Hindawi Publishing Corporation
Journal of Automated Methods and Management in Chemistry
Volume 2009, Article ID 291820, 6 pages
doi:10.1155/2009/291820

Research Article

An Improved Ensemble Method for Completely Automatic Optimization of Spectral Interval Selection in Multivariate Calibration

Xiao-Ping Yu,[1] Lu Xu,[1] and Ru-Qin Yu[2]

[1] *College of Life Sciences, China Jiliang University, Hangzhou 310018, China*
[2] *State Key Laboratory of Chemo/Biosensing and Chemometrics, College of Chemistry and Chemical Engineering, Hunan University, Changsha 410082, China*

Correspondence should be addressed to Ru-Qin Yu, rqyu@hnu.cn

Received 1 March 2009; Accepted 2 April 2009

Recommended by Peter Stockwell

In our recent work, Monte Carlo Cross Validation Stacked Regression (MCCVSR) is proposed to achieve automatic optimization of spectral interval selection in multivariate calibration. Though MCCVSR performs well in normal conditions, it is still necessary to improve it for more general applications. According to the well-known principle of "garbage in, garbage out (GIGO)", as a precise ensemble method, MCCVSR might be influenced by outlying and very bad submodels. In this paper, a statistical test is designed to exclude the ruinous submodels from the ensemble learning process, therefore, the combination process becomes more reliable. Though completely automated, the proposed method is adjustable according to the nature of the data analyzed, including the size of training samples, resolution of spectra and quantitative potentials of the submodels. The effectiveness of the submodel refining is demonstrated by the investigation of a real standard data.

1. Introduction

Mutivariate spectroscopic calibration is an old and yet ever-growing research field in chemometrics. Multivariate calibration technique is very comprehensive and a successful application of this technique requires practitioners' experience and expertise. Multivariate calibration modeling involves many steps, such as outlier diagnosis, selection of representative training samples, data preprocessing, model optimization and validation [1]. Due to the complexity and uncertainty of the data analyzed, each of the above processes has much to do with the success of calibration and thus should be performed properly. Moreover, with increasing needs for quickly quantifying sought-for components in various complicated chemical systems involved in different subjects, automatic optimization of multivariate calibration modeling will undoubtedly boost the applications of chemometrics to analytical chemistry.

Modern spectroscopic instruments can provide a spectrum measured at hundreds and even thousands of wavelengths in a few seconds. An important step in multivariate calibration is wavelength selection. Taking the most popular method, partial least squares (PLSs), for example, wavelength selection and model optimization are usually performed simultaneously. Determination of model complexity of PLS should be based on a best subset of the measured wavelengths. Moreover, it is supported by both practical experiences [2–5] and theoretical research that proper wavelength selection is necessary for multivariate spectroscopic calibration [6, 7]. There have been many literatures devoted to this problem; for a comprehensive review one can see [8, 9].

The present paper is oriented to interval selection. Firstly, for such spectral data like near infrared (NIR) ones, an important feature of the analytical channels is their continuity [10, 11]. Spectral continuity for calibration means that when a certain wavelength contains useful quantitative information or is contaminated, so very likely are its neighboring wavelengths. Therefore, different spectral intervals will have different data structures, namely, different optimized interval

PLS models are very likely to have different model complexity. Secondly, for spectral data with hundreds and even thousands of wavelengths, it makes the wavelength selection procedure simpler to tackle the wavelengths as intervals, because the number of intervals will be much smaller than that of total wavelengths. As two pioneer methods for interval selection, interval PLS (iPLS) models [10, 11] are built on evenly split spectral intervals, while moving-window PLS (MWPLS) [12] develops interval PLS models based on a spectral window moving along the total spectral range. Both of these two methods can present a graphical demonstration of the quantitative potential and complexity of local intervals and provide a straightforward tool for interval selection and model optimization. The original iPLS and MWPLS select the intervals with low errors and less model complexity. This strategy is very reasonable and intuitive, but the selection of intervals included in iPLS or determining interval borders in MWPLS still depends much on experiences. Some researchers have also contributed to improving and optimizing the iPLS or MWPLS [13–15]; however, many of these methods are computationally expensive or do not achieve the optimum models. Considering the local data structure, putting all the seemingly "good" intervals into one PLS model might not be the best choice. For the above reasons, combining and optimizing the proper small interval models by ensemble learning methods seems very attractive.

