Received: 27 March 2020 | Revised: 23 June 2020 | Accepted: 27 August 2020

DOI: 10.1002/cem.3323

RESEARCH ARTICLE

Journal of CHEMOMETRICS WILEY

A new concept based on ensemble strategy and derivative for the quantitative analysis of infrared data

Hong Yan | Guo Tang | Yanmei Xiong | Shungeng Min

College of Science, China Agricultural University, Beijing, 100193, China

Correspondence
Yanmei Xiong and Shungeng Min,
College of Science, China Agricultural University, Beijing 100193, China.
Email: xiongym@cau.edu.cn; minsg@cau.edu.cn

Abstract

Preprocessing and variable selection are the most widely used strategies to develop accurate predictive models based on infrared spectroscopy. In our study, a new conception that the derivative combined with ensemble strategy based on competitive adaptive reweighted sampling (CARS), stability competitive adaptive reweighted sampling (SCARS), Monte Carlo uninformative variables elimination (MCUVE), and bootstrapping soft shrinkage (BOSS) is put forward. The proposed concept makes the best of the derivative spectra information and successfully combines the strengths of derivative spectra, CARS, SCARS, MCUVE, BOSS, and ensemble submodels. Compared with other methods in this study, this new method can establish good calibration models without increasing the complexity from the perspective of an end user. Also, overfitting issues can be prevented. Derivative1st-ECARS and Derivative1st-ESCARS have shown significant improvements in partial least regression calibration based on the experiments of three datasets. The proposed concept shows great potential of the chemometrics approaches applied to infrared data in multivariate calibration.

KEYWORDS

derivative, ensemble, infrared, overfitting, variable selection

1 | INTRODUCTION

The combination of multivariate analysis with spectroscopy has been widely applied in chemistry and other fields, such as food,[1–3] pharmacy,[4] and pesticide industry.[5] Generally, the spectra contain not only efficient information but also noise and redundant, irrelevant information. It is noted that in one spectrum, the number (hundreds or thousands) of variables is much larger than that of samples. Partial least regression (PLS), as the most commonly used method in multivariate analysis, has remarkable performance on account of its ability to conduct the collinear and dimensionality in spectra data.[1,6,7] However, satisfying results still cannot be achieved by PLS because of the existence of overfitting and nonlinearity. Sometimes, a good PLS calibration model may have a poor prediction performance on the test set. What is more, high-noise information can affect the interpretation of PLS model. Therefore, how to remove the noise and extract effective information becomes the research hotspots in chemometrics. To deal with these problems, preprocessing and variable selection are two optional ways.

Guo Tang contributed equally as the first author to this paper.

[Correction added on 22 February 2021, after first online publication: Peer review history statement has been added.]

https://doi.org/10.1002/cem.3323

Preprocessing is a significant step prior to PLS modeling. There are several methods proposed, such as Savitizky–Golay (SG), wavelets,[8,9] standard normal variate (SNV), multiplicative signal correction (MSC), correlation optimized warping (COW), and derivative. Among these methods, the most commonly used techniques are scatter correction and derivative. Derivative can not only detect the subtle features but also give a better resolution by magnifying small differences.[10,11] The peaks can be sharped, the background can be eliminated, and the signal-to-noise (S/N) ratio of peak can be improved. At present, derivative has been successfully applied in extracting analytical information from spectra with overlapped bands. However, the derivative can make small errors into big errors of spectrum, which is a serious problem.

After preprocessing is undertaken, variable selection is the following important procedure before building PLS model. The primary objective of variable selection is to improve the predictive ability of calibration model by removing those useless or irrelevant variables. However, the developed variable selection methods have several parameters involving modeling that may arouse the increase of modeling complexity, and the risk of ignoring informative variables also needs to be considered.

Until now, a bunch of variable selection methods have been proposed successfully. Generally, variable selection can be categorized into three sections: filter methods, wrapper methods, and embedded methods. Filter methods include two steps: First, PLS model should be fitted to the data, and then a threshold is introduced to complete the variable selection. The most popular method is the variable importance in projection.[12] In wrapper methods, it usually employs filter methods with iterative algorithm, which is based on supervised learning. There are several commonly used algorithms, such as uninformative variable elimination (UVE-PLS),[13] genetic algorithm combined with PLS regression,[14,15] subwindow permutation analysis coupled,[16] backward variable elimination,[17] and interval PLS.[18] For embedded methods, variable selection and modeling can be done in one step. However, the proposed methods are not so common.

As variable selection, ensemble strategy is another approach to deal with the noise or irrelevant information, which relies on establishing a list of submodels and fusing all the submodels, for example, stacked regression and boosting. It is used to provide better prediction accuracy and stability of models. In chemometrics, an ensemble of Monte Carlo uninformative variable elimination (EMCUVE),[19] boosting kernel PLS,[20] has been successfully applied in multivariate calibration.

In the present study, we proposed the combination of derivative and ensemble variable selection (ECARS, ESCARS, EMCUVE, and EBOSS) strategy designed to improve the prediction and robustness of quantitative analysis in infrared spectroscopy. Three datasets were employed for the evaluation of accuracy as well as robustness. During this strategy, preprocessing is first to adjust the original spectrum, and then different PLS models will be acquired, as the jackknife approach. Afterward, variable selection (competitive adaptive reweighted sampling [CARS], stability competitive adaptive reweighted sampling [SCARS], Monte Carlo uninformative variables elimination [MCUVE], and bootstrapping soft shrinkage [BOSS]) was applied to shrink the variable space by fivefold cross-validation. This step was repeated, and models (from this step) with better performance are used to generate ensemble model. Different preprocessing methods (SNV, Smooth, and MSC) are set to join the comparison of the multivariate calibration results. Moreover, the results of full-spectrum PLS model, preprocessing methods combined with CARS, SCARS, MCUVE, BOSS, ECARS, ESCARS, EMCUVE, and EBOSS models, were reported.

In the following parts, it would be demonstrated how preprocessing and ensemble variable selection take effect separately and simultaneously, how they influence the model accuracy and stability apparently in calibration modeling, and how the selected variables corresponded to the chemical information with uncorrelated contents.

2 | THEORY AND ALGORITHM

2.1 | Preprocessing

SNV is probably one of the most widely used preprocessing methods for scatter correction of near-infrared (NIR) data. Similar to SNV, the purpose of MSC is to reduce variability between the samples that arouses by scatter in NIR spectroscopy. Moving window smoothing function is mainly used to remove noise without major loss of intensity.[21]

Derivative has the ability to remove baseline effects in the spectra. The first derivative can only remove the baseline, but the second derivative can also remove the linear trend. There are two derivative methods: SG and Norris–Williams

(NW) derivatives. In the following study, the results of NW and SG derivatives almost had no difference, so SG derivative was chosen to demonstrate our investigations.

