Front. Optoelectron. 2018, 11(3): 296–305
https://doi.org/10.1007/s12200-018-0804-2

RESEARCH ARTICLE

Discrete combination method based on equidistant wavelength screening and its application to near-infrared analysis of hemoglobin

Tao Pan (✉)[1], Bingren Yan[1], Jiemei Chen[2], Lijun Yao (✉)[1]

1 Department of Optoelectronic Engineering, Jinan University, Guangzhou 510632, China
2 Department of Biological Engineering, Jinan University, Guangzhou 510632, China

Abstract A wavelength selection method for discrete wavelength combinations was developed based on equidistant combination-partial least squares (EC-PLS) and applied to a near-infrared (NIR) spectroscopic analysis of hemoglobin (Hb) in human peripheral blood samples. An allowable model set was established through EC-PLS on the basis of the sequence of the predicted error values. Then, the wavelengths that appeared in the allowable models were sorted, combined, and utilized for modeling, and the optimal number of wavelengths in the combinations was determined. The ideal discrete combination models were obtained by traversing the number of allowable models. The obtained optimal EC-PLS and discrete wavelength models contained 71 and 42 wavelengths, respectively. A simple and high-performance discrete model with 35 wavelengths was also established. The validation samples excluded from modeling were used to validate the three models. The root-mean-square errors for the NIR-predicted and clinically measured Hb values were 3.29, 2.86, and 2.90 $g \cdot L^{-1}$, respectively; the correlation coefficients, relative RMSEP, and ratios of performance to deviation were 0.980, 0.983, and 0.981; 2.7%, 2.3%, and 2.4%; and 4.6, 5.3, and 5.2, respectively. The three models achieved high prediction accuracy. Among them, the optimal discrete combination model performed the best and was the most effective in enhancing prediction performance and removing redundant wavelengths. The proposed optimization method for discrete wavelength combinations is applicable to NIR spectroscopic analyses of complex samples and can improve prediction performance. The proposed wavelength models can be utilized to design dedicated spectrometers for Hb and can provide a valuable reference for non-invasive Hb detection.

Keywords near-infrared (NIR) spectroscopy, equidistant combination-partial least squares (EC-PLS), allowable model set discrete combination models, hemoglobin

1 Introduction

Hemoglobin (Hb) is an iron-containing compound allosteric protein with the functions of transporting oxygen and carbon dioxide, maintaining blood acid–base balance, and others [1]. Hb content is one of the most common clinical indicators, and it plays a major role in the diagnosis of anemic diseases (e.g., hemolytic anemia and iron-deficiency anemia) and also has reference values in the diagnosis of other important diseases (e.g., chronic mountain sickness, leukemia, cardiopulmonary diseases, oncological diseases, kidney diseases, etc) [2–7]. The standard clinical method of Hb determination is spectrophotometry based on haemiglobincyanide (HiCN), which requires chemical reagents and specialized laboratory measurements. However, this method is unsuitable for non-invasive detection of Hb in routine health screening and cannot meet the requirements of continuous and real-time monitoring during operation.

The near-infrared (NIR) spectrum primarily reflects the vibration absorption with overtones and combination frequencies for hydrogen-containing functional groups (e.g., C–H, O–H, and N–H). The NIR spectrum is nondestructive and possesses the advantages of short wavelength and high energy, which can penetrate the surface of a sample and return. Therefore, it has the potential to be used in non-invasive detection. The NIR spectrum has been used *in vivo* or *in vitro* hemoglobin

Received January 31, 2018; accepted April 2, 2018

E-mails: tpan@jnu.edu.cn, tylj@jnu.edu.cn

analyses, and it has elicited widespread attention due to its green (non-invasive or reagent-less) detection methods [8–12]. The related mechanism of non-invasive detection is unclear due to the weak target signal and low signal-to-noise ratio (SNR). Currently, the accuracy of non-invasive detection has not reached the standard of clinical application. Therefore, basic research needs to be carried out deeply.

This work further investigated the NIR analysis method for Hb in human peripheral blood samples and focused on optimizing the NIR wavelength model, overcoming noise interference, and improving prediction accuracy. This study can provide a valuable reference for further noninvasive detection.

For complex samples with multiple components (e.g., human blood), spectroscopic analysis of the target component must mitigate the disturbance of the other components. Appropriate wavelength selection is an important but difficult aspect and is essential for improving prediction performance, reducing model complexity, and designing dedicated spectrometers with high SNR. Moving-window partial least squares (MW-PLS) is a well-performed method for continuous waveband selection that employs the initial wavelength and number of wavelengths as parameters [13–20]. As a promotion of the MW-PLS method, the recently proposed equidistant combination PLS (EC-PLS) method [21,22] focuses on the selection of the combination of equidistant wavelengths by using initial wavelength, number of wavelengths, and number of wavelength intervals as parameters. This approach can more effectively overcome spectral co-linearity in adjacent wavelengths and enhance the model prediction effect. The EC-PLS method can select all ergodic wavelength combinations with equidistant wavelengths within a large range because of the low freedom degree of its parameters. However, the molecular absorption bands of measured samples are not always equidistant, and the equidistant wavelength combinations selected by EC-PLS may still contain redundant wavelengths. Therefore, EC-PLS needs to be improved further.

In the present study, an allowable model set was established based on an appropriate level of permissibility. The wavelengths in the equidistant model set were sorted and combined according to their frequencies, and an ideal discrete wavelength model was proposed. The high water content of whole-blood samples can lead to saturated absorption and noise interference. Therefore, high-absorption wavebands were removed, and the remaining wavebands were used for modeling.

Savitzky-Golay (SG) smoothing [13,23–25] is a commonly utilized multi-parameter spectral preprocessing method that can effectively eliminate spectral noise. In this study, all modes of SG smoothing were used for modeling, and the optimal mode was selected according to the prediction effects. On this basis, the SG smoothing spectra were used for further optimization with EC-PLS.