In our recent work, an improved ensemble learning method, Monte Carlo Cross Validation (MCCV) [16] Stacked Regression (MCCVSR) [17] is used to optimize interval selection. Unlike other common ensemble methods, which achieve model combination by averaging, selecting a median and so on, MCCVSR has its peculiar optimization objective, namely the lowest root mean squared error of MCCV (RMSEMCCV). Moreover, MCCVSR gracefully combines the MCCV of models on small spectral intervals with nonnegative least squares (NNLSs), which is very computationally economic. Optimization of interval selection is achieved by weighting the submodels according to the criterion of lowest RMSEMCCV.

MCCVSR performs very well when the submodels are reasonable or not very bad. Moreover, it can exclude poor models by giving them zero weights in NNLS. However, a concern with general use of this method is when it is applied to data sets with more uncertainty, very bad submodels might spoil the prediction results. According to the well-known "garbage in, garbage out" principle, just one outlying submodel with nonzero weight in the combination will lead to poor predictions in the final ensemble model. Moreover, if many outlying or very poor submodels exist, they can mask each other and have nonzero weights in the ensemble model. So, for the purpose of obtaining an automatic, and more importantly, a generally reliable algorithm, it is necessary to preselect the submodels before combination in MCCVSR. In this work, a statistical test is designed to preselect interval models to develop a completely automatic algorithm for interval selection and model optimization.

2. Theory

2.1. MCCVSR. Stacked regression (SR) [18, 19] is an interesting ensemble method to combine submodels without suffering of correlation. Considering the fact that a large number of combination coefficients can increase the model's degree of freedom and lead to overfitting, MCCV [16] is introduced into SR to improve it. Because MCCV allows a large number of sampling times and a high percent of leave-out samples, it can effectively reduce the risk of overfitting in both submodels and combination. MCCVSR optimizes the combination model as follows:

$$\mathbf{y}_{\mathrm{MCCV}} = [\hat{\mathbf{y}}_{\mathrm{MCCV},1}, \hat{\mathbf{y}}_{\mathrm{MCCV},2}, \hat{\mathbf{y}}_{\mathrm{MCCV},3}, \ldots, \hat{\mathbf{y}}_{\mathrm{MCCV},K}]\mathbf{w}, \quad (1)$$

where the column vector $\mathbf{y}_{\mathrm{MCCV}}$ contains the reference concentration values of leave-out samples during MCCV sampling and $\hat{\mathbf{y}}_{\mathrm{MCCV},i}$ contains the corresponding predicted values by submodel i ($i = 1, 2, \ldots, K$). The $K \times 1$ vector $\mathbf{w}$ contains the model combination coefficients and K is the number of submodels.

The combination coefficient vector, $\mathbf{w}$, in (1) is readily computed by NNLS, which has been proved to be more suitable for combination than normal least squares by avoiding too large weights of some submodels [19]. The prediction by combined model can be expressed as:

$$\hat{\mathbf{y}}_{\mathrm{un}} = [\hat{\mathbf{y}}_{\mathrm{un},1}, \hat{\mathbf{y}}_{\mathrm{un},2}, \hat{\mathbf{y}}_{\mathrm{un},3}, \ldots, \hat{\mathbf{y}}_{\mathrm{un},K}]\mathbf{w}, \quad (2)$$

where $\hat{\mathbf{y}}_{\mathrm{un},i}$ is the predicted concentrations of unknown samples by submodel i ($i = 1, 2, \ldots, K$). More details of MCCVSR can be found in [17].

