Chemometric Studies in Near-Infrared Spectroscopy

Hongle An[1,2], Li Han[1,2], Yan Sun[1,2], Wensheng Cai[1,2], and Xueguang Shao[1,2(✉)]

[1] Research Center for Analytical Sciences, Frontiers Science Center for New Organic Matter, College of Chemistry, Tianjin Key Laboratory of Biosensing and Molecular Recognition, State Key Laboratory of Medicinal Chemical Biology, Nankai University, Tianjin 300071, China
xshao@nankai.edu.cn

[2] Haihe Laboratory of Sustainable Chemical Transformations, Tianjin 300192, China

Abstract. Near-infrared (NIR) spectroscopy has been a powerful technique for both qualitative and quantitative analysis. Due to the highly overlapping of the spectral bands, however, it is difficult to extract structural and quantitative information from the spectral data. Therefore, chemometric methods have been widely applied to enhance the spectral resolution or extract the spectral information, including modeling techniques, spectral preprocessing, variable selection, outlier detection, modeling transfer, etc. These methods provided colorful approaches for improving the models in both quantitative and discrimination analysis, greatly enhanced the applicability of NIR spectroscopy. On the other hand, temperature-dependent near-infrared spectroscopy was developed for analyzing liquid mixtures or aqueous systems. Chemometric methods were also established to build the quantitative models and extract the temperature-induced spectral variations. The former provided powerful tools for predicting the temperature or the concentration of the components, and the latter provided efficient approaches for understanding the structures and the interactions in chemical and biological samples or processes.

Keywords: Near-infrared spectroscopy · Chemometrics · Qualitative analysis · Quantitative analysis

1 Introduction

Near-infrared (NIR) spectroscopy, as a technique for rapid and non-destructive analysis with little or no sample preparation, has been widely used in different fields, such as environment, food, agriculture and industries [1–3]. However, the weak absorption and overlapping of the NIR spectral peaks make it difficult to provide the required spectral information for qualitative and quantitative analysis. Chemometric methods, therefore, have been adopted to improve the qualitative analysis and quantitative prediction of the technique. The first work in our group was published in 2002, in which the models for predicting the chemical components of tobacco samples using NIR spectra were established [4]. Since then, chemometric methods for improving the models of NIR spectra

X. Chu et al. (Eds.): ICNIR 2021, *Sense the Real Change: Proceedings of the 20th International Conference on Near Infrared Spectroscopy*, pp. 47–58, 2022.
https://doi.org/10.1007/978-981-19-4884-8_4

were developed [5, 6]. Besides partial least square (PLS) regression, consensus or ensemble modeling, multi-block strategies, and non-linear methods were studied, which may provide better prediction and interpretability in specific cases. NIR spectra contain not only the information of the components in the samples but also the spectral interferences, such as varying background and noise signals. To eliminate the irrelevant information, spectral preprocessing methods were studied. It has been proved that spectral preprocessing can effectively eliminate the interferences and improve the calibration model. Continuous wavelet transform (CWT) has been a commonly used method in our works to remove the variant background and enhance the spectral resolution [7–9]. Moreover, an NIR spectrum is composed of hundreds or even thousands of variables. The uninformative variables may degrade the model. Therefore, variable selection methods were proposed to select informative variables for building the predictive models [10–12]. These methods have been proved helpful for building sparse models, particularly in the case that the number of calibration samples is not enough. Furthermore, methods for calibration transfer were also studied for practical applications of NIR spectroscopy in the case that more than one instruments are needed.

From 2010, temperature-dependent NIR spectroscopy was developed for analyzing the structures and interactions in aqueous or bio-systems using the spectral variation induced by temperature. A quantitative spectra-temperature relationship (QSTR) model between NIR spectra of water and temperature was established [13, 14], and chemometric methods to extract quantitative and structural information from the temperature-dependent spectra were developed [15, 16]. High order chemometric algorithms were employed to extract the information from the high-dimensional spectral data. To simplify the calculation, multilevel simultaneous component analysis (MSCA) was studied and mutual factor analysis (MFA) was proposed for analyzing the high-dimensional data with the two-dimensional algorithm. Besides, chemometric algorithms were proposed to improve the resolution of NIR spectra or separate the mixed spectra of multi-component systems. Works were concentrated on capturing the temperature-induced spectral change of water, which can reflect the structures and interactions of water and solutes in aqueous solutions. Through the variation of water structures, the structure and the structural transformation of functional molecules such as proteins and thermo-responsive polymers were analyzed, and the role of water in the chemical and biological processes was revealed.