The spectral data in the window were fitted by the least square method, and the window is calculated by the polynomial coefficient. The most front-end data in the window were removed, and the adjacent spectral data at the end of the window was added to make the smooth window move within the whole spectrum to obtain the smoothed spectrum after different derivative analysis.

2.2 | A brief introduction of variable selection methods

2.2.1 | MCUVE

Uninformative variable elimination (UVE) is a popular variable selection method based on the regression coefficient of PLS, which aims to select the informative variables from the original spectrum. Variables with values that are below the defined threshold are regarded as uninformative. In MCUVE, Monte Carlo is introduced to UVE instead of leave-one-out strategy to simplify the PLS model.

2.2.2 | CARS and SCARS

CARS is another variable selection method based on regression coefficients.[22] In this method, the regression coefficients are computed on full spectra. The exponentially decreasing function is then employed to remove the variables that have small absolute regression coefficients and enforce feature selection. Consecutively, adaptive reweighted sampling is undertaken to realize a competitive feature selection based on the regression coefficients. SCARS is a method based on CARS, a more informative criterion, that is, the variable stability was employed to select important variables in this study. The definition of stability is the absolute value of regression coefficient divided by its standard deviation (SD).[23]

2.2.3 | BOSS

Generally, BOSS is carried out for each sampling run with four steps: (1) Bootstrap sampling (BSS) is used to generate subsets. (2) PLS submodels are built and found out the best models with model population analysis (MPA). (3) New weights for variables are obtained. (4) Because of the new weights, weighted bootstrap sampling (WBS) is applied to produce new subsets. WBS and MPA are applied to remain informative variables. The details mentioned above can be found in reference.[24]

2.3 | Ensemble variable selection methods

2.3.1 | ECARS

With the ensemble strategy, first, CARS algorithm needs to be run for R times (e.g., 1000 times) to obtain the selected cumulative frequency for each variable. Obviously, we can regard the cumulative selected frequency of each variable as the basis for evaluating the importance of each variable; the general variable with higher selected frequency owns bigger significance. Then, by setting the threshold, the variable with higher accumulated selected frequency above the threshold value is included in the final selected variable set. It should be noted that the setting of the threshold generally needs to be optimized. If the set threshold is too small, some interference variables will be selected. If the set threshold is too large, some effective variables will be lost. Hence, in our study, the range of the threshold settings is 0 to R-1 (R is the maximum cumulative frequency) and the interval is 1. A subset of variables can be obtained with the set threshold in each time. The distribution is based on these subsets of variables, creating a series of submodels. In this paper, the threshold corresponding to the submodel with the smallest RMSECV value is regarded as the optimized threshold, and the corresponding variable set is used as the final selected variable set.

Although Qingjuan Han et al.[19] used a similar threshold optimization strategy, the difference is that it sets the interval for optimizing the threshold to 0.05 * *R*. For MCUVE, which reserves a large number of variables, the size of the threshold optimization interval generally does not have a significant influence on the final optimized threshold. However, in terms of CARS with a significantly smaller number of retained variables, the smaller the optimization interval, the better the optimization of the optimal threshold. Therefore, in our study, the optimization interval used is set to 1, that is, the threshold is increased by one frequency at a time.

As ECARS, the ensemble strategy is also applied in MCUVE, SCARS, and BOSS. Therefore, EMCUVE, ECARS, ESCARS, and EBOSS are the four ensemble variable methods in our study.

3 | EXPERIMENTAL SECTION

3.1 | Corn dataset

NIR datasets of corn were obtained from the website: http://www.eigenvector.com/data/Corn/index.html. The datasets contain 80 samples of corn. Each spectrum is in the range of 1100–2498 nm within 700 variables at intervals of 2 nm. The properties of protein and starch are analytical targets. Forty-eight samples were used to make up the calibration set, and the 32 samples were used as the independent test according to the K-Stone sampling.[25,26] The prediction accuracy of the oil content of the corn samples (%, w/w) was investigated using spectra measured on mp5 and the oil; starch content was explored on mp6.

3.2 | Diesel fuels dataset

One NIR, diesel fuel dataset was downloaded from the link: http://www.eigenvector.com/data/SWRI/index.html. Each spectrum includes 401 variables ranging from 750 to 1550 nm at 2-nm interval. The property is freezing temperature. One hundred forty-seven calibration samples and 98 test samples were grouped by K-Stone sampling.

3.3 | Software

MATLAB (V2016a, Mathworks, USA) was installed on my personal computer (SSD) with an Intel Core i5-4210U 2.4-GHz CPU and 8-GB RAM for analysis, the Unscrambler 9.7 (CAMO, Norway), Python 3.6, Anaconda3-4.1.1.

4 | RESULTS AND DISCUSSION

Tables S1 and S2 summarized RMSEP/RMSEC, etc values of four variable selection algorithms, four ensemble variable selection methods, and preprocessing PLS models of all datasets.

4.1 | Spectra preprocessing

In NIR, because of the interactions between sample particles and NIR radiation, the shift of absorbency levels caused by light scattering may be harmful for NIR linear calibration and spectra interpretation. Therefore, we employed SNV and MSC to reduce scattering effects and colinearity changes in NIR. First and second derivatives were employed to reinforce spectra resolution and remove the background absorption. SG algorithm was used for the derivative algorithm, and smoothing is also applied in the NIR dataset.

4.2 | Corn dataset

The number of repeated runs of the CARS is a parameter in ECARS algorithm. The corn sample system is used as an example to illustrate the effect of the parameter mentioned above on the results of the ECARS algorithm.

Figure 1 shows that the accumulated frequency distribution of each variable after CARS was repeated for 100, 500, and 1000 times on the corn dataset. From the comparison of the three subgraphs in Figure 1, although the total number of repeated operations *R* is different, the variable points that have accumulated higher frequency are still mainly concentrated in several information variable ranges, for example, the points around 1800 nm and the variables around 2200 nm are selected to be relatively high under the conditions of repeated operation times. We discover that ensemble strategy can achieve more stable variable importance information. What is more interesting is that the ratio of the accumulated selected frequency to the *R* of most effective variables in these variable ranges does not change significantly with the change of *R*; thus, repeated times of CARS algorithm have little influence on the final results.

The number of latent variables was 5, which was determined by fivefold cross-validation, and the number of selected variables on corn dataset was also reported in Table S1. The SD was provided in the Supporting Information. Figure 2 also can reflect the robustness of different methods and changes with the application of ensemble strategy and derivative.

