Robustness of models developed by multivariate calibration. Part II: The influence of pre-processing methods

M. Zeaiter, J.-M. Roger, V. Bellon-Maurel

Infrared (IR) spectroscopic techniques combined with multivariate calibration (MVC) methods are promising for on-line monitoring. In a previous article [M. Zeaiter, M. Roger, V. Belon-Maurel, D. Rutledge, Trends Anal. Chem. 23 (2004) 157], robustness of the calibration was defined and different ways to evaluate it were identified.

In order to improve the robustness of these calibration methods for industrial applications, an overview is presented of the existing methods, usually used to enhance prediction-model performance. The first part focuses on geometric spectral pre-processing methods, such as normalization methods, smoothing and derivatives. The second part discusses dimensionality-reduction methods, represented by orthogonalization and variable- selection methods. The impact of each method on the enhancement of the robustness of models developed by MVC is analyzed and discussed.

Keywords: Multivariate calibration; NIR spectroscopy; Orthogonal projection; Pre-processing methods; Robustness; Variable selection

Notation: Capital bold characters will be used for matrices (e.g., **X**); small bold characters for column vectors (e.g., $\mathbf{x}_j$ will denote the *j*th column of **X**); row vectors will be denoted by the transpose notation (e.g., $\mathbf{x}^{\mathrm{T}}{}_i$ will denote the *i*th row of **X**); non-bold characters will be used for scalars (e.g., matrix elements x_{ij}). When needed for the purpose of clarity, matrix dimensions are indicated as $\mathbf{X}(n \times p)$, where *n* is the number of lines and *p* the number of columns

Glossary: LMVC, Linear multivariate calibration; ILS, Inverse least squares; CLS, Classical least squares; PLS, Partial least squares; MLR, Multiple linear regression; PCR, Principal components regression; **X**, Matrix of *n* spectra and *p* wavenumbers; **y**, Reference vector of *n* rows; *f*, Multivariate function; **e**, Vector of error; $\hat{\mathbf{y}}$, Estimated reference vector of *n* rows; **b**, Vector of the *p* regression coefficients; $\mathbf{b}_0$, Intercept $\hat{\mathbf{y}} = \mathbf{X}\mathbf{b} + b_0$; $\|\mathbf{x}\|$, The Euclidian norm of **x**, i.e. $(\mathbf{x}^{\mathrm{T}}\mathbf{x})^{\frac{1}{2}}$; $\delta\mathbf{x}$, The vector of influence factors effect; $\delta\mathbf{x}_1$, The vector of systematic variations of the influence factors or structured effect; $\delta\mathbf{x}_2$, The vector of random variations of the influence factors effects or noise effect; PRESS, Predicted residual error sum of squares $\mathrm{PRESS} = \sum(\hat{\mathbf{y}} - \mathbf{y})^2$; SEV, The standard error of validation $\mathrm{SEC}^2 = \frac{1}{n_1 - 1}\|\hat{\mathbf{y}} - \mathbf{y}\|^2$; SEC, The standard error of calibration $\mathrm{SEC}^2 = \frac{1}{n - n_1 - 1}\|\hat{\mathbf{y}} - \mathbf{y}\|^2$; SEP, The standard error of prediction $\mathrm{SEP}^2 = \frac{1}{n}\|\hat{\mathbf{y}} - \mathbf{y}\|^2$; BS, The prediction bias $\mathrm{BS} = \overline{(\hat{\mathbf{y}} - \mathbf{y})}$; **w**, The weight vector; **T**, The scores matrix of **X**; **P**, The loadings matrix of **X**; $\vec{\mathscr{S}}$, The *p*-dimensional space of the spectra; $\vec{\mathscr{C}}$, The spectral space of the parameter of interest; $\vec{\mathscr{N}}$, The space of the rest of the spectral information; **X**, The spectral matrix of $\vec{\mathscr{S}}$; $\mathbf{X}^{+}$, The spectral matrix of $\vec{\mathscr{C}}$; $\mathbf{X}^{-}$, The matrix in $\vec{\mathscr{N}}$; $\mathbf{Z}^{-}$, The matrix of the a basis of $\vec{\mathscr{N}}$; $\vec{\mathscr{E}}$, The space of residuals; **E**, The matrix of residuals in $\vec{\mathscr{E}}$; SNR_j, The signal to noise ratio

M. Zeaiter*, J.-M. Roger, V. Bellon-Maurel
Cemagref of Montpellier
Research Unit of Information
and Technologies for
Agro-Processes (ITAP),
Rue J-F Breton, BP 5095,
F-34196 Montpellier Cedex 1,
France

*Corresponding author.
E-mail: magida_z@yahoo.com,
roger@montpellier.cemagref.fr,
bellon@montpellier.cemagref.fr.

1. Introduction

In order to promote spectroscopic techniques in real-life industrial applications, one must ensure the robustness of the calibration models. In a previous review article, focusing on the definition of robustness and the ways of assessing it, we defined robustness as "the stability of the predictive capacity of the calibration model against perturbations centered on standard conditions" [1]. In this article, we present different pre-processing methods used to improve the calibration model, and discuss their contributions to the robustness improvement of calibration models.