2 Materials and methods

2.1 Samples, instruments and reference methods

A chronological, two groups of human peripheral blood samples were collected from a hospital and placed in 0.2% ethylenediaminetetraacetic acid-containing tubes. The first group (180 samples) in the first day was used for modeling, whereas the second group (120 samples) in the second day was used for validation. The Hb values of all samples (300 samples) were measured with a BC-3000Plus automatic blood cell analyzer (Shenzhen Mairui, China) by using the full blood cell count method. Given that blood samples were collected and used in this study, the informed consent of all individual participants was obtained. Experiments were performed in compliance with relevant laws and institutional guidelines and approved by a local medical institution, which obtained informed consent from all subjects.

The measured values were used as reference values for the calibration, prediction, and validation of spectroscopic analysis. The measured values of the 300 samples ranged from 83 to 161 $g \cdot L^{-1}$, and the mean value and standard deviation were 125.5 and 14.6 $g \cdot L^{-1}$, respectively.

The spectra were collected by the XDS Rapid Content™ liquid grating spectrometer (FOSS, Denmark) equipped with a 2 mm cuvette transmission accessory. The scanning scope of the spectrum spanned 400–2498 nm with a 2 nm wavelength gap, including the entire NIR region and a large part of the visible region. Wavebands of 400–1100 nm and 1100–2498 nm were adopted for silicon and plumbous sulfide detection, respectively. Each sample was measured thrice, and the mean value of the three measurements was used for modeling. The spectra were measured at a temperature of (25±1)°C and relative humidity of (46±1)%.

2.2 Evaluation indicators

The 180 modeling samples were randomly divided into calibration (100 samples) and prediction (80 samples) sets. The corresponding root-mean-square (RMS) error and correlation coefficient for prediction were calculated and denoted as $RMSEP_M$ and $R_{P,M}$, respectively. The model parameters were optimized according to the minimum $RMSEP_M$.

Then, the 120 validation samples were utilized to validate the selected models. The corresponding RMS error and correlation coefficient for prediction in validation set were calculated and denoted as RMSEP and R_P, respectively. The relative RMSEP and the ratio of performance to deviation were further calculated and denoted as RRMSEP and RPD, respectively. Among them $\mathrm{RRMSEP} = \frac{\mathrm{RMSEP}}{C_{\mathrm{AVE}}}$, $\mathrm{RPD} = \frac{C_{\mathrm{SD}}}{\mathrm{RMSEP}}$, C_{AVE}, C_{SD} is the

mean values and the standard deviation of actual measured values for validation set, respectively.

Quantitative analyses of Hb was performed according to these processes.

2.3 PLS with SG smoothing

The parameters of the SG method include the order of derivatives (d), the degree of polynomial (p), and the number of smoothing points (m, odd). In the original work on SG [24], parameters d, p, and m were set to $d = 0, 1, 2, 3, 4, 5$; $p = 2, 3, 4, 5, 6$; and $m = 5, 7,..., 25$. Considering that the absolute values of the fourth and fifth derivatives are very small (which means a large amount of spectral information is missing), the SG modes using these derivatives were not used for screening in this study. The remaining 99 modes were adopted. Furthermore, if the wavelength gap and number of smoothing points are small, then the smoothing window is narrow and the information in the window for smoothing is insufficient, and it is difficult to get satisfactory preprocessing effects. Thus, it was necessary to expand the number of smoothing points (m).

In the present study, m was expanded to 5, 7,..., 51 (odd). The calculation formulas for the added SG smoothing modes were derived, and 264 modes were obtained. That is, the parameters were set to $d = 0, 1, 2, 3$; $p = 2, 3, 4, 5, 6$; and $m = 5, 7,..., 51$. The number of latent variables (F) was set to $F = 1, 2,..., 20$. A PLS model was established for each SG smoothing mode, and the optimal smoothing mode was selected according to the minimum $\mathrm{RMSEP_M}$.

2.4 EC-PLS method

The EC-PLS method selects an appropriate combination of equidistant wavelengths through PLS modeling. The parameters used in this study were 1) initial wavelength (I), 2) number of wavelengths (N), 3) number of wavelength intervals (G), and 4) number of latent variables (F). In a specific screening region, each combination of equidistant wavelengths, which corresponded to a parameter combination (I,N,G), was employed to establish PLS calibration and prediction models. The optimal F was also determined according to the prediction effect. The optimal parameter combination (I,N,G) was further screened according to the prediction effect of the PLS models.

2.4.1 Selection of number of PLS factors

In PLS regression, the number of latent variables (F) is an important parameter that corresponds to the number of integrated spectral variables. The selection of a reasonable F is necessary but difficult. In this study, based on the prediction effect ($\mathrm{RMSEP_M}$) of the calibration and the prediction set, the optimal F is determined as the follows:

$$\mathrm{RMSEP_M}(I,N,G) = \min_{F} \mathrm{RMSEP_M}(I,N,G,F), \quad (1)$$

where (I,N,G) is an any fixed parameter combination.

2.4.2 Global optimal model

The global optimal model was further selected according to the following equation:

$$\mathrm{RMSEP}_* = \min_{I,N,G} \mathrm{RMSEP_M}(I,N,G). \quad (2)$$

2.4.3 Local optimal model for a single parameter

In the manufacturing of a spectrometer, the adopted wavelengths (I, N, G) are usually restricted due to limits in cost and material properties. The selected global optimal model does not always meet actual conditions. Therefore, the local optimal model that corresponds to a single parameter (I, N or G) is significant.

For any fixed I, N, and G, the corresponding local optimal model is selected respectively according to the following equations:

$$\mathrm{RMSEP_M}(I) = \min_{N,G} \mathrm{RMSEP_M}(I,N,G), \quad (3)$$

$$\mathrm{RMSEP_M}(N) = \min_{I,G} \mathrm{RMSEP_M}(I,N,G), \quad (4)$$

and

$$\mathrm{RMSEP_M}(G) = \min_{I,N} \mathrm{RMSEP_M}(I,N,G). \quad (5)$$

Among the local optimal models, the models that are close to the global optimal model in terms of prediction performance and satisfied the actual constraints are expected to be appropriate selections.