2.2. Refining Submodels by Statistical Tests. Here, a statistical method is introduced to test the significance of correlation coefficient, r, between $\hat{\mathbf{y}}_{\mathrm{MCCV},i}$ ($i = 1, 2, \ldots, K$) and the corresponding reference values, $\mathbf{y}_{\mathrm{MCCV}}$. Only the submodels with a sigificantly sufficient correlation coefficient can be included for combination. Because the sample distribution of correlation coefficients is much more complex than that of means or mean differences, Fisher's approximately normal transformation [20, 21] of r to Z is used:

$$Z = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right). \quad (3)$$

The new approximately normal statistic Z has an expected standard deviation σ_z near to $\sqrt{1/(n-3)}$, where n is the length of sampling vector. The obtained Z-test value is referred to a normal distribution to test whether r is sigificantly larger than a threshold value. Considering the natures of different data sets, the significance levels and the threshold value of the above one-sided test should be adjustable. For instance, given the frequently used significance level, 0.05, when the spectral intervals are very effective for quantitative analysis, one can adopt a higher threshold value, and vice versa. In this paper, the default threshold value of correlation coefficient is 0.80.

2.3. Optimizing Interval Selection by Improved MCCVSR. In the original paper of MCCVSR, submodels built on evolving

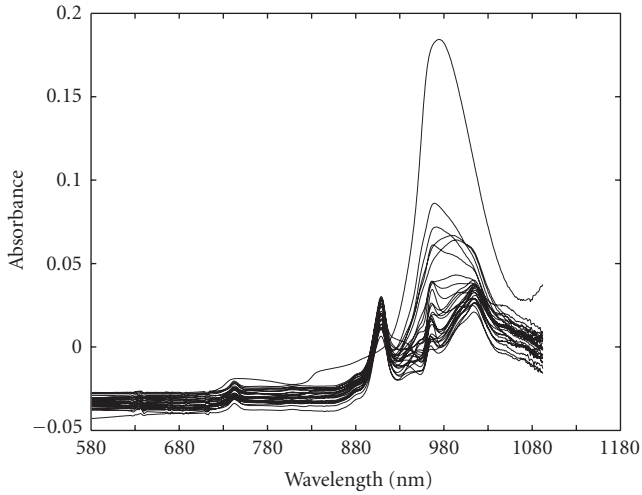


Figure 1: Some of the original spectra in the temperature data set.

spectral intervals [17] are combined. MCCVSR optimizes interval selection by weighting the submodels to achieve the lowest RMSEMCCV value among all combined models with nonnegative constraints. It is just necessary to do MCCV on small interval models and combine them by NNLS, which is very computationally economic.

In order to achieve more precision in interval selection, the idea of moving window is introduced into MCCVSR. The step of evolving interval models can be adjusted in terms of the resolution of spectral data. For example, the wavelength step can be 1, 2, 3, 4, 5 or other positive integers. A default wavelength step of 5 and a window width of 30 are adopted in this paper. Of course, for spectral data with very high resolution, it is wise to have a larger evolving step to save computation time.

3. Data Descriptions

To test the performances of the proposed method, a standard real data set is investigated.

Temperature data [22] Spectra of 19 mixtures of ethanol, water and isopropanol and the spectra of the pure compounds are recorded on an UV-VIS spectra HP 8453 spectrometer. Spectra ranged from 580–1091 nm with 1 nm increment are measured at 30, 40, 50, 60, and 70 degrees Celsius. Representative samples measured at the five temperatures are selected to form a training set to predict concentrations of the three components.

