	2
T3_12	3	1	2 and 3
T3_21	3	1 and 2	3
Q3_13	3	1	2, 3, and 4
Q3_22	3	1 and 2	3 and 4
Q3_31	3	1, 2, and 3	4
B4_11	4	1	2
T4_12	4	1	2 and 3
T4_21	4	1 and 2	3
Q4_13	4	1	2, 3, and 4
Q4_22	4	1 and 2	3 and 4
Q4_31	4	1, 2, and 3	4

[a]The first letter identifies the total number of components (B, binary; T, ternary; Q, quaternary), the first number the data order, and the final two numbers the number of calibrated analytes and number of interfering agents, respectively.

Noise Addition. In Monte Carlo studies, uncertainty can in principle be added in four different manners: (1) in calibration concentrations only, (2) in calibration signals only, (3) in test sample signals only, and (4) in all concentrations and signals. When focusing on the sensitivity parameter, only the test sample is considered to carry instrumental noise, in order to leave the concentration uncertainty as only depending on the sensitivity and the signal noise (see eq 8).[35−37] In the present case, all four possibilities were considered, which provides the opportunity of testing the adequacy of eq 8 for estimating the prediction uncertainty.

In each of the synthetic data sets, the value of the signal uncertainty, i.e., the value of $[\mathrm{var}(x)]^{1/2}$ was 0.002 units, and the value of $[\mathrm{var}(y_{cal})]^{1/2}$ was 0.001 units. They were selected so that the relative impact of these values of instrumental and concentration uncertainty is comparable. After creating each data set, noise was added in the different manners discussed above, and for each data set and Monte Carlo cycle, the calibration data were joined with each test sample data and submitted to PLS/RML. The calibration/prediction process followed the usual steps, already described in the literature and summarized in the Supporting Information for second-, third-, and fourth-order data sets. It was repeated 1 000 times using different random seeds for the signal and/or concentration uncertainty, depending on the manner in which noise was added to the synthetic data. Statistical analysis provides the variance in the estimated concentration of the analyte of interest (constituent **1** in all cases).

EXPERIMENTAL DATA

Equipment. Excitation−emission (EEM) fluorescence matrixes were recorded on a fast-scanning Varian Cary Eclipse spectrofluorometer (Melbourne, Australia) equipped with two Czerny-Turner monochromators, a xenon flash lamp, a quartz cell, and connected to a PC microcomputer via an IEEE 488 (GPIB) serial interface. Instrumental parameters were

excitation and emission slit widths, 5 nm; detector voltage, 600 V; scanning speed, 12 000 nm/min; cell temperature, 35 °C; excitation range, 244−312 nm each 4 nm; emission range, 311−491 nm each 2 nm; time range, 2−13.2 min each 0.8 min; pH values, 9.5, 9.8, 10.0, 10.2, and 10.8. Each sample gave an array of size 18 × 91 × 15 × 5. For further details see ref 19.

Calibration and Test Samples. Five calibration samples were prepared as aqueous solutions with the analyte carbaryl in the range 50−250 μg L^{-1} using borate buffer to adjust the pH. Nine test samples were also prepared, containing the analyte in the range 100−250 μg L^{-1} and fuberidazole or thiabendazole as interfering agents (samples **1**−**5**, thiabendazole 25 μg L^{-1}, samples **6**−**9**, fuberidazole 125 μg L^{-1}).

Experimental Data Sets. The original fourth-order data (arrays with four modes per sample: excitation, emission, time, and pH) were employed to determine the analyte in the presence of interfering agents using U-PLS/RQL. From these data, arrays with three modes per sample (excitation, emission, and time) were taken at the intermediate pH value of 10.0 and subjected to U-PLS/RTL analysis. Finally, excitation−emission data matrices were selected from the complete fourth-order data set at intermediate values of pH (10.0) and time (7.6 min) and were processed using U-PLS/RBL.

