frontiers | Frontiers in Chemistry

ORIGINAL RESEARCH
published: 07 April 2022
doi: 10.3389/fchem.2022.812567

How to Resolve the Maximum Valuable Information in Complex NIR Signal: A Practicable Method Based on Wavelet Transform

*Jing Chen and Xiaoquan Lu**

Key Lab of Bioelectrochemistry and Environmental Analysis of Gansu, College of Chemistry and Chemical Engineering, Northwest Normal University, Lanzhou, China

A key problem in the field of near infrared (NIR) spectrum study is to obtain the valuable information from the complex NIR signal. A maximum information extraction method based on Wavelet Transform (WT) is proposed in this paper for helping the relative researchers to resolve the signal. The results show that the method can serve as an effective tool for obtaining the maximum valuable information in NIR study.

Keywords: wavelet transform (CWT), near infrared spectrum, uninformative variable elimination, residual error sum of square, root mean square error

OPEN ACCESS

Edited by:
Erdal Dinç,
Ankara University, Turkey

Reviewed by:
Özgür Üstündağ,
Ankara University, Turkey
Agnieszka Martyna,
University of Silesia in Katowice, Poland

****Correspondence:***
Xiaoquan Lu
luxq@nwnu.edu.cn

Specialty section:
This article was submitted to Analytical Chemistry, a section of the journal Frontiers in Chemistry

Received: *10 November 2021*
Accepted: *25 February 2022*
Published: *07 April 2022*

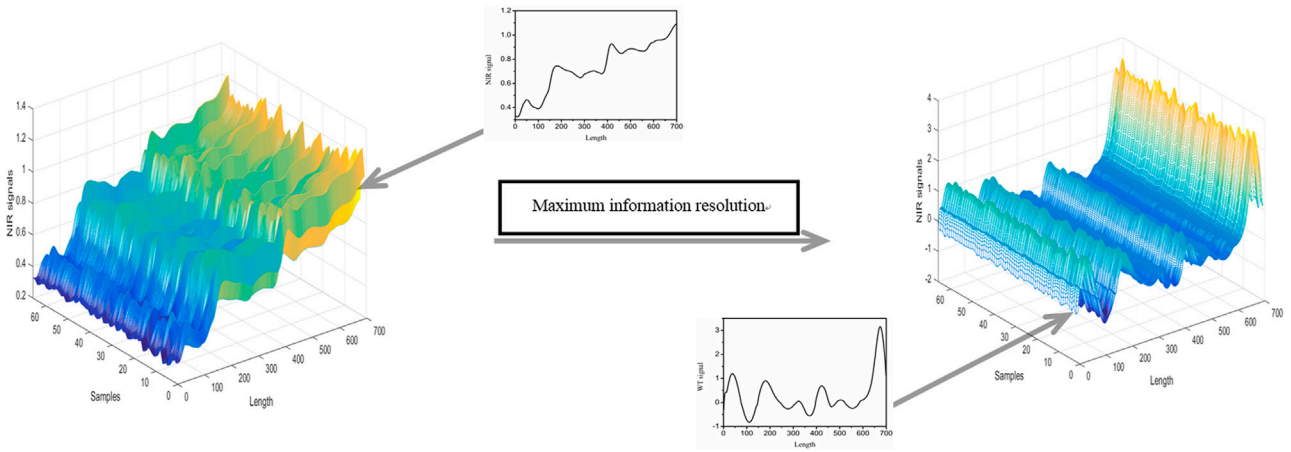
Citation:
Chen J and Lu X (2022) How to Resolve the Maximum Valuable Information in Complex NIR Signal: A Practicable Method Based on Wavelet Transform. Front. Chem. 10:812567. doi: 10.3389/fchem.2022.812567

INTRODUCTION

With the advantages of nondestructive measurement, rapidity and simplicity, near-infrared (NIR) spectroscopy has been widely applied to measure samples in the industries of food (Stenlund et al., 2009) and pharmaceutical (Abrahamsson et al., 2005) and the agricultural products (Pedro and Ferreira, 2005; Ozaki et al., 2006). However, the spectral signals of samples which are interfered by background and noise are always seriously overlapping and contain some variations irrelevant to concentration (Rutan et al., 1998). The key problem is how to extract valuable information from these complex spectral bands in the NIR region.