2 Modeling Techniques

PLS is the most commonly used linear calibration method due to its practicability and versatility. To improve the reliability of the models, particularly in the case that the calibration samples are not enough, ensemble modeling methods based on PLS were developed for modeling the NIR spectra [17–20]. For modeling the data set with large number of variables and making the model more interpretable, multiblock partial least squares (MB-PLS) method was proposed [18]. In the method, the spectra were decomposed into blocks by discrete wavelet transform (DWT), and the relative importance of the blocks was estimated by both the super-weights and the block-weights determined by the prediction error of the sub-models in cross validation. Two NIR data sets of tobacco

samples were investigated by the proposed method, and it was found that the weighted MB-PLS coupled with DWT gives a better predictive accuracy and interpretability compared with the ordinary PLS and MB-PLS methods. Besides, weighted-PLS regression method in multivariate calibration of NIR spectra was proposed [19]. In the approach, the spectra were decomposed into different scale blocks (or frequency components) by wavelet transform (WT) at first, and then PLS models were built with the decomposed components. A more precise prediction was obtained by the combined model compared with the ordinary PLS model. In weighted-PLS regression method, the spectra were split into groups of variables according to the statistic values of variables, i.e., the stability, which were used to evaluate the importance of variables in a calibration model [20]. Because the stability reflects the relative importance of the variables for modeling, these groups present different spectral information for the construction of PLS models. The two weighted-PLS methods were applied in modeling the NIR spectra of different complex samples. Compared with the results obtained with ordinary PLS method, the proposed methods are proved to be high-performance tools for multivariate calibration of NIR spectra.

In practical works, it was found that there is non-linear relationship between the NIR spectral response and the component content, which may make PLS model poor in predictive ability. To solve this problem, a local regression method based on WT was developed to improve the universality and prediction precision of the models [21]. In the algorithm, DWT was firstly utilized to compress the NIR spectra, and then, the calibration subsets were individually selected for each prediction sample according to the Euclidean distance in wavelet domain. The method was used to quantitative determination of chlorine in tobacco samples. The results obtained by the method are superior to that of the global PLS method and principal component analysis (PCA) based local regression method. Moreover, a consensus least squares support vector regression (LS-SVR) method was proposed based on the principle of consensus modeling [22]. In the proposed method, NIR spectra of tobacco lamina samples were firstly preprocessed using DWT for eliminating the spectral background and noise, then, consensus LS-SVR technique was used for building the calibration model. With an optimization of the parameters involved in the modeling, a satisfied model was achieved for predicting the content of reducing sugar in tobacco lamina samples.

3 Preprocessing

NIR spectra contain not only the information of the components to be analyzed but also the light scattering and fluctuating background. Therefore, appropriate spectral preprocessing is vital to eliminate the irrelevant information before multivariate calibration. Methods such as multiplicative scatter correction (MSC), standard normal variate (SNV), Savitzky-Golay (SG) smoothing and orthogonal signal correction (OSC), were developed to remove spectral interference [23–26]. The influence of scattering caused by uneven distribution of sample particles and different particle sizes can be effectively eliminated by these methods. In addition, WT is a useful signal processing method and widely applied in spectral analysis. WT with a wavelet function can be regarded as a smoothing and differentiation process. Owing to the characteristic of the double localization in position and frequency domains, WT can decompose a signal into localized

contributions representing the information of different frequencies. Among these contributions, there are contributions which represent the high resolution signals because their frequency is higher than the lower frequency background and lower than the high frequency noise. Based on this property, WT can be often employed for data compression, data smoothing, baseline correction, resolution of multicomponent overlapping signals, etc. In our previous works, WT has been proved to be an efficient tool for removing the variant background and noise. For example, a hybrid algorithm was developed based on the utilization of multi-resolution, which is one of the most important advantages of WT [27]. The spectral signals were split into different frequency components, and the number of spectral data points remained unchanged. The method was applied to the simulated data and experimental NIR spectral data. It was found that the method can be used to remove the low-frequency background and the high-frequency noise simultaneously with the help of WT. CWT was also applied as the pretreatment method to eliminate the background in the discrimination of pharmaceutical products [28]. Furthermore, CWT was found to be an efficient tool due to its advantage in smoothing and flexibility for calculating high order derivatives, and the fourth order derivative was proved to be a good choice for improving resolution as well as reducing the noise and sidelobe effects [29].

4 Variable Selection

NIR spectral data matrix is often quite large and some variables may be irrelevant to the multivariate calibration, which reduce the quality of the model. Eliminating uninformative variables can simplify calibration modeling and improve the accuracy and robustness of prediction models. In order to obtain better quantitative calibration models, several variable selection methods for selecting characteristic wavelengths of NIR spectra have been developed, such as uninformative variable elimination (UVE), competitive adaptive reweighted sampling (CARS) and genetic algorithm (GA) [30–32].