2.5 Allowable wavelengths model based on the EC-PLS

In order to eliminate the redundant wavelengths in the equidistant wavelength models with EC-PLS method, an allowable model set was proposed.

The equidistant wavelength models with EC-PLS were sorted according to the values of $\mathrm{RMSEP_M}$ from small to large. The first S_0 models were used and expressed as follows:

$$\Lambda^{(s)} = \left\{\lambda_1^{(s)},\lambda_2^{(s)},...,\lambda_{N^{(s)}}^{(s)}\right\}, \quad s = 1,2,...,S_0, \quad (6)$$

where $\Lambda^{(s)}$ is the sth combination of equidistant wavelengths, $N^{(s)}$ is the number of wavelengths of the sth combination and $\lambda_p^{(s)}$ is the pth wavelength of the sth combination, i.e., $p = 1,2,...,N^{(s)}$.

Noteworthy that the first model happens to be the global optimal EC-PLS model. When S_0 is small (e.g., $S_0 \leqslant 100$), the $RMSEP_M$ of the S_0 models are close to one of the optimal model. This result indicates that these S_0 models are similar to the optimal model in terms of prediction effect.

Given that these S_0 models may exhibit a containing relationship, they need to be further simplified. When a containing relationship exists between two wavelength combinations as follows:

$$\{\lambda_1^{(i)},\lambda_2^{(i)},...,\lambda_{N^{(i)}}^{(i)}\} \subset \{\lambda_1^{(j)},\lambda_2^{(j)},...,\lambda_{N^{(j)}}^{(j)}\},$$

$$1 \leqslant i \neq j \leqslant S_0, \qquad (7)$$

the latter contains redundant wavelengths and must be removed. The remaining models were established as the allowable models, and the total number of these allowable models was denoted as S ($S \leqslant S_0$).

2.6 Discrete wavelengths model based on the allowable wavelengths model

Assume that a total of K wavelengths appear in the S allowable models, the K wavelengths arranged in a descending order of their occurrence frequencies are expressed as follows:

$$\mu_1,\mu_2,...,\mu_K, \qquad (8)$$

where the wavelengths with same frequencies are sorted according to the natural order of the wavelengths from small to large. The wavelengths with high frequencies suggest high information, and the wavelengths with low frequencies may be redundant wavelengths.

The K combinations of discrete wavelengths are then proposed as follows:

$$\Omega_k = \{\mu_1,\mu_2,...,\mu_k\}, \quad k = 1,2,...,K. \qquad (9)$$

These combinations were used to establish PLS model. The corresponding $RMSEP_M$ and $R_{P,M}$ values were calculated, and the optimal combination of discrete wavelengths was selected according to the minimum $RMSEP_M$. The corresponding number of wavelengths is still denoted by N.

The above result is related to the number of allowable models (S). An appropriate S needs to be screened to obtain a better discrete combination model. In the current study, S was set to $S \in \{1\} \cup \{10,20,...,100\}$, and the above experiment was performed. The appropriate S value was selected according to the corresponding prediction effect ($RMSEP_M$) and wavelength number (N). An ideal discrete combination model was further obtained through a comparison.

The computer platform was developed with MATLAB R2009b software.

3 Results and discussion

The NIR spectra of 300 human peripheral blood samples in the entire scanning region (400–2498 nm) are shown in Fig. 1. It seems clear that the saturated absorption occurred in the wavebands around 1950 and 2400 nm, which caused high noise interference. In order to avoid strong absorbance, wavebands with absorbance higher than 4 (corresponding to 99.9% absorption rate) were eliminated, and the remaining were the combinations of 400–1880 and 2100–2300 nm.

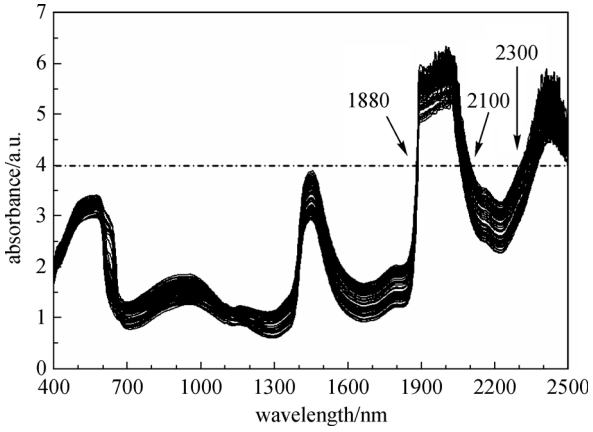

Fig. 1 NIR spectra of 300 human peripheral blood samples

3.1 PLS models

PLS models for Hb were established based on the entire scanning region (400–2498 nm) and the unsaturated region (400–1880 and 2100–2300 nm). The modeling effects ($RMSEP_M$, $R_{P,M}$) are summarized in Table 1. The results indicated that the prediction effect for the unsaturated region was significantly better than that for the entire scanning region. Therefore, it is necessary to remove the saturation waveband with high absorption, and the remaining waveband combination of 400–1880 and 2100–2300 nm will be carried out for further modeling.

Table 1 Parameters and prediction effects of PLS models with the entire scanning region and the unsaturated region

wavelength/nm	N	F	$RMSEP_M/(g \cdot L^{-1})$	$R_{P,M}$
400–2498	1050	6	4.39	0.952
400–1800 and 2100–2300	842	8	3.80	0.965

3.2 PLS with SG smoothing

The PLS models corresponding to 264 SG smoothing modes were established in the unsaturated region (400–1880 and 2100–2300 nm) and were called SG-PLS models. The modeling effects of the local optimal models

corresponding to each d (d = 0, 1, 2, 3) are summarized in Table 2. The global optimal SG mode was 2nd order derivative, 2nd degree polynomial and 31 smoothing points (d = 2, p = 2, 3, and m = 31). The corresponding $RMSEP_M$ was 3.14 $g \cdot L^{-1}$, which was obviously better than that without SG smoothing. The prediction effects of the global optimal SG-PLS model are also summarized in Table 3. The corresponding SG derivative spectra are shown in Fig. 2. The baseline deviations (drifts) of the spectra of the different samples significantly decreased.

