Analytica Chimica Acta 864 (2015) 1–8

Contents lists available at ScienceDirect

Analytica Chimica Acta

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

Calibration transfer employing univariate correction and robust regression

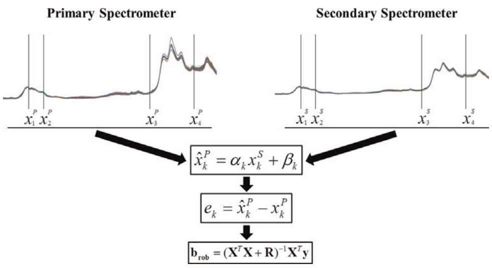
Roberto Kawakami Harrop Galvão[a], Sófacles Figueredo Carreiro Soares[b,c],
Marcelo Nascimento Martins[a], Maria Fernanda Pimentel[d],
Mário César Ugulino Araújo[b,*]

[a] *Instituto Tecnológico de Aeronáutica, Divisão de Engenharia Eletrônica, São José dos Campos,São Paulo 12228-900, Brazil*
[b] *Universidade Federal da Paraíba, CCEN, Departamento de Química, Caixa Postal 5093, João Pessoa, Paraíba CEP 58051-970, Brazil*
[c] *Universidade Federal da Paraíba, CT, Departamento de Engenharia Química, João Pessoa, Paraíba CEP 58051-900, Brazil*
[d] *Universidade Federal de Pernambuco, Departamento de Engenharia Química, Recife, Pernambuco 50740-521, Brazil*

HIGHLIGHTS

- Calibration transfer involving individual wavelengths.
- Suitable for dedicated instruments.
- Examples involving near infrared spectrometric analysis of gasoline and corn.
- Better results compared to piecewise direct standardization (PDS).

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 17 June 2014
Received in revised form 23 September 2014
Accepted 2 October 2014
Available online 8 October 2014

Keywords:
Multivariate calibration
Calibration transfer
Univariate correction
Robust regression
Variable selection

ABSTRACT

This paper proposes a new method for calibration transfer, which was specifically designed to work with isolated variables, rather than the full spectrum or spectral windows. For this purpose, a univariate procedure is initially employed to correct the spectral measurements of the secondary instrument, given a set of transfer samples. A robust regression technique is then used to obtain a model with low sensitivity with respect to the univariate correction residuals. The proposed method is employed in two case studies involving near infrared spectrometric determination of specific mass, research octane number and naphthenes in gasoline, and moisture and oil in corn. In both cases, better calibration transfer results were obtained in comparison with piecewise direct standardization (PDS). The proposed method should be of a particular value for use with application-targeted instruments that monitor only a small set of spectral variables.

1. Introduction

The development of multivariate calibration models involves several stages, typically including the collection of samples and recording of analytical signals, followed by the actual construction and validation of the model. All these stages are important to achieve good predictions when the resulting model is employed in the analysis of new samples. In particular, it would be desirable to eliminate or minimize the sources of data variability that are not related to the analytical property of interest. However, there are cases in which changes in the analytical conditions occur after the calibration has been carried out, with adverse effects on the prediction ability of the model [1,2]. Such changes may refer to the physical/chemical characteristics of the sample (such as viscosity, granularity, surface texture, and presence of interferent species),

---

[*] Corresponding author. Tel.: +55 83 3216 7438; fax: +55 83 3216 7437.
E-mail addresses: laqa@quimica.ufpb.br, mariougulino@gmail.com (M.C.U. Araújo).

http://dx.doi.org/10.1016/j.aca.2014.10.001
0003-2670/

environmental conditions (temperature and humidity, for instance), as well as the response function of the instrument itself. Issues associated to the instrument response typically arise because of general aging effects, deterioration of specific parts or maintenance interventions. Difficulties may also arise if the instrument employed for data acquisition is not the same used for building the calibration model [1].

Such changes can be taken into account by recalibrating the model using a data set acquired under the new analytical conditions. However, this procedure may be expensive, laborious and time-consuming, as the entire model-building process would need to be repeated. Alternatively, different methods have been developed to compensate for changes in the experimental conditions without the need for a full model recalibration. Examples include direct and piecewise direct standardization [3], baseline correction [4], multiplicative signal correction [5], finite impulse response filtering [6], orthogonal signal correction [7], wavelet decompositions [8], and transfer by orthogonal projections [9], among others. In this context, a typical application consists of transferring the calibration from a primary (or "master") instrument to a secondary (or "slave") one.

