Chemometrics and Intelligent Laboratory Systems ∎ (∎∎∎∎) ∎∎∎–∎∎∎

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Multi-way figures of merit in the presence of heteroscedastic and correlated instrumental noise: Unfolded partial least-squares with residual multi-linearization

Franco Allegrini, Alejandro C. Olivieri *

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

ARTICLE INFO

Article history:
Received 7 March 2016
Received in revised form
19 July 2016
Accepted 5 September 2016

Keywords:
Unfolded partial least-squares
Residual multi-linearization
Figures of merit
Heteroscedastic noise
Correlated noise

ABSTRACT

In the presence of correlated and/or heteroscedastic noise, i.e., for measurement noise which is not independent and identically distributed (iid), new expressions are required to estimate multi-way calibration figures of merit. They are derived in the present report, with focus towards a useful multi-way approach based on unfolded partial least-squares with residual multi-linearization. The expressions allow one to estimate figures of merit under a generalized noise propagation scenario, and to gain insight into the various uncertainty sources contributing to the overall prediction error and limit of detection. Through the study of both simulated and experimental data, it is shown that significant differences exist between the values estimated assuming an iid noise structure and when the underlying structure deviates from this classical paradigm.

1. Introduction

Multi-way calibration is becoming increasingly popular in the chemical analysis of complex samples, particularly for its ability to cope with uncalibrated interferents [1–3]. This leads to considerably simpler calibration strategies, thanks to the achievement of the second-order advantage, which is potentially inherent to data arrays with at least two different instrumental modes [1]. In the framework of multi-way calibration, research on analytical figures of merit (AFOMs) has made considerable progress in recent years, although the usual assumption has been to consider instrumental errors as independently and identically distributed (iid) [4–8].

Multiple causes may lead to instrumental noise structures which deviate from the simple iid condition [9]. Multi-way AFOM expressions which are valid under this general scenario are required, for a variety of reasons: (1) method development and optimization, (2) comparison of different methodologies, (3) uncertainty reporting along with prediction results, and (4) assessment of detection capabilities. Recently, equations were developed for the prediction uncertainty of first-order multivariate calibration in the presence of generalized noise structures [10], extending previous developments in the field [11].

In the context of multi-way calibration, an approximation has been proposed based on the mean square calibration error which can be achieved by processing second-order data [12]. This latter approach assumes that the measurement noise structure is the same both in calibration and prediction. Moreover, it only considers the overall effect of the noise, with no insight into each of the individual error sources. The present report intends to fill the gap between first-order and multi-way calibration AFOMs for generalized noise structures.

We should first consider the sensitivity, a relevant figure of merit affecting all calibration scenarios [13–15]. Recently, a general sensitivity expression has been discussed, which is able to cover from univariate to multi-way data processing [4]. The strategy to derive the general equation involved the study of the propagation of noise from a test sample to the prediction of the analyte concentration. A very small amount of iid noise was added to a test sample signal, to probe the relative magnitude of the propagation, regardless of the experimental noise structure [4]. Thus, it is reasonable to assume that the sensitivity definition will not change, even when the true noise structure is not iid.

On the other hand, other relevant figures of merit such as prediction uncertainty and detection capabilities may be significantly affected by the noise structure. These parameters should always be reported when developing new analytical protocols [11,16,17]. It might be argued that replicate sample analysis could in principle provide an experimental estimation of these figures. However, it is important to be able to dissect the overall

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

http://dx.doi.org/10.1016/j.chemolab.2016.09.001
0169-7439/

uncertainty into the different contributing error sources. This could allow one to identify the influence of specific errors, to limit and/or mitigate them, leading to improved analytical results.

In this work, a general scheme is presented to estimate sample dependent uncertainties in a multi-way calibration model based on unfolded partial least-squares with residual multi-linearization (U-PLS/RML). The latter has been widely employed in recent years to process multi-way data achieving the important second-order advantage [18]. To illustrate the usefulness of the proposed expressions, we describe different situations which depend on the structure of the measurement noise. The adequacy of the results was demonstrated through extensive noise addition simulations, and also by application to experimental data sets.

