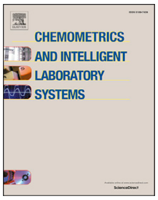
Chemometrics and Intelligent Laboratory Systems 185 (2019) 114–121

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemometrics



An improved weighted multiplicative scatter correction algorithm with the use of variable selection: Application to near-infrared spectra

Yifan Wu [a,b], Silong Peng [a,b,*], Qiong Xie [a], Quanjie Han [a,b], Genwei Zhang [c], Haigang Sun [d]

[a] Institute of Automation, Chinese Academy of Sciences, 100190, Beijing, China
[b] University of Chinese Academy of Sciences, 100190, Beijing, China
[c] State Key Laboratory of NBC Protection for Civilian, Beijing, 102205, China
[d] GWDC Mud Logging Company, CNPC, 124010, Panjin, China

ARTICLE INFO

Keywords:
Multiplicative scatter correction
Weighted least squares
Variable selection
Model population analysis

ABSTRACT

Multiplicative light scattering has posed great challenge in near-infrared (NIR) quantitative analysis. When estimating the scattering parameters, uninformative variables for scattering effects may bias the estimation. Weighted least squares (WLS) can be used to avoid the influence of the uninformative variables. In this work, we proposed an improved weighted multiplicative scatter correction algorithm with the use of variable selection (WMSCVS). Baseline is removed first and then variable selection is used to obtain the optimal weights of WLS in estimating multiplicative parameters. The variable selection algorithm, which is designed based on model population analysis (MPA), implements an iterative optimization process. In each iteration, weighted bootstrap sampling (WBS) is used to generate variable subsets and exponentially decreasing function (EDF) is used to control the number of sampled variables. The interpretability and stability of the variable selection results as well as the predictive performance of the corrected spectra were investigated by using two NIR datasets. The experimental results showed that the proposed WMSCVS could give better predictive performance than the state-of-art correction methods.

1. Introduction

In recent years, near-infrared (NIR) spectroscopy has been more and more widely used [1,2]. Partial least squares (PLS) is usually used to build calibration model for quantitative analysis [3]. When the samples to be analyzed are solid or emulsions and dispersions, multiplicative light scattering effects due to the variations of optical path lengths deteriorate the performance of PLS model. Multiplicative scatter correction (MSC) [4] and standard normal variate (SNV) [5] are the most widely-used scatter correction methods for NIR spectra. MSC attempts to estimate the coefficient which describes the scattering by regressing the spectrum to correct onto a reference spectrum while SNV subtracts the spectrum mean from each spectral variable and subsequently dividing that value by the standard deviation of the spectrum. However, when the raw spectra contain large chemical variations, the correction performance of MSC and SNV is poor.

The basic MSC model has been extended to include new parameters to account for the physical and chemical factors that affect the measured absorbance spectra, the reputed extended multiplicative signal correction (EMSC) [6,7]. The inverse MSC models are termed inverted scatter correction (ISC) [8] and extended inverted signal correction (EISC) [9]. The linear and quadratic terms of wavelengths and a quadratic term relating to the spectrum are involved in EISC model. The prior of pure spectra of sample components is also involved to extend the EISC model. When the prior is available, the EMSC and EISC can show good performance of scatter correction [10]. However, if the pure components spectra information is not available, the parameter estimation of the above methods tends to be biased [11].

A solution to the biased estimation is to use some of the variables, or wavelengths, for parameter estimation instead of using the whole spectrum. The reason is that the influence of uninformative variables for scattering effects can be avoided. Based on the framework of SNV, Bi et al. [12] proposed interference dominant region correction (IDRC), which selected a region dominated by scattering effects for estimating the SNV parameters. The selection of the interference dominant region is based on trial-and-error and the root mean square error of cross-validation (RMSECV) of PLS model is used as the measurement of the goodness of the selection. Based on the framework of EMSC,

* Corresponding author. Institute of Automation, Chinese Academy of Sciences, 100190, Beijing, China.
E-mail address: silong.peng@ia.ac.cn (S. Peng).

https://doi.org/10.1016/j.chemolab.2019.01.005
Received 23 May 2018; Received in revised form 7 January 2019; Accepted 10 January 2019
Available online 14 January 2019

Gallagher et al. [13] proposed to use weighted least squares (WLS) for parameter estimation, which sets the weights of spectral regions known to contain interesting features to zero. However, the prior of the location of interesting features is usually not available. Therefore, an automatic way of setting the weights is in urgent need.

Variable selection has been popular in chemometrics [14–16]. Model population analysis (MPA) is a general framework for designing new chemometrics algorithms and it can be employed to design variable selection algorithms [17–19]. MPA emphasizes that information should be extracted by analyzing a number of sub-models statistically since the results or parameters of one single model are not always reliable. In detail, MPA usually contains three stages: (1) sub-datasets generation procedure, where random sampling method is applied to obtain a series of sub-datasets from variable or sample space, such as jackknife sampling, weighted bootstrap sampling (WBS) [18] and binary matrix sampling (BMS) [20]; (2) modeling procedure, where a series of sub-models are established based on sub-datasets generated in the previous step; (3) statistical analysis procedure, where interested outputs (e.g., RMSECV value) of all these sub-models are analyzed statistically.

Inspired by variable selection's powerful ability to detect informative variables for a specific task, we proposed to use variable selection to obtain the weights of WLS in estimating multiplicative parameters, termed as weighted multiplicative scatter correction with variable selection (WMSCVS). The variable selection was performed on baseline-removed spectra. The variable selection algorithm was designed based on MPA. WBS was used to sample variable and exponentially decreasing function (EDF) [21] was used to control the number of sampled variables. There are two advantages when considering using variable selection based on MPA to set the weights of WLS. One is that the variable selection can ensure the optimality of the performance of scatter correction. The other is that MPA extracts information from a large number of sub-models, which is beneficial for the stability of variable selection.