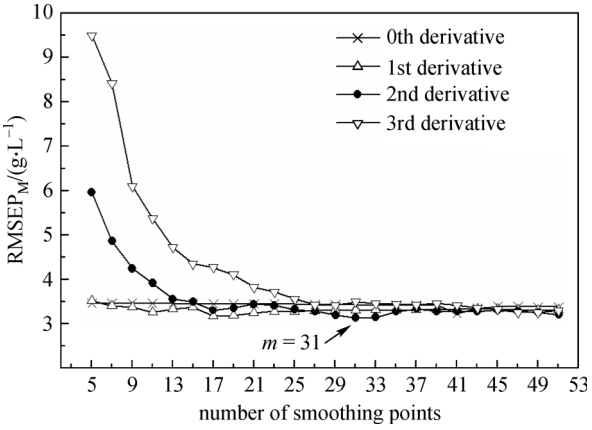

Fig. 2 SG derivative spectra (d = 2, p = 2, 3, m = 31) of all samples at 400–1880 and 2100–2300 nm

In addition, the $RMSEP_M$ of the local optimal models corresponding to each m (m = 5, 7, ..., 51) distinguished by different orders of the derivative is shown in Fig. 3. The optimal m was above 25 (m = 31), so it is necessary to extend the number of smoothing points. These results indicated that the SG method can further reduce spectrum noise and improve prediction performance. Thus, the spectra processed by the global optimal SG mode were used for further modeling.

3.3 Selection of equidistant wavelengths combination with EC-PLS

On the basis of the above mentioned SG derivative spectra, equidistant wavelength combinations were further determined using the EC-PLS method. The corresponding waveband region was 400–1880 and 2100–2300 nm. Parameters I, N, G, and F were set to $I \in \{780,782,...,1880\} \cup \{2100,2102,...,2300\}$, $N \in \{1,2,...,100\}$, $G \in \{1,2,...,10\}$ and $F \in \{1,2,...,20\}$, respectively.

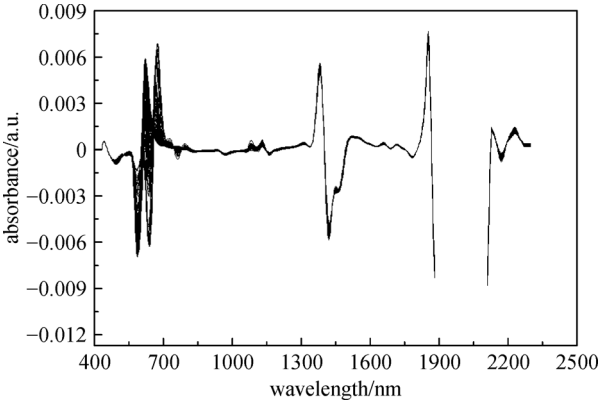

Fig. 3 $RMSEP_M$ of the local optimal models for each m distinguished by different orders of derivative

The obtained parameters of the global optimal model were I = 1230 nm, N = 71, G = 6, and F = 7. The parameters and prediction effects are summarized in Table 3. As shown in Table 3, the global optimal EC-PLS model was better than the optimal SG-PLS model; thus, the number of adopted wavelengths significantly decreased (N = 71).

The combination of parameters (I, N, G) corresponded to a continuous waveband when G = 1. In this case, EC-PLS is equivalent to MW-PLS. Therefore, EC-PLS is the extension of MW-PLS in terms of the algorithm.

In addition, the $RMSEP_M$ values of the local optimal models corresponding to each single parameter (I, N, and G) are shown in Fig. 4. The results of the global optimal model can also be observed in the local optimal model sequences. Figure 4 provides many available models, among which the models with prediction effects close to that of the global optimal model are still good options for practical application.

3.4 Discrete combination models

First, the EC-PLS models were sorted according to $RMSEP_M$ in an ascending order. Second, the evolution of their $RMSEP_M$ values was observed. The $RMSEP_M$ of the global optimal EC-PLS model was 2.67 $g \cdot L^{-1}$ (Table 3), which corresponded to the case of s = 1. The first 200 values of $RMSEP_M$ are given in Fig. 5 ($1 \leqslant s \leqslant 200$). When

Table 2 Effects of the local optimal SG-PLS model in 400–1880 and 2100–2300 nm corresponding to each order of derivative

d	p	m	F	$RMSEP_M/(g \cdot L^{-1})$	$R_{P,M}$
0	6	41	10	3.24	0.973
1	2	17	11	3.18	0.976
2	2	31	11	3.14	0.977
3	3	49	12	3.26	0.971

Notes: d, order of derivatives; p, degree of polynomial; m, number of smoothing points