It is hoped that the present report will stimulate further research concerning the estimation of multi-way analytical figures of merit for generalized noise structures when other data processing algorithms are applied, such as multi-linear decomposition [19] or multivariate curve resolution [20].

2. Theory

2.1. U-PLS/RML

The theory of U-PLS/RML is well-known [18]. In the case of three-way/second-order calibration, data matrices are measured for each experimental sample. The (unfolded) test sample signal $\mathbf{x}$ is modeled as the sum of two contributions: (1) the portion of the test signal modeled by the calibration, and (2) the signal from the interferents modeled by RML:

$$\mathbf{x} = \text{Calibration model of } \mathbf{x} + \text{RML model of } \mathbf{x} + e = \\ = \mathbf{P}\mathbf{t}^{\mathrm{T}} + \sum_{n=1}^{N\,\mathrm{int}} \mathbf{c}_{\mathrm{int},n} \otimes \mathbf{b}_{\mathrm{int},n} + \mathbf{e} \tag{1}$$

where $\mathbf{P}$ is the matrix of U-PLS calibration loadings, $\mathbf{t}$ is the test sample scores, the vectors $\mathbf{b}_{\mathrm{int},n}$ and $\mathbf{c}_{\mathrm{int},n}$ are the profiles in each data mode for the nth interferent, N_{int} is the number of interferents, $\otimes$ indicates the Kronecker product, and $\mathbf{e}$ is a vector of model errors (see Table 1 for details on vector and matrix sizes). In Eq. (1), the product $\mathbf{P}\ \mathbf{t}^{\mathrm{T}}$ represents the part of $\mathbf{x}$ which can be modeled by the calibration parameters, while the summation of Kronecker products represents the contribution from the interferents.

Table 1
Parameter symbols, size and details regarding the variables discussed in the present report.

Parameter	Size	Details
$\boldsymbol{\beta}_{\mathrm{eff}}$	$JK \times 1$	Effective U-PLS regression coefficients
$\mathbf{b}_{\mathrm{int},n}$	$J \times 1$	Profile for interferent in the first mode
$\mathbf{c}_{\mathrm{int},n}$	$K \times 1$	Profile for interferent in the second mode
$\mathbf{e}$	$JK \times 1$	Vector of second order RML residuals
$\mathbf{h}$	$1 \times I$	Sample leverage vector
$\mathbf{I}_J$	$J \times J$	Identity matrix
$\mathbf{I}_K$	$K \times K$	Identity matrix
$\mathbf{I}_{JK}$	$JK \times JK$	Identity matrix
$\mathbf{P}$	$JK \times A$	U-PLS loading matrix
$\mathbf{P}_{\mathrm{eff}}$	$JK \times A$	Effective U-PLS loading matrix
$\mathbf{P}_{\mathrm{Zint}}$	$JK \times JK$	Orthogonal projection matrix to $\mathbf{Z}_{\mathrm{int}}$
$\mathbf{t}$	$1 \times A$	Sample score vector
$\mathbf{T}$	$I \times A$	Calibration score matrix
$\mathbf{X}$	$I \times JK$	Calibration data matrix
$\mathbf{x}$	$JK \times 1$	Test data vector (after unfolding the data matrix)
$\mathbf{v}$	$A \times 1$	Vector of latent U-PLS regression coefficients
$\mathbf{y}_{\mathrm{cal}}$	$I \times 1$	Calibration concentrations
$\mathbf{Z}_{\mathrm{int}}$	$JK \times N_{\mathrm{int}}(J+K)$	Matrix spanning the interferent space
$\boldsymbol{\Sigma}^2_{\mathrm{x}}$	$JK \times JK$	Error covariance matrix for test sample
$\boldsymbol{\Sigma}^2_{\mathrm{X}}$	$JK \times JK$	Error covariance matrix for calibration samples
$\boldsymbol{\Sigma}^2_{\mathrm{X},i}$	$JK \times JK$	Error covariance matrix for calibration sample i
$\boldsymbol{\Sigma}^2_{\hat{\mathrm{X}},\mathrm{eff}}$	$JK \times JK$	Effective error covariance matrix for calibration

