Available online at www.sciencedirect.com

SCIENCE @ DIRECT®

Chemometrics and Intelligent Laboratory Systems 80 (2006) 50 – 56

Chemometrics and intelligent laboratory systems

www.elsevier.com/locate/chemolab

Comparison of variance sources and confidence limits in two PLSR models for determination of the polymorphic purity of carbamazepine

Jez Willian B. Braga, Ronei Jesus Poppi *

Instituto de Química, Universidade Estadual de Campinas, Caixa Postal 6154, 13084-971, Campinas-SP, Brazil

Received 25 August 2004; received in revised form 17 June 2005; accepted 29 June 2005
Available online 21 September 2005

Abstract

This paper presents a study of the variance sources and confidence limits in two PLSR models for the determination of the polymorphic purity of Carbamazepine, using near and mid infrared spectroscopy. The variance sources estimated and compared were reference values, instrumental responses and fit of the model. The variance of instrumental responses was estimated experimentally and theoretically, and the differences were discussed. The confidence limits at three confidence levels: 95%, 90% and 50% were determined, presenting a good agreement with the expected values. The predictive ability of the models was compared, showing that both present the same overall performances with RMSEP of 0.67% for near infrared and 0.62% for mid infrared spectroscopy. It was also verified that, for both models, the main variance source remains the error of the PLSR model.

Keywords: Prediction error; Confidence limits; Uncertainty sources; Partial least squares regression

1. Introduction

The quantitative determination of a property of interest in a chemical system is one of the most frequent practices in analytical chemistry. In the majority of cases, this property is the concentration of a compound present in the system. However, when instrumental methods are employed, concentration is a property that cannot be observed directly, being determined indirectly, through a relationship with a measured physical or chemical property, such as emission or absorption of radiation, conductivity or electrical potential, in a practice called calibration [1]. Different calibration methods are available to achieve this purpose, being classified following the complexity of the data that is treated as zero, first and second order calibration [2].

Independent of the calibration method that is employed, the measurement of the uncertainty present in the predictions using the model developed is an important characteristic. The uncertainty can be defined as a parameter, associated with the result of a measurement, that characterizes the dispersion of the values that could be attributed to the measurand [3,4]. In practice the uncertainty on the result may arise from different sources, such as sampling, environmental conditions, matrix effects and interferences, uncertainties of masses and volumetric equipment, reference values, approximations and assumptions incorporated in the measurement method and procedure, and random variation [5].

The uncertainty of a predicted property of interest for an unknown sample depends on all uncertainty sources involved in the measurement process. Their determination for a predicted value in multivariate calibration methods, such as Partial Least Squares Regression (PLSR), by a sample-specific standard error has been the focus of considerable research in resent years [4,6–17]. The latest contributions and applications have suggested that the approach of errors in variables (EIV), proposed by Faber and Kowalski [6], and simplified by Faber and Bro [7], can be applied for determination of confidence limits in a predicted concentration for an unknown sample. This last simplification was recently implemented in an integrated Matlab toolbox for

* Corresponding author. Tel.: +55 19 37883126; fax: +55 19 37883023.
E-mail address: ronei@iqm.unicamp.br (R.J. Poppi).

0169-7439/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2005.06.004

first-order multivariate calibration [8] and, when the uncertainty in the reference values can be neglected, it leads to the early proposal of Höskuldsson [9] that is already adopted by the American Standards for Testing Materials (ASTM) [18].

The aim of this paper is to compare the uncertainty sources and confidence limits for two PLSR models in the determination of the polymorphic purity of Carbamazepine III in binary mixtures of polymorphs I and III. The models were built by diffuse reflectance spectroscopy in the near (NIR) and mid infrared (MIR) regions. Variance sources from reference values, instrumental responses and fit of the model were investigated for both models and the confidence limits and the prediction errors were determined.

2. Theory

2.1. Regression model

The calibration models described in the next sections were built with the instrumental responses and reference values for the interest property mean centered, thus, in all equations introduced below this procedure will be implicit in the notation.

The standard linear regression model can be expressed as [10,12,17]:

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e} \tag{1}$$

where $\mathbf{e}$ $(I \times 1)$ is the unmodeled part of $\mathbf{y}$ $(I \times 1)$ the true predictand vector, $\mathbf{X}$ $(I \times J)$ is a matrix of instrumental responses (true predictor matrix); $\mathbf{b}$ is the regression vector. Where $\mathbf{e}$ is assumed to be identically and independently distributed (iid), and I and J denote the number of calibration samples and predictor variables (e.g., wave numbers), respectively.