Multivariate calibration models which have been successfully applied to analyze NIR spectral data have greatly developed NIR applications (Inácioa et al., 2013; de Oliveira et al., 2014; Goodarzi et al., 2015; Pan et al., 2015; Yun et al., 2015; Eskildsen et al., 2016). A reliable calibration model is created by sufficient spectral data to assure the predicting accuracy of test set. The weaker the analytical signal of calibration and prediction set is, the worse the model's predicting accuracy is. Some efforts have been explored to squeeze the complex NIR signal by eliminating "uninformative" signal points. Among them, uninformative variable elimination (UVE) (Centner et al., 1996) has been successfully applied as an classical method. A "stability" is defined in the method to estimate the significance of each signal point, and a cut-off threshold is generated by regression coefficients based on a random variable matrix with small amplitude. Many "uninformative" signal points are eliminated according to this cut-off threshold value. There is the strong possibility to miss some significant signal points because the signal is overlapped seriously.

As an effective mathematical microscope, Wavelet Transform (WT) is very helpful for enlarging the signal details. Here, it is used to extract the maximum information by resolving the original spectrum signal. Then, the signal is reconstructed by the resolved signal before constructing model. The method is a valuable tool for the relative researchers.



GRAPHICAL ABSTRACT |

METHODS

Theory and Algorithm

The Continuous Wavelet Transform (CWT) of the signal (or data) $f(x)$ is defined as:

$$W(a,b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} f(x)\psi^*\left(\frac{x-b}{a}\right)dx \qquad (1)$$

Where $W(a,b)$ is the CWT of $f(x)$, a (a>0; a∈R) is the scaling factor, b (b∈R) is the window factor, and ψ^* is the wavelet which is the dilation and translation of the mother wavelet (Chau et al., 2004; Kalteh, 2013; Subaie and Mourou, 2013; Yuan et al., 2014; de Yong et al., 2015; Martyna et al., 2015; Yu et al., 2015). With the progressive increasement or decreasement of the scale, the wavelet changes regularly. As shown in **Figure 1**, with the increasement of a from 1 to 40, the Mexh wavelet becomes shorter and wider.

It has been widely confirmed that the WT can resolve valuable information in the signal, such as the resolution of overlapping peaks and the cancellation of background and noise (Dinc et al., 2006; Jena et al., 2014; Fu et al., 2015; Lopes-dos-Santos et al., 2015; Dinç and Yazan, 2018). WT is actually the convolution of the initial signal $f(x)$ and a special wavelet at a scale value. Since the scale a can be a series of consecutive integers (PATHAK and SINGH, 2016), the WT results of the initial signal are spread into a three dimensional space to show the signal details more clearly.

When the wavelet maximum overlaps with the signal at a signal point, the convolution result maximum presents the point information. Our WT program obtained the same results with some commercial softwares. If the scale a is set as a fixed value, the wavelet cannot usually maximum overlap with the whole signal at each point. However, in a scale range, the wavelet can maximum overlap with each signal point by the change of wavelet. Therefore, the maximum and minimum WT value of the signal in a scale range are used here to reconstruct signal to present the maximum information at each signal point. The complex NIR signal will be used here. The troughs in the signal may contain some important information. So, the minimum WT values at some signal points are also considered.

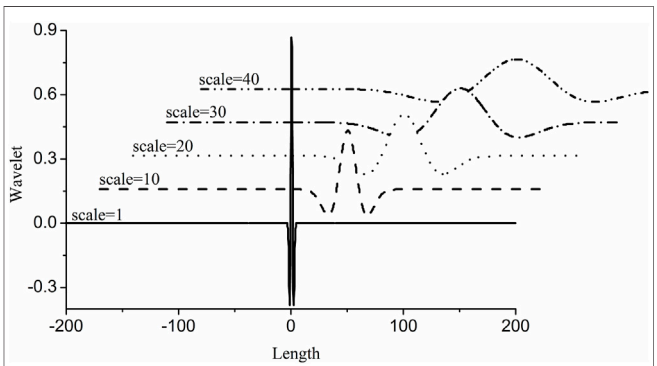


FIGURE 1 | The mexh wavelet with the increasement of *a* from 1 to 40.

Calculation Methods

For exploring the detailed information of a signal obtained by WT, the methods like WT, UVE and other analytical calculation method are develop, and the simulated signal was generated with Matlab which also has a WT command set integrated in the software. The figures are drew by Origin.