2. Method

2.1. Multiplicative light scattering

For relatively simple systems, the effects of light scattering can be approximated by the following EMSC model [7]:

$$\boldsymbol{x}_i = a_i 1 + b_i \boldsymbol{x}_{i,chem} + d_i \boldsymbol{\lambda} + e_i \boldsymbol{\lambda}^2 + \varepsilon_i \tag{1}$$

where the row vectors $\boldsymbol{x}_i$ and $\boldsymbol{x}_{i,chem}$ are the measured absorbance spectrum and the theoretical spectrum of the ith sample respectively. 1 is a row vector with its elements equal to unity. The parameters a_i and b_i denote the additive and multiplicative effects of light scattering. d_i and e_i are introduced to account for the smooth wavelength-dependent spectral variations. The wavelength row vector $\boldsymbol{\lambda}$ is a linear function of the wavelength, and the entries lie between −1 and +1. ε_i captures the unknown sources of spectral variation.

2.2. Baseline removal

The influence of the baseline offset (a_i) and the smooth wavelength-dependent spectral variations (d_i and e_i) can be removed by projecting the measured spectrum $\boldsymbol{x}_i$ onto the orthogonal complement of space spanned by the row vectors of $\boldsymbol{P} = [1; \lambda; \lambda^2]$ [11]:

$$\begin{aligned} \boldsymbol{z}_i &= \boldsymbol{x}_i (\boldsymbol{I} - \boldsymbol{P}^+ \boldsymbol{P}) \\ &= b_i \boldsymbol{z}_{i,chem} + \varepsilon_i^* \\ \boldsymbol{z}_{i,chem} &= \boldsymbol{x}_{i,chem} (\boldsymbol{I} - \boldsymbol{P}^+ \boldsymbol{P}), \varepsilon_i^* = \varepsilon_i (\boldsymbol{I} - \boldsymbol{P}^+ \boldsymbol{P}) \end{aligned} \tag{2}$$

where $\boldsymbol{z}_i$ is the orthogonal-projection preprocessed spectrum of the ith sample.

2.3. Weighted multiplicative scatter correction

If the multiplicative parameter b_i in Eq. (2) had been known theoretically, or estimated perfectly, then the following correction would remove the multiplicative scaling effect:

$$\boldsymbol{z}_{i,corr} = \frac{\boldsymbol{z}_i}{b_i} \tag{3}$$

where $\boldsymbol{z}_{i,corr}$ is the scattering-corrected spectrum of the ith sample.

In terms of the estimation of b_i, if we assume that the spectral variations only result from multiplicative effects and ignore the other factors such as the chemical variations between samples, the mean spectrum $\boldsymbol{m} = \sum_{i=1}^{I} \boldsymbol{z}_i$ can be used as the reference spectrum of an “ideal” sample and each sample's spectrum is then corrected so that all samples appear to have the same scatter level as the “ideal” [4]. Therefore, an estimation of b_i can be obtained by regressing $\boldsymbol{m}$ onto $\boldsymbol{z}_i$:

$$\hat{b}_i = \boldsymbol{z}_i \boldsymbol{m}^\intercal (\boldsymbol{m}\boldsymbol{m}^\intercal)^{-1} \tag{4}$$

where $\intercal$ denotes the matrix transposition operation.

However, in practice, the variations of the other factors are usually not ignorable. In some variables the variations of multiplicative effects are dominant and we term these variables as informative variables for multiplicative effects. While in some variables the variations of the other factors are dominant and we term these variables as uninformative variables for multiplicative effects. In order to avoid the influence of the uninformative variables, WLS is used to estimate the multiplicative parameters [13]:

$$\hat{b}_i = \boldsymbol{z}_i \boldsymbol{W} \boldsymbol{m}^\intercal (\boldsymbol{m}\boldsymbol{W}\boldsymbol{m}^\intercal)^{-1} \tag{5}$$

where $\boldsymbol{W}$ is a diagonal matrix with elements 0 or 1. Element 1 means that the corresponding variable is selected for estimation while element 0 means that it is eliminated.

2.4. Variable selection for weights setting

The critical problem is the setting of the weighting matrix $\boldsymbol{W}$, that is, the selection of informative variables for multiplicative parameters estimation. One may try to select the variables by visual inspection with the principle that the selected variables should not contain spectral shape change since multiplicative effects will not cause spectral shape change. However, there are two disadvantages. One is that the selection result is subjective, which varies from people to people. The other is that the selection result is not necessarily optimal.

In this study, we proposed to use variable selection to set the weights of WLS to obtain an optimal and objective solution in estimating multiplicative parameters, which is termed as WMSCVS. The variable selection algorithm is designed based on the framework of MPA. It is an iterative algorithm where WBS is used to generate variable subsets and EDF is used to control the number of sampled variables in each iteration. In terms of the modeling procedure of MPA, we build PLS model with the corrected spectra and the RMSECV value of the PLS is analyzed statistically. The details of the variable selection algorithm are introduced in the following.

2.4.1. Weighted bootstrap sampling

WBS is used for random sampling with replacement [22], which is derived from bootstrap sampling. The goal of WBS is to select a number (say R) of objects from a pool of objects with certain selection probability for each object. R trials will be conducted and one object will be selected in each trial. The probability of being selected for each object is determined by the sampling weights. The objects with larger weights have higher probabilities to be selected. It is noted that after one run of WBS, the selected objects are not necessary to be unique, that is, there may be some objects being selected for more than once while some not being selected at all. In this study, the objects selected more than once are included only once in the estimation step.

2.4.2. Exponentially decreasing function

In order to ensure the efficiency of the optimization process, EDF [21] is used to control the number of sampled variables in each iteration. In the *i*th iteration, the ratio of variables to be kept is computed using an EDF defined as:

$$r_i = ae^{-ki} \tag{6}$$

where a and k are two constants determined by specific conditions, which are calculated as:

$$a = \left(\frac{p}{2}\right)^{1/(N-1)} k = \frac{\ln(p/2)}{N-1} \tag{7}$$

where p is the number of wavelengths, N is the number of iterations, *ln* denotes the natural logarithm. The advantage of EDF is that the process of variable reduction can be roughly divided into two stages. In the first stage, variables are eliminated rapidly which performs a 'fast selection', whereas in the second stage, variables are reduced in a very gentle manner, which is instead called a 'refined selection' stage. Therefore, the variables with large chemical information can be removed in a stepwise and efficient way because of the advantage of EDF.