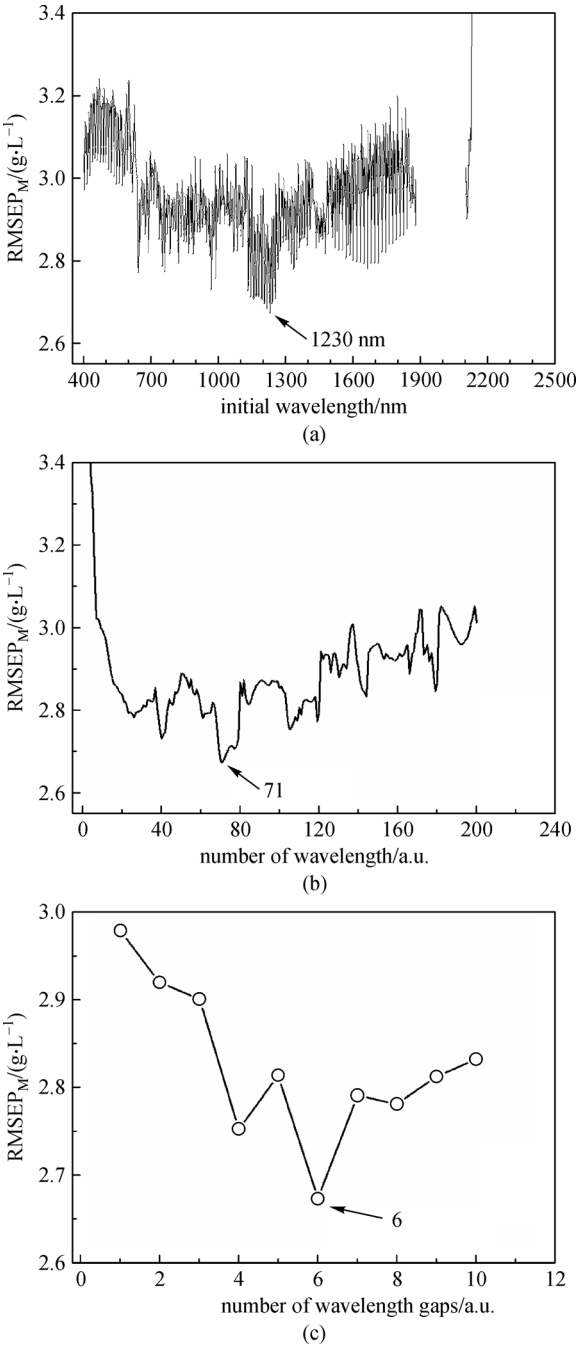

Fig. 4 $RMSEP_M$ of the local optimal models with EC-PLS for (a) initial wavelength, (b) number of wavelength, and (c) number of wavelength gaps

$s \leqslant 100$, the corresponding $RMSEP_M$ was less than or equal to 2.81 $g \cdot L^{-1}$. These 100 models almost had no difference with the optimal EC-PLS model in terms of prediction effect (Fig. 5) and were thus regarded as allowable. The number of allowable models (S) was set to within 100 in the experiments.

For the cases of $S = 10, 20, ..., 100$, the corresponding optimal combinations of discrete wavelengths were selected according to the method mentioned in Section 2.6. The number of adopted wavelengths (N) of the optimal model of discrete wavelengths corresponding to each number of allowable models (S) were further determined. The $RMSEP_M$ and N of the optimal model of discrete wavelengths for each S are shown in Fig. 6. The minimum $RMSEP_M$ ($RMSEP_M = 2.55$ $g \cdot L^{-1}$) was achieved when $S = 40$, and the corresponding N was also close to the minimum ($N = 42$). The corresponding parameters and modeling effects are also summarized in Table 3. The results indicated that the optimal discrete model was superior to the optimal EC-PLS model ($RMSEP_M = 2.67$ g $\cdot L^{-1}$, $N = 71$) in terms of prediction performance and complexity of the wavelength model. The wavelength combination (Ω_{42}) of the optimal discrete model was the follows: 1242, 1254, 1266, 1278, 1290, 1302, 1314, 1326, 1338, 1350, 1362, 1374, 1386, 1398, 1410, 1434, 1458, 1482, 1506, 1530, 1554, 1578, 1602, 1626, 1650, 1674, 1698, 1722, 1746, 1770, 1794, 1818, 1842, 1866, 2108, 2132, 2156, 2180, 2204, 2228, 2252, 2276 nm. The model is a non-equidistant discrete combination model. For easy observation, these wavelengths are labeled in the average spectrum of the samples in Fig. 7. For comparison, the equidistant wavelength combinations of the optimal EC-PLS model are also labeled in Fig. 7. Notably, the wavelength combination ($N = 42$) of the optimal discrete model was included in the equidistant wavelength combination ($N = 71$) of the optimal EC-PLS model. This result indicated that these two models possessed good consistency. The results also shown that the optimal equidistant model indeed contained many redundant wavelengths.

We further investigated the case of $S = 40$. A total of 260 wavelengths appeared in the 40 allowable models ($K = 260$). For each number of adopted wavelengths ($1 \leqslant k \leqslant 260$), the corresponding predicted effect ($RMSEP_M$) is shown in Fig. 8. The corresponding wavelength combination was as follows: 1242, 1266, 1290, 1314, 1338, 1362, 1386, 1410, 1434, 1458, 1482,

Table 3 Parameters and prediction effects of the SG-PLS model, optimal EC-PLS model, and two discrete combination models

method	wavelengths models	F	$RMSEP_M/(g \cdot L^{-1})$	$R_{P,M}$
SG-PLS	400–1880 and 2100–2300 nm	11	3.14	0.977
EC-PLS	$I = 1230$ nm, $N = 71$, $G = 6$	7	2.67	0.983
DC-PLS	$N = 42$	8	2.55	0.985
	$N = 35$	8	2.62	0.984