RESULTS AND DISCUSSION

Simulated Data. Results will only be discussed for U-PLS/RML models, since those for N-PLS/RML were similar to the former ones. For each simulated data system, the calibration data were employed to build a U-PLS model, focusing on the analyte of interest (component **1** in all cases). Depending on the number of data modes per sample, postcalibration RBL, RTL, or RQL was applied to obtain suitable sample scores for analyte prediction in each test sample. Repeating the calibration/prediction procedure, a number of times, using different seeds for the noise added to the data, allowed one to obtain the uncertainty in predicted concentration. This was done, as discussed above, in four different situations, including noise in calibration concentrations, in calibration signals, in test sample signals, and in all of them. This allowed one to check the performance of eq 8 and also of its three separate terms in the presence of the three different uncertainty sources.

Because of the large number of studied systems, a convenient way to summarize the results is by plotting the Monte Carlo concentration uncertainties vs those estimated with eq 8, identifying the different uncertainty sources using specific symbols. Figure 2A corresponds to all second-order systems, Figure 2B to all third-order systems and Figure 2C to all fourth-order systems (see captions to Figure 2 and Table 3). Each of these plots provides information on 960 different systems: 10 different overlapping situations, 4 different test samples, 6 different number of analytes and interfering agents, and 4 different uncertainty sources. It should be noticed that all systems required the second-order advantage for successful analyte quantitation, since in all cases uncalibrated interferences occurred in the test samples.

The overall results presented in Figure 2A−C suggest that the presently described approach to U-PLS/RML sensitivities and concentration uncertainties is appropriate, since the Monte Carlo uncertainties reasonably match those computed from eq 8, after inserting a suitable sensitivity parameter calculated through the relevant eq 6. This also means that the expression shown in Table 1 under the heading Jacobian approach, i.e., eq 6, is the correct formula for estimating the sensitivity in these

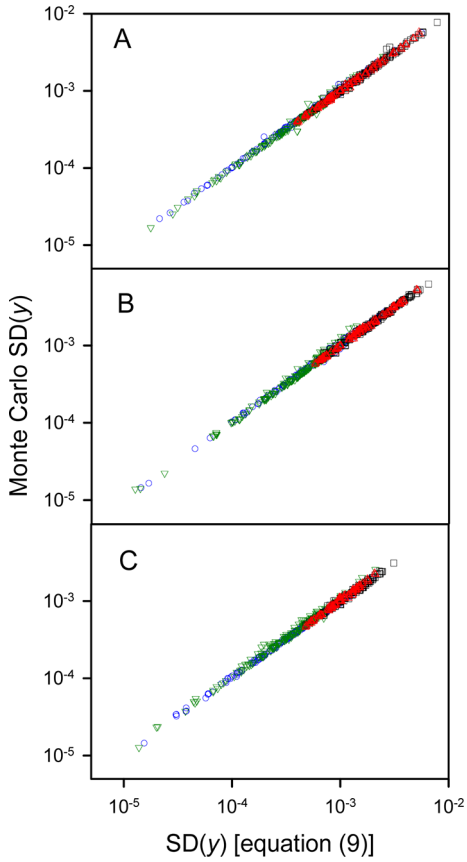


Figure 2. Plot of uncertainties in predicted concentration after Monte Carlo noise addition, as a function of estimations based on eq 8. (A) Results for all second-order data systems (B2_11, T2_12, T2_21, Q2_13, Q2_22, and Q2_31, see Table 2 for explanation of symbols). (B) Results for all third-order data systems (B3_11, T3_12, T3_21, Q3_13, Q3_22, and Q3_31). (C) Results for all fourth-order data systems (B4_11, T4_12, T4_21, Q4_13, Q4_22, and Q4_31). In the three plots, the symbols identify the following cases: blue circles, noise only in calibration concentrations; 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.

systems. Visual inspection of Figure 2A−C immediately indicates that the uncertainty transmitted by calibration (blue circles and green down triangles) is smaller than that propagated from the test sample (red up triangles), because the former ones are scaled by the sample leverage.

It is interesting to note that the intuitive net analyte signal approach, which leads to the NAS expressions shown in Table 1, is not adequate, in general, to cover the expected sensitivities (see the Supporting Information). Only in the case of PLS/RBL applied to second-order signals, the Jacobian and the NAS approaches agree, leading to numerically identical SEN values. In a more general framework, it is a strong indication that the intuitive higher-order NAS concept based on the direct removal of the interfering agent signals should be revisited to make it consistent with the correct Jacobian approach.

