2018 5th International Conference on Information Science and Control Engineering

A novel stacked regression algorithm based on slice transform for small sample size problem in spectroscopic analysis

Yifan Wu[1,2], Silong Peng[1,2]∗, Qiong Xie[1], Quanjie Han[1,2]
1.Institute of Automation, Chinese Academy of Sciences
100190 Beijing, China
2.University of Chinese Academy of Sciences, 100190 Beijing, China
Email:Yifan Wu: yfwu5216@ia.ac.cn, Silong Peng: silong.peng@ia.ac.cn

***Abstract*—In spectroscopic data analysis, small sample size (SSS) problem occurs. A solution is to perform variable selection, which has been proved to be critical to improve the performance of the regression model, such as partial least squares (PLS) regression. Stacked moving window partial least squares (SMWPLS) aims to combine variable sets instead of selecting a subset to improve the model robustness. In this study, we proposed a novel weighting strategy to calculate the combination weights. Slice transform (SLT) is used to map the cross-validation (CV) weights to new weights in a piecewise linear manner. The parameters of SLT are optimized with the least-square criterion. Experiments on two near-infrared (NIR) data sets demonstrated the efficiency of the proposed SLT weighting.**

***Keywords*-Small sample size problem; Variable selection; Stacked regression; Slice transform;**

I. Introduction

Spectroscopic dataset usually contains hundreds to thousands of variables, while the number of samples is less than a hundred [1]. Such data is characterized as ”high-dimension small-sample size (SSS)” or ”large p small n” for that it has a much larger dimension p than the sample size n [2]. Using spectroscopic data to perform quantitative analysis has become an important tool for modern analytical chemistry. In terms of machine learning, the task of the quantitative analysis is to build regression model, with the spectral variables as predictors and the concentration variable of target analyte as response. The regression model will be used to predict the concentration of the future sample. Usually there are redundant and irrelevant variables in spectral data and they will deteriorate the predictive performance of the regression [3], [4]. Dimension reduction techniques such as principle component analysis (PCA) or partial least squares (PLS) was often used because of the ability of overcoming both the dimensionality and the collinear problems [5], [6], [7]. However, there are still a number of obstacles with these methods, such as how to eliminate noise interference during use and how to interpret the reduced data.

In order to solve SSS problem, one approach is based on variable selection [8], [9], [10], [11]. The goal of variable selection is to obtain a small set of variables that permits less regression error and more interpretability than the original set of variables. Moving window strategy combined with PLS (MWPLS) is an efficient way for variables selection [12], [13]. In MWPLS, a series of PLS models in every window that moves over the whole spectral region are built, and then useful spectral intervals, i.e., informative regions, in terms of the least complexity of PLS models reaching a desired error level are located. Improvement may be seen in many variable selection methods, however, there is the risk that eliminating most of the variables may result in the loss of useful information.

Another approach to the SSS problem is to use a strategy termed stacked regression, which relies on the idea of combining variable sets instead of selecting the best one to improve the model robustness [14], [15]. In stacked moving window PLS (SMWPLS) [16], [17], a moving window is used to establish a set of overlapping intervals for building PLS sub-models. It is well known that combining sub-models is able to give better or at least not worse performance than selecting the single best. The remaining problem is how to combine given a finite data set. An efficient method is to use the reciprocal of the square cross-validation error [16], [18] as combination weights, which is intuitionistic and easy to compute but not necessarily optimal. Breiman et al. [19] proposed to use least-square linear regression to learn the combination weights under the constraint that all regression coefficients are non-negative. This non-negativity constraint was found to be crucial to guarantee that the stacked regression has good generalization.

In the task of quantitative analysis with spectroscopic data, the number of samples is small and the response vector $\boldsymbol{y}$ usually contains unexpected noise. Therefore, least-squares tends to fit the noise. In this study, we proposed an improved SMWPLS with a novel weighting rule. The proposed weighting rule is based on least-square linear regression with a constraint that the regression coefficient is a nonlinear mapping of the CV weights. Slice transform (SLT) is employed to perform the nonlinear mapping in a piecewise linear manner. Two near-infrared (NIR) spectra dataset were used to demonstrated the efficiency of the proposed weighting rule.