RESULTS AND DISCUSSION

The signal S in **Figure 2** is simulated by referring the actual NIR spectrum data (Shao et al., 2010) (http://www.idrc-chambersburg.org/shootout2010.html) to show the resolution ability of the method. The simulated signal S is formed by the signals a-g. If there were no effective methods to resolve the simulated signal, it will be easy to lose some valuable information

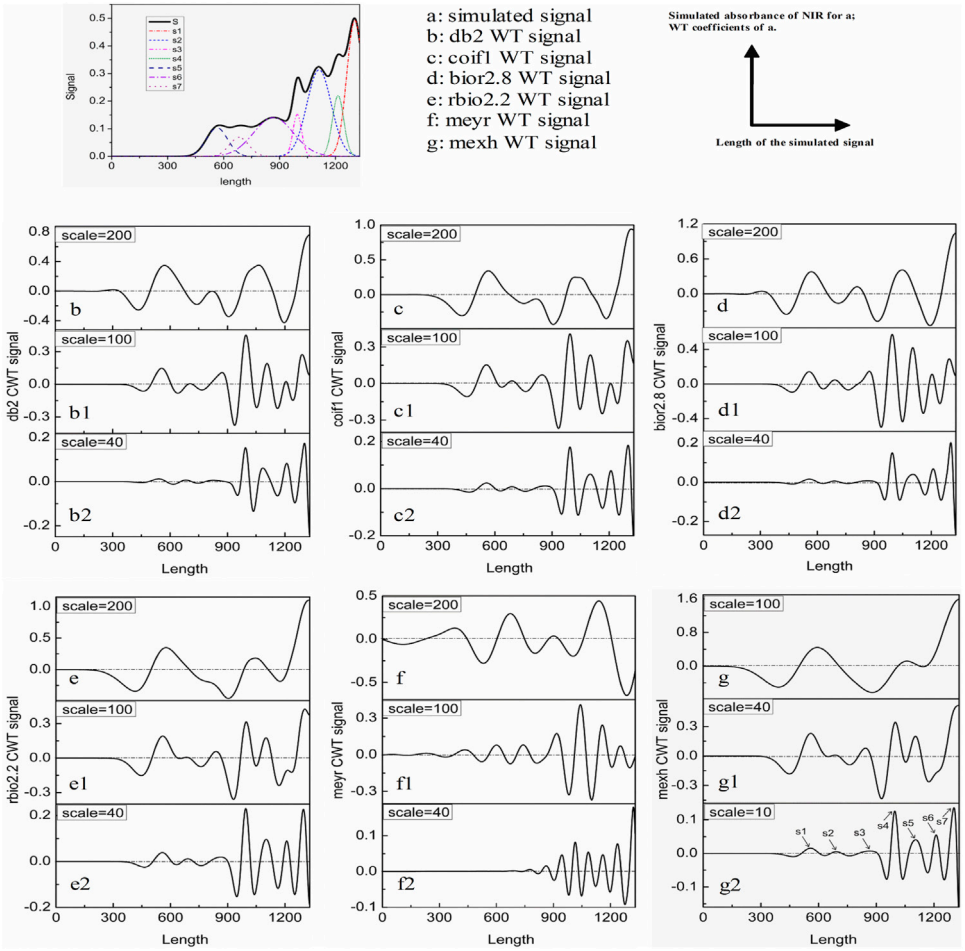


FIGURE 2 | a: the simulated signal S; b–g: the WT results of S by different wavelets with different scales.

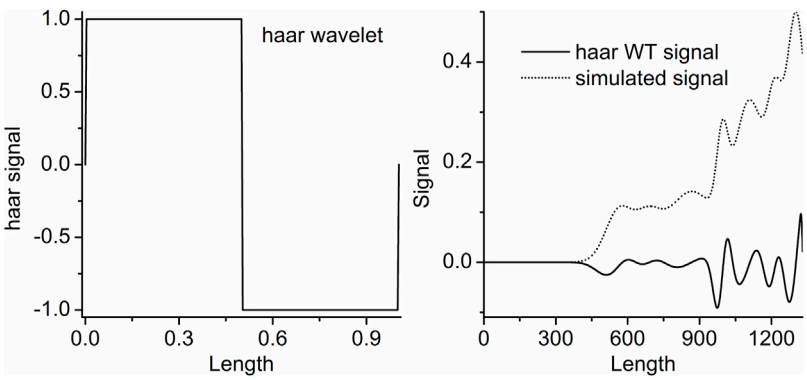


FIGURE 3 | Haar wavelet and the Haar WT result of the simulated signal.