The aim of the RML procedure is to find the score vector $\mathbf{t}$ minimizing the norm of the vector $\mathbf{e}$ in Eq. (1), rendering at the same time the interferent profiles in each data mode. Once $\mathbf{t}$ is found by RML, prediction of the analyte concentration $\hat{y}$ proceeds through:

$$\hat{y} = \mathbf{t}\,\mathbf{v} = \mathbf{t}\ \mathbf{T}^{+}\mathbf{y}_{\mathrm{cal}} \tag{2}$$

where $\mathbf{v}$ is the vector of latent regression coefficients provided by the U-PLS calibration model, $\mathbf{T}$ is the matrix of calibration scores, $\mathbf{y}_{\mathrm{cal}}$ the vector of analyte calibration concentrations and '+' indicates the pseudo-inverse operation. An analogous expression to Eq. (1) holds for higher-order data [18].

2.2. Prediction uncertainty

A general expression for prediction uncertainty using U-PLS/RML is derived in this section. It can be easily extended for further multi-way data systems. In the most general scenario, noise affects both calibration and test sample signals and calibration concentrations, and hence differentiation of Eq. (1) leads to:

$$d\mathbf{x} = d\mathbf{P}\,\mathbf{t}^{\mathrm{T}} + \mathbf{P}\,d\mathbf{t}^{\mathrm{T}} + d\left(\sum_{n=1}^{N\,\mathrm{int}} \mathbf{c}_{\mathrm{int},n} \otimes \mathbf{b}_{\mathrm{int},n}\right) \tag{3}$$

The last term of Eq. (3) can be shown to be the product of a matrix $\mathbf{Z}_{\mathrm{int}}$ representing the space spanned by the interferents and a column vector containing the differentials $d\mathbf{c}_{\mathrm{int},n}$ and $d\mathbf{b}_{\mathrm{int},n}$ (see Appendix), i.e.:

$$d\mathbf{x} = d\mathbf{P}\,\mathbf{t}^{T} + \mathbf{P}d\mathbf{t}^{T} + \mathbf{Z}_{\mathrm{int}}\left[d\mathbf{b}_{\mathrm{int},1};\ d\mathbf{c}_{\mathrm{int},1};\ d\mathbf{b}_{\mathrm{int},2};\ d\mathbf{c}_{\mathrm{int},2};\ \ldots\right] \tag{4}$$

where the usual MATLAB notation ';' is employed to append column vectors on top of each other [21], and $\mathbf{Z}_{\mathrm{int}}$ is given by:

$$\mathbf{Z}_{int} = [\mathbf{c}_{\mathrm{int},1} \otimes \mathbf{I}_J,\ \otimes \mathbf{I}_K \otimes \mathbf{b}_{\mathrm{int},1},\ \mathbf{c}_{\mathrm{int},2} \otimes \mathbf{I}_J,\ \mathbf{I}_K \otimes \mathbf{b}_{\mathrm{int},2},\ \ldots] \tag{5}$$

where the ',' appends matrices adjacent to each other, and $\mathbf{I}_J$ and $\mathbf{I}_K$ are $J \times J$ and $K \times K$ identity matrices respectively. This suggest that the last term in Eq. (4) can be removed by multiplication by a suitable projection matrix, orthogonal to $\mathbf{Z}_{\mathrm{int}}$:

$$\mathbf{P}_{\mathrm{Zint}}d\mathbf{x} = \mathbf{P}_{\mathrm{Zint}}d\mathbf{P}\,\mathbf{t}^{T} + \mathbf{P}_{\mathrm{Zint}}\mathbf{P}d\mathbf{t}^{T} \tag{6}$$