2.4.3. Procedure of the proposed variable selection algorithm

Assume that the number of iterations is set to N, in each iteration, the procedure of the proposed variable selection algorithm can be summarized in the following steps:

(1) WBS is performed K (e.g., 500) times on the variable space so that K variable subsets are generated. The number of variables for each subset is calculated by EDF (Eq. (6)). Note that the initial weight of each variable is set to be equal with each other.
(2) For each variable subset, the multiplicative parameters are estimated with the selected variables (Eq. (5)) and the corrected spectra were obtained (Eq. (3)). PLS model is built with the corrected spectra and RMSECV value of the PLS is calculated by five-fold cross-validation.
(3) Extract a ratio α (e.g., 10%) of best variable subsets with lowest RMSECV values.
(4) Weights updating for WBS. Count the appearance frequency of each variable in the best variable subsets and the new weight of variable m can be calculated as following:

$$w_m = \frac{f_m}{k_{best}} \tag{8}$$

where f_m is the appearance frequency of the m-th variable in the best variable subsets, k_{best} is the number of best variable subsets.

In each iteration, the variable subset with the lowest RMSECV value is recorded. After all iterations, the subset with the lowest RMSECV value among all recorded ones will be considered as the optimal variable subset. Fig. 1 shows the flowchart of the variable selection algorithm.

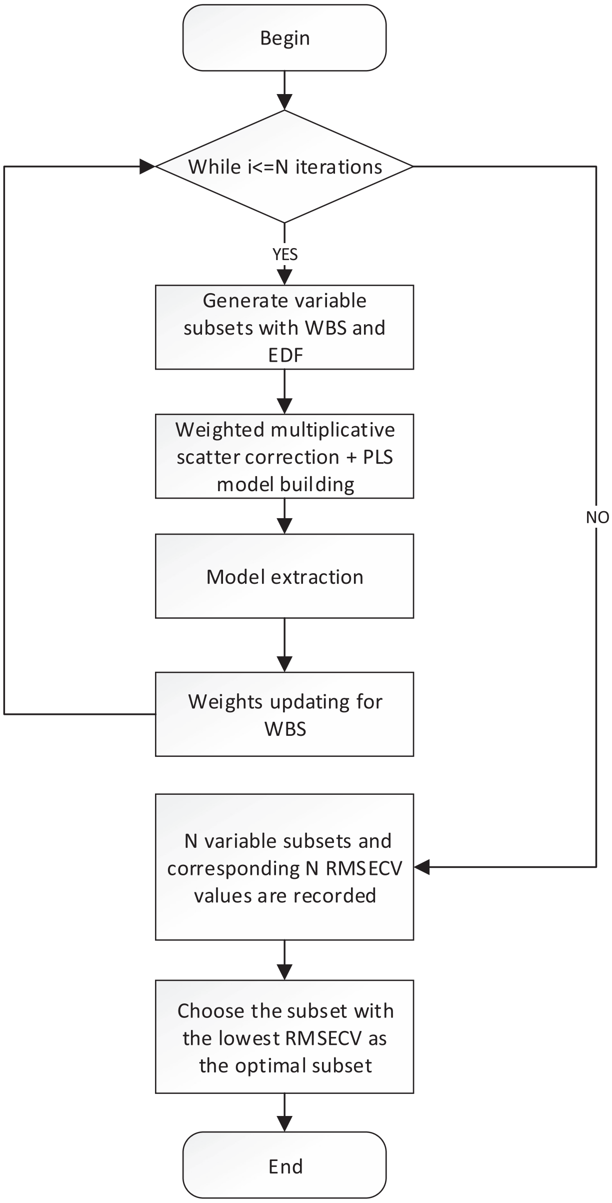

Begin
While i<=N iterations
YES
Generate variable subsets with WBS and EDF
Weighted multiplicative scatter correction + PLS model building
Model extraction
Weights updating for WBS
NO
N variable subsets and corresponding N RMSECV values are recorded
Choose the subset with the lowest RMSECV as the optimal subset
End

Fig. 1. Flowchart of the variable selection of WMSCVS.

3. Experimental

3.1. Dataset

The first data set (meat data set), termed Tecator, originated from the food industry [23] (available from http://lib.stat.cmu.edu/datasets/tecator). It consisted of 215 near infrared absorbance spectra of meat samples, recorded on Tecator Infratec food and feed analyzer working in the wavelength range of 850–1050 nm with an interval of 2 nm by the NIR transmission principle. Each sample contained finely chopped pure meat with different contents. The contents of moisture (39.3–76.6%), fat (0.9–49.1%) and protein (11–21.8%) were investigated in the following analysis. As indicated by data owner, the calibration set contains 129 samples and other 86 samples were selected as the prediction set.

The second data set was a four-component suspension system provided by Steponavicius and Thennadil [24]. The data composed of three fully miscible absorbing species of water, deuterium oxide (D_2O), ethanol and a species that both absorbed and scattered light (that is, a particulate species of polystyrene). A total of 45 samples were prepared using various combinations of the concentrations of the four components and spectral data were collected in the wavelength region of 1500–1880 nm at 2 nm intervals with an integration time of 10 s, resulting in measurements at 191 discrete wavelengths per spectrum. The concentrations of the water (32.6–76.9 vol%) and deuterium oxide (20.1–58.0 vol%) were investigated in the following analysis, and the total diffuse transmittance spectra were transformed into absorbance spectra before modeling. For dataset division, the Kennard-Stone (KS) algorithm [25] was used, resulting in 33 (75%) calibration samples and 12 (25%) prediction samples.