The results of preprocessing combined variable selection and the ensemble variable selection methods on the corn dataset (oil and starch) were reported in Figure 2 and Table S1. In oil system, Derivative1st7W-ESCARS gave the minimal RMSEP (0.0567), which was much smaller in comparison with full-spectrum model results and nonensemble variable selection algorithm. It is clear that Derivative-ECARS and Derivative-ESCARS have greatly improved the prediction performance. Derivative-ECARS and Derivative-ESCARS exhibited not only the best prediction ability in terms of the RMSEC and RMSEP values but also the best stability based on corn dataset. The number of retained variables was also the smallest, indicating that the model was the most parsimony and simplified.

In CARS algorithm series, Derivative1st7W-ECARS only retained seven variables to achieve the smallest RMSEP of CARS series (0.0625). Through Figure 2 and Table S1, the RMSEP value of Derivative1st7W-ECARS was 13.07% less than that of Derivative1st7W-CARS. Similarly, the RMSEP value of Derivative1st3W-CARS drops from 0.1260 to 0.0841 (Derivative1st3W-ECARS). Through Figure 4, it can be found that after derivative (1st3w, 1st5w, 1st7w, and 2st5w), the key variables around 1700 nm were reserved by ECARS, which were closely related with the first and second overtone absorption peaks of C–H. All other methods in CARS series only retained the variables around 2300 nm, which was attributed to the C–H vibration absorption peak.[22–24,27] It is also a critical reason for the good prediction performance of combination of derivative and ECARS. Compared with ECARS, CARS reserved uninformative variables around 1200 nm and variables between 2200 and 2400 nm based on eight pretreatment methods and did not retain the variables at 1700 nm, which represented the first and second double-frequency absorption peaks of C-H.[28–30] These two factors led to the poor performance of CARS, which also proved that CARS have poor anti-noise ability and cannot guarantee the retention of informative variables. Compared with CARS, ECARS combined with the first derivative method can not only reserve the variables with chemical significance efficiently but also highly eliminate uninformative variables. Therefore, by using the model, we acquired the best prediction ability and stability, which fully demonstrated the effectiveness of ensemble strategy in multivariate linear regression.

The red-dotted line in Figure 3 represented the optimized threshold value of ECARS algorithm in the corn oil system. In the first derivative, as the width of the smooth window increased, some uninformative variables were

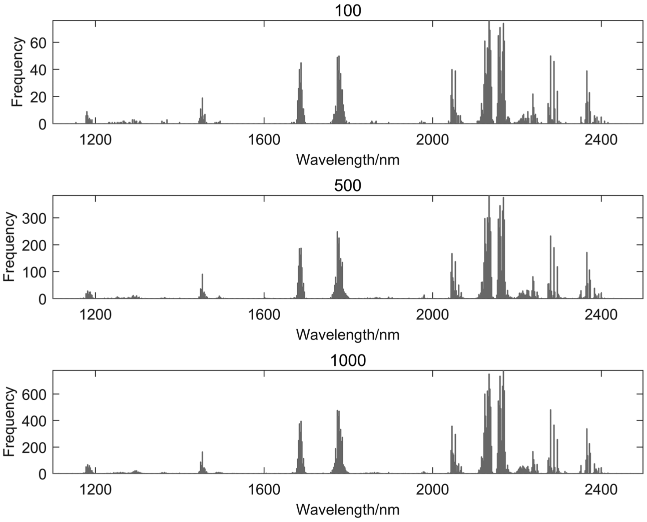

FIGURE 1 Accumulated selected frequencies of each wavelength by CARS algorithm after 100, 500, and 1000 repeated runs in corn samples

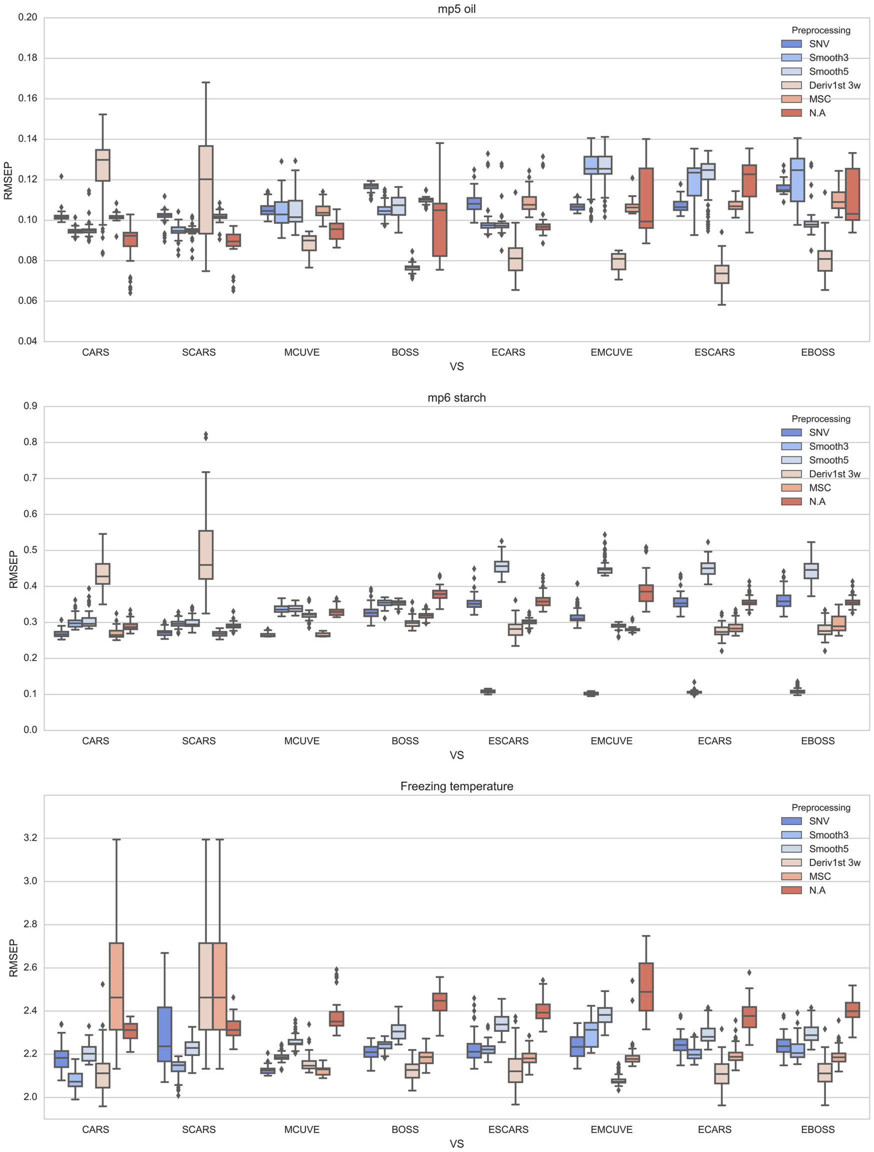

FIGURE 2 Boxplot of RMSEP on three datasets

gradually eliminated (marked by the blue box), which further confirmed that when the width of the smooth window was 7, informative variables were almost retained, and the prediction result was the best. To verify the validity of the ensemble strategy, we also verified on SCARS, MCUVE, and BOSS (Figures 4, 5, and 6).