(a-g), and no benefit to qualitative and especially quantitative analysis.

Haar wavelet can be used to resolve the overlapped signal (Chen et al., 2015). However, it is easy to result in an error qualitative analysis result. That is because Haar WT is same as the first derivation of the signal. The transformed results of the peaks and troughs are zero (**Figure 3**). In order to assure the ability of intuitive and accurate qualitative and quantitative analysis of the

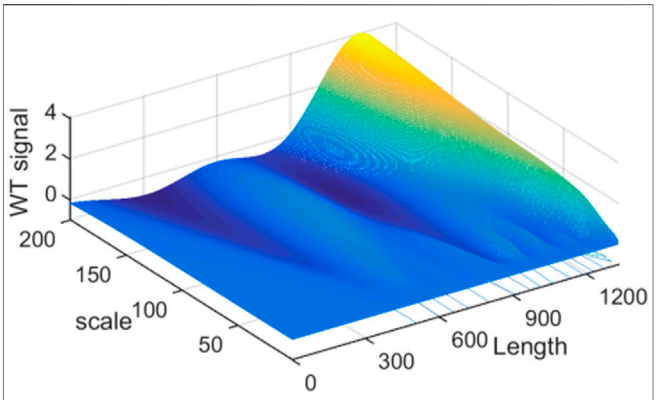


FIGURE 4 | The three dimensional mexh WT figure of the simulated signal.

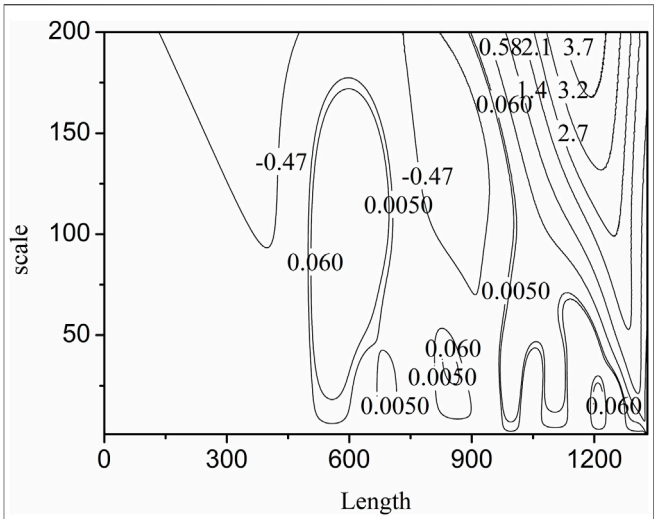


FIGURE 5 | The contour figure of the mexh WT results of the simulated signal.

method, the wavelets can obtain the WT results like the second derivative results of the signal are utilized. **Figure 2**. b–g show the resolution ability of some wavelets at different scales. All WT results are obtained by boundary extension. As is known that the WT results present the background and noise of the signal when the scale *a* is set as a small value, because a higher and narrower wavelet is easy to overlap the subtle background and noise, such as **Figure 1**. $a = 1$. But some valuable information is easy to be neglected if the scale *a* is set as a too large value. It can be seen from **Figure 2**. b–g or **Figure 4**. Therefore, we just select the maximum and minimum WT values in some scale range. Our aim is to afford a useful method to the relative researchers. So, we compared the resolution ability of the wavelets in **Figure 2**. The relative researchers can select suitable wavelet according to them.

Figure 2. g is the Mexh WT of the simulated signal. It is same as the second derivation of the signal. It is clear that some information are cancelled when the scale is set as a larger value 100, such as the sub signal s2, s3, s4 and s6. If the scale is set as a suitable value such as 10, all the valuable information can be resolved. By examining the three dimensional Figure with the scale less than 40, if the scale is set as a certain value, some sub signal points maybe occur the maximum WT values, but it is not for others. This is also clear in the contour **Figure 5**. We just show some sections in this figure. From above analysis, the maximum WT values of the signal in the scale range 40 or some near value can present the maximum information of sub signals.

By using different wavelets, the above method is utilized to resolve the protein signal in the corn dataset (http://www.eigenvector.com/data/Corn/index.html). The results of the regression analysis for this signal are shown in **Table 1**. Factor for partial least square analysis is selected by the predicted residual error sum of square (PRESS) values. The relationship between Factors and PRESS values for the regression analysis of

TABLE 1 | The results of the regression analysis for the protein NIR signal.