3.2. Computation

The proposed WMSCVS was compared with other scattering correction methods including MSC, SNV, EMSC, EISC and IDRC. The PLS method was used to train calibration model with the use of the corrected spectra. Five-fold cross-validation was performed to the calibration set to calculate the RMSECV value of PLS. An F-test based on RMSECV was involved to select the optimal number of latent variables [26]. The significance level was set to 0.25 as suggested previously. The EMSC and EISC without the terms of pure component spectra were used for comparison. For IDRC, the maximum number of regions for equal segmentation was set to 10. For WMSCVS, the number of iterations of EDF was set to 50. In each iteration, 500 variable subsets were generated by WBS. The ratio of the model extraction, α, was set to 10%.

3.3. Evaluation of the methods

In this work, the root mean square error of prediction (RMSEP) from the test set was used as a measure of model performance. RMSEP is defined as follows:

$$RMSEP = \sqrt{\frac{\sum \left(y_{pre} - y_{ref}\right)^2}{N_{test}}} \tag{9}$$

where y_{pre} is the predicted value, y_{ref} is the reference value, and N_{test} is the number of samples in the test set.

The stability of the variable selection of WMSCVS is measured by the similarity of the variable sets selected in 50 runs [27,28]. First, the similarity of any two different selections is calculated as follows:

$$S_{ij} = \frac{|v_i \cap v_j| - E\left(|v_i \cap v_j|\right)}{\sqrt{|v|_i \times |v|_j} - E\left(|v_i \cap v_j|\right)} \tag{10}$$

where v_i and v_j denote the variable sets in ith and jth runs of variable selection respectively. $|v_i \cap v_j| \left(0 \le |v_i \cap v_j| \le \sqrt{|v_i| \times |v_j|}\right)$ represents the number of common variables between v_i and v_j. Then, the stability of the variable selection method is calculated with the average of the pair-wise similarity for all possible pairs as follows:

$$stability = \frac{2}{r \times (r-1)} \sum_{i=1}^{r-1} \sum_{j=i+1}^{r} S_{ij} \tag{11}$$

where r is the total number of runs. The greater the value of stability, the more stable is the method.

3.4. Software implementation

All computations were performed in Matlab 2015a (Mathworks, Inc., Natick, MA, USA) and run on a personal computer with 2.20-GHz Intel Core 2 processor, 4 GB RAM, and Windows 7 operating system. The programs were written in-house using Matlab language. The Matlab codes for implementing WMSCVS have been made public, which can be downloaded from https://github.com/jkk544/WMSCVS.

4. Results and discussion

4.1. Tecator data

Firstly, we presented the result of variable selection of WMSCVS. The variable selection result of IDRC was used for comparison. Fig. 2 shows the results of variable selection of IDRC and WMSCVS, with the raw spectra of Tecator data and the baseline-removal spectra as background respectively. It can be seen that the raw spectra suffered from serious additive baseline shift and multiplicative effects (left part of Fig. 2). Though the additive baseline effects and possible wavelength-dependent spectral variations could be readily removed by orthogonal projection preprocessing (right part of Fig. 2), the removal of multiplicative effects was more than orthogonal projection. It was obvious that variables of 900–940 nm contained large spectral shape change, which indicated that these variables contained information mostly not of multiplicative effects.

Therefore, they were uninformative for estimating the multiplicative parameters and they might bias the estimation since they represented information from other factors of spectral variations. WLS could minimize the influence of the uninformative variables by setting their weights to zero. The problem was that an optimal solution of weights for WLS was more than visual inspection. In this study, the proposed WMSCVS used a variable selection algorithm designed based on the framework of MPA to determine the optimal solution. The black bars in Fig. 2 indicate the locations of the selected variables for both IDRC and WMSCVS. There are two key differences between IDRC and WMSCVS. One is that IDRC is based on region search while WMSCVS is based on combination optimization of individual variables. The other is that IDRC performs variable selection on the raw spectra while WMSCVS performs variable selection on the baseline-removed spectra. It was difficult to interpret the selection result of IDRC since various factors of spectral variations were mixed together in the raw spectra. In contrast, the selection result of WMSCVS was more interpretable. It could be seen that WMSCVS selected variables of three different locations containing small spectral variations. In addition to variables of 900–940 nm, variables such as 960–1000 nm

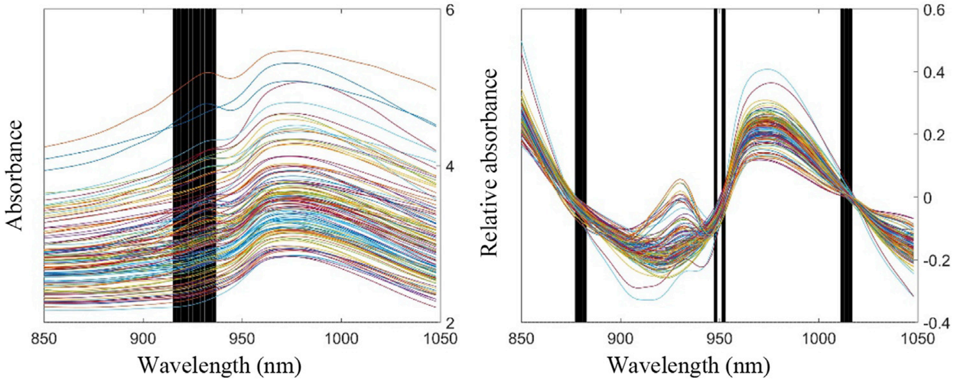


Fig. 2. The results of variable selection on Tecator (moisture) data. Left part: IDRC with the raw spectra as background, right part: WMSCVS with the baseline-removed spectra as background.

were also considered to be uninformative for WMSCVS, which was not obvious with visual inspection.