A variable selection method, named as MC-UVE, was proposed based on the principle of Monte Carlo (MC) and UVE, which is applied in the quantitative analysis of NIR spectral data. Firstly, a large number of regression models were established with corresponding training subsets selected by MC technique, and then each variable was sorted based on the stability of the PLS coefficients in these models. The variables with poor stability were removed as uninformative variables. The method was used to analyze the contents of nicotine and sugar in tobacco samples. The results showed that this method can select important wavelengths and make the prediction more precise compared with UVE-PLS and conventional PLS [10]. Moreover, MC-UVE combined with successive projections algorithm (SPA) was proved to be an effective way to generate variable subsets by removing the uninformative variables from NIR spectra of complex lamina samples [33]. In addition, the silver substrate can reduce spectral interference and enhance the spectral response, and the detection performance can be further enhanced with the help of chemometric methods. To obtain satisfactory PLS model between the spectra and concentration of lysozyme solutions, MC-UVE method was used to eliminate uninformative variables [34].

An effective variable selection method named as RT-PLS was proposed based on randomization test for NIR spectral analysis. The importance of the variables can be

evaluated by a statistic P, and the variable will be removed as an uninformative variable if P value is greater than the threshold. With applications of RT-PLS for the quantitative analysis of corn and tobacco samples, it was proved that RT may be a good alternative for multivariate analysis [11]. RT can also be employed as a gene selection method to deal with gene expression data [35]. Besides, latent projective graph (LPG) was adopted in variable selection for NIR spectral analysis [36]. The method was based on the assumption that collinear variables may have the same contribution to the modeling. An LPG was calculated by performing PCA on an NIR spectral data firstly, and then informative variables were selected from LPG. The method was applied in three NIR datasets of pharmaceutical tablets, blood and tobacco samples, and it was shown that accurate models are built by using only the variables at the inflections of LPG. Moreover, a method for variable selection was developed based on the detection of the influential variables (IVs) [37]. PCA was used to insight into the clustering of the models built with different combination of the spectral wavelengths. The IVs can be distinguished by the frequency number of variables. From the results of five datasets, it was proved that parsimonious and precise models can be obtained by the proposed method. Furthermore, a variable importance criterion named as C was developed to determine the importance of variables in the multivariate calibration model. The value of C was a measurement of the average contribution for a variable to the prediction error of a model. The multi-step shrinkage strategy can further improve the variable selection effect of the C value. The criterion was applied in the diesel fuel dataset and two blood datasets, and the results demonstrated that the value of C is an effective parameter for selecting the important wavelengths compared with MC-UVE, RT and CARS [12].

5 Outlier Detection

Outliers contained in the calibration data set are usually caused by the instrument, operation, and samples preparation and the outliers may interfere with the prediction accuracy of the multivariate calibration model in NIR spectral analysis. In order to detect outliers and improve the quality of the models, a large number of alternative approaches were proposed for outlier detection. A useful algorithm named as CWT-mIPOW-PLS for simultaneous outlier detection and variable selection was proposed for NIR spectra [38]. The mIPOW-PLS was proposed to remove both the useless wavelengths and the multiple outliers in CWT domain. The calibration models based on NIR spectra of sugar aqueous solutions and tobacco samples were built accurately by using this method. Moreover, an extension of the Kennard-Stone (EKS) algorithm was used to extract the optimal samples in the CWT domain, which can effectively reduce spectral interference and the number of samples for obtaining high-quality calibration model [39]. There is another algorithm to detect outliers in NIR spectra analysis. A large number of PLS models were constructed by random test cross validation firstly, and then the models were sorted by the prediction residual error sum of squares. The outliers can be detected by the plot of the accumulative probability which can be gotten from the sorted PLS models [40]. Applying this method to wheat, gasoline, corn and tobacco lamina samples datasets, the method was proved to be a more efficient method for detecting outliers compared with the conventional LOOCV method. In addition, an improved boosting

PLS was proposed to enhance the prediction ability of the quantitative model when the outliers existed [17]. A robust step was added to modify sampling weight for weakening the effect of the outliers in the models.

6 Model Transfer

Model transfer is essential for practical applications of NIR spectroscopy because differences may exist between the spectra measured using different instruments. Model transfer methods for correcting prediction error without using the spectra standard samples were developed. Based on the assumption that spectral differences of different instruments and the prediction error were linear, a useful method without standard samples was developed according to the linear relationship [41]. The linear model correction (LMC) method was proved to be an efficient way for transferring the models between different instruments in NIR spectra analysis when standard samples are not available [42]. The method assumes that the NIR spectra measured on different instruments are linearly correlated if the samples are similar. The coefficients of the master model can be applied in the slave model by using the constrained optimization method. Furthermore, a modified method based on LMC was proposed for improving the model transfer accuracy and computational efficiency [43]. This method was used to analyze the NIR spectra data of pharmaceutical tablets from different instruments. It was found that the method can achieve efficient transfer and the calculation can be simplified by the modification.