In spectroscopy, the goal of calibration is to replace slow, expensive measurement of the property of interest, *y*, by a spectroscopic one that is cheaper or faster, nevertheless still sufficiently accurate [2]. For IR spectroscopy, MVC is defined as "A process for creating a model, *f*, that relates sample properties, *y*, to the intensities or absorbencies, *x*, at more than one wavelength or frequency of a set of known reference samples" [3].

0165-9936/$ - see front matter © 2005 Elsevier Ltd. All rights reserved. doi:10.1016/j.trac.2004.11.023

Theory indicates that a linear form of the function, *f*, is to be adopted, since the Lambert–Beer's law represents the linear relationship between concentration and absorbance [4–6], so linear MVC (LMVC) models are used, such as inverse least squares (ILS), classical least squares (CLS), multiple linear regression (MLR), principal component regression (PCR) and partial least squares regression (PLSR) [7]. This linear model relating **y** (*n* values of the property of interest) to **X** (*n* spectra) is presented in the following equation:

$$\mathbf{y} = \mathrm{b}_0 + \mathbf{X}\mathbf{b} + \mathbf{e}. \tag{1}$$

In the following, only linear models are considered. LMVC aims at estimating b_0 and **b** (i.e. the regression parameters of the model), and **e** is the matrix of residuals supposedly due to random noise of the zero mean [8,9].

The development of the regression model comprises three stages:

(1) the calibration model is built and validated using a training set $(\mathbf{X}_0, \mathbf{y}_0)$ and a validation set $(\mathbf{X}_1, \mathbf{y}_1)$; the result is an error of validation, SEV, that is used to set up the model;
(2) both $(\mathbf{X}_0, \mathbf{y}_0)$ and $(\mathbf{X}_1, \mathbf{y}_1)$ are used to compute the SEC of the model; and,
(3) an independent test set $(\mathbf{X}_2, \mathbf{y}_2)$ is used to evaluate the model performance with an indicator criterion, namely the error of prediction, SEP.

Mostly, the first and second steps are merged together using the cross-validation technique (e.g., leave one out (LOO) method, contiguous blocks, randomization [10,11] or the bootstrap [12]), so the standard error of calibration (SEC) and the standard error of validation (SEV) are computed simultaneously.

The robustness problem is due to variations in the measurement conditions caused by variations in influence factors that affect the spectral measurement by adding a perturbation, $\delta\mathbf{x}$. This perturbation is represented in the prediction responses of Equation (1) as an error, $\delta\hat{\mathbf{y}}$, such that

$$\delta\hat{\mathbf{y}} = \delta\mathbf{x}^{\mathrm{T}}\mathbf{b}$$

which yields

$$|\delta\hat{\mathbf{y}}| = \|\delta\mathbf{x}\|\,\|\mathbf{b}\|\,|\cos(\delta\mathbf{x}, \mathbf{b})|. \tag{2}$$

To minimize $|\delta\hat{\mathbf{y}}|$, the minimization of one or more of the three terms of the right-hand part of Equation (2) is required.

The first part of this article deals with geometric, spectral data-pre-processing methods. The second part presents the processing methods used to extract from the spectral space the subspace that holds the informative features.

The contribution of these methods to enhancing the robustness of the calibration model is discussed according to Equation (2).

2. Geometric spectral pre-processing methods

Geometric pre-processing methods are widely carried out to correct spectral data from drift in baseline, nonlinearity, curvilinearity, as well as additive and multiplicative effects.

They can be divided into two different categories with respect to the intended corrections:

(1) one corrects for the shifts and the trends in baseline and curvilinearity, and for multiplicative interference, mainly due to scattering; these are the normalization methods; and,
(2) the smoothing used to reduce noise and differentiation to correct peak overlap and constant or linear baseline drift.

2.1. Spectral normalization

The normalization pre-processing method consists of giving the same weight to all absorbencies. Although spectral normalization methods are applied to each individual spectrum, only some of them require the whole data set to compute correction factors. In [4], the following, different methods used for spectral normalization are presented.

2.1.1. The standard normal variate (SNV) transformation

Light scattering due to the interactions between IR radiation and sample particles, often creates a shift of absorbency levels that could be harmful for spectral interpretation and linear calibration of NIR diffuse reflectance spectra. It results in path-length variations that lead to a background signal level that varies with wavelengths. This background effect, responsible for the baseline shift and curvature, may vary greatly between and within samples [13–15].

The SNV transformation was introduced by Barnes et al. [16,17] to reduce multiplicative effects of scattering, particle size and multicolinearity changes over all the NIR spectra. Each spectrum is centered and then scaled by its standard deviation. Its disadvantage remains in the assumption that multiplicative effects are uniform over the whole spectral range, which is not always fulfilled, so artifacts could be introduced by this transformation.

2.1.2. Robust normal variate (RNV) transformation

Guo et al. [18] tackled the artifacts created by SNV transformation by solving the ''Closure'' problem. Closure is the statistical term indicating that the sum of the data is necessarily equal to a certain amount, so that if one of the variables changes in one direction, the other variables must change into the opposite direction in order to compensate for the change and ensure the constancy of the sum. The variable are ''closed'' using SNV because the corrected spectral values are of zero