Secondly, we investigated the process of the variable selection. Moisture was used for illustration and the results of fat and protein were similar. Fig. 3 shows the changing trend of the number of sampled variables and five-fold RMSECV values with the increasing of iterations. As expected, the number of sampled variables decreased fast at the beginning and became more and more slowly. Such characteristic was due to EDF and it guaranteed the efficiency of the selection process. The RMSECV values first descend quickly from iterations 1–25 which should be ascribed to the elimination of the uninformative variables, then changes in a gentle way from sampling runs 26–50 corresponding to the phase that the sampled variables do not change obviously. In this study, the variable subset with the lowest RMSECV value was selected as the best subset for WLS.

Thirdly, the proposed WMSCVS was compared with other correction methods (MSC, SNV, EMSC, EISC and IDRC) on PLS performance (RMSEP from test set). We conducted WMSCVS 50 runs independently and calculated the mean and the standard deviation of the PLS results. The number of latent variables (LVs) used in PLS model and the RMSEP values of various methods are summarized in Table 1.

It can be seen that the PLS model with raw spectra could not provide satisfactory predictive performances primarily due to the existence of scattering effects on the spectra. MSC and SNV gave better predictive performance than PLS with no scatter correction. However, the corrections were considered to be at low efficiency. The RMSEP values of EMSC were significantly large. The poor performance of EMSC was due to the bias in estimating the correction parameters. This occurred when the pure component spectra were unavailable. Conversely, the EISC method provided good performance for moisture and fat. The results also showed that these methods were unstable without the pure component spectra. In consideration that the information of pure component spectra was usually not available in practice, both IDRC and the proposed WMSCVS tried to select part of the variables to estimate the scattering parameters so that the influence of the uninformative variables could be reduced. IDRC used the region selection strategy while WMSCVS used the individual variable selection strategy. Therefore, WMSCVS was more flexible in terms of combination optimization. For each component, it could be seen that WMSCVS achieved the lowest RMSEP among all methods for the three components. Moreover, the standard deviations of different runs of WMSCVS were almost negligible in relative to the corresponding means. The RMSEP-LV plot of the comparison methods are shown in Fig. 4. The results for baseline correction (BC), but no multiplicative treatment was also provided in Table 1 to demonstrate the improved performance due to the proposed variable selection strategy.

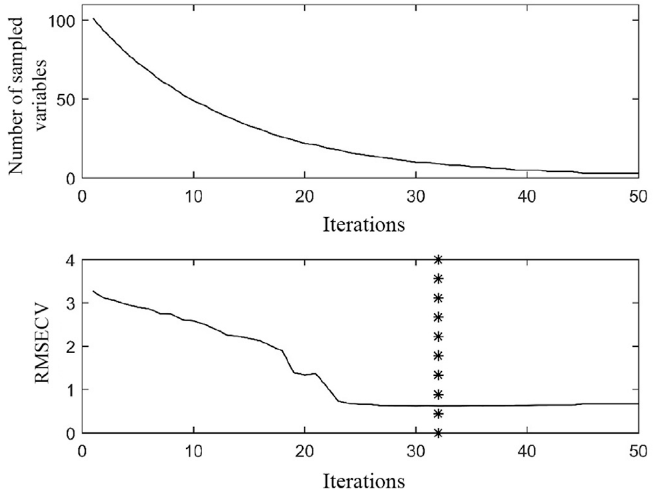


Fig. 3. The changing trend of the number of sampled variables (upper part) and RMSECV values (lower part) with the increasing iterations on Tecator (moisture) data. The vertical asterisk line denotes the optimal point where the RMSECV values achieve the lowest.

Table 1
Results of PLS models with different scatter correction methods on Tecator data.The number in parentheses is the standard deviation of results in 50 independent runs.

Method	Moisture		Fat		Protein	
	LVs	RMSEP	LVs	RMSEP	LVs	RMSEP
RAW	13	2.70	13	2.81	12	0.72
BC	10	2.44	10	2.76	9	0.88
MSC	5	2.05	6	2.23	10	0.60
SNV	7	1.85	4	1.98	12	0.61
EMSC	4	3.98	8	4.67	9	1.08
EISC	8	1.31	13	0.95	7	0.94
IDRC	11	1.08	8	1.07	11	0.55
WMSCVS	11(0)	**0.62** (< 0.001)	10(0)	**0.54** (0.009)	10(0)	**0.53** (0.001)

Bold means the lowest value among all methods.

Finally, the stability of the variable selection of WMSCVS was investigated. For each component, the variable selection results of 50 WMSCVS runs were shown in Fig. 5. It can be seen that WMSCVS showed good stability on Tecator data. For each component, WMSCVS selected the variables from the same three locations in each run. Especially for moisture and protein, WMSCVS almost selected the same variables in each run. Moreover, different components offered the same selection result, which demonstrated the robustness of WMSCVS. Table 2 shows the quantitative index of the stability of the variable selection.

4.2. *Four component data*

Firstly, we investigated the result of the variable selection of WMSCVS. IDRC was used for comparison. Fig. 6 shows the selected variables of IDRC and WMSCVS respectively. The total transmittance spectra of the four-component suspension samples are presented in the left part of Fig. 6. Baseline shift and multiplicative effects can be observed because of the variation of polystyrene particle and concentration. The broad peak in the wavelength region of 1600–1750 nm because of the formation of $H_2O - D_2O$ dimers is a nonlinear spectral response [24], which leads to more latent variables used when PLS modeling on the target components. After orthogonal-projection preprocessing (right part of Fig. 6), the baseline effect was readily removed. It can be seen that variables of 1760–1820 nm contained large spectral shape change and they should not be included when estimating multiplicative parameters. The black bars in Fig. 6 indicate the locations of the selected variables for both IDRC and WMSCVS. The selection result of IDRC was difficult to interpret. However, it was readily interpretable that WMSCVS selected variables from four different locations and avoided the variable region of 1760–1820 nm which contained large spectral shape change. In addition, variable regions such as 1560–1600 nm and 1650–1690 nm were also considered to be uninformative for WMSCVS.