978-1-5386-5500-9/18/$31.00 ©2018 IEEE
DOI 10.1109/ICISCE.2018.00026

IEEE computer society

II. Method

A. Stacked moving window partial least square

Firstly, a window with fixed size is moved along the full spectra to generate overlapping intervals. Then, PLS sub-model is built on each interval. Finally, all PLS sub-models are combined linearly with different weights:

$$\boldsymbol{\beta}_{stacked} = \sum_{k=1}^{K} w_k \boldsymbol{\beta}_k \tag{1}$$

where $\boldsymbol{\beta}_{stacked}$ is the stacked regression coefficient, w_k and $\boldsymbol{\beta}_k$ is the weight and the regression coefficient of the k-th sub-model.

B. Slice transform

SLT is a linear representation of a signal $\boldsymbol{x} \in \Re^{K\times 1}$ where the basis functions are linear splines [20]. Let $x_i \in [a, b), i = 1, ..., K$ be one of the elements of $\boldsymbol{x}$, the interval $[a, b)$ is divided into M equidistant bins and the endpoint vector is denoted as $\boldsymbol{q} = [q_0, q_1, \cdots, q_M]^{\intercal}(q_0 = a < q_1 < q_2 < \cdots < q_M = b)$.
Apparently, x_i must fall and can only fall in one of the bins, we denote it as:

$$r(x_i) = \frac{x_i - q_{\pi(x_i)-1}}{q_{\pi(x_i)} - q_{\pi(x_i)-1}} \tag{2}$$

where $r(x_i) \in [0, 1)$. At the same time, x_i can be expressed as:

$$x_i = r(x_i)q_{\pi(x_i)} + (1 - r(x_i))q_{\pi(x_i)-1} \tag{3}$$

Therefore, the SLT of $\boldsymbol{x}$ can be defined as:

$$\boldsymbol{x} = \boldsymbol{S}_q(\boldsymbol{x})\boldsymbol{q} \tag{4}$$

where $\boldsymbol{S}_q(\boldsymbol{x})$ is a matrix with K rows and $M + 1$ columns. The critical property of SLT is the substitution property. That is, the piecewise linear mapping of $\boldsymbol{x}$ to $\boldsymbol{z}$ can be implemented by substituting a new node vector $\boldsymbol{p}$ for $\boldsymbol{q}$ in Eq.(4):

$$\boldsymbol{z} = \boldsymbol{S}_q(\boldsymbol{x})\boldsymbol{p} \tag{5}$$

Figure 1 illustrates the substitution operation [20].

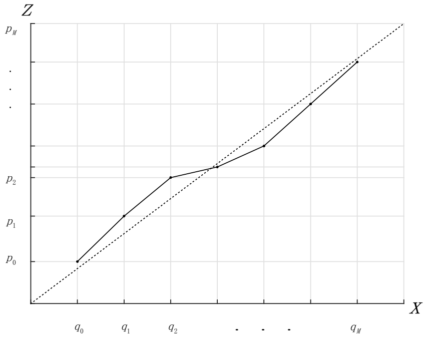

Figure 1: Illustration of the substitution property of SLT.

C. SLT based weighting strategy

To calculate the combination weights for SMWPLS, a simple but efficient approach is to use cross-validation error as a sign to measure the contributions of each sub-interval model:

$$\begin{aligned} \boldsymbol{w}_{cv} &= \tfrac{1}{m}\left[\tfrac{1}{s_1^2}, ..., \tfrac{1}{s_k^2}, ..., \tfrac{1}{s_K^2}\right] \\ m &= \sum_{k=1}^{K} \tfrac{1}{s_k^2} \end{aligned} \tag{6}$$

where s_k is the cross-validation error of the k-th sub-model. The CV weights are optimal under the condition that the sub-models are uncorrelated [21]. However, in practice, the uncorrelated condition is hard to fulfill since the sub-model are used to predict the same target. The CV weights are widely used because they are easy to compute. More importantly, they are consistent with our intuition that the better performance, the higher weight.

