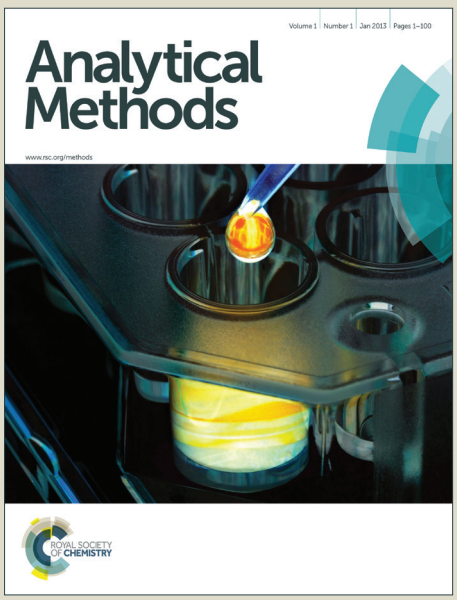
Analytical Methods

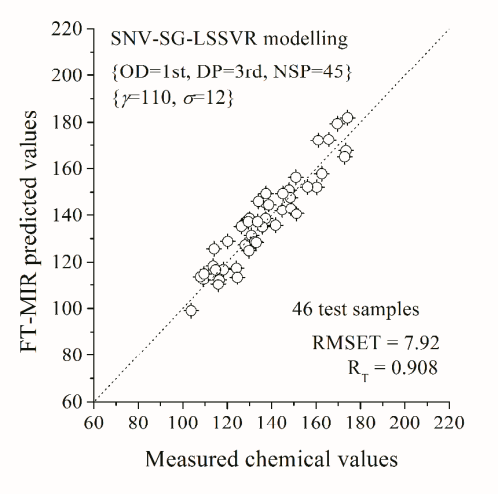
Accepted Manuscript



ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/methods

Graphical Abstract



We enhance the Fourier transform mid-infrared (FT-MIR) modelling performance by establishing nonlinear calibration models for the quantitative analysis of Haemoglobin (HGB) in Human Blood (complex analyte). We use the grid-search technique to have the parameters of LSSVR models tuning in a moderate range and combined optimizing in conjunction with different pre-processing modes. To obtain the stable modelling parameters, we find the grid-search optimal LSSVR models with the best pre-processing mode based on the average predictive results of 30 different divisions of calibration and validation sets. And the optimized SNV-SGS-LSSVR model is found in the HGB fingerprint region. The optimal model is evaluated much satisfactory for an independent test sample set. Our study shows that the combined optimization of grid-search LSSVR modelling with the SNV-SGS pre-processing has the potential of improving predictive abilities of FT-MIR spectroscopic analysis of HGB in human blood.

Cite this: DOI: 10.1039/c0xx00000x

www.rsc.org/xxxxxx

ARTICLE TYPE

FT-MIR Modelling Enhancement for the Quantitative Determination of Haemoglobin in Human Blood by Combined Optimization of Grid-Search LSSVR Algorithm with Different Pre-Processing Modes

Hua-Zhou Chen*, Wu Ai, Quan-Xi Feng, Guo-Qiang Tang*

Received (in XXX, XXX) Xth XXXXXXXXX 20XX, Accepted Xth XXXXXXXXX 20XX
DOI: 10.1039/b000000x

Haemoglobin (HGB) is an important factor to determine anaemia and iron nutrition for human health. Quantitative determination for HGB in human blood is established by the rapid analytical tool of Fourier transform mid-infrared (FT-MIR) spectrometry with its chemometric algorithms. Least squares support vector regression (LSSVR) is utilized for the nonlinear modelling. For modelling enhancement, we propose the grid-search technique applied to the parameter tuning of LSSVR modelling. Also, we constructed the framework for discussing the separate and combined use of the spectral pre-processing methods of multiplicative scatter correction (MSC), standard normal variate (SNV) and Savitzky-Golay smoother (SGS), in which the SGS parameters were set tunable in a certain designated range. The performances of different pre-processing modes were evaluated in combination of grid-search LSSVR modelling. To get stable results, the grid-search LSSVR models and pre-processing modes were established based on the average predictive results of 30 different calibration-validation divisions. These analytical methods are executed on the FT-MIR fingerprint region of human blood HGB, with comparison to the full-scanning region. Results show that the optimized model appears in the fingerprint region. In the evaluation part for test samples, the designated optimal model outputs a root mean square error of testing (RMSET) no more than 6% of the mean chemical value and a correlation coefficient higher than 0.9. This study shows that the combined optimization of grid-search LSSVR algorithm with different pre-processing modes has the potential of improving predictive abilities of FT-MIR spectroscopic analysis of HGB in human blood.