In ESCARS series, similar with CARS, the lowest RMSEP (0.0567) was acquired with 10 retained variables, which was 29.48% lower than that of Derivative1st7W-SCARS. The RMSEP of Derivative1st3W was decreased from 0.1176 (SCARS) to 0.0608 (ESCARS), and the RMSEP of Derivative1st5W dropped from 0.0818 (SCARS) to 0.0705 (ESCARS). We noticed that ESCARS not only showed great superiority in model prediction ability and stability but also performed well in model simplification.

All SCARS retained about 150 variables (Table S1) based on first derivative pretreatment method, whereas ESCARS reserved several variables. Variable selection in ESCARS was similar to that in ECARS. Please refer to the explanation of ECARS variable selection for details.

According to the performance of ECARS and ESCARS, we can draw the conclusion that the ensemble strategy is very effective in improving the antinoise capability of CARS and SCARS. Meanwhile, combining with the first derivative, the informative variables can be retained, the uninformative variables can be eliminated, and the optimal and simplified model can be built.

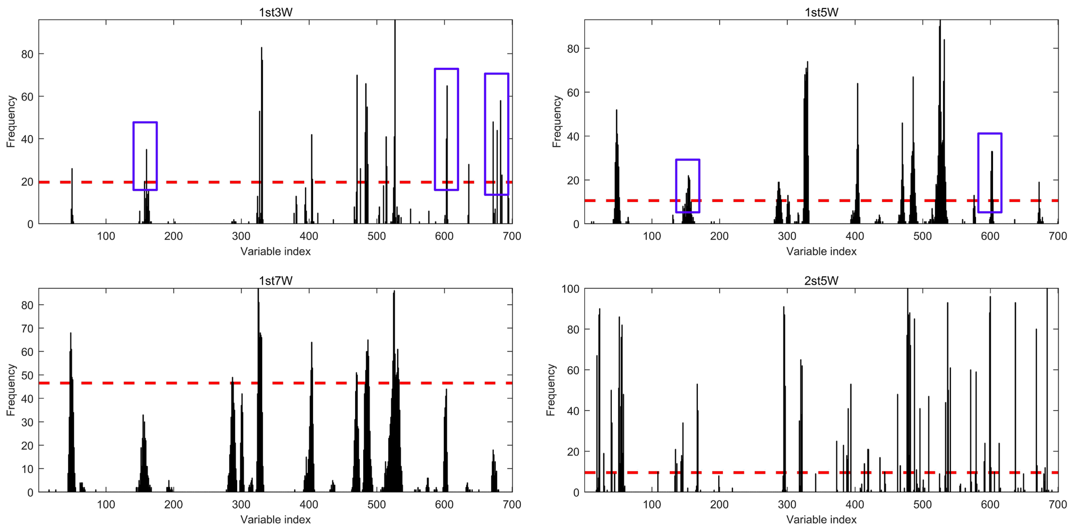

FIGURE 3 Accumulated selected frequencies of each wavelength after smoothing by moving windows with different width in mp5 oil samples on ECARS (the width of moving window 3, 5, and 7, respectively; the dash line in red represents finally selected threshold)

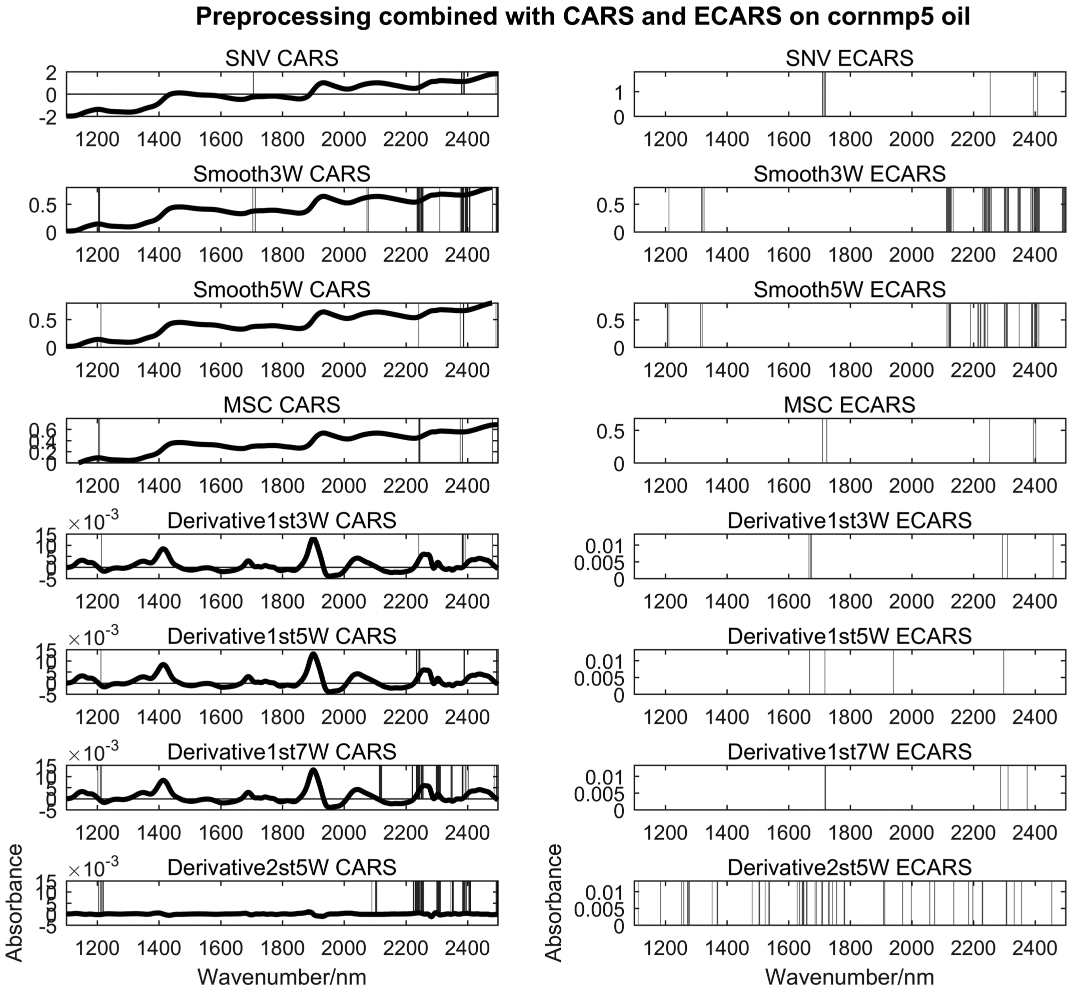

FIGURE 4 The variable selection of CARS and ECARS based on preprocessing on cornmp5 oil dataset