Methods	Wavelets	Data: Corn/Protein				
The proposed method		Factor	RMSEC(Selected a)	R_c	RMSEP	R_p
	Rbio2.2	6	0.2161 (15)	0.8985	0.2249	0.9125
	Rbio2.4	6	0.2185 (20)	0.8959	0.2212	0.9166
	Rbio2.6	7	0.2017 (20)	0.9132	0.1938	0.9349
	Rbio2.8	7	0.2095 (20)	0.9055	0.2153	0.9166
	Bior2.2	6	0.2214 (15)	0.8929	0.2533	0.8840
	Bior2.4	5	0.2562 (15)	0.8536	0.2691	0.8669
	bior2.6	5	0.2574 (15)	0.8521	0.2706	0.8657
	Bior2.8	6	0.2201 (25)	0.8941	0.2345	0.9082
	Mexh	8	0.2038 (30)	0.9121	0.2279	0.8989
	Meyr	6	0.2295(25)	0.8881	0.2099	0.9189
	Sym2	6	0.2139 (15)	0.9002	0.2300	0.9095
	Db2	6	0.2139 (15)	0.9002	0.2300	0.9095
	Coif1	6	0.2153 (25)	0.9004	0.2248	0.9069
	Gaus2	6	0.2260 (40)	0.8882	0.2587	0.8602
pls			0.2458			
UVE-pls			0.2349			