Considering both the advantage and disadvantage of the CV weights, a reasonable solution is to modify the CV weights. With the substituting property, the SLT can describe a family of nonlinear mapping functions in a matrix form. This, in turn, enables a simple optimization of the mapping functions as a solution to a linear set of equations. Therefore, we can obtain the optimal (piecewise linear) map to approximate general nonlinear maps. In this study, the substitution property of SLT is utilized to map the CV weight vector to a new one:

$$\boldsymbol{w}_{slt} = \boldsymbol{S}_q(\boldsymbol{w}_{cv})\boldsymbol{p} \tag{7}$$

In this study, the optimal piecewise mapping function is obtained by minimizing the following least squares:

$$\begin{aligned} \hat{\boldsymbol{p}} &= \underset{p}{\arg\min} \quad \|\boldsymbol{y} - \boldsymbol{L}\boldsymbol{p}\|^2 \\ \boldsymbol{L} &= \boldsymbol{Y}\boldsymbol{S}_q(\boldsymbol{w}_{cv}) \\ \boldsymbol{Y} &= \begin{bmatrix} \hat{y}_{11} & \hat{y}_{21} & \cdots & \hat{y}_{K1} \\ \hat{y}_{12} & \hat{y}_{22} & \cdots & \hat{y}_{K2} \\ \vdots & \vdots & \vdots & \vdots \\ \hat{y}_{1n} & \hat{y}_{2n} & \cdots & \hat{y}_{Kn} \end{bmatrix} \end{aligned} \tag{8}$$

where $\hat{y}_{ki}$ is the cross-validation prediction of the i-th sample from the k-th regression sub-model, $\boldsymbol{y}$ is the response vector. The explicit solution of Eq.(8) can be derived:

$$\hat{\boldsymbol{p}} = (\boldsymbol{L}^T\boldsymbol{L} + \lambda\boldsymbol{I})^{-1}\boldsymbol{L}^T\boldsymbol{y} \tag{9}$$

For $\lambda\boldsymbol{I}$ in Eq. (9), this additional term is added to avoid the singularity of matrix inversion operation where $\boldsymbol{I}$, an identity matrix, and λ, a small scale value, provide robust computation. The λ is configured to 10^{-6} in all of the experiments. With the new node vector obtained, the SLT weights can be calculated by Eq. (7). On the basis of the flexibility of SLT, the weight can handle some nonlinear situations. When we restrict the mapping function to a piecewise linear mapping, the optimal mapping function can be solved by the matrix of the predictions $\boldsymbol{Y}$ and the response vector $\boldsymbol{y}$ adaptively.

In addition, another view for interpreting the SLT weighting is:

$$\begin{aligned} \boldsymbol{w}_{slt} &= \arg\min \left\|\boldsymbol{y} - \boldsymbol{Y}\boldsymbol{w}\right\|^2 \\ s.t. \quad \boldsymbol{w} &= \boldsymbol{S}_q\left(\boldsymbol{w}_{cv}\right)\boldsymbol{p} \end{aligned} \tag{10}$$

In the task of quantitative analysis in analytical chemistry, the response vector $\boldsymbol{y}$ may contain unexpected noise due to factors such as the complexity of the mixture, the experience of the operator and so on. In this case, least-square solution without any prior (such as NNLS) tends to fit the noise. The least-square optimization of Eq.(10) can be viewed as using the CV weights as prior, which reduce the risk of fitting noise .

III. Materials

The first dataset was Corn data collected in Cargill, which can be downloaded from http://www.eigenvector.com/data/Corn/index.html. The dataset consisted of 80 corn samples. The spectra of corn samples were corrected with a wavelength range of 1100-2498 nm at 2 nm resolution, which results in 700 variables. In this study, the spectra data of mp5 and the concentration data of starch were used. After two outlier samples were removed, both the training set and the test set consisted of 39 samples . Detail of sample division can be found in [22].

The second dataset was Ternary Mixture data, which was described by Wulfert et al. [23]. Near-infrared spectra were measured with a ternary mixture of ethanol, water and isopropanol. In this study, iso-propanol was selected as the target component. Wavelength range of 850-1049 nm at 2 nm resolution were used, resulting in 100 variables. The training set consisted of 65 spectra and the test set consisted of 30 spectra.

For the two NIR training data, the condition number of the spectra matrixes were very large due to multicollinearity, which indicated that the original feature space is not suitable for linear least-squares. Table I shows the properties of the spectra matrixes.

Table I: Properties of the spectra matrix of the training set.

Data set	Variables	Samples	Condition number
Corn	700	39	1.06×10^5
Ternary Mixture	100	65	1.73×10^4

IV. Experimental results

The programs are written in house in Matlab Version R2015a and run in a personal computer with a 2.20 GHz Intel Core 2 processor, 4 GB RAM, and a Windows 7 operating system.

