Spectral interval combination optimization (ICO) on rapid quality assessment of *Solanaceae* plant: a validation study

Qianqian Li^{1,3} · Yue Huang² · Xiangzhong Song³ · Jixiong Zhang³ · Shungeng Min³

Revised: 26 February 2019 / Accepted: 6 March 2019 / Published online: 14 March 2019
© Association of Food Scientists & Technologists (India) 2019

Abstract A novel spectral variable selection method, named as interval combination optimization (ICO), was proposed in the previous study of us. In the present study, ICO coupled with near infrared (NIR) spectroscopy was applied to the rapid determination of four primary constituents including total sugar, reducing sugar, total nitrogen and nicotine in *Nicotiana* plant. Partial least squares regressions was performed after ICO algorithm. The full spectrum was divided into forty equal-width intervals, and the interval with lower root mean squared error of cross-validation was selected for further analysis. As a result, only 155 variables were retained from 1555 variables for each constituent. Particularly, as a variables selection method, ICO improved the prediction accuracy of calibration model and obtained a satisfactory result compared with full-spectrum data. Results revealed that NIR combined with ICO could be efficiently used for rapid analysis of quality associated constituents of *Nicotiana* plant. Moreover, this study provided a supplementary verification of the proposed variable selection method for the further applications.

Keywords Interval combination optimization · Weighted bootstrap sampling · Partial least squares · *Nicotiana* plant

Introduction

Generally, the quality associated constituents of *Nicotiana* plant are mainly involving total sugar, reducing sugar, total nitrogen and nicotine and so on, which are usually mentioned as the routine constituents of the industrial crop. Since the determination of the routine constituents is complicated, time-consuming and laborious (Shao et al. 2007), it is necessary to find a fast and accurate method to determine the routine chemical constituents in *Nicotiana* plant so as to guarantee the product quality. Meanwhile, as a powerful analytical method, NIR spectroscopy has been widely applied to the fields of petroleum (Syunyaev et al. 2009), food (Ritota and Manzi 2018), medicine (Zontov et al. 2016) and tobacco (Duan et al. 2012) due to its features of rapidity, simplicity, non-destruction and so on. The NIR region (between 4000 and 12,500 cm^{-1}) contains information of C–H, N–H and O–H bonds which are the combinations and overtones of the fundamental vibrations (Tan et al. 2008).

Quantitative NIR measurements are usually based on chemometrics to relate the spectrum with the concentration of a particular component. Typically, the quantitative analysis methods such as partial least squares (PLS) (Höskuldsson 1988) and principal component regression (PCR) (Mevik and Cederkvist 2004) are usually executing the full-spectrum for analysis, which will decrease the predict performance with uninformative variables included. In fact, one of the primary problem in multivariate calibration analysis is to decide which variables bring about a robust model with higher accuracy (Abrahamsson and Johansson 2003). As is known, the number of variables in NIR analysis is usually much more than the sample number. So, the collinearity and redundancy are two common phenomena among the data matrix, which would affect the

✉ Yue Huang
orange07@126.com

¹ School of Marine Science, China University of Geosciences, Beijing 100086, People's Republic of China

² College of Food Science and Nutritional Engineering, China Agricultural University, Beijing 100083, People's Republic of China

³ College of Science, China Agricultural University, Beijing 100193, People's Republic of China

accuracy and robustness of the model, for the interferences generated from background, noise and overlapping bands have great impacts on the performance of the model (Zhang et al. 2013).

Variable selection aims to find informative spectral and yield better predictive performance with reliable and interpretable information that relating the property of measured component for quantitative analysis. Many chemometric variable selection approaches have been developed, such as genetic algorithms (GA) (Broadhurst et al. 1997), interval PLS (iPLS) (Pereira et al. 2008), backward interval partial least-squares (biPLS) (Leardi and Nørgaard 2004), synergy interval partial least-squares (siPLS) (Zuo et al. 2014), moving-window partial least-squares (MWPLS) (Fu et al. 2007), uninformative variable elimination (UVE) (Ye et al. 2008), Monte Carlo based UVE (MCUVE) (Han et al. 2008).