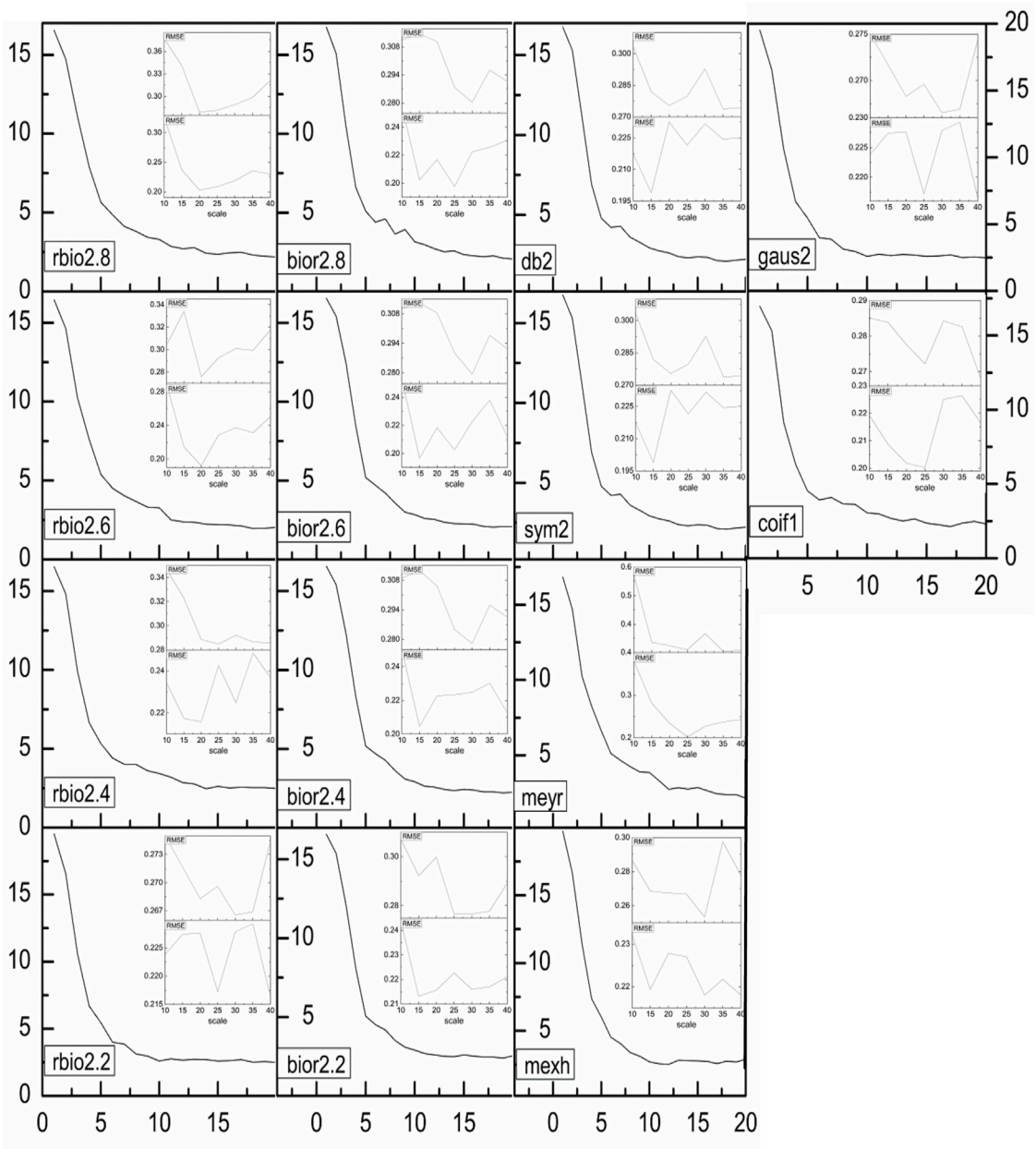


FIGURE 6 | The relationship between Factors and PRESS values for the regression analysis of the signals reconstructed by different WT. The upper is the RMSE for prediction, and the lower is the RMSE for calibration.

the signals reconstructed by different WT are shown in **Figure 6**. With the gradually increasement of the factor value from 1 to 20 by step 1, if the ratio between the present PRESS value and the former PRESS value is more than 0.9, the former factor value is used to construct regression model.

RMSEC is the root mean square error (RMSE) value for calibration. Selected *a* is the selected WT scale for modeling. The scale *a* value that is corresponding to the minimum RMSEC is selected when *a* changes from 10 to 40 by step 5. R_C is the corresponding coefficient for calibration, and R_P is the corresponding coefficient for prediction. We can also easy to compare the method with PLS from the results in the Table.

UVE is utilized to select valuable signal points after WT, and some selected results are shown in **Figure 7** (others in supplementary figures). The curves in **Figure 7** are the WT signals, circles are the selected signal points which can generate the minimum PRESS value by UVE for 100 times. If a signal point is selected in all UVE repetitions, a dot is set in the

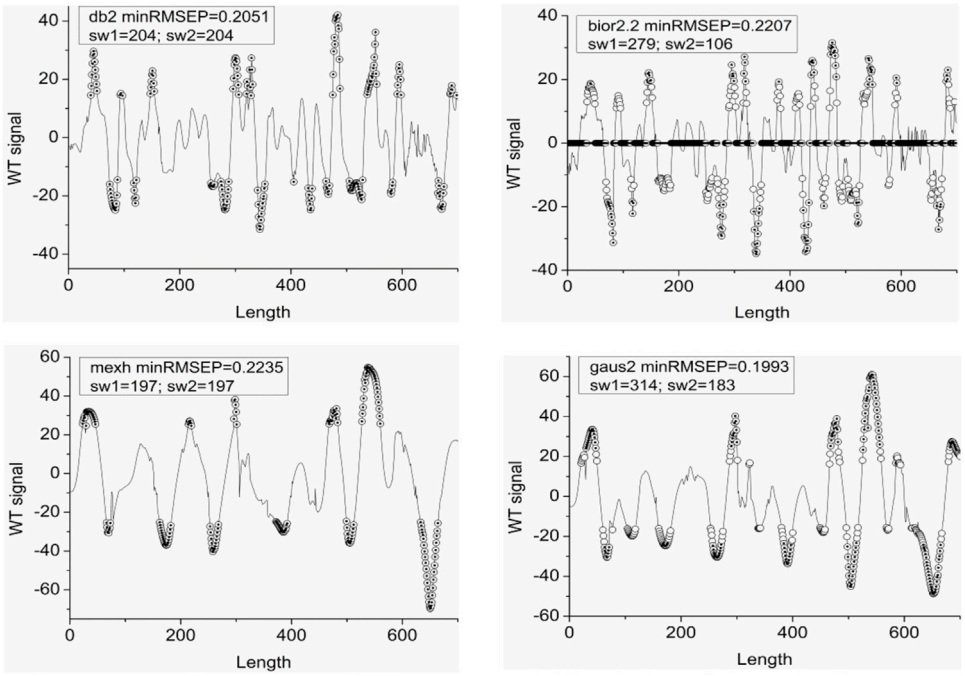


FIGURE 7 | Some select valuable signal points by UVE after WT. minRmsep is the minimum RMSEP value after 100 times UVE. sw is the selected spectrum point number.

circle. From the results, the peaks and troughs in the WT signal are the valuable information. As mentioned above, the selected troughs in the complex NIR signal may contain some important information.