where $\mathbf{P}_{\mathrm{Zint}} = (\mathbf{I}_{JK} - \mathbf{Z}_{\mathrm{int}}\ \mathbf{Z}_{\mathrm{int}}{}^{+})$ and $\mathbf{I}_{JK}$ is an identity matrix of size $JK \times JK$. From Eq. (6):

$$d\mathbf{t} = [d\mathbf{x}^{T} - \mathbf{t}d\mathbf{P}^{T}]\mathbf{P}_{\mathrm{Zint}}\left(\mathbf{P}_{\mathrm{Zint}}\mathbf{P}\right)^{+T} \tag{7}$$

Since $\mathbf{P}^{\mathrm{T}} = \mathbf{T}^{+}\ \mathbf{X}$, where $\mathbf{X}$ is the matrix of calibration (unfolded) signals, differentiation of $\mathbf{P}$ and replacement in Eq. (7) gives:

$$d\mathbf{t} = \left\{ d\mathbf{x}^{T} - \mathbf{t}\left[\mathbf{T}^{+}d\mathbf{X} + d\left(\mathbf{T}^{+}\right)\mathbf{X}\right]\right\}\mathbf{P}_{\mathrm{Zint}}\left(\mathbf{P}_{\mathrm{Zint}}\mathbf{P}\right)^{+T} \tag{8}$$

We now focus attention on the expression for the differential change in predicted concentration, starting from Eq. (2):

$$d\hat{y} = d(\mathbf{t}\mathbf{T}^{+}\mathbf{y}_{\mathrm{cal}}) = \mathbf{t}d(\mathbf{T}^{+})\mathbf{y}_{\mathrm{cal}} + (d\mathbf{t})\mathbf{T}^{+}\mathbf{y}_{\mathrm{cal}} + \mathbf{t}\mathbf{T}^{+}d\mathbf{y}_{\mathrm{cal}} \tag{9}$$

And inserting in the latter equation $d\mathbf{t}$ from Eq. (8):

$$d\hat{y} = \mathbf{t}d(\mathbf{T}^{+})\mathbf{y}_{\mathrm{cal}} + (d\mathbf{x}^{\mathrm{T}})\mathbf{P}_{\mathrm{Zint}}(\mathbf{P}_{\mathrm{Zint}}\mathbf{P})^{+\mathrm{T}}\mathbf{T}^{+}\mathbf{y}_{\mathrm{cal}} \\ - \mathbf{t}\mathbf{T}^{+}d\mathbf{X}\mathbf{P}_{\mathrm{Zint}}(\mathbf{P}_{\mathrm{Zint}}\mathbf{P})^{+\mathrm{T}}\mathbf{T}^{+}\mathbf{y}_{\mathrm{cal}} - \\ -\mathbf{t}d(\mathbf{T}^{+})\mathbf{X}\mathbf{P}_{\mathrm{Zint}}(\mathbf{P}_{\mathrm{Zint}}\mathbf{P})^{+\mathrm{T}}\mathbf{T}^{+}\mathbf{y}_{\mathrm{cal}} + \mathbf{t}\mathbf{T}^{+}d\mathbf{y}_{\mathrm{cal}} \tag{10}$$

In the latter equation, two important changes can be made: (1) the factor $\mathbf{P}_{\mathrm{Zint}}\ (\mathbf{P}_{\mathrm{Zint}}\ \mathbf{P})^{+\mathrm{T}}$ can be condensed as $\mathbf{P}_{\mathrm{eff}}{}^{+\mathrm{T}}$, where

Please cite this article as: F. Allegrini, A.C. Olivieri, Multi-way figures of merit in the presence of heteroscedastic and correlated instrumental noise: Unfolded partial least-squares with residual multi-linearization, Chemometrics and Intelligent Laboratory Systems (2016), http://dx.doi.org/10.1016/j.chemolab.2016.09.001