1. Introduction

Haemoglobin (HGB) concentration in human blood is an important determination indicator for anaemia and iron nutritional status. The direct rapid determination method of human blood HGB concentration was a significant research branch in human health monitoring systems.[1-3] Infrared (IR) spectroscopy is effective determination method for analyzing the substantial molecular structure as well as measuring the component content. As the water responses strongly in the IR spectra, many samples contain a lot of water in biology, medicine, agriculture, food and other fields, thus making the direct quantitative analysis in use of IR spectroscopy limited.[4-6] With the development of Fourier transform mid-infrared (FT-MIR) spectroscopy and modern analytical methodology, the difficulties mentioned above have been overcome. FT-MIR spectroscopy with its chemometric algorithms has been widely used in biology, medicine, agriculture, food and other fields,[7-10] as a new quantitative analytical technology with the advantages of reagent-free, on-line, real-time and in-situ.

The rapid quantitative analysis without reagents for blood primary components (such as haemoglobin, albumin, globulin, glucose, cholesterol, triglycerides, etc.) is performed by near-infrared and mid-infrared spectroscopy.[11-14] The quantitative analysis of blood HGB by near-infrared and mid-infrared spectroscopy is feasible, but its predictive accuracy is not high enough and needs further improvement.[15-16] As far as we know, the research on quantitative analysis of HGB concentration in human blood samples by FT-MIR spectroscopy has been little reported. Lee[17] output the preliminary experiments confirming the feasibility of quantitation for HGB in human blood by using FT-MIR spectroscopy. Perez-Guaita[18] utilized the attenuated total reflectance technology to study the quantitative determination of HGB concentration in whole blood. But none of them worked on the modelling stability for the rapid, reagent-free spectroscopic technology. Therefore, collecting a large number of samples for FT-MIR experiments, and establishing stable and reliable analysis models are further work to be completed.

Extracting information and eliminating noise in spectroscopy analysis are very important for improving model predictive effect, especially for analytes with multi-components.[19-21] Partial least squares (PLS) is a common linear method widely used for comprehensively screening spectroscopic data, extracting information variables and overcoming spectral colinearity.[22-24]

But linear approaches cannot meet the quantitative modelling accuracy because the spectroscopic analysis of a single component in complex systems (e.g. human blood) is influenced by the responses of other components and noises.[25] Thus the discussion of nonlinear chemometric method is indispensible.

Least squares support vector regression (LSSVR) is a popular nonlinear analytical method. The fundamental concept in LSSVR is to map the original data onto a high dimensional space by utilizing kernels, followed by a linear regression between the dependent variable and the high dimensional data.[26] The distribution of the feature samples in high dimensional space depends on the selection of the kernel and the corresponding parameters. The Gaussian radial basis function (RBF) kernel is the most commonly-used kernel for regression applications due to its effectiveness and faster training process.[27] The regularization extension and the kernel width (degree of generalization) of the kernel are two vital factors affecting the nonlinear models. LSSVR has a global optimum and exhibits model accuracy in nonlinear and nonstationary data, and the kernel parameters can well determine its resistance to the noise components. There are increasing evidences showing that the RBF kernel has moderate robustness and stability to enable nonlinear modelling for FT-MIR data.

FT-MIR spectra are easily affected by physical properties of the analysed products and other interference. It is necessary to perform mathematical pre-process to reduce the systematic noise, enhancing the contribution of the chemical signals. The aim of this study was to develop a grid-search mode for the quantitative determination of Haemoglobin, based on the LSSVR algorithm, for the comparison between raw and pre-processed spectra. Some chemometric tools were used for spectral smoothing, including multiplicative scatter correction (MSC), standard normal variate (SNV), and Savitzky-Golay smoother (SGS) and etc.[28] The performances of different spectral pre-processing methods lead to variable predictive results. MSC is an averaging regression method for noise reduction.[29] The diffuse reflection scattering effect could be corrected based on the average spectrum. SNV is a centre transformation for the spectra used to remove slope variation and to correct for scattering effects.[30] SGS is on the basis of least square regression.[31] It includes raw data smoothing and derivative smoothing, which are determined by tuning 3 parameters (i.e. order of derivatives, degree of polynomial and number of smoothing points). We compared the separate and combined use of these different pre-processing methods in this work, to search for the best noise reduction mode.[32-33] Nevertheless, the selection of the appropriate pre-processing mode can only be based on the predictive results of the calibration model. Thus, combined with LSSVR modelling, the combined optimization of pre-processing mode will be more effective.