Experimental Data. The experimental data corresponds to the quantitation of the fluorescent pesticide carbaryl, which hydrolyzes in alkaline media to fluorescent 1-naphthol. The kinetics of the reaction is pH-dependent, providing the opportunity to measure fourth-order excitation−emission fluorescence matrixes as a function of time and pH. The calibration samples only contain the analyte carbaryl, but the

Table 3. Analytical Results and Figures of Merit for the Experimental Example Using U-PLS/RBL, U-PLS/RTL, and U-PLS/RQL

sample	nominal	U-PLS/RBL	U-PLS/RTL	U-PLS/RQL
		Analytical Results[a]		
1	100	86(1.6)	85(0.6)	98(0.5)
2	125	108(1.5)	107(0.5)	120(0.4)
3	150	134(1.5)	136(0.5)	149(0.4)
4	200	172(1.6)	171(0.5)	186(0.4)
5	250	218(1.6)	223(0.6)	245(0.6)
6	100	85(1.6)	85(0.6)	89(0.5)
7	100	91(1.6)	92(0.5)	94(0.5)
8	200	184(1.5)	185(0.5)	197(0.4)
9	250	202(1.6)	253(0.7)	241(0.6)
		Figures of Merit[b]		
RMSEP/μg L^{-1}		25	18	7.3
REP/%		16	12	4.9
SEN/AFU L μg^{-1}		1.3	5.5	12
γ/L μg^{-1}		0.7	3.1	6.7
LOD/μg L^{-1}		5.3	2	1.5
LOQ/μg L^{-1}		16	6	4.5

[a]All concentrations in μg L^{-1}. Standard deviations from eq 8 in parentheses. In all cases, data were processed after mean centering, with one U-PLS latent variable and one RML component. [b]RMSEP = root-mean-square error of prediction, REP = relative error of prediction based on the mean calibration concentration, SEN from eq 6), and γ, LOD, and LOQ from eqs 10−12, $[\mathrm{var}(y_{cal})]^{1/2}$ = 1 μg L^{-1}, $[\mathrm{var}(x)]^{1/2}$ = 2 AFU (arbitrary fluorescence units).

test samples contain, in addition to the analyte, another fluorescent pesticide as an interfering agent (thiabendazole or fuberidazole). Hence the second-order advantage is required for successful analyte determination.

For the analysis of the experimental data, fourth-order data are available, and thus it is possible to explore all the different possibilities provided by second-, third-, and fourth-order calibration. The first step was thus to select, from the complete fourth-order data for each sample, a second-order subarray which would correspond to measuring excitation−emission fluorescence matrixes at fixed values of reaction time and pH. These data were analyzed using U-PLS/RBL. The use of cross-validation for assessing the optimum number of calibration latent variables rendered A = 1. Although two chemical components occur in calibration, since they are mutually correlated because one component is hydrolyzed to yield the second one, a single U-PLS latent variable is understandable. In order to model the presence of the interfering agents in the test signals, on the other hand, RBL required a single component, which is also the expected result in view of the composition of the test samples. Figure 3 shows the emission and excitation profiles which were obtained by RBL in two typical test samples (one containing fuberidazole as the interfering agent and another one containing thiabendazole), in comparison with the known profiles for pure carbaryl (the analyte of interest) and of 1-naphthol (its hydrolysis product). The success in retrieving these profiles is directly tied to the success in achieving the second-order advantage, allowing the RBL procedure to remove the contribution of the interfering agent from the total test signal for each sample. The prediction results are shown in Table 3, along with the corresponding figures of merit estimated using eqs 8−12, which will be compared below with those corresponding to third- and fourth-order data analysis.