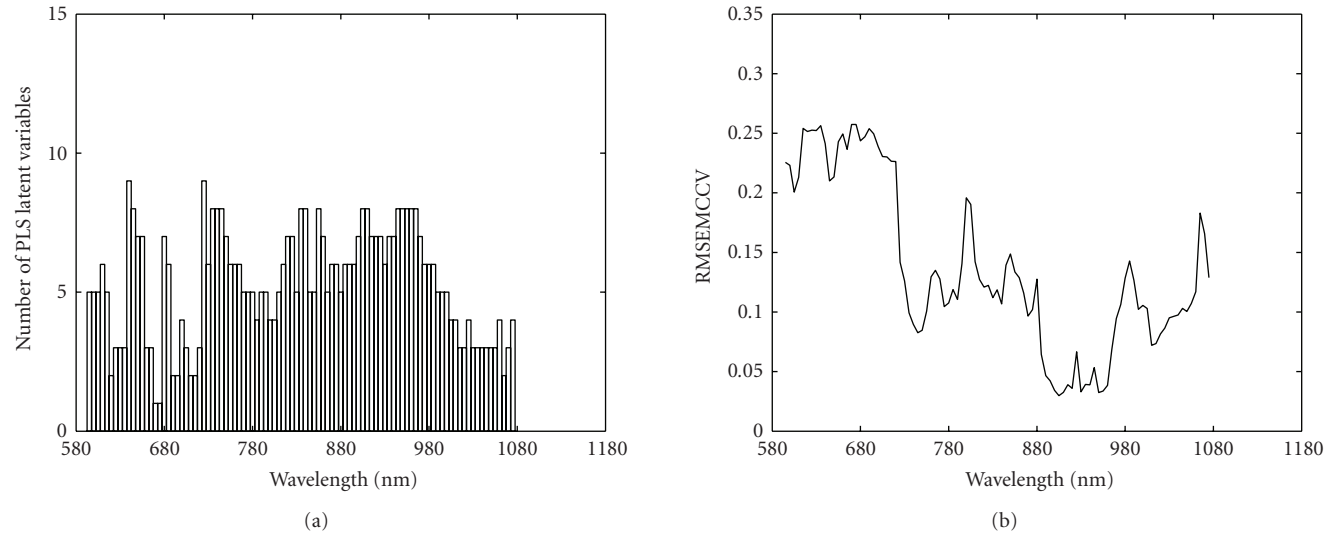
4. Results and Discussions

The data set has 19 mixtures of 3 components, ethanol, water and isopropanol, together with pure components measured at 5 different temperatures, so we have totally 110 samples at hand. To develop global calibration models for predicting percentages of the 3 components, at each temperature, DUPLEX method [23] is used to uniformly select 16 samples for training and 6 samples for test. So we have a training set of 80 samples and a test set of 30 samples. Some of the original training spectra are plotted in Figure 1.

For each component, PLS model with total spectral range, MCCVSR model and improved MCCVSR model with refining step are built. The complexity of PLS model and PLS interval models is determined by MCCV, where the sampling time is 50, and each time 50 percents of the training samples are left out for prediction. The numbers of latent variables are such determined that the RMSEMCCV value is minimized. The root mean squared error of calibration (RMSEC) and the root mean squared error of prediction (RMSEP) are used to evaluate the quality of models. The results of PLS models with total spectral range are listed in Table 1. It can be seen from Table 1 that the numbers of PLS latent variables in these models are much larger than 3, indicating the high complexity of the data. Essentially, influenced by temperature variations and other factors, the spectra are far from the expected ones of a common 3-component system. Moreover, it should be noted that the RMSEP values are much higher than RMSEC values. It is very clear that some spectral intervals are complicated and the global models contain many non-concentration-correlated variations; therefore, it is very necessary to perform wavelength selection.