Secondly, the variable selection process of WMSCVS was investigated with the use of water as example. WMSCVS implemented variable selection in an iterative manner based on the framework of MPA. Fig. 7 shows the changing trend of the number of sampled variables and the five-fold RMSECV value with the increasing of sampling runs. With the elimination of the uninformative variables, the RMSECV value decreased at the beginning and then increased because of the elimination of some essential variables. The weights of WBS during the iterations are shown in Fig. 8. It can be seen that the optimal variable subset was searched in a soft shrinkage manner. In detail, uninformative variables were not eliminated directly, but were assigned to smaller sampling weights, which can help to lower the risk of removing informative variables by mistake. The optimal variable subset is obtained in iteration 39. Therefore, the informative variables for estimating multiplicative parameters are thus obtain around 1506 nm, 1610 nm, 1700 nm and 1850 nm.

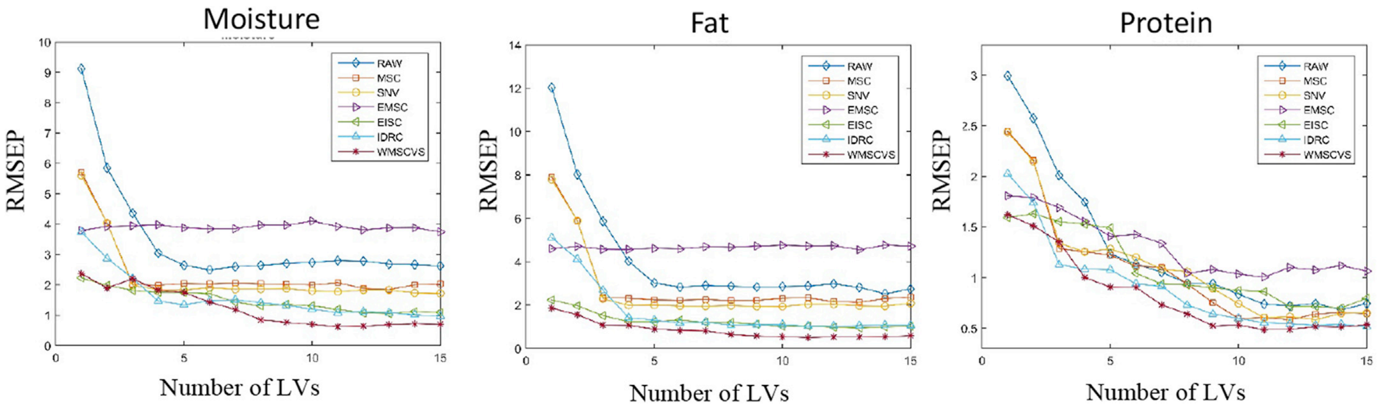


Fig. 4. The RMSEP vs LVs plot of PLS models with different scatter correction methods on Tecator data.

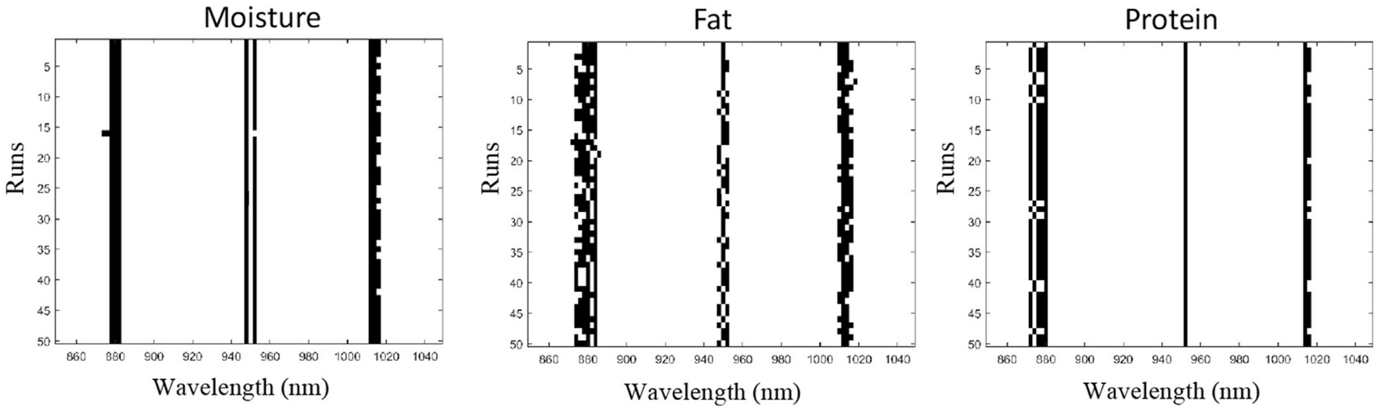


Fig. 5. Variable selection results of WMSCVS in 50 runs on Tecator data.

Table 2
Quantitative index of the stability of the variable selection of WMSCVS.

Data set	Component	Stability
Tecator	Moisture	0.965 ± 0.050
	Fat	0.707 ± 0.150
	Protein	0.860 ± 0.208
Four Comp.	Water	0.244 ± 0.208
	D_2O	0.361 ± 0.172

Finally, we investigated the predictive performance of the corrected spectra. Table 3 shows the results of PLS models with different scattering correction methods. PLS of the raw spectra used more than four latent variables because of both the scattering effects and response nonlinearity. The scattering effects and the nonlinear response had become considerable challenges to the correction methods. The conventional scatter correction methods, MSC and SNV, which performed well for Tecator data, failed to correct this data. After checking the corrected spectra, we found that baseline additive effects remained. EMSC and EISC gave similar predictive performance since they gave similar corrected spectra. Compared with IDRC, WMSCVS was more flexible due to the