A new wavelength interval selection algorithm, named as interval combination optimization (ICO), was first proposed by Song and Huang in our research group (Song et al. 2016). ICO was proposed coupled with weighted bootstrap sampling (WBS) under the framework of model population analysis (MPA) (Li et al. 2010).

WBS is a random sampling technique with replacement, different objects are assigned with different weights with respect to the frequency of appearance. MPA, as a general framework for variable selection and assessment, could provide a comprehensive insight from a large quantity of sub-models (Huang et al. 2013). The kernel idea of MPA is to establish data-analysis methods via statistically to analyze the distribution of sub-models (Li et al. 2012). Interval selection methods are expected to be more promising than single wavelength methods. As an interval selection method, ICO searches the interval weights that go through the entire variable space, and then chooses the useful sub-intervals.

In this study, ICO coupled with WBS and MPA was utilized for spectral variables selection, wherein WBS was assigned for variable selection according to the weight. With the purpose of improving prediction accuracy, ICO was applied to decomposing NIR data of total sugar, reducing sugar, total nitrogen and nicotine. One hundred plant samples from Yunnan province was quantitatively analyzed on considering the weight of each interval. Necessarily, the results obtained from the final ICO models were compared with those from the full-spectrum data.

Materials and methods

Sample preparation

All 100 plant leaf samples from Yunnan province were labeled, dried in an oven at 60 °C and then grounded by the

whirl wind grinding machine. Chemical concentration of total sugar, reducing sugar, total nitrogen and nicotine were measured by a continuous flow analyzer. The concentration of total sugar ranged from 10.59 to 41.63% with the mean value of 25.81%; reducing sugar ranged from 7.34 to 32.13% with the mean value of 18.73%; total nitrogen ranged from 1.17 to 2.21% with the mean value of 1.67% and nicotine ranged from 0.81 to 2.99% with the mean value of 1.56%, respectively.

NIR analysis

The NIR diffuse reflectance spectra were collected by a FT-NIR spectrometer (Matrix-I, Brucker, Germany). Spectra covered from 10,000 to 4000 cm^{-1} at a resolution of 4 cm^{-1} . Each spectrum was an average of 64 scans.

Partial least-square regression modeling (PLS)

The PLS algorithm (Wold et al. 1989) was a supervised analysis in view of the relationship between the signal intensity and the sample characteristics (Geladi and Kowalski 1985). In this study, the optimization and the predictive ability of modeling were mainly presented by determination coefficient (R^2) and root mean square error of prediction (RMSEP). Models were all established by five-fold cross validation and parameters of latent variables were optimized based on the root mean square error of cross-validation (RMSECV) generated from the calibration set.

Kennard–Stone algorithm (KS)

KS algorithm (Kennard and Stone 1969) intended to choose a representative subset from a pool of N samples. On the purpose of ensuring a uniform distribution of subsets, KS followed the steps of new selections would be far from the samples already chosen. The algorithm employed the Euclidean distances $d_x(p, q)$ between the x -vectors of each pair (p, q) of samples calculated as Eq. 1

$$d_x(p, q) = \sqrt{\sum_{j=1}^J [x_p(j) - x_q(j)]^2} \quad (1)$$

where J denoted the number of samples. At each subsequent iteration, the algorithm selected the sample that exhibited the largest minimum distance with respect to any sample that had been chosen. Such a procedure was repeated until the calibration and validation set were generated. As a result, 75 samples were set as the calibration set, and the remaining 25 samples were set as the validation set to verify the performance of the model.

Weighted bootstrap sampling (WBS)

WBS, derived from the bootstrap method (BSS), was a random sampling technique (Lee and Young 2003). According to this algorithm, different objects were assigned with different weights. In this strategy, all variables had the same chance of being selected in the subsets. And the objects being chosen no less than once were extracted. The selected objects were not unique, some may had great chance of being selected more than once, yet others might have no chance of being selected. Theoretically, the ratio of objects being selected no less than once was about 0.632, which could be obtained from Eqs. 2 and 3, where n was the sampling number and p was the sampling probability of not being selected.

$$p = \left(1 - \frac{1}{n}\right)^n \quad (2)$$

$$\lim_{n \rightarrow \infty} p = \lim_{n \rightarrow \infty} \left(1 - \frac{1}{n}\right)^n = e^{-1} = 0.368 \quad (3)$$

Model population analysis (MPA)

MPA (Li et al. 2010) was a general framework for designing new forms of chemometrics algorithms. The kernel idea of MPA was to extract information from numerous sub-models. The procedure of MPA can be summarized as followings: firstly, a lot of sub-models were generated according to Monte Carlo sampling (MCS). Secondly, RMSECV was considered as the evaluation index of each sub-model. Thirdly, the outputs of all the sub-models were statistically analyzed. The risk of ignoring essential variables was reduced by applying soft shrinkage during the whole processes.