7 Discrimination Analysis

Discrimination analysis of samples, especially complex actual products, is a significant subject in many fields. PCA is one of the most important pattern recognition methods, which can reduce spectral data dimensionality and extract distinguishing features information from NIR spectra. In order to discriminate azithromycin tablets from four manufacturers, PCA was performed after CWT preprocessing and variable selection of NIR diffuse reflectance spectra [44]. The results showed that the classification of manufacture sites was acceptable. A method was developed to discriminate the Chinese patent medicines using NIR spectroscopy and principal component discriminant transformation [45]. The optimal set of orthogonal discriminant vectors was designed by maximizing Fisher's discriminant function. The discrimination models were very accurate for the discrimination of different types of medicines. To distinguish different brands of tobacco products, methods were proposed using the NIR spectra datasets. PCA combined with hierarchical cluster analysis (HCA) can accurately determine the brand of tobacco products [46]. In addition, the principal component accumulation (PCAcc) method was used for the multiclass problem of tobacco samples [47]. It was found that the method can discriminate different parts of tobacco leaves and different brands of cigarettes accurately. Moreover, PCA combined other chemometric methods is often applied in the rapid disease diagnosis using NIR spectra of human serum samples [48, 49]. For example, PCA, DWT, linear discriminant analysis (LDA) and partial least squares discriminant analysis (PLSDA) were used for discrimination of the sera from healthy and possible kidney patients, and satisfactory classification was obtained.

8 Extracting Information from Temperature-Dependent Near-Infrared Spectra

Temperature-dependent NIR spectroscopy is a technique for measuring the NIR spectra of a sample at different temperatures. Temperature effect can be considered as a source of information for multivariate spectroscopic analysis. A QSTR model between NIR spectra and the temperature was established based on PLS regression and applied to the quantitative determination of the compositions in aqueous solutions of methanol, ethanol, n-hexane, and their mixtures [13, 14]. The temperature-dependent NIR spectra measured at different conditions generally generate high-dimensional data. Thus, high order chemometric algorithms are attractive for dealing with such data. N-way principal component analysis (NPCA), parallel factor analysis (PARAFAC) and alternating trilinear decomposition (ATLD) were adopted to explore the spectral information from the temperature-dependent NIR spectra of a binary water-ethanol and a ternary water-ethanol-isopropanol mixtures [50]. The temperature- and concentration-induced spectral variations were obtained and the quantitative model was successfully built. In order to simplify the calculation, two-dimensional algorithms were developed for analyzing the high-dimensional data. Multilevel simultaneous component analysis (MSCA), which unfolds a data array into a two-dimensional matrix, was used to study temperature-dependent NIR spectra. A two-level MSCA model was employed to capture the temperature- and concentration-induced spectral variations of aqueous solutions and real serum samples [51, 52]. Moreover, quantitative analysis of the NIR spectra for aqueous proline solution under multiple perturbations was investigated by a three level MSCA [53]. The spectral changes induced by pH, concentration and temperature were discribed by the three level models, respectively. A new method, mutual factor analysis (MFA), was also developed [54, 55]. The method unfolds a high-dimensional data array into a combined data matrix and then extracts the common spectral feature contained in the spectra of different temperatures or different concentrations by PCA. The relative quantity of the extracted spectral feature can be used to build the quantitative model. The method was employed for the quantitative determination of glucose solutions and serum samples. A calibration model with a good correlation coefficient was obtained for the measurement of the glucose content.

The structure of water has been an interesting subject in chemistry and biology for decades due to the complexity and flexibility of hydrogen bonding. With the help of chemometric methods, the spectral features of different water structures can be obtained. To investigate the effect of the temperature on the NIR spectra of water, a method combined CWT and MC-UVE was proposed for the selection of the temperature-dependent variables (wavenumbers) from the NIR spectra measured at different temperatures [56]. Seven variables with a significant temperature dependency can be found and the variables for different solutions are not identical, indicating that the variables can be used to discriminate different solutions. To understand water structures in liquid water and aqueous solutions, Gaussian fitting was adopted to analyze the temperature-dependent NIR spectra of water [57]. The spectral components corresponding to the nine water structures with different hydrogen were obtained from the spectra of water and glucose solutions by Gaussian fitting with a knowledge-based genetic algorithm [58]. The knowledge about the variation of these water structures with temperature was included

in the method. Through the structural variation with temperature and glucose concentration, the dissociation of the water clusters and the enhancement of tetrahedral water structure was observed, respectively. Alcohols and amines can be regarded as models for investigating the hydrogen-bonded interactions. The spectral features of eight alcohol structures were identified from the resolution-enhanced spectra calculated by CWT [59]. The stability of the aggregates was obtained from the temperature effect and a sequence of the stability was deduced. Independent component analysis (ICA) was adopted for analyzing the temperature effect on the spectra of primary aliphatic amines [60]. Three independent components (ICs) corresponding to the spectral information of the free, linearly and cyclically hydrogen-bonded NH groups, respectively, were obtained. With the reconstructed spectra from the ICs, the variation of the three forms of NH groups with concentration and temperature was observed.

