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

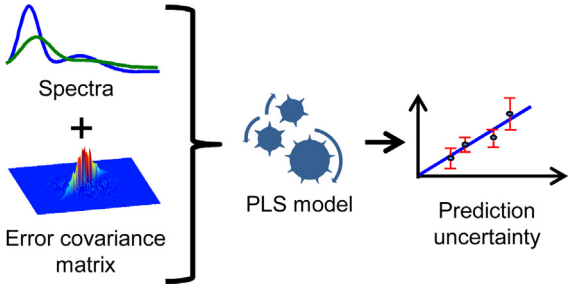
Franco Allegrini [a], Peter D. Wentzell [b, **], Alejandro C. Olivieri [a, *]

[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
0003-2670/

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:

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

where $\hat{y}$ 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