Interval combination optimization (ICO) algorithm

ICO was executed under the framework of MPA (Song et al. 2016). Procedure of ICO could be outlined as follows: firstly, the spectra were generated into N equal-width intervals, and M random combinations via WBS. Secondly, the weight of each interval was calculated by the appearance of frequency. Thirdly, PLS algorithm and five-fold cross validation were used to extract the ratio (α) of optimal intervals with lower RMSECV. Finally, the intervals with lower RMSECV were selected as the optimal intervals.

Software and scripts

The programs were coded in software Matlab (Version 2016a, Mathworks, USA) and were run on a personal

computer. Matlab scripts used in this study are available upon request.

Result and discussion

Parameters on algorithms

KS algorithm was applied to generate representative the calibration and validation sets. In order to distribute the subset uniformly, KS followed the steps of new selections located far from the selected ones. The spectra of the four routine constituents, total sugar, reducing sugar, total nitrogen and nicotine were separated into two sets according to Eq. 1. As a result, 75 samples were set as calibration set and 25 samples as validation set. The calibration set was executed for interval selection firstly and then developed the PLS model. The maximum number of latent variables (LVs) was limited to ten, and the optimal number of LVs was determined by fivefold cross validation technique.

There were three parameters in ICO, i.e., the number of equal-width interval (N), the number of random combinations intervals (M) and the ratio of best interval combinations (α). A larger interval width would resulted in some narrow peaks being omitted, whereas, a smaller interval width would increase the calculation burden as well as have the risk of plunging into local optimum. Therefore, in order to get a stable result, N , M and α were set to 40, 1000 and 0.1, respectively. The advantage of the ICO strategy was to analyze a few sub-intervals with large weights to obtain overall information of the model. The root mean square error of cross validation (RMSECV), determination coefficient (R^2) and root mean square error of prediction (RMSEP) were used to assess model performance.

Determination of the four routine constituents by ICO

The spectral range of *Nicotiana* samples was from 10,000 to 4000 cm^{-1} with an interval of 4 cm^{-1} and each spectrum had 1555 data points. Normally, no significant differences could be observed over the raw spectra (Fig. 2) due to high overlapping. Differentiating the four constituents from the total absorbance spectra was impossible, so multivariate calibration model was carried out for quantitative analysis. Thus, the spectra from the original data matrix were decomposed by ICO and PLS algorithms. Prior to establish the model, the spectra were standardized with center treatment so as to increase the orthogonality of the base vectors. Selecting informative variables before developing the model would bring lots of benefits, such as increasing the model prediction accuracy, reducing

calculation time and getting better understanding of the relationship between characteristic peaks and the data.

The whole spectrum was divided into forty equal sub-intervals and the optimal interval combinations of the four routine constituents were searched by soft shrinkage manner. As a random sampling technology, MPA had the ability of improving predictive performance with the results obtained from multiple models rather than one single model. Besides, MPA was applicable to discrete variables, and resulted in more stable results with meaningful variables.