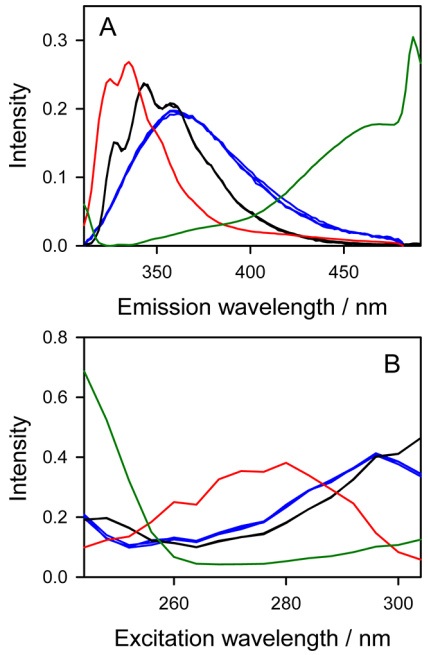


Figure 3. Excitation (A) and emission (B) profiles for the various components of the experimental example. Green and red lines correspond to the experimental spectra for the analyte carbaryl and its hydrolysis product 1-nahphtol, respectively. Blue and black lines (three similar traces) indicate the profiles for the interfering agents fuberidazole and thiabendzaole, as retrieved from the test samples **1** and **6** respectively, by U-PLS/RML analysis of second-, third-, and fourth-order data.

Sensitivity should improve on recording and processing third-order data corresponding to the measurement of the above data matrices as a function of time. This can be studied by selecting, from the fourth-order data for each sample, third-order data at a fixed pH value (10.0). When this data set was submitted to U-PLS calibration, cross-validation suggested again that A = 1 is a reasonable choice, even when two responsive chemical components occur in the calibration set. The result is understandable in view of the mutual correlation of these two components due to the kinetics of the reaction being monitored. For each of the test samples containing interferences, RTL allowed to model the corresponding interfering agent in each sample. A comparison of the excitation and emission profiles is provided in Figure 3, where an excellent match to the results from RBL applied to the second-order data set is observed. The time profile (not shown) retrieved by RTL from the third-order data set is a constant profile (as expected since the interfering agent is pH-stable). The specific prediction results and figures of merit are shown in Table 3 (see below for a full comparison with other data orders).

The complete fourth-order data set was finally submitted to U-PLS/RQL, with similar qualitative results in comparison with the above analyses, but with an additional profile in the pH mode for the interfering agent profiles. The excitation and emission fluorescence profiles retrieved by this algorithm are comparable to those provided by U-PLS/RBL and U-PLS/RTL, as can be seen in Figure 3. The quantitative determination of the analyte carbaryl is reported in Table 3 for comparison with the previous methodologies.

In comparing the relative figures of merit presented in Table 3 for second-, third-, and fourth-order data for the studied experimental system, increasing sensitivities and analytical

dx.doi.org/10.1021/ac302996d | Anal. Chem. 2012, 84, 10823−10830

sensitivities are apparent on increasing the data order. A steady improvement in the average concentration error indicators (root-mean-square error (RMSE) and relative error of prediction (REP)) is also observed as well as in uncertainty in predicted concentrations and detecting capabilities (LOD and LOQ). However, the improvement in SD, LOD, and LOQ is not directly proportional to the gain in sensitivity. This is expected on inspection of eq 8, where two of the three terms are directly affected by the sensitivity parameter. The last term, however, depends on the sample leverage and on the uncertainty in calibration concentrations but not on the sensitivity. This implies a milder effect at high sensitivities, because the uncertainty may be mainly controlled by the calibration concentration uncertainty, which is constant across all data orders. Figure 4 compares the LOD values with the rough approximation which ignores the sample leverage and considers only the first term in eq 8 in estimating SD(y_0), i.e.,