The division of all samples into calibration, validation and test sample set is an important part for model establishment and evaluation. Calibration samples were used to establish analytical models, validation samples used to optimize the modelling parameters, and test samples used to reveal the model's predictive performance.[34] Many experimental results showed that different divisions of calibration set and validation set would cause fluctuations of optimization effects, and that the corresponding modelling parameters (such as SG smoothing, LSSVR parameters, etc.) were also changed. Namely, the optimal model for each division was unstable for all divisions.[35] In order to establish stable models, it is necessary to make many different divisions and establish calibration models for each division. To search for the steady modelling parameters, the model predictive results were averaged based on different divisions and the optimal model with its parameters can be selected based on the averages.

In this paper, enhanced chemometric methods were applied for the FT-MIR spectroscopic analysis of HGB in human blood. Pre-processing methods were compared and the selection of the best pre-processing mode is designed in the combination with the grid-search LSSVR parameter tuning. This chemometric strategy is expected to the application in an extended wider scope. To get stable results, all grid-search LSSVR parameters were used to establish calibration models based on 30 different divisions of calibration and validation sets respectively, the selection processes for the optimal model were based on the average predictive results of the 30 divisions. For comparison, we have the optimal stable pre-processed LSSVR model compare to the commonly-used linear PLS regression, to verify the feasibility of this enhanced methodology performing on the FT-MIR spectrum of HGB concentration of human blood samples.

2. Experiments and methods

2.1 Materials and measurement

One hundred and eighty-one samples of human blood were collected in a fair hospital, from the routine human health examination. The HGB data of all samples were measured by BC-2800 Blood Cell Analyzer. The measured data were as reference chemical values for the spectroscopic calibration. The HGB chemical values of 181 samples ranged from 103.8 to 174.2 (g/L), the mean value and the standard deviation are 134.6 and 18.4 (g/L) respectively.

The FT-MIR spectra were collected using the Nicolet 360 FT-MIR Spectrometer (Thermo Fisher, USA) with its Smart OMNI cell. Samples were dripped on the glass film, having a thickness of 0.1mm. The full scanning range was from 600 to 4000 cm^{-1}. Scans of symmetrical interferograms with the resolution of 4 cm^{-1} were added for each spectrum. The surrounding conditions were controlled at the temperature of 25±1°C and the humidity of 46%RH. Each sample was measured 5 times and the average spectrum was for FT-MIR analysis.

2.2 Modelling preparation

Model indicators include root mean squared error (RMSE) and correlation coefficient (R).[36-37] They are calculated as

$$\mathrm{RMSE} = \sqrt{\frac{1}{N-1}\sum_{j=1}^{N}(C'_j - C_j)^2},$$

$$\mathrm{R} = \frac{\sum_{j=1}^{N}(C_j - C_m)(C'_j - C'_m)}{\sqrt{\sum_{j=1}^{N}(C_j - C_m)^2 \sum_{j=1}^{N}(C'_j - C'_m)^2}},$$

where C'_j and C_j are FT-MIR predicted value and measured

chemical value of the *j*-th sample, C'_m and C_m are respectively the mean predicted value and the mean chemical value of all samples, and *N* is the total number of samples in the designated sample set.

At a rough ratio of 2:1:1, all 181 human blood samples were divided into calibration, validation and test sets at fully random. The numbers of samples for calibration, for validation and for test were 90, 45 and 46 respectively. To get stable modelling parameters, the test sample set was independent out of the calibration-validation process, and the division of calibration and validation sets was performed for 30 times, and the calibration models established for each division. The RMSE and R of validation set (of test set) were denoted as RMSEV and R_V (RMSET and R_T) respectively.

For each combination of modelling parameters, the validation results (RMSEV and R_V) in 30 different divisions were obtained. And then the average value of RMSEV and R_V were further calculated and denoted by $RMSEV_{Ave}$, $R_{V,Ave}$ respectively. In this paper, $RMSEV_{Ave}$ was chosen as the indicator for the optimization of model parameters. Namely, according to the minimum $RMSEV_{Ave}$, the optimal pre-processing mode was selected, and the corresponding $R_{V,Ave}$ was found.