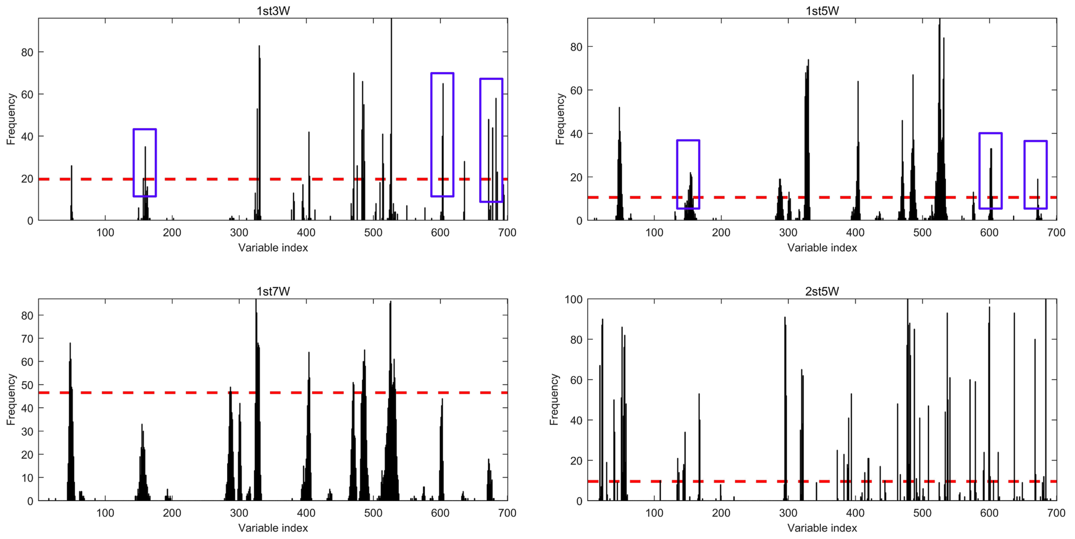

FIGURE 5 Accumulated selected frequencies of each wavelength after smoothing by moving windows with different width in mp5 oil samples on ESCARS (the width of moving window 3, 5, and 7, respectively; the dash line in red represents finally selected threshold)

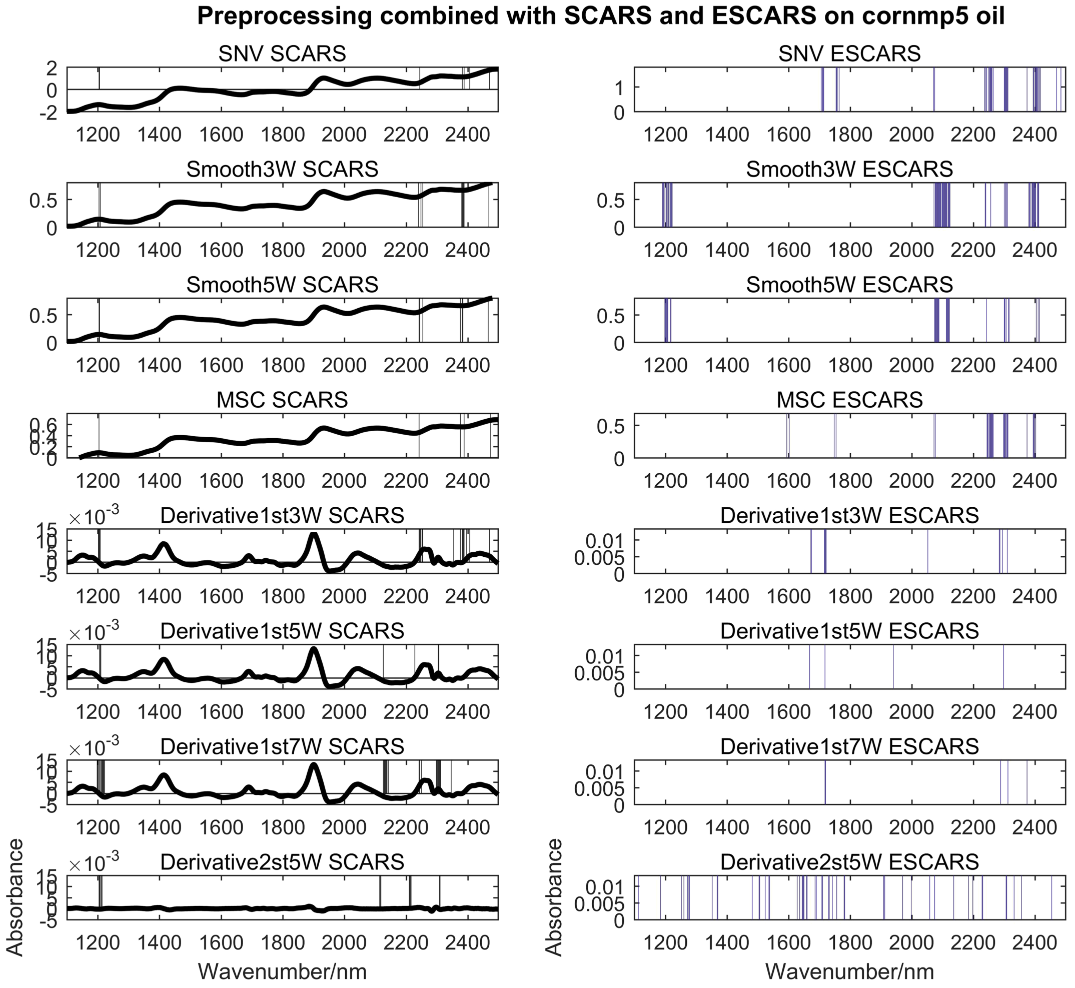

FIGURE 6 The variable selection of SCARS and ESCARS based on preprocessing on cornmp5 oil dataset

As a commonly used pretreatment method, derivative can eliminate background interference effectively, resolve overlapping peak, and improve the resolution and sensitivity; however, it may expand the noise signal or reduce the S/N ratio. The second derivative is more sensitive to noise, which will amplify noise information severely; in this case, noise could be selected easily, and the informative variables may be eliminated. However, the first derivative combined with the ensemble variable selection algorithm can make full use of the advantages of the two methods and finally get the optimized model.

