Chemometrics and Intelligent Laboratory Systems 171 (2017) 234–240

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Sampling Error Profile Analysis for calibration transfer in multivariate calibration

Feiyu Zhang [a], Wanchao Chen [a], Ruoqiu Zhang [a], Boyang Ding [a], Heming Yao [b], Jiong Ge [b], Lei Ju [b], Wuye Yang [a], Yiping Du [a,*]

[a] Shanghai Key Laboratory of Functional Materials Chemistry, School of Chemistry & Molecular Engineering, East China University of Science and Technology, Shanghai 200237, PR China

[b] Shanghai Tobacco Group Company Ltd., Shanghai 200082, China

ARTICLE INFO

Keywords:
Sampling Error Profile Analysis (SEPA)
Calibration transfer
PDS
Near-infrared

ABSTRACT

A new strategy named Sampling Error Profile Analysis (SEPA) is proposed in the optimization for some parameters in piecewise direct standardization (PDS), such as the number of principal components and window size, and the evaluation for the calibration transfer. Partial least squares (PLS) with mean-centering is used in PDS for calibration transfer. Random re-sampling is carried out in SEPA to obtain a series of subsets and build same number sub-models that produce corresponding number root mean square errors (RMSE), of which the mean value and standard deviation are calculated. To take both accuracy and stability into account, the sum of the mean value and standard deviation are used for parameter optimization and model evaluation. The performance of the proposed strategy has been tested on two data sets: a ternary mixture dataset and a corn dataset. Compared with PDS, SEPA-PDS obtained lower prediction errors, indicating that the transfer model would be more robust and effective when using the parameters optimized by SEPA. Compared with other two commonly used calibration transfer methods of slope and bias correction (SBC) and spectral space transformation (SST), SEPA-PDS acquired more satisfactory results.

1. Introduction

Near-infrared (NIR) spectroscopy has been widely used in agricultural [1,2], petrochemical [3,4] and pharmaceutical [5–8] in the past few decades, and has been proved to be a rapid, low-cost and non-destructive analysis method [9]. Building a multivariate calibration model is a key step in NIR spectroscopic analysis and the performance of quantitative and qualitative analysis depends almost entirely on the calibration model. The development of a reliable multivariate calibration model tends to be time-consuming and costly. Therefore, it is satisfactory if the calibration models can be used for an extended period. However, there exist many worrying situations in which the multivariate calibration model can become inapplicable because of the differences in the spectra measured, including baseline drift, wavelength drift and absorbance fluctuations. Large prediction errors would be caused when a calibration model developed on one instrument (primary instrument) has to be applied to the spectra collected on another instrument (secondary instrument), or when the spectra are measured on aging or repaired equipment.

To solve the problem mentioned above, various methods of calibration transfer have been developed. In general, calibration transfer methods can be classified as prediction correction and spectral transfer methods. Slope and bias correction (SBC) [10] is one of the most widely used methods for correcting predicted values. Since SBC is a univariate approach, it can only be used for correction when the differences between the instrumental responses are simple. Compared with prediction correction methods, spectral transfer methods are used more frequently. Spectral space transformation (SST) [11] is one of the spectral transfer methods that eliminates the spectral differences between different instruments through the transformation between two spectral spaces spanned by the corresponding spectra of a subset of standardization samples measured on two instruments. Calibration transfer methods based on canonical correlation analysis (CCA) [12,13] have been proposed. Recently, a Transfer via Extreme learning machine Auto-encoder Method (TEAM) [14] has been proposed to solve the spectra standardization problem. Calibration transfers based on Tikhonov Regularization (TR) [15–17] have been proposed to maintain spectral calibration models. Direct standardization (DS) [18,19] and piecewise direct

* Corresponding author.
E-mail address: yipingdu@ecust.edu.cn (Y. Du).