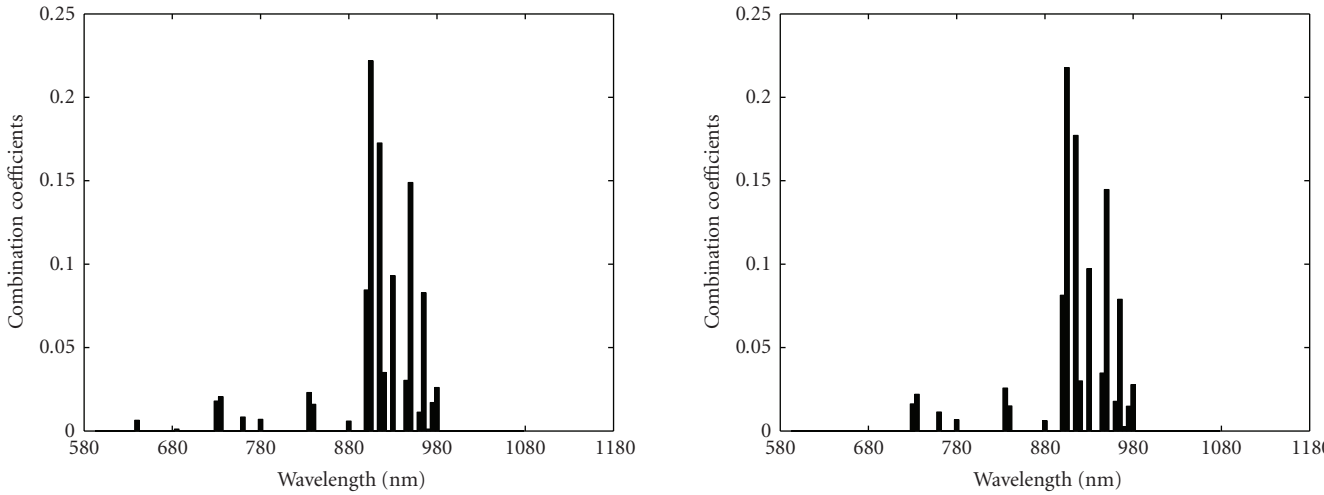
For MCCVSR models, PLS submodels are built on a spectral interval moving along the spectral range. The interval contains 30 wavelengths and its step is set to be 5 wavelengths, so we have 97 interval models in all. The complexity of all the interval models is determined by MCCV described above. For each interval model, the number of PLS latent variables is determined to obtain the lowest RMSEMCCV value. Submodels are then combined by $\mathbf{w}$, as in (1). As an example, Figure 2 presents the optimized complexity of interval models and their RMSEMCCV values for the prediction of ethanol. As shown in Figure 2, the local data structures are very complicated, because some interval models with lower complexity have higher RMSEMCCV values while many interval models with higher complexity present better quantitative potentials. Therefore, an intuitive selection of intervals in terms of lower complexity and errors is not easy and automation of this procedure is necessary. The combination coefficients of MCCVSR and MCCVSR with submodel refining for prediction of ethanol are plotted in Figure 3. Here, the significance level of the test is set to be 0.05 and the threshold value of correlation coefficients is 0.80. From Figure 3, it can be seen that with submodel refining, some interval models are excluded from the final combination, including two submodels that have nonzero weights in MCCVSR. This change might seem too trivial but should not be overlooked. Considering the nature of NNLS, when most submodels (predictors) are very accurate, the power of MCCVSR against bad submodels is very strong, which is the case as above. However, when the spectral intervals generally have poor quantitative potentials, MCCVSR is prone to include very bad models.

The calibration results of the three components obtained by combination models are listed in Table 2. Compared with the PLS models with total spectral range, the combination models demonstrate improved training and predicting



(a) (b)

Figure 2: (a) Model complexity and (b) RMSEMCCV values of interval models for prediction of ethanol.



(a) (b)

Figure 3: The combination coefficients of (a) MCCVSR and (b) MCCVSR with submodel refining for predicting ethanol.

Table 1: The results of PLS model with total spectral range for the temperature data.

Component	LVn[a]	RMSEMCCV	RMSEC	RMSEP
Ethanol	12	0.0098	0.0287	0.0257
Water	13	0.0038	0.0094	0.0136
Isopropanol	12	0.0082	0.0225	0.0212

[a]The number of PLS latent variables.

Table 2: The calibration results of the three components obtained by combination models.

Component	Nm[a]		RMSEMCCV		RMSEC		RMSEP	
	1[b]	2[c]	1	2	1	2	1	2
Ethanol	97	67	0.0137	0.0138	0.0103	0.0104	0.0183	0.0185
Water	97	81	0.0094	0.0094	0.0077	0.0078	0.0121	0.0120
Isopropanol	97	72	0.0136	0.0137	0.0109	0.0109	0.0170	0.0164

[a]The number of submodels for combination.
[b]Results obtained by MCCVSR.
[c]Results obtained by MCCVSR with submodel refining.

performances in terms of RMSEC and RMSEP. With interval models refined, the number of submodels for combination is reduced but the precision is maintained.