Because the theory of MCUVE and BOSS is different from the two methods above, it is necessary to select MCUVE and BOSS in order to investigate the effectiveness of the ensemble strategy.

In EMCUVE series, Derivative1st7W-EMCUVE still obtained the minimum RMSEP (0.0754). Compared with MCUVE, EMCUVE prediction accuracy had been improved, but not by much. According to Figure S1, after the first derivative, EMCUVE still reserved the variables concentrated in the 1700 and 2300 nm; however, as the weak ability in eliminating uninformative variables of MCUVE, it also retained the variables between around 1200 and 1400 nm, which were uninformative; it is also the reason why the performance of EMCUVE is stable, but not as good as ECARS and ESCARS.

In BOSS series, as the BOSS is a variable selection algorithm developed based on MPA, its stability and antinoise capability are superior to CARS and SCARS, so the first derivative combined with BOSS also gives a small RMSEP, Derivative1st5W-BOSS (0.0667) and Derivative1st7W-BOSS (0.0649). Derivative1st5W-EBOSS obtained the minimal RMSEP (0.0634); as shown in Figure S3, the region of variables selected for the BOSS and EBOSS did not change significantly, which also explained the small derivation of RMSEP. Hence, the ensemble strategy does not have a super advantage in BOSS algorithm, only slightly improved.

According to the results of corn oil system, the first derivative combined with ECARS and ESCARS can effectively select the informative variables, eliminate the uninformative variables, and finally achieve the optimal model.

The ensemble variable selection algorithm also performed well in the corn starch dataset. In CARS series, Derivative1st3W-ECARS gave the minimum RMSEP (0.2425); similarly, the first derivative combined with ECARS showed a good prediction ability, and the RMSEP increased by more than 30% compared with CARS. Meanwhile, the robustness of the ECARS model is also confirmed by Figure 2. It can be found from the variable selection diagram (-Figure S4) that the first derivative combined with ECARS retained the absorption peaks around 1700 and 1800 nm, which represented the first and second overtones of C–H and C–H vibration combination linked to 2300 nm. The second derivative will make the noise amplification multiple higher, which is not conducive to the variable selection algorithm to reserve informative variables, and the selected variable region is also relatively dispersed. In comparison, CARS retained some uninformative variables (noise) besides the selected variables in ECARS, which leads to poor prediction accuracy and stability of the model directly. The RMSEP of Derivative1st3W-CARS was 0.4939, which was obviously overfitting, and the boxplot (Figure 2) exhibited the poor stability of CARS after derivative pretreatment. It is consistent with the corn oil dataset.

In addition to derivative pretreatment, SNV, smoothing, and MSC did not have much difference in the models of CARS and ECARS, but the number of selected variables by ECARS decreased.

In SCARS series, the RMSEP of Derivative1st3W-ESCARS was 0.2297, which was much smaller than that of full-spectrum PLS model (0.3933). According to Figure S6, the selected variable region of first derivative combined with ESCARS was more concentrated, whereas the other methods are relatively dispersive. Details were similar to ECARS discussed before. At the same time, EMCUVE also has a similar situation with ESCARS. After the first derivative, the selected variables are mainly concentrated around 1800 and 2200 nm (Figure S8).

4.3 | Diesel freezing temperature dataset

In order to further investigate the reason of Derivation1st-CARS, Derivation1st-SCARS and the difference between models take the diesel freezing temperature dataset for example. The results of diesel freezing temperature were displayed in Table S2 and Figure 2. The selected variables were showed in Figures 7, 8 and in the Supporting Information. In all CARS-based methods, Derivative1st7W-ECARS performed best both in accuracy and robustness (Figure 2). The combination of deviation and ensemble strategy also obtained the best results in SCARS-, MCUVE-, and BOSS-based methods (Table S2 and Figure 2). The selected variables were concentrated in the regions, which were 950–1100 nm, 1200–1300 nm, and 1450–1550 nm. Compared with ECARS, Derivative1st7W retained more informative variables in the regions 1000–1100 nm and 1200–1300 nm and eliminated some irrelevant information around 750 nm (Figure 8).

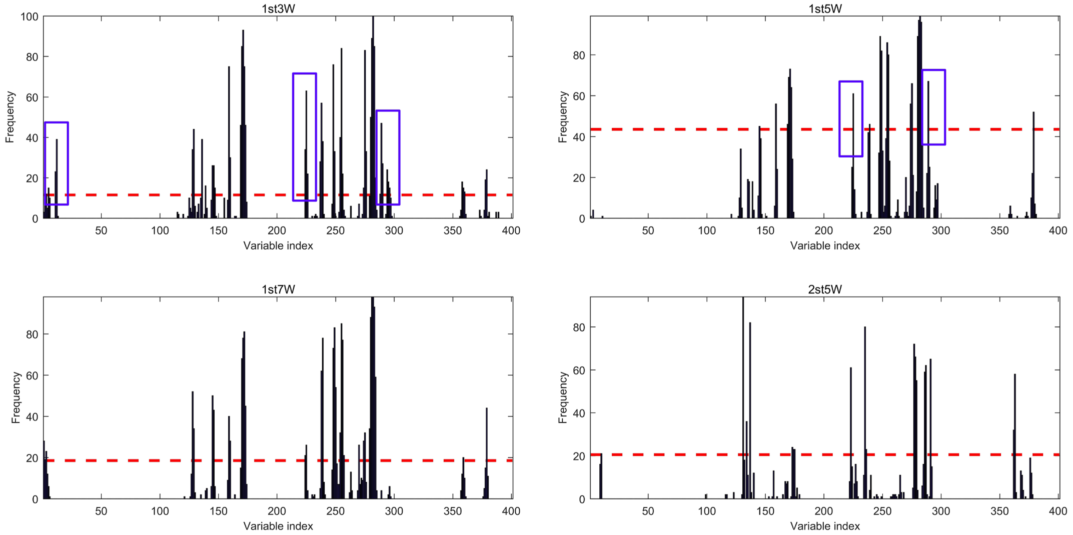

FIGURE 7 Accumulated selected frequencies of each wavelength after smoothing by moving windows with different width in diesel freezing temperature samples on ECARS (the width of moving window 3, 5, and 7, respectively; the dash line in red represents finally selected threshold)