The interaction of water and solutes is of great significance for understanding the properties of aqueous solutions or bio-systems. The interaction in alcohol-water mixtures has been studied using temperature-dependent NIR spectroscopy. A useful method was proposed based on the rotation of the loadings in PCA [61]. The spectra of ethanol and water were calculated from the spectra of the mixtures and the calculated spectra were found to reflect the structure in the mixture. Moreover, the structure of water at low temperatures and the mechanism of the cryoprotectant dimethyl sulfoxide (DMSO) in reducing the freezing point were investigated [62, 63]. From the resolution-enhanced spectra by CWT, the hydrogen bonding of DMSO and water was found and the interaction was weakened by the existence of FA (a model compound of protein) exists in the mixture. However, the variation of spectral feature with the content of FA was not found by ATLD, implying that, although FA may slightly reduce the anti-freezing effect, DMSO is still the key component to prevent water from icing. The interaction of water and protein or polymer plays a key role in the phase transition of the macromolecules in aqueous solutions or bio-organisms. Combined with two-dimensional correlation NIR spectroscopy and Gaussian fitting, the spectral variations of different water structures during the gelation of ovalbumin were observed and the denaturation mechanism was elucidated [64]. The variation of hydration water during the aggregation of the core fragment of tau, R2/wt, induced by heparin, was also investigated [65]. The spectral features of water structures around NH and CH groups were found in the loadings of PCA, and the variation of the structures during the aggregation was analyzed by 2D correlation spectroscopy. To understand the function of water on the thermal stability of the proteins in a confined environment, the water structures in reverse micelles (RMs) were studied [66]. The spectral feature of bridging water, which connects protein and the inner surface of RMs, was found by PCA. The bridging water may be the reason for enhancing the thermal stability of the protein in RMs. Temperature-sensitive polymers exhibit phase separation in aqueous solutions above the lower critical solution temperature (LCST). The interaction of water and the polymer is supposed to be the key factor in driving the aggregation. The interactions during the aggregation of poly(N,N-dimethylaminoethyl methacrylate) (PDMAEMA) and poly(N-isopropyl acrylamide) (PNIPAM) were investigated [67, 68]. NPCA was used to extract the information of spectral variation with temperature and concentration. It was found that, during the phase transition of PDMAEMA, the water molecules connecting the polymer chains in the loose hydrophobic structure by two

hydrogen bonds are important for forming intermediate state. The water structure with three hydrogen bonds plays a key role in the stabilization of PNIPAM, which may connect the NH and CO groups in the polymer. When urea is added, the water species are destroyed, thus leading to a phase transition at a lower temperature.

9 Conclusion

NIR spectroscopy is a powerful tool for quantitative determination and structural analysis in complex systems. Chemometric methods must be applied in the spectral analysis due to the complexity of the spectra. The methods of modeling techniques, spectral preprocessing, variable selection, outlier detection, modeling transfer and discrimination analysis were developed for NIR spectra modeling. In these methods, PLS is the most widely applied modeling method. Besides, CWT, MC-UVE, RT and PCA are also powerful tools to perform spectral analysis. Furthermore, quantitative and structural information can be extracted from temperature-dependent NIR spectra by chemometric methods. Quantitative models between the spectra and the temperature or concentration can be established. Using the spectral variation with temperature, the changes of structure and interaction in liquid mixtures or aqueous systems can be observed. With the development of chemometrics, NIR spectroscopy can be a promising technique for understanding the properties or functions of compounds in chemical and biological systems.

Acknowledgments. This study was supported by the National Natural Science Foundation of China (No. 22174075), the Natural Science Foundation of Tianjin, China (No. 20JCYBJC01480), the Frontiers Science Center for New Organic Matter, Nankai University (No. 63181206), the Fundamental Research Funds for the Central Universities, Nankai University (No. 63211019) and the Haihe Laboratory of Sustainable Chemical Transformations.

References

1. Moros, J., Garrigues, S., de la Guardia, M.: Vibrational spectroscopy provides a green tool for multi-component analysis. Trends Anal. Chem. **29**, 578–591 (2010)
2. Blanco, M., Villarroya, I.: NIR spectroscopy: a rapid-response analytical tool. Trends Anal. Chem. **21**, 240–250 (2002)
3. Pasquini, C.: Near infrared spectroscopy: a mature analytical technique with new perspectives - a review. Anal. Chim. Acta **1026**, 8–36 (2018)
4. Wang, F., Chen, D., Shao, X.G.: Study on model of near-infrared spectroscopy and chemical components of cigarettes. Tob. Sci. Technol. **5**, 23–26 (2002)
5. Shao, X.G., Bian, X.H., Liu, J.J., Zhang, M., Cai, W.S.: Multivariate calibration methods in near infrared spectroscopic analysis. Anal. Methods **2**, 1662–1666 (2010)
6. Zhang, J., Cai, W.S., Shao, X.G.: New algorithms for calibration transfer in near infrared spectroscopy. Prog. Chem. **29**, 902–910 (2017)
7. Shao, X.G., Cai, W.S.: Wavelet analysis in analytical chemistry. Rev. Anal. Chem. **17**, 235–285 (1998)
8. Shao, X.G., Leung, A.K.M., Chau, F.T.: Wavelet: a new trend in chemistry. Acc. Chem. Res. **36**, 276–283 (2003)