Note: DC-PLS, discrete combination-PLS

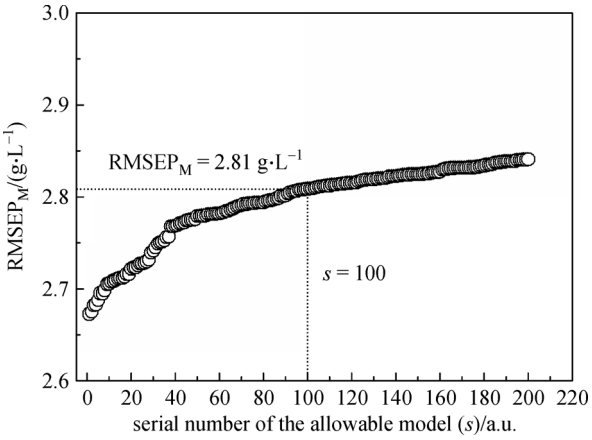

Fig. 5 First 200 values of RMSEP$_M$ with EC-PLS

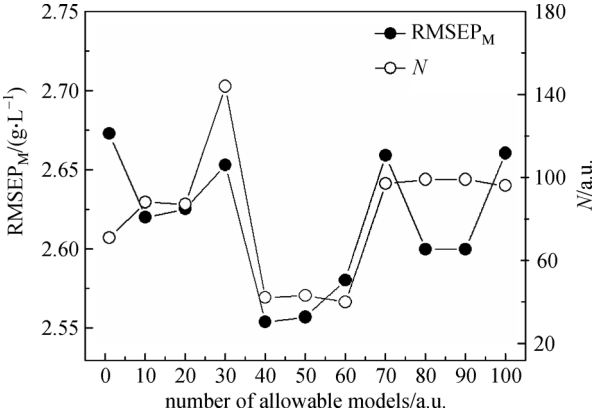

Fig. 6 RMSEP$_M$ and number of adopted wavelengths of the optimal discrete combination model corresponding to each allowable model set

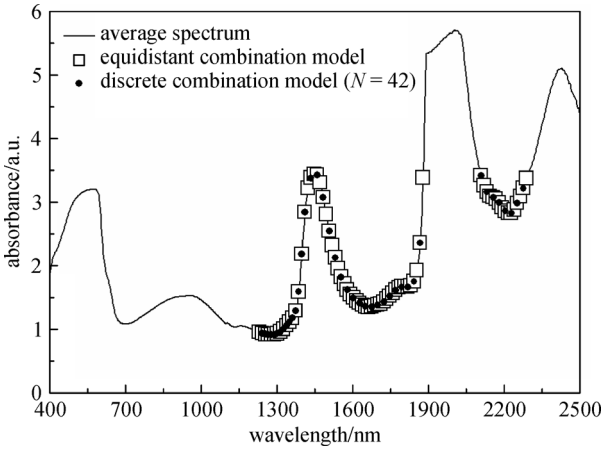

Fig. 7 Wavelength combinations of the optimal EC-PLS model and optimal discrete combination model labeled in the average spectrum of the samples

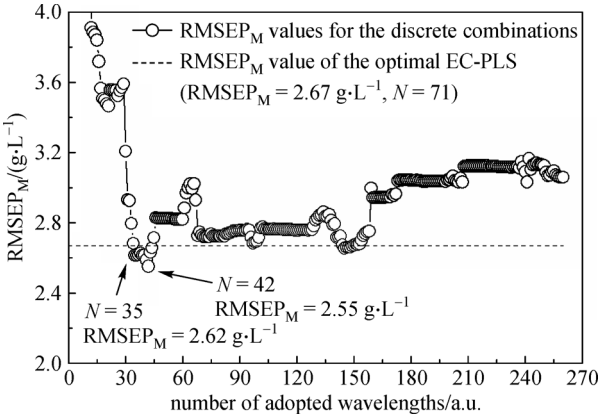

Fig. 8 RMSEP$_M$ of each discrete combination model in the case of S = 40

1506, 1530, 1554, 1578, 1602, 1626, 1650, 1674, 1698, 1722, 1746, 1770, 1794, 1818, 1842, 1866, 2108, 2132, 2156, 2180, 2204, 2228, 2252, 2276 nm. This model is also a non-equidistant discrete combination model. The wavelength combination (Ω_{35}) was included in the wavelength combination (Ω_{42}). Among all the models that are superior to the optimal equidistant model, this model (Ω_{35}) is the simplest. Therefore, it also has a reference value.

3.5 Independent validation

The validation samples excluded from the modeling optimization process were used to validate the optimal EC-PLS model and the two discrete models. The PLS regression coefficients were then calculated using the smoothing spectra and measured values of all modeling samples depending on the corresponding parameters. The predicted Hb values of the validation samples were then calculated using the obtained regression coefficients and the smoothing spectra of the validation samples.

The prediction effects for validation of the three models (RMSEP, R_P, RRMSEP, and RPD) are summarized in Table 4. The three models achieved good validation effects, while the optimal discrete waveband model performed better. The relationship between the predicted and clinically measured Hb values of the 120 validation samples for Hb are shown in Figs. 9 and 10. High correlations were observed between the prediction and clinically measured values.

Given that NIR spectra are flat and overlapping, the wavelength selection for modeling is an important and albeit difficult aspect. Establishing the global optimal discrete model by using an exhaustive method is impossible due to the large number of NIR wavelengths and large amount of computation. A proper search strategy has always been of great concern. By using the initial wavelength, number of wavelengths, and number of

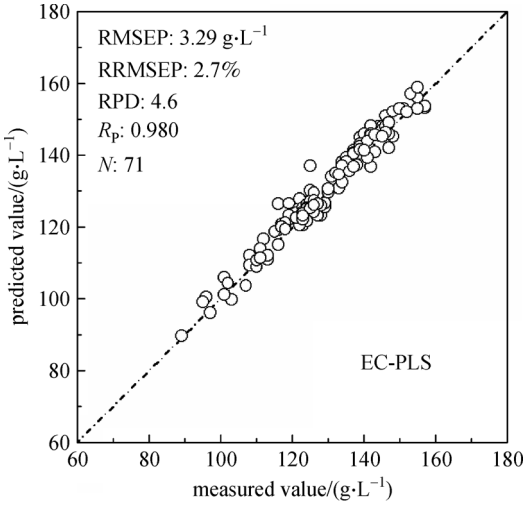

Fig. 9 Relationship between the predicted and measured values of the validation samples for the optimal EC-PLS model

wavelength intervals as the parameters, EC-PLS can identify the equidistant ergodic wavelength combination within a large range. In the present study, in order to eliminate the redundant wavelengths in the equidistant model, the models were sorted. The front models, which were equivalent to the optimal equidistant model, were regarded as the allowable models. Then, the wavelengths that appeared in the allowable models were sorted, combined, and used for modeling, and the local optimal discrete model was selected. Finally, the global optimal discrete model was obtained by traversing the number of allowable models (S). The results indicated that the above method is effective in improving the prediction effect and avoiding redundant wavelengths. Therefore, the proposed strategy for the discrete combination model is appropriate.