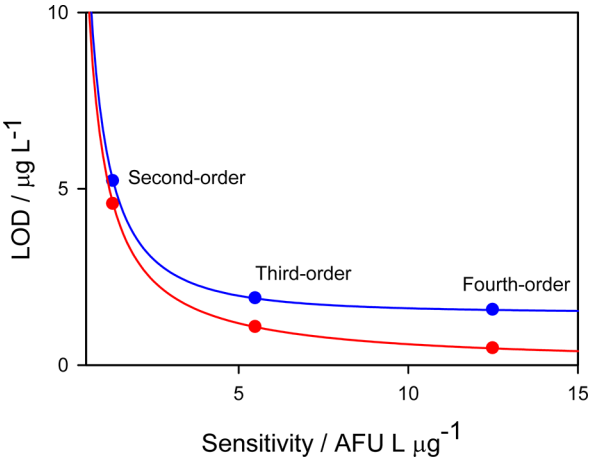


Figure 4. Limits of detection for the determination of the analyte carbaryl in the experimental example. Red circles, LOD values as obtained with U-PLS/RML for the different experimental data orders, from the approximate expression LOD = 3.3 $[\mathrm{var}(x)]^{1/2}/\mathrm{SEN_J}$, with the solid red line showing the continuous variation of LOD as a function of $\mathrm{SEN_J}$. Blue circles, LOD values from the IUPAC's recommended expression (12), inserting in the equation for SD(y_0) the value of h corresponding to the test sample **1**: for second-order data, h = 0.23, for third-order data, h = 0.24 and for fourth-order data, h = 0.25. The solid blue line corresponds to eq 11 assuming h = 0.24. AFU = arbitrary fluorescence units.

$$\mathrm{LOD} = 3.3[\mathrm{var}(x)]^{1/2}/\mathrm{SEN} \quad (13)$$

It is apparent that this latter approximation seriously overestimates the detection capability (Figure 4), while the values shown in Table 3, based on the complete eq 8, provide a more realistic estimation.

CONCLUSIONS

The new expressions derived in this report for second-, third-, and fourth-order multivariate calibration using partial least-squares regression with residual multilinearization provide valuable information for advanced data processing users by (1) allowing important figures of merit to be estimated and reported for multivariate calibration of all data orders, (2) offering new insights into the intuitively useful concept of net analyte signal, challenging the traditional definition and triggering further research on this subject, and (3) paving the way to a future sensitivity expression applicable to all multivariate algorithms.

ASSOCIATED CONTENT

Supporting Information

Additional information as noted in text. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*E-mail: olivieri@iquir-conicet.gov.ar.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

Universidad Nacional de Rosario, CONICET (Consejo Nacional de Investigaciones Científicas y Técnicas, Project No. PIP 1950) and ANPCyT (Agencia Nacional de Promoción Científica y Tecnológica, Project No. PICT-2010-0084) are gratefully acknowledged for financial support. F.A. thanks CONICET for a doctoral fellowship.