https://doi.org/10.1016/j.chemolab.2017.10.016
Received 19 May 2017; Received in revised form 3 October 2017; Accepted 18 October 2017
Available online 7 November 2017

standardization (PDS) [20–23] aim to find a transfer matrix to standardize the spectra of the samples measured on the secondary instrument into the spectra as measured on the primary instrument. DS directly relates the response of samples obtained on the primary instrument to that obtained on the secondary instrument. The spectra of all wavelengths on the secondary instrument are used to fit each spectral point on the primary instrument in DS. Among the existing spectral transfer methods, PDS is probably the most widely used. In PDS, the transformation matrix is estimated by a moving window, which enables better modeling of possible nonlinearities. However, the window size has a significant effect on the performance of PDS, which means that the selection of window size should be careful. During the regression to obtain the transformation matrix, PCA and PLS are often used, of which the number of principal components is also an important parameter in PDS. Generally speaking, the larger the number of standardization samples, the higher probability that the process of calibration transfer is valid. Nevertheless, it would be preferable if satisfactory results can be achieved with fewer standardization samples, since fewer standardization samples require decreased analysis time with lower associated costs [11]. To sum up, the number of principal components, window size and the number of standardization samples need to be carefully optimized in PDS. However, the three parameters mentioned above tended to be determined just by a fixed validation set, easily causing the problem that the parameters can't be perfectly optimized.

In this paper, a new strategy named Sampling Error Profile Analysis (SEPA) was proposed to solve the problem mentioned above. Random re-sampling is the core of SEPA, to overcome the disadvantages of sampling only once with a fixed validation set. The random re-sampling can produce a series of sub-models to obtain corresponding number prediction errors. The mean and standard deviation of root mean square errors (RMSE) are used to optimize the parameters in PDS, which means that both mean and standard deviation of the errors are taken into consideration thus the transfer model is more accurate and stable. In this work, SEPA coupled with PDS (SEPA-PDS) is used for calibration transfer in two NIR datasets, and the performance of SEPA-PDS is compared to that of SBC, SST and PDS.

2. Theory and algorithm

2.1. Notations

Assume the spectral matrices $\boldsymbol{X}_1(m \times p_1)$ and $\boldsymbol{X}_2(m \times p_2)$ are the spectra of the same samples measured on the primary and secondary instruments respectively, where m signifies the number of samples, while p_1 and p_2 are the wavelength numbers of the spectra measured on the two instruments. $\boldsymbol{Wave}_1(1 \times p_1)$ and $\boldsymbol{Wave}_2(1 \times p_2)$ are the wavelength vectors of on the two instruments.

2.2. PDS algorithm

In $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ there are different resolutions normally, thus $\boldsymbol{Wave}_1$ and $\boldsymbol{Wave}_2$ contain different values of wavelength. For the ith wavelength in $\boldsymbol{X}_1$, its closest point of jth wavelength in $\boldsymbol{X}_2$ is found, where the absolute value of $\boldsymbol{Wave}_1(i) - \boldsymbol{Wave}_2(j)$ attains a minimum. Then, a window in $\boldsymbol{X}_2$ with a center of $\boldsymbol{Wave}_2(j)$, is used to build a regression model to fit the ith wavelength in $\boldsymbol{X}_1$. The window size is an odd number, smaller than the wavelength number of $\boldsymbol{X}_2$. To fit the ith column in $\boldsymbol{X}_1$ ($\boldsymbol{r}_i$), the regression model is built as follows:

$$\boldsymbol{r}_i = \boldsymbol{R}_i \boldsymbol{b}_i \tag{1}$$

where, $\boldsymbol{R}_i$ is the localized response matrix of the transfer samples and $\boldsymbol{b}_i$ is the vector of transformation coefficients for the ith wavelength in $\boldsymbol{X}_1$. The regression vectors can be calculated by PCR or PLS. In this work, PLS was used for regression after column mean-centering the spectra matrices. If the spectra points contained in the window do not exist, they are ignored. For example, if $\boldsymbol{Wave}_1(1) - \boldsymbol{Wave}_2(1)$ attains a minimum, to fit the 1st wavelength in $\boldsymbol{X}_1$ with a window size of 5, only the first three wavelengths in $\boldsymbol{X}_2$ are used to build the regression model. It is worth mentioning that the resolution and range of wavelengths are allowed to be different here, but $\boldsymbol{Wave}_1$ must be included in $\boldsymbol{Wave}_2$. The calculated regression vectors can then be assembled to form a banded diagonal transformation matrix $\boldsymbol{F}$ as follows:

$$\boldsymbol{F} = \mathrm{diag}\left(\boldsymbol{b}_1^T, \boldsymbol{b}_2^T, \cdots, \boldsymbol{b}_j^T, \cdots \boldsymbol{b}_k^T\right) \tag{2}$$

where, k is the number of wavelengths in $\boldsymbol{X}_1$.

2.3. SEPA-PDS

The strategy of SEPA-PDS aims to fully optimize the parameters in PDS by random re-sampling. The Kennard and Stone algorithm [24] will be used to select standardization samples in this work. The steps of SEPA-PDS are as follows:

Step 1: Optimization of the number of principal components (PCs) in PLS and window size for the first time. ***Ns*** samples are randomly picked out as the set of standardization samples to calculate the transfer matrix ***F***. The remaining ***Nv*** samples are treated as the validation set, and a value of root mean square error of prediction for the validation set ($\mathrm{RMSEP_v}$) is obtained after transformation of the spectra in $\boldsymbol{X}_2$ of validation set using ***F***. The ratio of ***Ns*** and ***Nv*** is about 2:1. The process of random sampling is repeated ***N*** times, and at each time the number of PCs changes from 1 to 5, while the window size from 1 to 99. The upper limits of the number of PCs and window size are adjustable. Hence, under each combination of the number of PCs and window size, ***N*** values of $\mathrm{RMSEP_v}$ are acquired. The ***N*** errors exhibit a profile like a normal distribution normally, with which the mean value and standard deviation are calculated that are accurate and robust estimation of $\mathrm{RMSEP_v}$. In this article, ***N*** is set to 1000 and the sum of mean and standard deviation is used to optimize the number of PCs and window size, to take both accuracy and stability into consideration.

Step 2: Optimization of the number of standardization samples. After sorting the spectra of samples in $\boldsymbol{X}_1$ by the Kennard and Stone algorithm, ***Ns*** samples, those at the front, are picked out as the standardization set, while the remaining ***Nv*** samples as the validation set. ***Nt*** samples are randomly picked out from the standardization set to obtain a transfer matrix ***F***. ***Nt*** varies from 5 to ***Ns***, and under each value the random sampling is repeated ***N*** times to obtain a series of errors of $\mathrm{RMSEP_v}$ for the validation set, of which the mean value is used to optimize the number of standardization samples.

Step 3: Optimization of the number of PCs and window size for the second time. Theoretically, the parameters optimized by ***Step 1*** are stable, but considering the possible effects caused by the number of standardization samples, they are optimized once again here. ***Step 1*** is repeated after ***Ns*** is replaced by the optimized number of standardization obtained in ***Step 2***.

The steps above are illustrated in Fig. 1.

3. Dataset descriptions

3.1. Ternary mixture dataset

103 samples composed of decahydronaphthalene, butyl acetate and hexyl alcohol were prepared, according to a ternary mixture design. The mass fraction of decahydronaphthalene was observed as **y**, ranging from 0 to 100 (%). The NIR spectra were acquired on a INSION (Germany) spectrometer and a BWTEK (America) spectrometer. The spectra in the range of 894.2–1942.6 nm were acquired using the INSION spectrometer with 128 points contained. The spectra acquired on the BWTEK spectrometer were in the range of 876.04–1711.2 nm, containing 511 points. In this study, the INSION spectrometer was selected as the primary