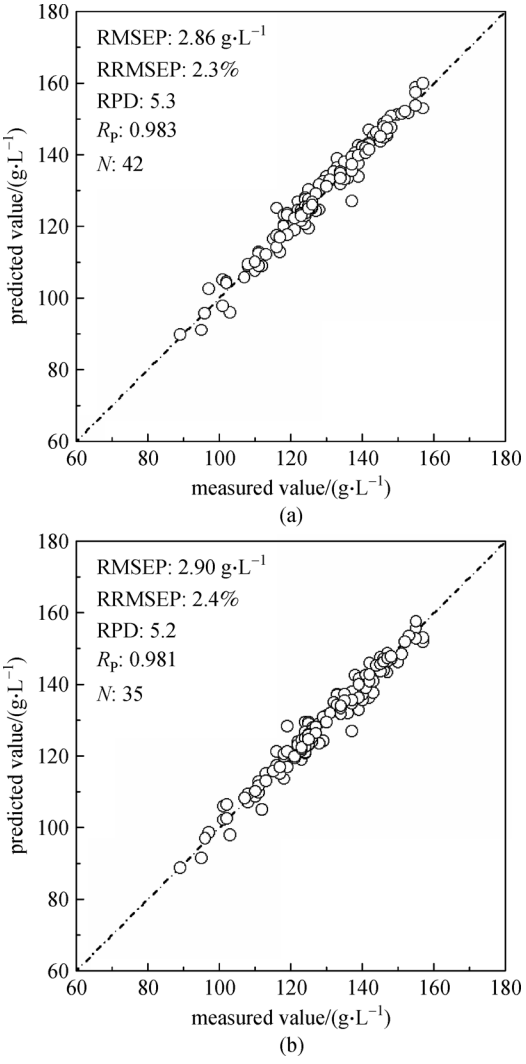

Fig. 10 Relationship between the predicted and measured values for the selected discrete combination models with (a) $N = 42$ and (b) $N = 35$

4 Conclusion

A wavelength selection method for discrete wavelength combinations was proposed based on equidistant wavelength screening. With EC-PLS, the first round of wavelength screening was achieved by traversing equidistant wavelength combinations in a large range using three parameters (I, N, and G). The information wavelengths were aggregated based on the frequencies of the occurring wavelengths in the allowable model set. Then, the second round of wavelength screening was achieved by traversing two dimensions: number of allowable models (S) and number of wavelengths in the combination (k). The redundant wavelengths were eliminated, and the prediction performance was further improved. Finally, the ideal discrete combination model was obtained.

The optimal discrete combination model was validated by NIR analysis of Hb in human peripheral blood samples. The results indicated that the optimal discrete model was superior to the optimal EC-PLS model in terms of prediction performance and complexity of the wavelength

Table 4 Validation effects of the optimal EC-PLS model and two discrete combination models

method	N	RMSEP/(g·L^{-1})	R_P	RRMSEP	RPD
EC-PLS	71	3.29	0.980	2.7%	4.6
DC-PLS	42	2.86	0.983	2.3%	5.3
	35	2.90	0.981	2.4%	5.2

Note: DC-PLS, discrete combination-PLS

model. Notably, the obtained RRMSEP for Hb detection was less than 3%, which is expected for clinical application. The proposed wavelength model can be utilized to design dedicated spectrometers for Hb and provides a valuable reference for further non-invasive Hb detection.

Acknowledgements This work was supported by the National Natural Science Foundation of China (Grant No. 61078040), the Science and Technology Project of Guangdong Province of China (Nos. 2014A020213016, and 2014A020212445).

Compliance with ethics guidelines All procedures followed were in accordance with the ethical standards of the responsible committee on human experimentation (institutional and national) and with the Helsinki Declaration of 1975, as revised in 2000 (5). Informed consent was obtained from all patients for being included in the study.