9. Ma, C.X., Shao, X.G.: Continuous wavelet transform applied to removing the fluctuating background in near-infrared spectra. J. Chem. Inf. Comput. Sci. **44**, 907–911 (2004)
10. Cai, W.S., Li, Y.K., Shao, X.G.: A variable selection method based on uninformative variable elimination for multivariate calibration of near-infrared spectra. Chemom. Intell. Lab. Syst. **90**, 188–194 (2008)
11. Xu, H., Liu, Z.C., Cai, W.S., Shao, X.G.: A wavelength selection method based on randomization test for near-infrared spectral analysis. Chemom. Intell. Lab. Syst. **97**, 189–193 (2009)
12. Zhang, J., Cui, X., Cai, W., Shao, X.: A variable importance criterion for variable selection in near-infrared spectral analysis. Sci. China Chem. **62**(2), 271–279 (2018). https://doi.org/10.1007/s11426-018-9368-9
13. Shao, X.G., Kang, J., Cai, W.S.: Quantitative determination by temperature dependent near-infrared spectra. Talanta **82**, 1017–1021 (2010)
14. Kang, J., Cai, W.S., Shao, X.G.: Quantitative determination by temperature dependent near-infrared spectra: a further study. Talanta **85**, 420–424 (2011)
15. Cui, X., Sun, Y., Cai, W., Shao, X.: Chemometric methods for extracting information from temperature-dependent near-infrared spectra. Sci. China Chem. **62**(5), 583–591 (2019). https://doi.org/10.1007/s11426-018-9398-2
16. Sun, Y., Cai, W.S., Shao, X.G.: Chemometrics: an excavator in temperature-dependent near-infrared spectroscopy. Molecules **27**, 452 (2022)
17. Shao, X.G., Bian, X.H., Cai, W.S.: An improved boosting partial least squares method for near-infrared spectroscopic quantitative analysis. Anal. Chim. Acta **666**, 32–37 (2010)
18. Jing, M., Cai, W.S., Shao, X.G.: Multiblock partial least squares regression based on wavelet transform for quantitative analysis of near infrared spectra. Chemom. Intell. Lab. Syst. **100**, 22–27 (2010)
19. Liu, Z.C., Cai, W.S., Shao, X.G.: A weighted multiscale regression for multivariate calibration of near infrared spectra. Analyst **134**, 261–266 (2009)
20. Xu, H., Cai, W.S., Shao, X.G.: Weighted partial least squares regression by variable grouping strategy for multivariate calibration of near infrared spectra. Anal. Methods **2**, 289–294 (2010)
21. Shi, X., Cai, W.S., Shao, X.G.: Local regression method in wavelet domain and its application in near-infrared quantitative analysis, Chinese. J. Anal. Chem. **36**, 1093–1096 (2008)
22. Li, Y.K., Shao, X.G., Cai, W.S.: A consensus least squares support vector regression (LS-SVR) for analysis of near-infrared spectra of plant samples. Talanta **72**, 217–222 (2007)
23. Helland, I.S., Naes, T., Isaksson, T.: Related versions of the multiplicative scatter correction method for preprocessing spectroscopic data. Chemom. Intell. Lab. Syst. **29**, 233–241 (1995)
24. Barnes, R.J., Dhanoa, M.S., Lister, S.J.: Standard normal variate transformation and detrending of near-infrared diffuse reflectance spectra. Appl. Spectrosc. **43**, 772–777 (1989)
25. Savitzky, A., Golay, M.J.E.: Smoothing and differentiation of data by simplified least squares procedures. Anal. Chem. **36**, 1627–1639 (1964)
26. Sjoblom, J., Svensson, O., Josefson, M., Kullberg, H., Wold, S.: An evaluation of orthogonal signal correction applied to calibration transfer of near infrared spectra. Chemom. Intell. Lab. Syst. **44**, 229–244 (1998)
27. Chen, D., Shao, X.G., Hu, B., Su, Q.D.: A background and noise elimination method for quantitative calibration of near infrared spectra. Anal. Chim. Acta **511**, 37–45 (2004)
28. Shan, R.F., Mao, Z.Y., Yin, L.H., Cai, W.S., Shao, X.G.: Discrimination of Chinese patent medicines using near-infrared spectroscopy and principal component accumulation method. Anal. Methods **6**, 4692–4697 (2014)
29. Shao, X.G., Cui, X.Y., Wang, M., Cai, W.S.: High order derivative to investigate the complexity of the near infrared spectra of aqueous solutions. Spectrochim. Acta Part A **213**, 83–89 (2019)
30. Centner, V., Massart, D.L.: Elimination of uninformative variables for multivariate calibration. Anal. Chem. **68**, 3851–3858 (1996)