Some parameters involved in MCCVSR should be discussed. When performing MCCV, two important parameters are the percentage of left-out samples and the sampling time. Generally speaking, as soon as outliers are removed and the computation time permits, a larger sampling time and a higher percentage of left-out samples are helpful to reduce the risk of overfitting in both single submodels and the combination. On the other hand, the percentage of left-out samples can be adjusted according to the size of training samples in order to have enough representative samples for modeling. The sizes of spectral interval and evolving step are also adjustable. Firstly, an interval should contain enough wavelengths (at least 20 channels) to build a stable calibration model. Secondly, the evolving step can be larger to save time when the spectral resolution is high. When performing the statistical test, given a significance level of 0.05, in order to have enough models for combination, the threshold value of the correlation coefficient can be adjusted according to the quantitative potentials of submodels. An empirical value of 0.80 is recommended, which is enough to eliminate the outlying models.

5. Conclusions

In our recent work, MCCVSR has been proved to be a computationally economic and effective method for wavelength selection. In order to make the MCCVSR algorithm to be more reliable and completely automated for wavelength selection, a statistical test is designed to exclude the outlying submodels from the final ensemble learning with no or little degradation of the model precision. By studying a real data set, the improved MCCVSR method performs almost as well as the original algorithm in terms of training and prediction. Moreover, with less and refined submodels, the final combination is sure to be more reliable. Moreover, the algorithm is completely automated and adjustable according to the nature of the data analyzed. Though just the problem of wavelength selection is tackled, it is evident that the idea of refining the submodels before ensemble combination is generally beneficial to multivariate calibration with ensemble methods like bagging [24]. Finally, the proposed method can perform reliable wavelength selection automatically and is robust against poor interval models but not outliers in reference concentrations ($\mathbf{y}$) or spectra ($\mathbf{X}$). So, it is not a robust multivariate calibration method like robust principal component regression [25, 26] and robust PLS [27], outliers should be weeded before the calibration.

Acknowledgments

This work was financially supported by Ministry of Agriculture of the People's Republic of China (Grants no. 2009ZX08012-013B). The authors thank Zi-Hong Ye and Xin-Da Lin for their discussions

References

[1] H. Martens and T. Næs, *Multivariate Calibration*, John Wiley & Sons, Chichester, UK, 1989.

[2] C. W. Brown, P. F. Lynch, R. J. Obremski, and D. S. Lavery, "Matrix representations and criteria for selecting analytical wavelengths for multicomponent spectroscopic analysis," *Analytical Chemistry*, vol. 54, no. 9, pp. 1472–1479, 1982.

[3] S. D. Frans and J. M. Harris, "Selection of analytical wavelengths for multicomponent spectrophotometric determinations," *Analytical Chemistry*, vol. 57, no. 13, pp. 2680–2684, 1985.

[4] J. H. Kalivas, N. Roberts, and J. M. Sutter, "Global optimization by simulated annealing with wavelength selection for ultraviolet-visible spectrophotometry," *Analytical Chemistry*, vol. 61, no. 18, pp. 2024–2030, 1989.

[5] D. Jouan-Rimbaud, B. Walczak, D. L. Massart, I. R. Last, and K. A. Prebble, "Comparison of multivariate methods based on latent vectors and methods based on wavelength selection for the analysis of near-infrared spectroscopic data," *Analytica Chimica Acta*, vol. 304, no. 3, pp. 285–295, 1995.

[6] C. H. Spiegelman, M. J. McShane, M. J. Goetz, M. Motamedi, Q. L. Yue, and G. L. Coté, "Theoretical justification of wavelength selection in PLS calibration: development of a new algorithm," *Analytical Chemistry*, vol. 70, no. 1, pp. 35–44, 1998.

[7] B. Nadler and R. R. Coifman, "The prediction error in CLS and PLS: the importance of feature selection prior to multivariate calibration," *Journal of Chemometrics*, vol. 19, no. 2, pp. 107–118, 2005.

[8] A. Höskuldsson, "Variable and subset selection in PLS regression," *Chemometrics and Intelligent Laboratory Systems*, vol. 55, no. 1-2, pp. 23–38, 2001.