References

1. Cantor C R, Schimmel P R. Biophysical Chemistry.New York: W. H. Freeman and Company, 1980
2. Anand I, McMurray J J V, Whitmore J, Warren M, Pham A, McCamish M A, Burton P B. Anemia and its relationship to clinical outcome in heart failure. Circulation, 2004, 110(2): 149–154
3. Reeves J T, Leon-Velarde F. Chronic mountain sickness: recent studies of the relationship between hemoglobin concentration and oxygen transport. High Altitude Medicine & Biology, 2004, 5(2): 147–155
4. Weatherall D J, Edwards J A, Donohoe W T. Haemoglobin and red cell enzyme changes in juvenile myeloid leukaemia. British Medical Journal, 1968, 1(5593): 679–681
5. Machovec K A, Jaquiss R D B, Kaemmer D D, Ames W A, Homi H M, Walczak R J Jr, Lodge A J, Jooste E H. Cardiopulmonary bypass strategy for a cyanotic child with hemoglobin SC disease. The Annals of thoracic surgery, 2016, 101(6): 2373–2375
6. Messina A, Fogliani A M. Alexithymia in oncological patients: the role of hemoglobin, malignancy type and tumor staging. European Neuropsychopharmacology, 2011, 21(8): S174–S175
7. Phrommintikul A, Haas S J, Elsik M, Krum H. Mortality and target haemoglobin concentrations in anaemic patients with chronic kidney disease treated with erythropoietin: a meta-analysis. Lancet, 2007, 369(9559): 381–388
8. Vályi-Nagy I, Kaffka K J, Jákó J M, Gönczöl E, Domján G. Application of near infrared spectroscopy to the determination of haemoglobin. Clinica Chimica Acta, 1997, 264(1): 117–125
9. Lee Y, Lee S, In J, Chung S H, Yon J H. Prediction of plasma hemoglobin concentration by near-infrared spectroscopy. Journal of Korean Medical Science, 2008, 23(4): 674–677
10. Shan X, Chen L, Yuan Y, Liu C, Zhang X, Sheng Y, Xu F. Quantitative analysis of hemoglobin content in polymeric nanoparticles as blood substitutes using Fourier transform infrared spectroscopy. Journal of Materials Science: Materials in Medicine, 2010, 21(1): 241–249
11. Macknet M R, Allard M, Applegate R L 2nd, Rook J. The accuracy of noninvasive and continuous total hemoglobin measurement by pulse CO-Oximetry in human subjects undergoing hemodilution. Anesthesia and Analgesia, 2010, 111(6): 1424–1426
12. Butwick A, Hilton G, Carvalho B. Non-invasive haemoglobin measurement in patients undergoing elective Caesarean section. British Journal of Anaesthesia, 2012, 108(2): 271–277
13. Jiang J H, Berry R J, Siesler H W, Ozaki Y. Wavelength interval selection in multicomponent spectral analysis by moving window partial least-squares regression with applications to mid-infrared and near-infrared spectroscopic data. Analytical Chemistry, 2002, 74 (14): 3555–3565
14. Du Y P, Liang Y Z, Jiang J H, Berry R J, Ozaki Y. Spectral regions selection to improve prediction ability of PLS models by changeable size moving window partial least squares and searching combination moving window partial least squares. Analytica Chimica Acta, 2004, 501(2): 183–191
15. Chen H Z, Pan T, Chen J M, Lu Q P. Waveband selection for NIR spectroscopy analysis of soil organic matter based on SG smoothing and MWPLS methods. Chemometrics and Intelligent Laboratory Systems, 2011, 107(1): 139–146
16. Pan T, Chen Z H, Chen J M, Liu Z Y. Near-infrared spectroscopy with waveband selection stability for the determination of COD in sugar refinery wastewater. Analytical Methods, 2012, 4(4): 1046–1052
17. Pan T, Liu J M, Chen J M, Zhang G P, Zhao Y. Rapid determination of preliminary thalassaemia screening indicators based on near-infrared spectroscopy with wavelength selection stability. Analytical Methods, 2013, 5(17): 4355–4362
18. Pan T, Li M M, Chen J M, Xue H Y. Quantification of glycated hemoglobin indicator HbA1c through near-infrared spectroscopy. Journal of Innovative Optical Health Sciences, 2014, 7(4): 1350060
19. Chen J M, Ai T, Pan T, Yao L J, Xia F G. AO–MW–PLS method applied to rapid quantification of teicoplanin with near-infrared spectroscopy. Journal of Innovative Optical Health Sciences, 2017, 10(1): 1650029
20. Yao L J, Xu W Q, Pan T, Chen J M. Moving-window bis-correlation coefficients method for visible and near-infrared spectral discriminant analysis with applications. Journal of Innovative Optical Health Sciences, 2018, 11(2): 1850005
21. Pan T, Li M, Chen J. Selection method of quasi-continuous wavelength combination with applications to the near-infrared spectroscopic analysis of soil organic matter. Applied Spectroscopy, 2014, 68(3): 263–271
22. Yao L, Lyu N, Chen J, Pan T, Yu J. Joint analyses model for total cholesterol and triglyceride in human serum with near-infrared spectroscopy. Spectrochimica Acta, Part A: Molecular and Biomolecular Spectroscopy, 2016, 159: 53–59
23. Xie J, Pan T, Chen J M, Chen H Z, Ren X H. Joint optimization of Savitzky-Golay smoothing models and partial least squares factors for near-infrared spectroscopic analysis of serum glucose. Chinese Journal of Analytical Chemistry, 2010, 38(3): 342–346
24. Savitzky A, Golay M J E. Smoothing and differentiation of data by simplified least squares procedures. Analytical Chemistry, 1964, 36 (8): 1627–1639
25. Guo H S, Chen J M, Pan T, Wang J H, Cao G. Vis–NIR wavelength selection for non-destructive discriminant analysis of breed screening of transgenic sugarcane. Analytical Methods, 2014, 6(21): 8810–8816

Tao Pan is a professor and Ph.D. supervisor in Department of Optoelectronic Engineering at Jinan University. He received his B.S. degree of mathematics from Sichuan University, China, and his Ph.D. degree of biological information engineering from Mie University, Japan. He is director of Applied Spectroscopy Laboratory at Jinan University, College of Science and Engineering. He is engaged in studies of spectroscopy, biomedical information, chemometrics, pattern recognition and partial differential equations, and so on. He has published more than 90 peer reviewed papers. He has received four academic awards issued by the Ministry of Personnel of the People's Republic of China and Guangxi Province, and won the honors of "First batch of 100 outstanding overseas students" issued by the Ministry of Education of the People's Republic of China, etc.

Bingren Yan is a master student majored in optoelectronic engineering from the Department of Optoelectronic Engineering at Jinan University, Guangzhou, China.

Jiemei Chen is an associate professor in Department of Biological Engineering at Jinan University. She received her B.S. and M.S. degrees of microbiology from Sichuan University and Guangxi University, China, and her Ph.D. degree of biology from Mie University, Japan. She is engaged in studies of microbiology, spectroscopy, biomedical information, etc. She has published more than 60 peer reviewed papers.

Lijun Yao is a lecturer in Department of Optoelectronic Engineering at Jinan University. He graduated from the First Aviation Academy of Chinese Air Force, China. He received his M.S. degree of condensed matter physics and Ph.D. degree of biological information technology from Jinan University, China. He is engaged in studies of spectroscopy, biomedical information, chemometrics. He has published more than 40 peer reviewed papers.