A. Evaluation measures

In this study, the root mean square error of prediction (RMSEP) from the test set was used as the criterion of evaluating different models. RMSEP is defined as follows:

$$RMSEP = \sqrt{\left(\frac{\sum\left(y_{pred} - y_{true}\right)^2}{N_{test}}\right)} \tag{11}$$

where N_{test} is the number of samples in the test set, y_{true} is the reference value, and y_{pred} is the predicted value provided by different models.

B. Experimental results and discussion

PLS with full variables, MWPLS with the best variable interval, SMWPLS with CV weights (SMWPLS_CV), SMWPLS with NNLS weights (SMWPLS_NNLS) and SMWPLS with the proposed SLT weights (SMWPLS_SLT) were investigated in this study. Three parameters should be tuned, which are the number of latent variables for PLS, the window size for moving-window (MW) strategy and the number of bins for SLT. The number of latent variables was selected automatically with a statistical test [24]. The window size influences the performance of the sub-model. If the window size is too large, redundant variables or noise variables may be included. If the window size is too small, the information will not be sufficient for multi-component discrimination. The number of bins of SLT controls the flexibility of SLT. The greater the number of bins, the more flexible the generated mapping functions are, but at the same time, the risk of overfitting is increased. The parameters for MW and SLT were tuned manually with multiple experiments, which are shown in Table II.

Table II: Parameter settings for MW and SLT.

Data set	Window size	No. of bins
Corn	61	4
Ternary Mixture	61	5

We implemented the moving-window strategy by finding the center of each window. That is, if K sub-models, or windows, were desired, the full variable length was divided into $(K + 1)$ equal segmentations with the K middle endpoints as the window centers. For SMWPLS, different ensemble sizes (number of sub-models) were tested.

Firstly, we investigated the advantage of SLT weighting at a specific ensemble size. The ensemble sizes of 60 and 30 were selected for Corn data and Ternary Mixture data respectively. Table III summarizes the RMSECV and RMSEP results of different models. For Corn data, the the root mean square error of cross-validation (RMSECV) of PLS is larger than RMSEP, which seemed abnormal. With variable selection, the RMSECV of MWPLS was smaller than its RMSEP. Moreover, the RMSEP of MWPLS was smaller than that of PLS. The above results demonstrated that variable selection helped to improve model interpretability as well as predictive performance. SMWPLS solves the problem by using all the information of sub-models instead of choosing the best one. The NNLS weighting yielded larger RMSEP than the CV weighting. The reason might be that NNLS weighting assigned high weights to sub-models with poor performance mistakenly (upper part of Figure 2). It can be seen that NNLS assigned the highest weights to the 18th and the 51th sub-models, whose weights were low in the CV weighting. In contrast, the SLT weights seemed more reasonable since they were modified from the CV weights with nonlinear mapping. It can be seen that SMWPLS_SLT yielded the lowest RMSEP among all models.

For Ternary Mixture data, MWPLS yielded lower RMSEP than PLS due to variable selection. However, SMWPLS_CV yielded higher RMSEP than MWPLS since the CV weighting was empirical instead of optimal. With least-squares optimization, both NNLS weighting and SLT weighting enhanced the weight of the 13th sub-model (lower part of Figure 2), which was the best sub-model in terms of RMSECV. It can be seen that the RMSEPs of SMWPLS_NNLS and SMWPLS_SLT were smaller than that of MWPLS.

Table III: Predictive results of different models on the test set.

Model	Corn		Ternary Mixture	
	RMSECV	RMSEP	RMSECV	RMSEP
PLS	0.328	0.253	1.250	1.382
MWPLS	0.216	0.229	1.012	1.122
SMWPLS_CV	-	0.224	-	1.308
SMWPLS_NNLS	-	0.258	-	1.083
SMWPLS_SLT	-	**0.170**	-	**0.962**

Bold means the lowest RMSEP value of the corresponding data set

Secondly, we investigated the predictive performance of SMWPLS with different weightings at different ensemble sizes. Results of Corn are shown in the upper part of Figure 3. It can be seen that the proposed SLT weighting yielded the lowest RMSEPs at all the ensemble sizes. The RMSEPs of the NNLS were similar to those of the CV in this data, except the size of 60, the reason of which has been explored in the previous discussion. Results of Ternary Mixture (lower part of Figure 3) also demonstrated the efficiency of the proposed SLT weighting. It can be seen that the performance of the CV weighting was relatively stable with the increase of the ensemble size. With optimization, both the NNLS and the SLT achieved lower RMSEPs than the CV. However, the performance of the SLT was less stable than the CV. For the SLT, it can be seen that the RMSEPs of ensemble sizes 25 and 40 were much larger than the other ensemble sizes.