CONCLUSION

The proposed valuable information extraction method can effectively extract the maximum valuable information from NIR signal. All the information in the sub signals of the simulated one are successfully resolved by the method. By resolving actual protein dataset, the detail information in it is totally emerged. After further UVE study, obviously comparable results are obtained. The method will be very helpful for the relative researchers.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: http://www.eigenvector.com/data/Corn/index.html.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

FUNDING

This study is supported by National Natural Science Foundation of China (Nos.21565022,22174110), the Industrial support plan of Gansu Provincial Department of Education (2021cyzc-01) and Special Fund Project for Guiding Local Scientific and Technological Development by the Central Government (No. 2020-2060503-17). Graduate Cultivating and Curriculum Reform Program of Northwest Normal University (2021KGLX01005) and Undergraduate Course Design.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fchem.2022.812567/full#supplementary-material

REFERENCES

Abrahamsson, C., Johansson, J., Andersson-Engels, S., Svanberg, S., and Folestad, S. (2005). Time-Resolved NIR Spectroscopy for Quantitative Analysis of Intact Pharmaceutical Tablets. *Anal. Chem.* 77 (4), 1055–1059. doi:10.1021/ac0487754

Centner, V., Massart, D.-L., de Noord, O. E., de Jong, S., Vandeginste, B. M., and Sterna, C. (1996). Elimination of Uninformative Variables for Multivariate Calibration. *Anal. Chem.* 68, 3851–3858. doi:10.1021/ac960321m

Chau, F. T., Liang, Y. Z., Gao, J. B., and Shao, X. G. (2004). *Chemometrics: From Basics to Watelet Transform*. John Wiley & Sons.

Chen, S.-C., Hayden, D. M., and Young, S. S. (2015). The Wavelet Transforms and Statistical Models for Near Infrared Spectra Analysis. *J. Math. Chem.* 53, 551–572. doi:10.1007/s10910-014-0434-x

de Oliveira, R. R., de Lima, K. M. G., Tauler, R., and de Juan, A. (2014). Application of Correlation Constrained Multivariate Curve Resolution Alternating Least-Squares Methods for Determination of Compounds of Interest in Biodiesel Blends Using NIR and UV-Visible Spectroscopic Data. *Talanta* 125, 233–241. doi:10.1016/j.talanta.2014.02.073

de Yong, D., Bhowmik, S., and Magnago, F. (2015). An Effective Power Quality Classifier Using Wavelet Transform and Support Vector Machines. *Expert Syst. Appl.* 42 (15-16), 6075–6081. doi:10.1016/j.eswa.2015.04.002

Dinc, E., Baleanu, D., and Üstündağ, Ö. (2006). An Approach to Quantitative Two-Component Analysis of a Mixture Containing Hydrochlorothiazide and Spironolactone in Tablets by One-Dimensional Continuous Daubechies and Biorthogonal Wavelet Analysis of UV-Spectra. *Spectrosc. Lett.* 36 (4), 341–355.

Dinç, E., and Yazan, Z. (2018). Wavelet Transform-Based UV Spectroscopy for Pharmaceutical Analysis. *Front. Chem.* 6 (1-12), 503. doi:10.3389/fchem.2018.00503

Eskildsen, C. E., Hansen, P. W., Skov, T., Marini, F., and Nørgaard, L. (2016). Evaluation of Multivariate Calibration Models Transferred between Spectroscopic Instruments: Applied to Near Infrared Measurements of Flour Samples. *J. Near Infrared Spectrosc.* 24, 151–156. doi:10.1255/jnirs.1210

Fu, M., Tian, Y., and Wu, F. (2015). Step-wise Support Vector Machines for Classification of Overlapping Samples. *Neurocomputing* 155, 159–166. doi:10.1016/j.neucom.2014.12.035

Goodarzi, M., Sharma, S., Ramon, H., and Saeys, W. (2015). Multivariate Calibration of NIR Spectroscopic Sensors for Continuous Glucose Monitoring. *Trac Trends Anal. Chem.* 67, 147–158. doi:10.1016/j.trac.2014.12.005

Inácioa, M. R. C., de Limaa, K. M. G., Lopesb, V. G., Pesso, J. D. C., and de Almeida Teixeira, G. H. (2013). Total Anthocyanin Content Determination in Intact Açaíand Palmitero-Juçara Fruit Using Near Infrared Spectroscopy (NIR) and Multivariate Calibration. *Food Chem.* 136 (3–4), 1160–1164.