The PLSR model, following SIMPLS formalism [19], is given by:

$$\hat{\mathbf{y}} = \tilde{\mathbf{X}}\hat{\mathbf{b}} \tag{2}$$

where $\hat{\mathbf{y}}$ is a vector with the estimated values for a property of interest; $\tilde{\mathbf{X}}$ is the matrix of measured instrumental responses; and $\hat{\mathbf{b}}$ is a vector with the estimated regression coefficient, calculated as:

$$\hat{\mathbf{b}} = \hat{\mathbf{R}}\hat{\mathbf{T}}^{\mathrm{T}}\tilde{\mathbf{y}} \tag{3}$$

where $\hat{\mathbf{R}}$ and $\hat{\mathbf{T}}$ are the estimated weight and score matrices, $\tilde{\mathbf{y}}$ is the measured reference values and “T” superscript indicates transpose operation.

2.2. Sample-specific standard error of prediction

The approach proposed by Faber and Kowalski [6] is based on the Errors in Variables (EIV) theory. It attempts to account for errors in reference values $\tilde{\mathbf{y}}$ and in instrumental responses $\tilde{\mathbf{X}}$, assuming that the errors are independently and identically distributed (iid) in both calibration and prediction data. After some simplifications made when the models explain a substantial part of the variance in $\tilde{\mathbf{X}}$, the expression to obtain the variance of the prediction errors for the PLSR model ($\hat{V}(y_\mathrm{i} - \hat{y}_\mathrm{i})$) can be expressed as [4]:

$$\hat{V}(y_i - \hat{y}_i) = \left(h_i + \frac{1}{I}\right)(\hat{V}(\mathbf{e}) + \hat{V}(\mathbf{y}) + ||\hat{\mathbf{b}}^2||\hat{V}(\mathbf{X})) + \hat{V}(\mathbf{e_i}) + ||\hat{\mathbf{b}}^2||\hat{V}(\mathbf{x_i}) \tag{4}$$

where $\hat{V}(\mathbf{y})$, $\hat{V}(\mathbf{X})$ and $\hat{V}(\mathbf{x_i})$ are the variances of the errors of the reference method, and of the instrumental responses in the calibration set and the prediction sample, respectively. $\hat{V}(\mathbf{e})$ and $\hat{V}(\mathbf{e_i})$ are the variances of the residuals for the calibration set and prediction samples and h is the leverage, defined as:

$$h_i = \tilde{\mathbf{x}}_\mathbf{i}^{\mathrm{T}}\left(\hat{\mathbf{R}}_A\hat{\mathbf{R}}_A^{\mathrm{T}}\right)\tilde{\mathbf{x}}_\mathbf{i} \tag{5}$$

where “A” is the number of latent variables (LV) used in the model. Eq. (4) is a general expression, which can be simplified. According to Faber and Kowalski [10] the mean square error of the calibration samples (MSEC) will not estimate $\hat{V}(\mathbf{e})$ directly, instead it estimates: $\hat{V}(\mathbf{e}) + \hat{V}(\mathbf{y}) + ||\hat{\mathbf{b}}||^2$ $\hat{V}(\mathbf{X})$. Assuming that there is no significant difference between the variance of the residuals and instrumental responses for the calibration step and in prediction samples, ($\hat{V}(\mathbf{e}) \approx \hat{V}(\mathbf{e_i})$ and $\hat{V}(\mathbf{X}) \approx V(\mathbf{x_i})$), and Eq. (4) reduces to the simplified form proposed by Faber and Bro [7], which can be written as:

$$\hat{V}(y_1 - \hat{y}_i) = \left(1 + h_i + \frac{1}{I}\right)\mathrm{MSEC} - \hat{V}(\mathbf{y}) \tag{6}$$

where MSEC is estimated as:

$$\mathrm{MSEC} = \frac{\sum_{i=1}^{I}(\tilde{y}_i - \hat{y}_i)^2}{I - v} \tag{7}$$

where v denotes the number of degrees of freedom lost in the model built. A representative estimate of v in PLSR models that leads to an unbiased estimate of MSEC is problematic. In a rigorous study, Van der Voet [20] shows that an estimate of v can be achieved for PLSR models by calculation of pseudo-degrees of freedom, using the results from leave-one-out cross-validation.

An estimate of an unbiased MSEC can also be obtained by a K-dimensional PCR model, since no relevant information is left out of the model [10]. The calculation of the optimal number of principal components is facilitated here, since an estimate for $\hat{V}(\mathbf{X})$ is assumed to be available, and it should be closest to the value calculated by:

$$\hat{V}(\mathbf{X}) = \frac{\sum_{k=K+1}^{\min(I,J)} \tilde{\lambda}_{\mathrm{K}}}{(I - K - 1)(J - K)} \tag{8}$$

where $\hat{\lambda}_{\mathrm{K}}$ denotes the K-th eigenvalue of $(\tilde{\mathbf{X}}^{\mathrm{T}}\tilde{\mathbf{X}})$.