REFERENCES

(1) Wold, S.; Hellberg, S. T. L.; Sjostrom, M.; Wold, H. PLS modeling with latent variables in two or more dimensions. In *International Symposium on PLS Model Building: Theory and Applications*, Frankfurt am Main, Germany, September 23−25, 1987.
(2) Wold, S.; Sjöström, M.; Eriksson, L. *Chemom. Intell. Lab. Syst.* **2001**, *58*, 109−130.
(3) Wold, S.; Trygg, J.; Berglund, A.; Antti, H. *Chemom. Intell. Lab. Syst.* **2001**, *58*, 131−150.
(4) Booksh, K. S.; Kowalski, B. R. *Anal. Chem.* **1994**, *66*, 782A−791A.
(5) Escandar, G. M.; Faber, N. M.; Goicoechea, H. C.; Muñoz de la Peña, A.; Olivieri, A. C.; Poppi, R. J. *Trends Anal. Chem.* **2007**, *26*, 752−765.
(6) Escandar, G. M.; Damiani, P. C.; Goicoechea, H. C.; Olivieri, A. C. *Microchem. J.* **2006**, *82*, 29−42.
(7) Olivieri, A. C. *Anal. Chem.* **2008**, *80*, 5713−5720.
(8) Olivieri, A. C.; Escandar, G. M.; Muñoz de la Peña, A. *Trends Anal. Chem.* **2011**, *30*, 607−617.
(9) Bro, R. *Crit. Rev. Anal. Chem.* **2006**, *36*, 279−293.
(10) Gómez, V.; Callao, M. P. *Anal. Chim. Acta* **2008**, *627*, 169−183.
(11) Ni, Y.; Gu, Y.; Kokot, S. *Anal. Lett.* **2012**, *45*, 933−948.
(12) Arancibia, J. A.; Damiani, P. C.; Escandar, G. M.; Ibañez, G. A.; Olivieri, A. C. *J. Chromatogr., B* **2012**, *910*, 22−30.
(13) Olivieri, A. C. *Anal. Meth.* **2012**, *4*, 1876−1886.
(14) Wold, S.; Geladi, P.; Esbensen, K.; Öhman, J. *J. Chemom.* **1987**, *1*, 41−56.
(15) Bro, R. *J. Chemom.* **1996**, *10*, 47−61.
(16) Öhman, J.; Geladi, P.; Wold, S. *J. Chemom.* **1990**, *4*, 79−90.
(17) Olivieri, A. C. *J. Chemom.* **2005**, *19*, 253−265.
(18) Arancibia, J. A.; Olivieri, A. C.; Bohoyo Gil, D.; Muñoz de la Peña, A.; Durán-Merás, I.; Espinosa Mansilla, A. *Chemom. Intell. Lab. Syst.* **2006**, *80*, 77−86.
(19) Maggio, R. M.; Muñoz de la Peña, A.; Olivieri, A. C. *Chemom. Intell. Lab. Syst.* **2011**, *109*, 178−185.
(20) Danzer, K.; Currie, L. A. *Pure Appl. Chem.* **1998**, *70*, 993−1014.
(21) Olivieri, A. C.; Faber, N. M.; Ferré, J.; Boqué, R.; Kalivas, J. H.; Mark, H. *Pure Appl. Chem.* **2006**, *78*, 633−661.
(22) Faber, K.; Kowalski, B. R. *J. Chemom.* **1997**, *11*, 181−238.
(23) Denham, M. C. *J. Chemom.* **1997**, *11*, 39−52.
(24) Faber, N. M. *Chemom. Intell. Lab. Syst.* **2000**, *52*, 123−136.
(25) Serneels, S.; Lemberge, P.; Van Espen, P. J. *J. Chemom.* **2004**, *18*, 76−80.
(26) Fernández Pierna, J. A.; Jin, L.; Wahl, F.; Faber, N. M.; Massart, D. L. *Chemom. Intell. Lab. Syst.* **2003**, *65*, 281−291.

(27) Faber, N. M.; Song, X. -H.; Hopke, P. K. *Trends Anal. Chem.* **2003**, *22*, 330−334.
(28) Phatak, A.; Reilly, P. M.; Penlidis, A. *Anal. Chim. Acta* **1993**, *277*, 495−501.
(29) Berger, A. J.; Feld, M. S. *Appl. Spectrosc.* **1997**, *51*, 725−732.
(30) Olivieri, A. C.; Faber, N. M. Validation and error. In *Comprehensive Chemometrics*; Brown, S., Tauler, R., Walczak, B., Eds.; Elsevier: Amsterdam, The Netherlands, 2009; Vol. 3, pp 91−120.
(31) Lorber, A. *Anal. Chem.* **1986**, *58*, 1167−1172.
(32) Faber, K.; Lorber, A.; Kowalski, B. R. *J. Chemom.* **1997**, *11*, 419−461.
(33) Messick, N. J.; Kalivas, J. H.; Lang, P. M. *Anal. Chem.* **1996**, *68*, 1572−579.
(34) Ho, C. N.; Christian, G. D.; Davidson, E. R. *Anal. Chem.* **1980**, *52*, 1071−1079.
(35) Olivieri, A. C.; Faber, N. M. *J. Chemom.* **2005**, *19*, 583−592.
(36) Olivieri, A. C. *Anal. Chem.* **2005**, *77*, 4936−4946.
(37) Olivieri, A. C.; Faber, N. M. *Anal. Chem.* **2012**, *84*, 186−193.
(38) Faber, N. M.; Bro, R. *Chemom. Intell. Lab. Syst.* **2002**, *61*, 133−149.
(39) Bro, R.; Rinnan, Å; Faber, N. M. *Chemom. Intell. Lab. Syst.* **2005**, *75*, 69−76.
(40) Cabezón, M.; Olivieri, A. C. *Chem. Educator* **2006**, *11*, 394−401.
(41) *MATLAB 7.10*; The MathWorks Inc.: Natick, MA, 2010.