Figure 1 showed the sampling weights of forty intervals by ICO algorithm of total sugar, reducing sugar, total nitrogen and nicotine. As shown in Fig. 1, the sampling weights were represented by different colors, the color bar varied from dark blue to dark red indicating the weights from 0 percent to 100 percent. Took an interval for example, if it was in dark red at the last iteration, this interval would be retained. On the other hand, if it was in dark blue at the last iteration, it would be excluded. The optimal interval combinations were searched from forty equal sub-intervals by soft shrinkage manner of the four routine constituents. In Fig. 1a, the first, 3rd, 5th and 6th sub-intervals were finally selected after seven iterations. All the sampling weights were set to 1 at the first iteration and the above four intervals were chosen after ICO algorithm. Many intervals (such as 2nd, 9th, 11th, 15th) were

assigned to 1 initially but the weights became smaller with iterations going on and finally these intervals were displayed in blue color and being excluded finally. WBS was a random sampling method with replacements and played an important role in ICO. Applying WBS could avoid regression coefficients colinearity and make the variable ranking more reliable. In soft shrinkage strategy, the variables considered as less significant were assigned to smaller weights, thus, ICO could ensure each informative interval having a chance to be evaluated, which could avoid iteration termination due to local minimum (Song et al. 2016). In Fig. 1b, four sub-intervals of the 3rd, 4th, 6th, 7th were retained as useful information to develop PLS model after thirteen iterations. According to Fig. 1b, the first interval was assigned to a higher sampling weight at the first iteration, however, the weight became smaller after four iterations and finally being excluded, which indicated the uninformative intervals were not eliminated immediately but were assigned to smaller weight. Consequently, the algorithm reduced the risk of removing variables by accident. The 6th interval was excluded after two iterations but was chosen again after twelve iterations and finally retained, indicating the ICO algorithm could ensure the excluded intervals still having a chance to be retained according to WBS. Both the 10th and 23rd intervals were chosen at first, but the color turned to blue at the 13th iteration, illustrating the intervals had the possibility to be

Fig. 1 The sampling weights of forty intervals by ICO algorithm. **a** Total sugar; **b** reducing sugar; **c** total nitrogen; **d** nicotine

excluded even in the last iteration. The attribution of the wavebands corresponding to selected intervals will be described in details below. As shown in Fig. 1c, the 2nd, 6th, 12th and 18th sub-intervals were finally chosen after eleven iterations. The 2nd and 6th intervals were assigned to higher sampling weights from the beginning to the end, whereas, the 18th interval was excluded during iteration but retained finally. In Fig. 1d, the 2nd, 4th, 5th and 14th intervals were selected after twenty iterations. The first interval was chosen initially, but the color turned to blue after eleven iterations and eventually turned to dark blue after twenty iterations, which revealed the first interval made little contribute in developing the model. The 2nd, 4th, 5th and 14th intervals were assigned to higher sampling weights after twenty iterations, and were finally selected to develop the model for the determination of nicotine.

The ICO algorithm of four routine constituents were all retained the intervals that located in $4000\text{--}7000\text{ cm}^{-1}$ (the combination and the first overtone of X–H). Figure 2 was the retained intervals of four routine constituents selected by ICO, and the chosen intervals directly demonstrated the relationship of samples and their composition. Furthermore, almost all the selected intervals were related to the chemical bonds of X–H. Therefore, it was necessary to

identify useful spectral regions before developing the PLS model in order to maximize the investigated components. The useful spectral regions laid the foundations to give a correct interpretation of the regression models.

As shown in Fig. 2a, three intervals located at $4000\text{--}4150\text{ cm}^{-1}$, $4301\text{--}4449\text{ cm}^{-1}$ and $4599\text{--}4899\text{ cm}^{-1}$ were selected for total sugar. The second interval was the combination of the stretching and bending vibrations of C–H and the third one were ascribed to the combination of stretching and bending vibrations of N–H (the left dash area) and O–H (the right dash area). When developing the model, the optimum LVs number was ten (Fig. 3) with the minimum value of the RMSECV. RMSEP (the evaluated index) and R^2 were summarized in Table 1, indicating the ICO algorithm resulted in a good predictive ability upon determination of total sugar. Four intervals were selected by ICO for reducing sugar, which were located at $4301\text{--}4150\text{ cm}^{-1}$, $4451\text{--}4597\text{ cm}^{-1}$ and $4749\text{--}5049\text{ cm}^{-1}$ in Fig. 2b. The first interval and third interval were corresponding to the combination of the stretching and bending vibrations of C–H and O–H, respectively. However, the spectral trough ($4451\text{--}4597\text{ cm}^{-1}$) was retained. The explanation was that the coordination of baseline interval was essential to acquire the elimination criterion. Thus, this region made some efforts to increase the