[9] L. Xu and W.-J. Zhang, "Comparison of different methods for variable selection," *Analytica Chimica Acta*, vol. 446, no. 1-2, pp. 477–483, 2001.

[10] L. Nørgaard, A. Saudland, J. Wagner, J. P. Nielsen, L. Munck, and S. B. Engelsen, "Interval partial least-squares regression (iPLS): a comparative chemometric study with an example from near-infrared spectroscopy," *Applied Spectroscopy*, vol. 54, no. 3, pp. 413–419, 2000.

[11] L. Munck, J. Pram Nielsen, B. Møller, et al., "Exploring the phenotypic expression of a regulatory proteome-altering gene by spectroscopy and chemometrics," *Analytica Chimica Acta*, vol. 446, no. 1-2, pp. 171–186, 2001.

[12] J.-H. Jiang, R. James Berry, H. W. Siesler, and Y. Ozaki, "Wavelength interval selection in multicomponent spectral analysis by moving window partial least-squares regression with applications to mid-infrared and near-infrared spectroscopic data," *Analytical Chemistry*, vol. 74, no. 14, pp. 3555–3565, 2002.

[13] R. Leardi and L. Nørgaard, "Sequential application of backward interval partial least squares and genetic algorithms for the selection of relevant spectral regions," *Journal of Chemometrics*, vol. 18, no. 11, pp. 486–497, 2004.

[14] Y. P. Du, Y. Z. Liang, J. H. Jiang, R. J. Berry, and 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*, vol. 501, no. 2, pp. 183–191, 2004.

[15] J. A. Cramer, K. E. Kramer, K. J. Johnson, R. E. Morris, and S. L. Rose-Pehrsson, "Automated wavelength selection for spectroscopic fuel models by symmetrically contracting repeated

unmoving window partial least squares," *Chemometrics and Intelligent Laboratory Systems*, vol. 92, no. 1, pp. 13–21, 2008.

[16] Q.-S. Xu and Y.-Z. Liang, "Monte Carlo cross validation," *Chemometrics and Intelligent Laboratory Systems*, vol. 56, no. 1, pp. 1–11, 2001.

[17] L. Xu, J.-H. Jiang, Y.-P. Zhou, H.-L. Wu, G.-L. Shen, and R.-Q. Yu, "MCCV stacked regression for model combination and fast spectral interval selection in multivariate calibration," *Chemometrics and Intelligent Laboratory Systems*, vol. 87, no. 2, pp. 226–230, 2007.

[18] D. Wolpert, "A mathematical theory of generalization: part I, part II," *Complex Systems*, vol. 4, pp. 151–200, 1990.

[19] L. Breiman, "Stacked regressions," *Machine Learning*, vol. 24, no. 1, pp. 49–64, 1996.

[20] R. A. Fisher, "Frequency distributions of the values of the correlation coefficient in samples from an indefinitely large population," *Biometrika*, vol. 10, no. 4, pp. 507–521, 1915.

[21] D. L. Hawkins, "Using *U* statistics to derive the asymptotic distribution of Fisher's *Z* statistic," *The American Statistician*, vol. 43, no. 4, pp. 235–237, 1989.

[22] F. Wülfert, W. Th. Kok, and A. K. Smilde, "Influence of temperature on vibrational spectra and consequences for the predictive ability of multivariate models," *Analytical Chemistry*, vol. 70, no. 9, pp. 1761–1767, 1998.

[23] R. D. Snee, "Validation of regression models: methods and examples," *Technometrics*, vol. 19, no. 4, pp. 415–428, 1977.

[24] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.

[25] M. Hubert and S. Verboven, "A robust PCR method for high-dimensional regressors," *Journal of Chemometrics*, vol. 17, no. 8-9, pp. 438–452, 2003.

[26] M. Hubert and K. Vanden Branden, "Robust methods for partial least squares regression," *Journal of Chemometrics*, vol. 17, no. 10, pp. 537–549, 2003.

[27] S. Serneels, C. Croux, P. Filzmoser, and P. J. Van Espen, "Partial robust M-regression," *Chemometrics and Intelligent Laboratory Systems*, vol. 79, no. 1-2, pp. 55–64, 2005.