V. Conclusion

In this paper, we have addressed the problem of high-dimension small-sample size regression. To achieve better predictive performance, we proposed the SLT weighting for SMWPLS. Experimental results on two NIR data sets

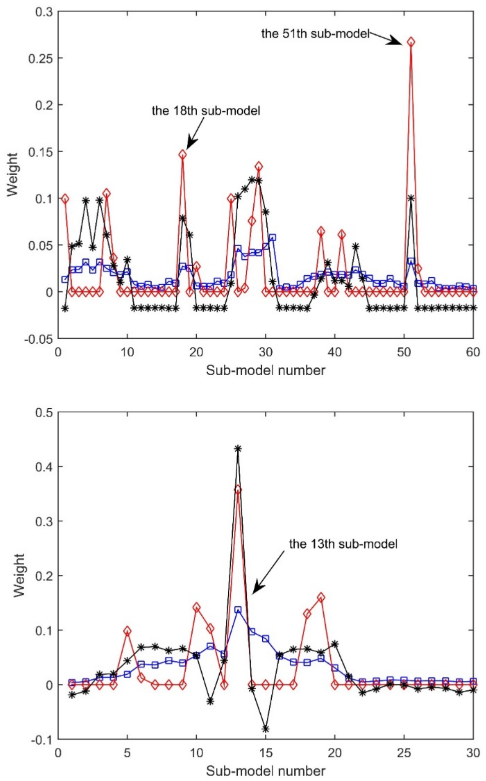

Figure 2: Illustration of weights for SMWPLS. Upper part: Corn data at the ensemble size of 60, lower part: Ternary Mixture data at the ensemble size of 30. Blue Square : CV weighting, diamond line: NNLS weighting, star line: SLT weighting.

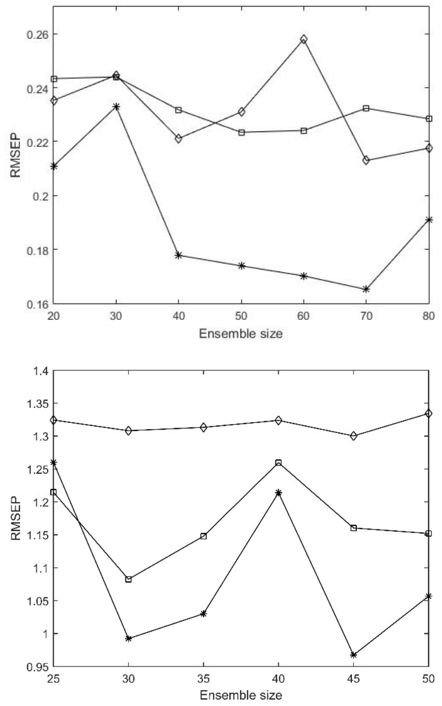

Figure 3: Predictive performance of SMWPLS at different ensemble sizes. Upper part: Corn data, lower part: Ternary Mixture data. Square line: CV weighting, diamond line: NNLS weighting, star line: SLT weighting.

demonstrated the efficiency of the SLT weighting. It can also be seen that different ensemble sizes led to different performance for SMWPLS_SLT. Therefore, we will focus on the optimization of the ensemble size in the next study.

Acknowledgment

The authors would like to thank Dr. Chen Chen for her valuable advice which leaded to substantial improvement to this work. This work is financially supported by the National Natural Science Foundation of China (Grant No.61571438) and the National Natural Science Foundation of China (Grant No.61601104).

References

[1] Z. Xiaobo, Z. Jiewen, M. J. Povey, M. Holmes, M. Hanpin, Variables selection methods in near-infrared spectroscopy, Anal Chim Acta 667 (1-2) (2010) 14–32.

[2] P. J. Bickel, E. Levina, Some theory for fisher's linear discriminant function,naive bayes, and some alternatives when there are many more variables than observations, Bernoulli 10 (6) (2004) 989–1010.