Fig. 2 The retained variables of four routine constituents selected by ICO. **a** Total sugar; **b** reducing sugar; **c** total nitrogen; **d** nicotine

Fig. 3 The RMSECV values of total sugar, reducing sugar, total nitrogen and nicotine versus the number of latent variables after ICO

prediction performance of reducing sugar model. As illustrated in Fig. 3b in the red line, the optimum LVs was eight. After ICO algorithm, 155 variables were chosen to develop the model. As a result, R^2 , RMSECV and RMSEP were 0.9408, 1.2352 and 1.4464, respectively. The results above indicated that the regression model was effective in determining the reducing sugar of plant samples. Intervals of $4150\text{--}4301\text{ cm}^{-1}$, $4749\text{--}4899\text{ cm}^{-1}$, $5648\text{--}5798\text{ cm}^{-1}$ and $6547\text{--}6698\text{ cm}^{-1}$ were selected for total nitrogen. The first and second interval were associated with the combination of the stretching and bending vibrations of C–H and O–H. Intervals of $5648\text{--}5798\text{ cm}^{-1}$ and $6547\text{--}6698\text{ cm}^{-1}$ were ascribed to the first overtone of C–H and N–H, respectively. Totally, 155 variables were retained after ICO algorithms as described by the dash area in Fig. 2c. The optimum number of LVs was ten shown in Fig. 3c in the blue line. It could be seen from Table 1 that the ICO model delivered a good performance with satisfying chemometric indicators R^2 , RMSECV and RMSEP, which revealed that the ICO algorithms resulted in a good predictive ability upon determination of total nitrogen. In Fig. 2d, three intervals $4150\text{--}4301\text{ cm}^{-1}$, $4451\text{--}4749\text{ cm}^{-1}$ and

$5947\text{--}6097\text{ cm}^{-1}$ were selected by ICO of nicotine. The first and second intervals were corresponding to the combination of stretching and bending vibrations of C–H and N–H. However, the spectral trough of $5947\text{--}6097\text{ cm}^{-1}$ was also included in calculation, and this interval was considered as the criterion of elimination in ICO calculation in coordination with the baseline variables. The results of nicotine after ICO were outlined in Table 1, which demonstrated the regression model was effective for quantifying nicotine.

ICO was an important approach in extracting useful chemical information with WBS as a core technology to generate random combinations. Results adequately suggested NIR combined with ICO algorithms could be effectively used in the determination of total sugar, reducing sugar, total nitrogen and nicotine. Moreover it also revealed that the retained intervals of four routine constituents directly demonstrated the relationship of the samples and their composition. Therefore, it was necessary to identify useful spectral regions before developing PLS models in order to maximize the investigated components.

Comparison of ICO selection and full-spectrum modeling

The performance of full-spectrum PLS models of four routine chemical constituents with optimal parameters were listed in Table 1. The optimal number of LVs determined by the RMSECV of the PLS models were 4, 5, 8 and 10 for total sugar, reducing sugar, total nitrogen and nicotine, respectively.

In total, 1555 variables were used in the full-spectrum whereas 155 variables were retained by ICO for each routine chemical constituent. As shown in Table 1, ICO generated better results than full-spectrum PLS. Linear curve of predicted versus measured value of the calibration set of four constituents were shown in Fig. 4. Obviously, the results of ICO were superior to the full-spectrum PLS

Table 1 Results of ICO and full-spectrum of the four constituents of total sugar, reducing sugar, total nitrogen and nicotine

Constituents	Methods	Ranges of selected variables (cm^{-1})	nVAR	nLVs	RMSECV (%)	R^2	RMSEP (%)	F-test
Total sugar	ICO	4899–4599, 4448–4301, 4150–4000	155	10	0.8573	0.985	0.8672	2.4360
	Full spectrum	10,000–4000	1555	4	1.1181	0.9672	1.3535	
Reducing sugar	ICO	5049–4749, 4597–4451, 4301–4150	155	8	1.2352	0.9408	1.4464	1.1460
	Full spectrum	10,000–4000	1555	5	1.513	0.9111	1.5484	
Total nitrogen	ICO	6698–6547, 5798–5648, 4899–4749, 4301–4150	155	10	0.061	0.937	0.0432	2.5808
	Full spectrum	10,000–4000	1555	8	0.0657	0.9271	0.0694	
Nicotine	ICO	6097–5947, 4749–4451, 4301–4150	155	10	0.1037	0.9648	0.1078	2.0329
	Full spectrum	10,000–4000	1555	10	0.1588	0.9175	0.1537