Within the scope of spectrometric techniques, calibration transfer is usually based on mathematical transformations involving the overall spectrum or windows of variables within the spectrum. For this purpose, a set of N_{trans} "transfer samples" measured at both instruments can be used to build a standardization model of the form

$$\mathbf{X}^P = \mathbf{X}^S F \tag{1}$$

where matrices $\mathbf{X}^P$ $(N_{\text{trans}} \times p)$ and $\mathbf{X}^S$ $(N_{\text{trans}} \times p)$ comprise the spectra of the transfer samples acquired in the primary and secondary instruments, respectively. The transformation matrix $\mathbf{F}$ $(p \times p)$ is typically obtained by principal component regression or partial least squares from the data in $\mathbf{X}^P$ and $\mathbf{X}^S$ [1]. This matrix can be used to standardize a new spectrum $\mathbf{x}^S$ $(1 \times p)$ acquired at the secondary instrument as

$$\hat{\mathbf{x}}^P = \mathbf{x}^S \mathbf{F} \tag{2}$$

in order to generate an estimate $\hat{\mathbf{x}}P$ $(1 \times p)$ of the spectrum $\mathbf{x}^P$ that would be obtained if the measurements were carried out at the primary instrument. Eq. (2) can also be written as

$$\hat{x}_k^P = \mathbf{x}^S \mathbf{f}_k, k = 1, 2, \ldots, p \tag{3}$$

where $\hat{x}_k^P$ denotes the kth variable of the standardized spectrum and $\mathbf{f}_k$ is the kth column of the transformation matrix $\mathbf{F}$.

An example of this calibration transfer procedure is the piecewise direct standardization (PDS) method [10,11], which is often used as a benchmark in comparative studies. In PDS, the standardization model relates each spectral variable in the primary spectrum with a window of variables in the secondary spectrum. In this case, the transformation matrix $\mathbf{F}$ has a sparse structure, because each column $\mathbf{f}_k$ will only have nonzero values within a window around the kth element.

At this point, one may argue that methods such as PDS would not be appropriate for use with application-targeted instruments that monitor only a small set of spectral variables using filters [12,13] and/or light emitting diodes [14,15], for example. In such a case, it would not be possible to derive a standardization model based on spectral windows, because the measurements would be related to isolated wavelengths. Within this scope, this paper proposes a new method for calibration transfer, which was specifically designed to work with isolated variables, rather than spectral windows. For this purpose, a univariate procedure is initially employed to correct the spectral measurements of the secondary instrument, given a set of transfer samples [1]. A robust regression technique is then used to build a new multivariate calibration model with low sensitivity with respect to the residuals of the univariate correction procedure. This new model is specifically designed for use with the corrected spectra of the secondary instrument. The use of robust regression in this context is the main contribution of the present work.

Two data sets are employed to illustrate the proposed method. The first data set consists of gasoline samples with FT-NIR spectra from two spectrometers, as well as reference values of specific mass, RON and naphthenes. Specific mass is routinely used to monitor fuel quality, as well as to check conformity with the standards issued by the regulatory agencies. The other two parameters were included in the study to represent the classes of composition-related and global physical properties, as in [16]. The second data set, which is publicly available, consists of corn samples with NIR spectra from two spectrometers and reference values of moisture and oil.

In both case studies (gasoline and corn), the transfer of calibration is concerned with a subset of spectral variables previously selected using a genetic algorithm (GA) [17]. The results are compared with those obtained using a full-spectrum partial least squares (PLS) model with PDS calibration transfer.

2. Background and theory

2.1. Robust regression

The mathematical development in this section is adapted from the formulation presented in [18]. Initially, let us assume that the property of interest y is to be related to p spectral variables x_1, x_2, . . . , x_p by a linear empirical model of the form $y = \mathbf{xb} + e$, where $\mathbf{x} = [x_1, x_2, \ \ldots, x_p]$, $\mathbf{b} = [b_1, b_2, \ \ldots, b_p]^T$ is a vector of coefficients to be determined, and e denotes the model residual. If a matrix $\mathbf{X}$ $(N \times p)$ of instrumental responses and a vector $\mathbf{y}$ $(N \times 1)$ of reference property values are available for $N \geq p$ calibration samples, the least-squares (LS) estimate for the vector of model coefficients $\mathbf{b} = [b_1, b_2, \ \ldots, b_p]^T$ is the one that minimizes the following cost function:

$$J(\mathbf{b}) = \| \mathbf{Xb} - \mathbf{y} \|^2 \tag{4}$$

which corresponds to the squared 2-norm of the model residuals. The solution is given by

$$\mathbf{b}_{\text{LS}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{5}$$

assuming that the variables were appropriately selected so that $\mathbf{X}^T\mathbf{X}$ is invertible. This is the standard formulation for multivariate calibration by multiple linear regression (MLR).

Now suppose that the instrumental response vector $\mathbf{x}$ may be affected by a stochastic perturbation $\Delta\mathbf{x}$ of zero mean and a known covariance. In this case, a more robust solution to the regression problem can be obtained by minimizing the following cost function:

$$J(\mathbf{b}) = E \| (\mathbf{X} + \Delta\mathbf{X})\mathbf{b} - \mathbf{y} \|^2 \tag{6}$$

where E denotes the expectancy operator for random variables. In this equation, $\Delta\mathbf{X}$ $(N \times p)$ is a perturbation term with $E[\Delta\mathbf{X}] = \mathbf{0}$ and $E[\Delta\mathbf{X}^{\mathrm{T}}\Delta\mathbf{X}] = \mathbf{R}$, where $\mathbf{R}$ is a known $(p \times p)$ matrix. This cost function can be re-written as