31. Li, H.D., Liang, Y.Z., Xu, Q.S., Cao, D.S.: Key wavelengths screening using competitive adaptive reweighted sampling method for multivariate calibration. Anal. Chim. Acta **648**, 77–84 (2009)
32. Lucasius, C.B., Kateman, G.: Genetic algorithms for large-scale optimization in chemometrics: an application. Trends Anal. Chem. **10**, 254–261 (1991)
33. Du, G.R., Cai, W.S., Shao, X.G.: A variable differential consensus method for improving the quantitative near-infrared spectroscopic analysis. Sci. China: Chem. **55**, 1946–1952 (2012)
34. Wang, C.C., Wang, S.Y., Cai, W.S., Shao, X.G.: Silver mirror for enhancing the detection ability of near-infrared diffuse reflectance spectroscopy. Talanta **162**, 123–129 (2017)
35. Mao, Z.Y., Cai, W.S., Shao, X.G.: Selecting significant genes by randomization test for cancer classification using gene expression data. J. Biomed. Inf. **46**, 594–601 (2013)
36. Shao, X.G., Du, G.R., Jing, M., Cai, W.S.: Application of latent projective graph in variable selection for near infrared spectral analysis. Chemom. Intell. Lab. Syst. **114**, 44–49 (2012)
37. Shao, X.G., Zhang, M., Cai, W.S.: Multivariate calibration of near-infrared spectra by using influential variables. Anal. Methods **4**, 467–473 (2012)
38. Chen, D., Shao, X.G., Hu, B., Su, Q.D.: Simultaneous wavelength selection and outlier detection in multivariate regression of near-infrared spectra. Anal. Sci. **21**, 161–166 (2005)
39. Chen, D., Cai, W.S., Shao, X.G.: An adaptive strategy for selecting representative calibration samples in the continuous wavelet domain for near-infrared spectral analysis. Anal. Bioanal. Chem. **387**, 1041–1048 (2007)
40. Liu, Z.C., Cai, W.S., Shao, X.G.: Outlier detection in near-infrared spectroscopic analysis by using Monte Carlo cross-validation. Sci. China Ser. B Chem. **51**, 751–759 (2008)
41. Li, X.Y., Cai, W.S., Shao, X.G.: Correcting multivariate calibration model for near infrared spectral analysis without using standard samples. J. Near Infrared Spectrosc. **23**, 285–291 (2015)
42. Liu, Y., Cai, W.S., Shao, X.G.: Linear model correction: a method for transferring a near-infrared multivariate calibration model without standard samples. Spectrochim. Acta Part A **169**, 197–201 (2016)
43. Zhang, J., Cui, X.Y., Cai, W.S., Shao, X.G.: Modified linear model correction: a calibration transfer method without standard samples. NIR News **29**, 24–27 (2018)
44. Li, P., Du, G.R., Cai, W.S., Shao, X.G.: Rapid and nondestructive analysis of pharmaceutical products using near-infrared diffuse reflectance spectroscopy. J. Pharm. Biomed. Anal. **70**, 288–294 (2012)
45. Xu, Z.H., Liu, Y., Li, X.Y., Cai, W.S., Shao, X.G.: Discriminant analysis of Chinese patent medicines based on near-infrared spectroscopy and principal component discriminant transformation. Spectrochim. Acta Part A **149**, 985–990 (2015)
46. Liu, J.J., Xu, H., Cai, W.S., Shao, X.G.: Discrimination of industrial products by on-line near infrared spectroscopy with an improved dendrogram. Chin. Chem. Lett. **22**, 1241–1244 (2011)
47. Wang, Y., Ma, X., Wen, Y.D., Liu, J.J., Cai, W.S., Shao, X.G.: Discrimination of plant samples using near-infrared spectroscopy with a principal component accumulation method. Anal. Methods **4**, 2893–2899 (2012)
48. Fan, M., Liu, X., Yu, X., Cui, X., Cai, W., Shao, X.: Near-infrared spectroscopy and chemometric modelling for rapid diagnosis of kidney disease. Sci. China Chem. **60**(2), 299–304 (2016). https://doi.org/10.1007/s11426-016-0092-6
49. Cui, X.Y., Yu, X.M., Cai, W.S., Shao, X.G.: Water as a probe for serum-based diagnosis by temperature-dependent near-infrared spectroscopy. Talanta **204**, 359–366 (2019)
50. Cui, X.Y., Zhang, J., Cai, W.S., Shao, X.G.: Chemometric algorithms for analyzing high dimensional temperature dependent near infrared spectra. Chemom. Intell. Lab. Syst. **170**, 109–117 (2017)