Fig. 4 Scatter plot of predicted versus actual value in calibration and validation set. **a** Total sugar; **b** reducing sugar; **c** total nitrogen; **d** nicotine

according to the evaluated index RMSEP. Furthermore, the R^2 of the four constituents of full-spectrum PLS were less satisfactory than ICO the result (Fig. 5b). As expected, ICO yielded better results than full-spectrum PLS might be attributed to taking away the irrelevant spectral information.

F-test was performed to test the difference of two models. The criterion of the statistical testing approach could be obtained via the F-test distribution table. If the F-test value was larger than the criterion, meaning the two models had significantly difference. The F-test values for the ICO and full-spectrum PLS of the four constituents were outlined in Table 1. The criterion to determine the models was 1 in the F-test distribution table ($\alpha = 0.05$, $F(\infty, \infty)$). As obtained from Table 1, F-test for the four constituents were all larger than 1. The overall results indicated that there were significant differences between ICO and full-spectrum PLS for the four constituents. Specifically, the interval selection methods had improved the predictive performance compared with full-spectrum PLS.

ICO could improve the effectiveness, the predictive performance as well as the robustness of the model by useful intervals with relationship to the four routine constituents. The results showed that NIR combined with ICO

performed better in detecting the four routine constituents than with full-spectrum PLS and would be further potentialized in terms of its capacity for fast and accurate determination.

Conclusion

Applying ICO to determine four routine quality associated constituents of total sugar, reducing sugar, total nitrogen and nicotine was studied in this work. Results showed that the performance of the four routine constituents after ICO were all superior to that of the full-spectrum PLS. Furthermore, the F-test showed that there were significant differences between ICO and full-spectrum PLS for the four constituents, and manifested ICO had improved the predictive performance significantly compared with full-spectrum PLS.

The results adequately suggested that the NIR combined with ICO algorithms could be effectively used in the determination of four routine chemical constituents in *Nicotiana* plant leaf samples. Moreover it also revealed that the retained variables of four routine constituents selected by ICO could directly demonstrate the relationship of the

Fig. 5 Comparison of the RMSEP and R^2 by ICO and full-spectrum PLS. **a** RMSEP; **b** R^2

samples and their composition. Conclusively, ICO algorithm eliminated the uninformative variables and improved the prediction ability of the model. As a variable selection method, ICO will be a promising tool for NIR analysis of complex samples and it would be prospective to monitor industrial crops quality during on-site production process in the future.

Acknowledgements The authors would like to thank Prof. Suqin Sun (Tsinghua University, Beijing, China) for the critical review of the manuscript. The authors are also grateful to Yunnan Tobacco Industry for providing samples used in this work. This research was supported by the National Natural Science Foundation of China (NSFC) and Fundamental Research Funds for the Central Universities.

Funding This study was funded by the National Natural Science Foundation of China (NSFC) (No. 31301685), and Fundamental Research Funds for the Central Universities (2652015164; 3142017100).

Compliance with ethical standards

Conflict of interest The authors declare that there is no conflict of interests.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

Informed consent Informed consent was obtained from all individual participants included in the study.