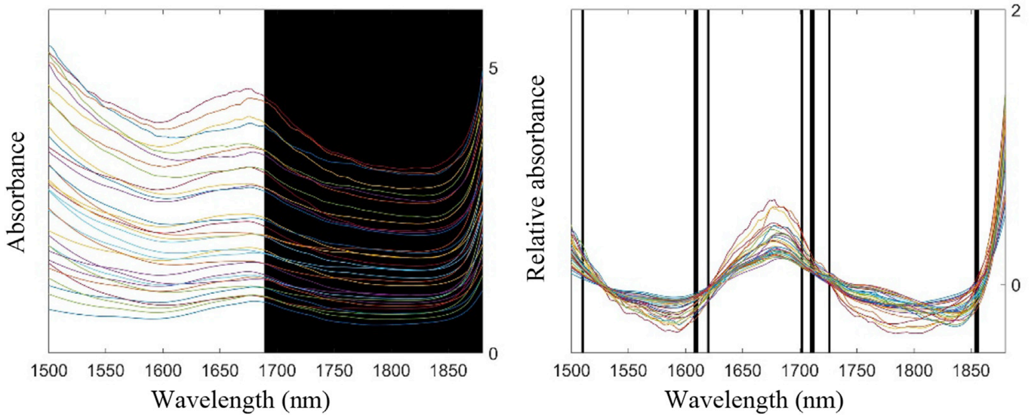


Fig. 6. The results of variable selection on Four Component (water) data. Left part: IDRC with the raw spectra as background, right part: WMSCVS with the orthogonal-projection preprocessing spectra as background.

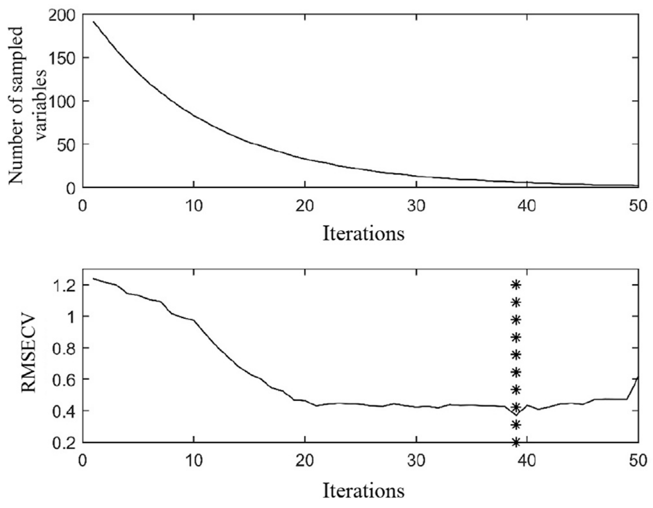


Fig. 7. The changing trend of the number of sampled variables (upper part) and RMSECV values (lower part) with the increasing iterations on Four Component (water) data. The vertical asterisk line denotes the optimal point where the RMSECV values achieve the lowest.

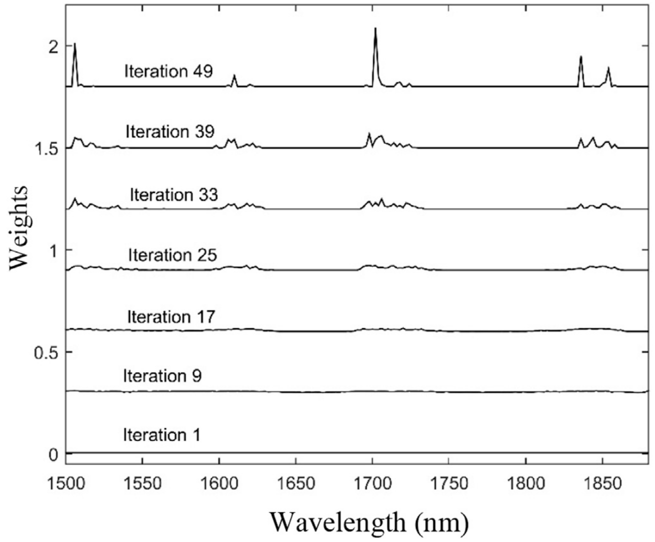


Fig. 8. The evolution of weights of WBS on the Four Component (water) data.

Table 3
Results of calibration models with different scatter correction methods on Four Component data. The number in parentheses is the standard deviation of results in 50 independent runs.

Method	Water		D_2O	
	LVs	RMSEP	LVs	RMSEP
RAW	6	1.69	6	3.60
MSC	1	3.65	4	3.73
SNV	4	2.64	5	2.37
EMSC	3	1.28	2	2.24
EISC	4	1.57	4	2.02
IDRC	5	1.34	6	1.55
WMSCVS	6.9(0.3)	**0.71**(0.127)	6(0)	**1.35**(0.104)

Bold means the lowest value among all methods.

individual variable selection strategy. It could be seen that WMSCVS achieved the lowest RMSEP among all methods. However, the nonlinearity also posed a challenge to the stability of the variable selection of WMSCVS. It can be seen from Table 2 that, compared with Tecator data, the stability results of 50 independent runs were smaller (0.244 for water and 0.361 for D_2O).

5. Conclusion

In this paper, we proposed to use variable selection for setting weights of WLS in estimating multiplicative parameters. From the experiments of two NIR datasets, we could see that an optimal and easy-to-interpret solution of the weights could be obtained by the proposed variable selection algorithm. Moreover, WMSCVS could provided better predictive performance compared with other correction methods. In this work, MSC based correction was used in this study, however, the idea can also extend to other methods. Our future work will include finding more comprehensive measurement for variable selection and applying the proposed method to wider applications.

Acknowledgements

Appreciation is given to the editor and the two anonymous referees for their valuable comments that have led to a substantial improvement in this paper, as well as to Tecator Company, Steponavicius and Thennadil for the availability of data sets. The authors would like to thank Dr. Chen Chen for her valuable advice to this work. This work is financially supported by the National Key R&D Program of China (Grant No.2017YFC0803505).

References

[1] C.A.T. dos Santos, R.N.M.J. Pscoa, J.A. Lopes, A review on the application of vibrational spectroscopy in the wine industry: from soil to bottle, Trac. Trends Anal. Chem. 88 (2017) 100–118.

[2] B.C. De, P.F. Chavez, J. Mantanus, R. Marini, P. Hubert, E. Rozet, E. Ziemons, Critical review of near-infrared spectroscopic methods validations in pharmaceutical applications, J. Pharmaceut. Biomed. Anal. 69 (8) (2012) 125–132.