51. Cui, X.Y., Liu, X.W., Yu, X.M., Cai, W.S., Shao, X.G.: Water can be a probe for sensing glucose in aqueous solutions by temperature dependent near infrared spectra. Anal. Chim. Acta **957**, 47–54 (2017)
52. Shan, R.F., Zhao, Y., Fan, M.L., Liu, X.W., Cai, W.S., Shao, X.G.: Multilevel analysis of temperature dependent near-infrared spectra. Talanta **131**, 170–174 (2015)
53. Han, L., Cui, X.Y., Cai, W.S., Shao, X.G.: Three-level simultaneous component analysis for analyzing the near-infrared spectra of aqueous solutions under multiple perturbations. Talanta **217**, 121036 (2020)
54. Shao, X.G., Cui, X.Y., Yu, X.M., Cai, W.S.: Mutual factor analysis for quantitative analysis by temperature dependent near infrared spectra. Talanta **183**, 142–148 (2018)
55. Wang, M.Y., Cui, X.Y., Cai, W.S., Shao, X.G.: Temperature-dependent near-infrared spectroscopy for sensitive detection of glucose. Acta Chim. Sinica **78**, 125–129 (2020)
56. Cui, X.Y., Zhang, J., Cai, W.S., Shao, X.G.: Selecting temperature-dependent variables in near-infrared spectra for aquaphotomics. Chemom. Intell. Lab. Syst. **183**, 23–28 (2018)
57. Cui, X.Y., Cai, W.S., Shao, X.G.: Glucose induced variation of water structure from temperature dependent near infrared spectra. RSC Adv. **6**, 105729–105736 (2016)
58. Tan, J.H., et al.: Knowledge-based genetic algorithm for resolving the near-infrared spectrum and understanding the water structures in aqueous solution. Chemom. Intell. Lab. Syst. **206**, 104150 (2020)
59. Sun, Y., Cui, X.Y., Cai, W.S., Shao, X.G.: Understanding the complexity of the structures in alcohol solutions by temperature-dependent near-infrared spectroscopy. Spectrochim. Acta Part A **229**, 117864 (2020)
60. Zhu, X.W., Cui, X.Y., Cai, W.S., Shao, X.G.: Temperature dependent near infrared spectroscopy for understanding the hydrogen bonding of amines. Acta Chim. Sinica **76**, 298–302 (2018)
61. Shao, X.G., Cui, X.Y., Liu, Y., Xia, Z.Z., Cai, W.S.: Understanding the molecular interaction in solutions by chemometric resolution of near-infrared spectra. ChemistrySelect **2**, 10027–10032 (2017)
62. Zhao, H.T., Sun, Y., Guo, Y.C., Cai, W.S., Shao, X.G.: Near infrared spectroscopy for low-temperature water structure analysis. Chem. J. Chin. Univ. **41**, 1968–1974 (2020)
63. Su, T., Sun, Y., Han, L., Cai, W.S., Shao, X.G.: Revealing the interactions of water with cryoprotectant and protein by near-infrared spectroscopy. Spectrochim. Acta Part A **266**, 120417 (2022)
64. Ma, L., Cui, X.Y., Cai, W.S., Shao, X.G.: Understanding the function of water during the gelation of globular proteins by temperature-dependent near infrared spectroscopy. Phys. Chem. Chem. Phys. **20**, 20132–20140 (2018)
65. Sun, Y., Ma, L., Cai, W.S., Shao, X.G.: Interaction between tau and water during the induced aggregation revealed by near-infrared spectroscopy. Spectrochim. Acta Part A **230**, 118046 (2020)
66. Wang, S.Y., Wang, M., Han, L., Sun, Y., Cai, W.S., Shao, X.G.: Insight into the stability of protein in confined environment through analyzing the structure of water by temperature-dependent near-infrared spectroscopy. Spectrochim. Acta Part A **267**, 120581 (2022)
67. Wang, L., Zhu, X.W., Cai, W.S., Shao, X.G.: Understanding the role of water in the aggregation of poly(N, N-dimethylaminoethyl methacrylate) in aqueous solution using temperature-dependent near-infrared spectroscopy. Phys. Chem. Chem. Phys. **21**, 5780–5789 (2019)
68. Ma, B., Wang, L., Han, L., Cai, W.S., Shao, X.G.: Understanding the effect of urea on the phase transition of poly(N-isopropylacrylamide) in aqueous solution by temperature-dependent near-infrared spectroscopy. Spectrochim. Acta Part A **253**, 119573 (2021)