Jena, D. P., Sahoo, S., and Panigrahi, S. N. (2014). Gear Fault Diagnosis Using Active Noise Cancellation and Adaptive Wavelet Transform. *Measurement* 47, 356–372. doi:10.1016/j.measurement.2013.09.006

Kalteh, A. M. (2013). Monthly River Flow Forecasting Using Artificial Neural Network and Support Vector Regression Models Coupled with Wavelet Transform. *Comput. Geosciences* 54, 1–8. doi:10.1016/j.cageo.2012.11.015

Lopes-dos-Santos, V., Panzeri, S., Kayser, C., Diamond, M. E., and Quian Quiroga, R. (2015). Extracting Information in Spike Time Patterns with Wavelets and Information Theory. *J. Neurophysiol.* 113, 1015–1033. doi:10.1152/jn.00380.2014

Martyna, A., Michalska, A., and Zadora, G. (2015). Interpretation of FTIR Spectra of Polymers and Raman Spectra of Car Paints by Means of Likelihood Ratio Approach Supported by Wavelet Transform for Reducing Data Dimensionality. *Anal. Bioanal. Chem.* 407, 3357–3376. doi:10.1007/s00216-015-8558-9

Ozaki, Y., McClure, W. F., and Christy, A. A. (2006). *Near-infrared Spectroscopy in Food Science and Technology*. Hoboken, NJ: John Wiley & Sons.

Pan, W., Zhao, J., and Chen, Q. (2015). Classification of Foodborne Pathogens Using Near Infrared (NIR) Laser Scatter Imaging System with Multivariate Calibration. *Sci. Rep.* 5. doi:10.1038/srep09524

Pathak, R. S., and Singh, A. (2016). Wavelet Transform of Generalized Functions in $K'\{M\ P\}$ Spaces. *Proc. Math. Sci.* 126, 213–226. doi:10.1007/s12044-016-0281-8

Pedro, A. M. K., and Ferreira, M. M. C. (2005). Nondestructive Determination of Solids and Carotenoids in Tomato Products by Near-Infrared Spectroscopy and Multivariate Calibration. *Anal. Chem.* 77 (8), 2505–2511. doi:10.1021/ac048651r

Rutan, S. C., de Noord, O. E., and Andréa, R. R. (1998). Characterization of the Sources of Variation Affecting Near-Infrared Spectroscopy Using Chemometric Methods. *Anal. Chem.* 70 (15), 3198–3201. doi:10.1021/ac9712052

Shao, X., Bian, X., and Cai, W. (2010). An Improved Boosting Partial Least Squares Method for Near-Infrared Spectroscopic Quantitative Analysis. *Anal. Chim. Acta* 666 (1-2), 32–37. doi:10.1016/j.aca.2010.03.036

Stenlund, H., Johansson, E., Gottfries, J., and Trygg, J. (2009). Unlocking Interpretation in Near Infrared Multivariate Calibrations by Orthogonal Partial Least Squares. *Anal. Chem.* 81 (1), 203–209. doi:10.1021/ac801803e

Subaie, R. F. A., and Mourou, M. A. (2013). The Continuous Wavelet Transform for A Bessel Type Operator on the Half Line. *ms* 1 (4), 196–203. doi:10.13189/ms.2013.010404

Yu, S., Li, B., Zhang, Q., Liu, C., and Meng, M. Q.-H. (2015). A Novel License Plate Location Method Based on Wavelet Transform and EMD Analysis. *Pattern Recognition* 48, 114–125. doi:10.1016/j.patcog.2014.07.027

Yuan, T., Wang, Z., LiNi, Z. W. D., Ni, W., and Liu, J. (2014). A Partial Least Squares and Wavelet-Transform Hybrid Model to Analyze Carbon Content in Coal Using Laser-Induced Breakdown Spectroscopy. *Analytica Chim. Acta* 807, 29–35. doi:10.1016/j.aca.2013.11.027

Yun, Y.-H., Wang, W.-T., Deng, B.-C., Lai, G.-B., Liu, X.-b., Ren, D.-B., et al. (2015). Using Variable Combination Population Analysis for Variable Selection in Multivariate Calibration. *Analytica Chim. Acta* 862, 14–23. doi:10.1016/j.aca.2014.12.048

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.