[3] C. H. Spiegelman, M. J. McShane, M. J. Goetz, M. Motamedi, Q. L. Yue, G. L. Cote, Theoretical justification of wavelength selection in pls calibration: Development of a new algorithm, Anal Chem 70 (1) (1998) 35–44.

[4] V. Centner, D. L. Massart, O. E. de Noord, S. de Jong, B. M. Vandeginste, C. Sterna, Elimination of uninformative variables for multivariate calibration, Anal Chem 68 (21) (1996) 3851–8.

[5] S. Wold, K. Esbensen, P. Geladi, Principal component analysis, Chemometrics and intelligent laboratory systems 2 (1-3) (1987) 37–52.

[6] H. Martens, T. Naes, Multivariate Calibration, John Wiley and Sons, Chichester and New York, 1989.

[7] S. Wold, M. Sjostrom, L. Eriksson, Pls-regression: a basic tool of chemometrics, Chemometrics and Intelligent Laboratory Systems 58 (2) (2001) 109–130.

[8] F. Lindgren, P. Geladi, S. R$\ddot{a}$nnar, S. Wold, Interactive variable selection (ivs) for pls. part 1: Theory and algorithms, Journal of Chemometrics 8 (5) (1994) 349–363.

[9] R. Leardi, A. L. Gonzalez, Genetic algorithms applied to feature selection in pls regression: how and when to use them, Chemometrics and Intelligent Laboratory Systems 41 (2) (1998) 195–207.

[10] H. Li, Y. Liang, Q. Xu, D. Cao, Key wavelengths screening using competitive adaptive reweighted sampling method for multivariate calibration, Analytica Chimica Acta 648 (1) (2009) 77–84.

[11] B. C. Deng, Y. H. Yun, D. S. Cao, Y. L. Yin, W. T. Wang, H. M. Lu, Q. Y. Luo, Y. Z. Liang, A bootstrapping soft shrinkage approach for variable selection in chemical modeling, Anal Chim Acta 908 (2016) 63–74.

[12] L. Xu, I. Schechter, Wavelength selection for simultaneous spectroscopic analysis. experimental and theoretical study, Analytical Chemistry 68 (14) (1996) 2392–2400.

[13] Y. P. Du, Y. Z. Liang, J. H. Jiang, R. J. Berry, Y. Ozaki, Spectral regions selection to improve prediction ability of pls models by changeable size moving window partial least squares and searching combination moving window partial least squares, Analytica Chimica Acta 501 (2) (2004) 183–191.

[14] Z.-H. Zhou, Ensemble methods: foundations and algorithms, CRC press, 2012.

[15] D. V. Poerio, S. D. Brown, Stacked interval sparse partial least squares regression analysis, Chemometrics and Intelligent Laboratory Systems 166 (2017) 49–60.

[16] W. Ni, S. D. Brown, R. Man, Stacked partial least squares regression analysis for spectral calibration and prediction, Journal of Chemometrics 23 (10) (2009) 505–517.

[17] L. Xu, J. H. Jiang, Y. P. Zhou, H. L. Wu, G. L. Shen, R. Q. Yu, Mccv stacked regression for model combination and fast spectral interval selection in multivariate calibration, Chemometrics and Intelligent Laboratory Systems 87 (2) (2007) 226C230.

[18] W. Ni, S. D. Brown, R. Man, Data fusion in multivariate calibration transfer, Anal Chim Acta 661 (2) (2010) 133–42.

[19] L. Breiman, Stacked regressions, Machine Learning 24 (1) (1996) 49–64.

[20] Y. Hel-Or, D. Shaked, A discriminative approach for wavelet denoising, IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society 17 (4) (2008) 443–57.

[21] L. N. C. Michael P. Perrone, When networks disagree ensemble methods for hybrid neural networks.

[22] R. N. Feudale, H. Tan, S. D. Brown, Piecewise orthogonal signal correction, Chemometrics and Intelligent Laboratory Systems 63 (2) (2002) 129–138.

[23] F. W$\ddot{u}$lfert, W. T. Kok, O. E. d. Noord, A. K. Smilde, Correction of temperature-induced spectral variation by continuous piecewise direct standardization, Analytical chemistry 72 (7) (2000) 1639–1644.

[24] D. M. Haaland, E. V. Thomas, A. Chem., Partial least-sqares methods for spectral analyses. 1. relation to other quantitative calibration methods and the extraction of quantitative information, Analytical Chemistry 60 (11) (1988) 1193–1202.