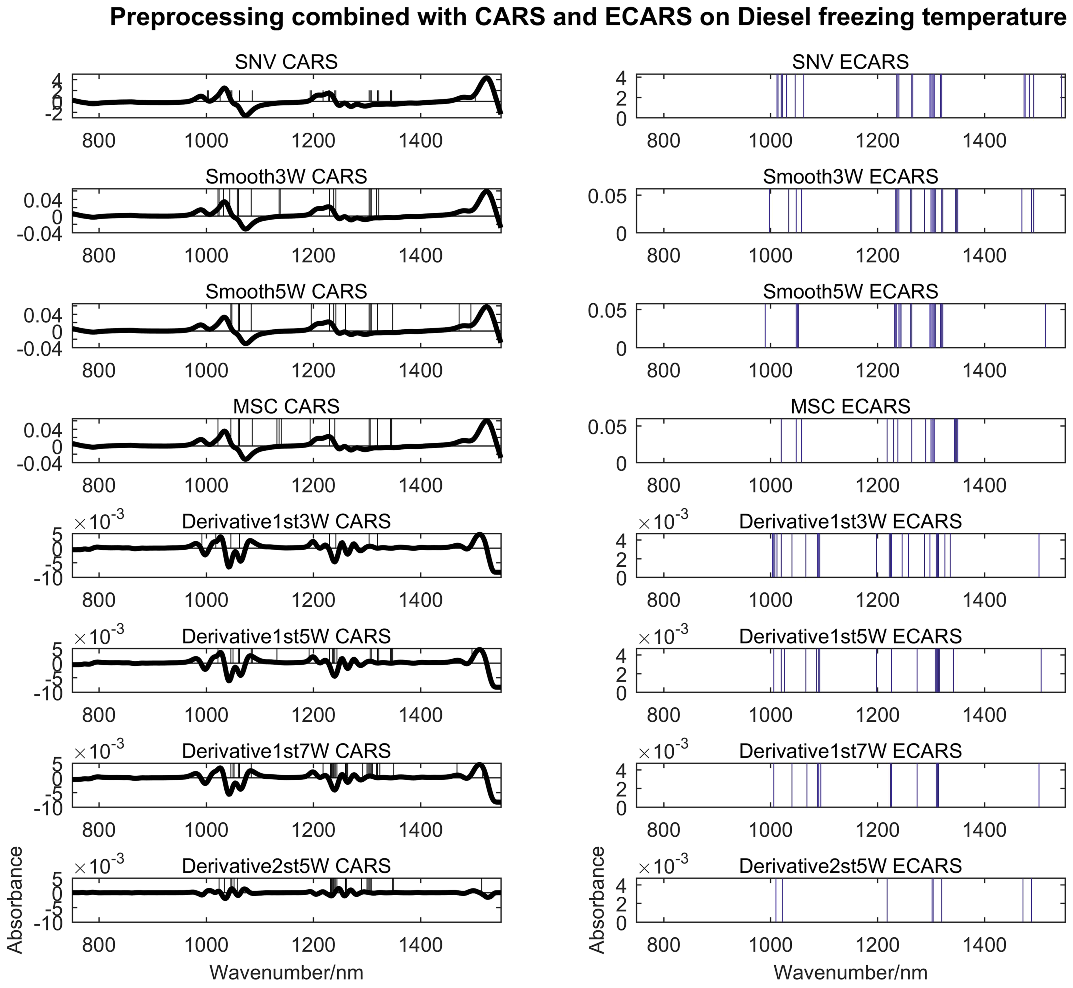

FIGURE 8 The variable selection of CARS and ECARS based on preprocessing on diesel freezing temperature dataset

We also noticed that the combination of deviation and ESCAS, EMCUVE, and EBOSS make the variables selected more concentrated and seek out the informative variables accurately (Supporting Information).

For all the three datasets, derivative aroused terrible prediction accuracy and stability on CARS and SCARS (Figure 2). However, Derivative1st7W-ECARS and Derivative1st7W-ESCARS obtained the lowest RMSEP and good stability. We can assume that the advantages of derivative and CARS and SCARS got the most utmost out by applying ensemble strategy. In terms of derivative, it can not only correct the background as other preprocessing methods can do but also decrease the S/N ratio by magnifying small differences.[9,21,31] Therefore, it is useful for variable selection methods to extract feature efficiently from overlapped spectrum. However, derivative can make small errors into big errors of spectrum, such as noise, which may be easily selected by CARS and SCARS because of their bad antinoise ability after derivative. For spectrum with good quality, CARS and SCARS can select the informative variables accurately and retain less variables.[22.23] Here, we introduced the ensemble strategy to improve the prediction accuracy and robustness. What is more, the risk of loss of information can be prevented. In conclusion, derivative combined with ECARS and ESCARS can achieve remarkable performance in prediction accuracy and robustness with sampler PLS model on NIR.

5 | CONCLUSION

With the purpose in improving the prediction accuracy and stability of multivariate regression analysis, we proposed ECARS, ESCARS, EMCUVE, and EBOSS procedure, which was combined with CARS, SCARS, MCUVE, BOSS, and ensemble concept, for variable selection in infrared spectra. Then, in a new conception that derivative combined with ensemble strategy based on CARS, SCARS is put forward. The new concept makes full use of the derivative spectra information and successfully combines the superiority of derivative information, CARS, SCARS, MCUVE, and BOSS and ensemble models. Compared with other methods employed in this study, this new method can establish good calibration models without increasing the complexity. The results based on the three datasets demonstrated that Derivative1st-ECARS or Derivative1st-ESCARS had much better prediction results compared with other mentioned modeling methods, especially with seven-moving window. Therefore, the performance of PLS calibration can be remarkably improved according to this new strategy. Additionally, it was demonstrated that the derivative (especially first-order derivative) was more useful than the original or smoothing, SNV, and MSC spectra for some datasets. However, there were still some issues that need to be further discussed. Our future work is to move forward to explore more comprehensive evaluation criteria for the submodels. We also expect that model ensemble concepts are promising in chemometric and are developed for novel implementations in other studies.

ACKNOWLEDGEMENTS

This work has been helped with Supervisor Shungeng Min and Guo Tang.

CONFLICT OF INTEREST

All authors declare that they have no conflict of interest.

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1002/CEM.3323.

ORCID

Hong Yan https://orcid.org/0000-0002-4059-6625
Shungeng Min https://orcid.org/0000-0003-4235-3413

REFERENCES

1. Karoui R, Downey G, Blecker C. Mid-infrared spectroscopy coupled with chemometrics: a tool for the analysis of intact food systems and the exploration of their molecular structure−quality relationships—a review. *Chem Rev*. 2010;110(10):6144-6168.
2. Downey G. Food and food ingredient authentication by mid-infrared spectroscopy and chemometrics. *TrAC Trends Anal Chem*. 1998;17(7):418-424.
3. Granato D, Putnik P, Kovačević DB, et al. Trends in chemometrics: food authentication, microbiology, and effects of processing. *Compr Rev Food Sci Food Saf*. 2018;17(3):663-677.