References

- Abrahamsson C, Johansson J (2003) Comparison of different variable selection methods conducted on NIR transmission measurements on intact tablets. *Chemom Intel Lab Syst* 69:3–12
- Broadhurst D, Goodacre R, Jones A, Rowland JJ, Kell DB (1997) Genetic algorithms as a method for variable selection in multiple linear regression and partial least squares regression, with applications to pyrolysis mass spectrometry. *Anal Chim Acta* 348:71–86
- Duan J, Huang Y, Li ZH, Zheng B, Li QQ, Wu LJ, Min SG (2012) Determination of 27 chemical constituents in Chinese southwest tobacco by FT-NIR spectroscopy. *Ind Crops Prod* 40:21–26
- Fu HY, Huang SY, Xu L, Tang LJ, Jiang JH, Yu RQ (2007) Moving window partial least-squares discriminant analysis for identification of different kinds of bezoar samples by near infrared spectroscopy and comparison of different pattern recognition methods. *J Near Infrared Spectrosc* 15:291–295
- Geladi P, Kowalski BR (1985) Partial least-squares regression: a tutorial. *Anal Chim Acta* 185:1–17
- Han QJ, Wu HL, Cai CB, Xu L (2008) An ensemble of Monte Carlo uninformative variable elimination for wavelength selection. *Anal Chim Acta* 612:121–125
- Höskuldsson A (1988) PLS regression methods. *J Chemom* 2:211–228
- Huang JH, Yan J, Wu QH, Duarte FM, Yi LZ, Lu HM, Xu QS, Liang YZ (2013) Selective of informative metabolites using random forests based on model population analysis. *Talanta* 117:549–555
- Kennard RW, Stone LA (1969) Computer aided design of experiments. *Technometrics* 11:137–148
- Leardi R, Nørgaard L (2004) Sequential application of backward interval partial least squares and genetic algorithms for the selection of relevant spectral regions. *J Chemom* 18:486–497
- Lee SS, Young GA (2003) Pre pivoting by weighted bootstrap iteration. *Biometrika* 90:393–410
- Li H, Liang Y, Xu Q, Cao D (2010) Model population analysis for variable selection. *J Chemom* 24:418–423
- Li HD, Liang YZ, Cao DS, Xu QS (2012) Model-population analysis and its applications in chemical and biological modeling. *TrAC* 38:154–162
- Mevik BH, Cederkvist HR (2004) Mean squared error of prediction (MSEP) estimates for principal component regression (PCR) and partial least squares regression (PLSR). *J Chemom* 18:422–429
- Pereira FC, Pontes JC, Neto FG, Santos RB, Galvão KH, Araújo CU (2008) NIR spectrometric determination of quality parameters in vegetable oils using iPLS and variable selection. *Food Res Int* 41:341–348
- Ritota M, Manzi P (2018) Melamine detection in milk and dairy products: traditional analytical methods and recent developments. *Food Anal Methods* 11:128–147
- Shao YN, Yong H, Wang YY (2007) A new approach to discriminate varieties of tobacco using vis/near infrared spectra. *Eur Food Res Technol* 224:591–596
- Song XZ, Huang Y, Yan H, Xiong YM, Min SG (2016) A novel algorithm for spectral interval combination optimization. *Anal Chim Acta* 948:19–29
- Syunyaev RZ, Balabin RM, Akhatov IS, Safieva JO (2009) Adsorption of petroleum asphaltenes onto reservoir rock sands studied by near-infrared (NIR) spectroscopy. *Energy Fuels* 23:1230–1236

- Tan C, Qin X, Li ML (2008) An ensemble method based on a self-organizing map for near-infrared spectral calibration of complex beverage samples. *Anal Bioanal Chem* 392:515–521
- Wold S, Kettaneh-Wold N, Skagerberg B (1989) Nonlinear PLS modeling. *Chemom Intell Lab Syst* 7:53–65
- Ye SF, Wang D, Min SG (2008) Successive projections algorithm combined with uninformative variable elimination for spectral variable selection. *Chemom Intell Lab Syst* 91:194–199
- Zhang X, Li W, Yin B, Chen W, Kelly DP, Wang X, Du Y (2013) Improvement of near infrared spectroscopic (NIRS) analysis of caffeine in roasted arabica coffee by variable selection method of stability competitive adaptive reweighted sampling (SCARS). *Spectrochim Acta A* 114:350–356
- Zontov YV, Balyklova KS, Titova AV, Rodionova OY, Pomerantsev AL (2016) Chemometric aided NIR portable instrument for rapid assessment of medicine quality. *J Pharm Biomed Anal* 131:87–93
- Zuo XB, Fang S, Liang XL (2014) Synergy interval partial least square (siPLS) with potentiometric titration multivariate calibration for the simultaneous determination of amino acids in mixtures. *Adv J Food Sci Technol* 6:1209–1218

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.