[3] S. Wold, M. Sjostrom, L. Eriksson, Pls-regression: a basic tool of chemometrics, Chemometr. Intell. Lab. Syst. 58 (2) (2001) 109–130.

[4] P. Geladi, D. Macdougall, H. Martens, Linearization and scatter-correction for near-infrared reflectance spectra of meat, Appl. Spectrosc. 39 (3) (1985) 491–500.

[5] R. Barnes, M. Dhanoa, S. Lister, Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra, Appl. Spectrosc. 43 (5) (1989) 772–777.

[6] H. Martens, E. Stark, Extended multiplicative signal correction and spectral interference subtraction: new preprocessing methods for near infrared spectroscopy, J. Pharmaceut. Biomed. Anal. 9 (8) (1991) 625–635.

[7] H. Martens, J.P. Nielsen, S.B. Engelsen, Light scattering and light absorbance separated by extended multiplicative signal correction. application to near-infrared transmission analysis of powder mixtures, Anal. Chem. 75 (3) (2003) 394–404.

[8] I.S. Helland, T. Naes, T. Isaksson, Related versions of the multiplicative scatter correction method for preprocessing spectroscopic data, Chemometr. Intell. Lab. Syst. 29 (2) (1995) 233–241.

[9] D.K. Pedersen, H. Martens, J.P. Nielsen, S.B. Engelsen, Near-infrared absorption and scattering separated by extended inverted signal correction (eisc): analysis of near-infrared transmittance spectra of single wheat seeds, Appl. Spectrosc. 56 (9) (2002) 1206–1214.

[10] N.K. Afseth, A. Kohler, Extended multiplicative signal correction in vibrational spectroscopy, a tutorial, Chemometr. Intell. Lab. Syst. 117 (2012) 92–99.

[11] Z.P. Chen, J. Morris, E. Martin, Extracting chemical information from spectral data with multiplicative light scattering effects by optical path-length estimation and correction, Anal. Chem. 78 (22) (2006) 7674–7681.

[12] Y. Bi, L. Tang, P. Shan, Q. Xie, Y. Hu, S. Peng, J. Tan, C. Li, Interference correction by extracting the information of interference dominant regions: application to near-infrared spectra, Spectrochim. Acta Mol. Biomol. Spectrosc. 129 (2014) 542–550.

[13] N.B. Gallagher, T.A. Blake, P.L. Gassman, Application of extended inverse scatter correction to mid-infrared reflectance spectra of soil, J. Chemometr. 19 (5–7) (2005) 271–281.

[14] W.-m. Shi, W. Kong, Q.-b. Tao, J.-j. Guo, M.-j. Xia, Q. Shen, B.-x. Ye, An adaptive wavelength interval selection by modified particle swarm optimization algorithm: simultaneous spectral or differential pulse voltammetric determination of multiple components with overlapping peaks, J. Anal. Chem. 68 (7) (2013) 630–638.

[15] C. Cernuda, E. Lughofer, P. Hintenaus, W. Märzinger, Enhanced genetic operators design for waveband selection in multivariate calibration based on nir spectroscopy, J. Chemometr. 28 (2) (2014) 123–136.

[16] F. Allegrini, A.C. Olivieri, A new and efficient variable selection algorithm based on ant colony optimization. applications to near infrared spectroscopy/partial least-squares analysis, Anal. Chim. Acta 699 (1) (2011) 18–25.

[17] B.-C. Deng, Y.-H. Yun, Y.-Z. Liang, Model population analysis in chemometrics, Chemometr. Intell. Lab. Syst. 149 (2015) 166–176.

[18] B.C. Deng, Y.H. Yun, D.S. Cao, Y.L. Yin, W.T. Wang, H.M. Lu, Q.Y. Luo, Y.Z. Liang, A bootstrapping soft shrinkage approach for variable selection in chemical modeling, Anal. Chim. Acta 908 (2016) 63–74.

[19] X. Song, Y. Huang, H. Yan, Y. Xiong, S. Min, A novel algorithm for spectral interval combination optimization, Anal. Chim. Acta 948 (2016) 19–29.

[20] B.-c. Deng, Y.-h. Yun, Y.-z. Liang, L.-z. Yi, A novel variable selection approach that iteratively optimizes variable space using weighted binary matrix sampling, Analyst 139 (19) (2014) 4836–4845.

[21] H. Li, Y. Liang, Q. Xu, D. Cao, Key wavelengths screening using competitive adaptive reweighted sampling method for multivariate calibration, Anal. Chim. Acta 648 (1) (2009) 77–84.

[22] P. Barbe, P. Bertail, in: The Weighted Bootstrap, vol. 98, Springer Science and Business Media, 2012.

[23] C. Borggaard, H.H. Thodberg, Optimal minimal neural interpretation of spectra, Anal. Chem. 64 (5) (1992) 545–551.

[24] R. Sdteponavicius, S.N. Thennadil, Extraction of chemical information of suspensions using radiative transfer theory to remove multiple scattering effects: application to a model multicomponent system, Anal. Chem. 83 (6) (2011) 1931–1937.

[25] R.W. Kennard, L.A. Stone, Computer aided design of experiments, Technometrics 11 (1) (1969) 137.

[26] D.M. Haaland, E.V. Thomas, A. Chem, Partial least-squares methods for spectral analyses. 1. relation to other quantitative calibration methods and the extraction of quantitative information, Anal. Chem. 60 (11) (1988) 1193–1202.

[27] B.C. Deng, Y.H. Yun, P. Ma, C.C. Lin, D.B. Ren, Y.Z. Liang, A new method for wavelength interval selection that intelligently optimizes the locations, widths and combinations of the intervals, Analyst 140 (6) (2015) 1876–1885.

[28] H.W. Lee, A. Bawn, S. Yoon, Reproducibility, complementary measure of predictability for robustness improvement of multivariate calibration models via variable selections, Anal. Chim. Acta 757 (2012) 11–18.