4. Rohman A, Dzulfianto A, Riswanto FDO. The employment of UV-spectroscopy combined with multivariate calibration for analysis of paracetamol, propyphenazone and caffeine. *IND j Pharm*. 2017;28(4):191-197. https://doi.org/10.14499/indonesianjpharm28iss4pp191
5. Yan H, Song X, Tian K, Chen Y, Xiong Y, Min S. Quantitative determination of additive Chlorantraniliprole in Abamectin preparation: investigation of bootstrapping soft shrinkage approach by mid-infrared spectroscopy. *Spectrochim Acta a Mol Biomol Spectrosc*. 2018;191: 296-302.
6. Wold H. Soft modeling by latent variables: the nonlinear iterative partial least squares approach. *Perspectives in Probability and Statistics*, Papers in Honour of M. S. Bartlett 1975.
7. Martens H, Naes T. Chichester, UK; 1989.
8. Barclay VJ, Bonner RF, Hamilton IP. Application of wavelet transforms to experimental spectra: smoothing, denoising, and data set compression. *Anal Chem*. 1997;69(1):78-90.
9. Savitzky A, Golay MJE. Smoothing and differentiation of data by simplified least squares procedures. *Anal Chem*. 1964;36(8):1627-1639.
10. Gerhard T. *Derivative Spectrophotometry: Low and Higher Order*. New York: VCH Publishers, Inc; 1994:101-169. https://doi.org/10.1002/3527601570.ch4
11. Kharintsev SS, Kamalova DI, Salakhov MK. Resolution enhancement of composite spectra with fractal noise in derivative spectrometry. *Appl Spectrosc*. 2000;54(5):721-730.
12. Wold S, Jonsson J, Sjörström M, Sandberg M, Rännar S. DNA and peptide sequences and chemical processes multivariately modelled by principal component analysis and partial least-squares projections to latent structures. *Anal Chim Acta*. 1993;277(2):239-253.
13. Centner V, Massart D-L, de Noord OE, de Jong S, Vandeginste BM, Sterna C. Elimination of uninformative variables for multivariate calibration. *Anal Chem*. 1996;68(21):3851-3858.
14. Leardi R, Gonzalez AL. Genetic algorithms applied to feature selection in PLS regression: how and when to use them. *Chemom Intel Lab Syst*. 1998;41(2):195-207.
15. Leardi R, Seasholtz MB, Pell RJ. Variable selection for multivariate calibration using a genetic algorithm: prediction of additive concentrations in polymer films from Fourier transform-infrared spectral data. *Anal Chim Acta*. 2002;461(2):189-200.
16. Li H-D, Zeng M-M, Tan B-B, Liang Y-Z, Xu Q-S, Cao D-S. Recipe for revealing informative metabolites based on model population analysis. *Metabolomics*. 2010;6(3):353-361.
17. Pierna JAF, Abbas O, Baeten V, Dardenne P. A backward variable selection method for PLS regression (BVSPLS). *Anal Chim Acta*. 2009;642(1-2):89-93.
18. Norgaard L, Saudland A, Wagner J, Nielsen JP, Munck L, Engelsen SB. Interval partial least-squares regression (iPLS): a comparative chemometric study with an example from near-infrared spectroscopy. *Appl Spectrosc*. 2000;54(3):413-419.
19. Han Q-J, Wu H-L, Cai C-B, Xu L, Yu R-Q. An ensemble of Monte Carlo uninformative variable elimination for wavelength selection. *Anal Chim Acta*. 2008;612(2):121-125.
20. Shinzawa H, Jiang JH, Ritthiruangdej P, Ozaki Y. Investigations of bagged kernel partial least squares (KPLS) and boosting KPLS with applications to near-infrared (NIR) spectra. *J Chemometr*. 2006;20(8-10):436-444.
21. Engel J, Gerretzen J, Szymańska E, et al. Breaking with trends in pre-processing? *TrAC Trends Anal Chem*. 2013;50:96-106.
22. Li H, Liang Y, Xu Q, Cao D. Key wavelengths screening using competitive adaptive reweighted sampling method for multivariate calibration. *Anal Chim Acta*. 2009;648(1):77-84.
23. Zheng K, Li Q, Wang J, et al. Stability competitive adaptive reweighted sampling (SCARS) and its applications to multivariate calibration of NIR spectra. *Chemom Intel Lab Syst*. 2012;112:48-54.
24. Deng B-C, Yun Y-H, Cao D-S, et al. A bootstrapping soft shrinkage approach for variable selection in chemical modeling. *Anal Chim Acta*. 2016;908:63-74.
25. Wu W, Massart DL. Artificial neural networks in classification of NIR spectral data: selection of the input. *Chem Intell Lab Syst*. 1996;35(1):127-135.
26. Rajer-Kanduč K, Zupan J, Majcen N. Separation of data on the training and test set for modelling: a case study for modelling of five colour properties of a white pigment. *Chem Intell Lab Syst*. 2003;65(2):221-229.
27. Li H-D, Xu Q-S, Liang Y-Z. Random frog: an efficient reversible jump Markov chain Monte Carlo-like approach for variable selection with applications to gene selection and disease classification. *Anal Chim Acta*. 2012;740:20-26.
28. Deng B-c, Yun Y-h, Liang Y-z, Yi L-z. A novel variable selection approach that iteratively optimizes variable space using weighted binary matrix sampling. *Analyst*. 2014;139(19):4836-4845.
29. Yun Y-H, Wang W-T, Tan M-L, et al. A strategy that iteratively retains informative variables for selecting optimal variable subset in multivariate calibration. *Anal Chim Acta*. 2014;807:36-43.
30. Wang W, Yun Y, Deng B, Fan W, Liang Y. Iteratively variable subset optimization for multivariate calibration. *RSC Adv*. 2015;5(116): 95771-95780.
31. Zhao N, Wu Z, Cheng Y, Shi X, Qiao Y. MDL and RMSEP assessment of spectral pretreatments by adding different noises in calibration/validation datasets. *Spectrochim Acta a Mol Biomol Spectrosc*. 2016;163:20-27.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Yan H, Tang G, Xiong Y, Min S. A new concept based on ensemble strategy and derivative for the quantitative analysis of infrared data. *Journal of Chemometrics*. 2021;35:e3323. https://doi.org/10.1002/cem.3323