2.3 Combined pre-processing mode construction

MSC and SNV are fundamental pre-processing methods in spectroscopic analysis. MSC uses the average data to regress the spectra. Suppose we have *N* samples and *P* wavelengths, and denote the *j*-th spectrum as A_j, we could, based on the averaging spectrum, have the MSC-corrected spectrum as

$$A_j^{\mathrm{M}} = m_j \frac{1}{N}\sum_{i=1}^{N} A_i + b_j,\ j=1, 2\ldots N,$$

where the m_j and b_j are obtained by least square regression. This corrected A_j^{M} is expected to eliminate the diffuse reflection scattering effects.

Unlike MSC, SNV uses the central spectral data to remove slope variation. Each spectrum is corrected individually by centring the spectral values, and then the centred spectrum is scaled by the standard deviation calculated from individual spectral values. We could have the SNV formula as follows,

$$A_{i,j}^{\mathrm{S}} = \frac{A_{i,j} - \overline{A}_j}{\sqrt{\sum_{i=1}^{P}(A_{i,j} - \overline{A}_j)^2 / (P-1)}}\quad i=1, 2\ldots P,\ j=1, 2\ldots N,$$

where $A_{i,j}$ and $\overline{A}_j$ represents the *i*-th value and the central value of the *j*-th spectrum, respectively.

SGS includes the raw smoothing and derivative smoothing, the pre-processed data are uniquely determined by order of derivatives (OD), degree of polynomial (DP) and number of smoothing points (NSP).[31] It is difficult to get prospective smoothing effects when the interval between spectral points is too small and NSP set as small values. Hence, in this paper, the parameters of SGS were set tuning in a certain range, where OD={1, 2, 3, 4}, DP={2, 3, 4, 5}, NSP={5, 7… 75 (odd numbers only)}. The corresponding groups of SGS smoothing coefficients were computed.[30] The appropriate smoothing parameters can be selected according to specific analytical objects.

In this paper, we compared the separate and combined use of these pre-processing methods and selected the best pre-processing mode. Before developing the human blood HGB calibration models, the spectra data were pretreated by MSC, SNV, SGS, combination of MSC and SGS (MSC-SGS), combination of SNV and SGS (SNV-SGS), in which SGS performed smoothing and derivatives with parameter tuning. The performance of these pre-processing modes was evaluated in combination of LSSVR modelling, aiming to find out the minimum value of the RMSE's and R's.

2.4 LSSVM algorithm and the grid-search design

In the process of LSSVR methodology,[27,38-39] the predictive value c'_{jv} of the *j*-th validation sample is expressed in the following manner,

$$c'_{jv} = \sum_{i=1}^{n} \alpha_i K(A_j^{\mathrm{V}}, A^{\mathrm{C}}) \text{ and } \alpha_i = \left((A_i^{\mathrm{C}})^{\mathrm{T}} A_i^{\mathrm{C}} + \frac{1}{2\gamma} \right)^{-1},$$

where α_i is the Lagrange multiplier which depends on the regularization parameter γ,[18] $K(x_j, x_i)$ is the kernel function, A_j^{V} is the FT-MIR spectrum of the *j*-th sample in the validation set, and A^{C} is a linear combination of all the calibration spectra (the FT-MIR spectra with *n* wavenumbers), weighted by the concentration values.

The effect of LSSVR depends on the selection of the kernel and corresponding parameters. The Gaussian radial basis function (RBF) kernel has moderate robustness and stability to enable nonlinear modelling for the acquired FT-MIR spectral data,[18-19,26-27] and it is expressed as follows,

$$K(A_j^{\mathrm{V}}, A_i^{\mathrm{C}}) = \exp\left(-\frac{(A_j^{\mathrm{V}} - A_i^{\mathrm{C}})^2}{2\sigma^2} \right),$$

where σ^2 represents the kernel width and is used to tune the degree of generalization. When we select RBF as kernel, the performance of LSSVR primarily depends on the selection of parameters γ and σ^2, where γ determines the trade-off between the training error (which can be thought of as the model accuracy in calibration dataset) and the model robustness.

To optimize these two parameters, a multiscale grid-search is performed to enable the development of suitable calibration models. We designed tuning the parameters γ and σ respectively changed in a certain variable range, consecutively, or with a reasonable step. A grid-search of these two parameters is particularly necessary for a smooth subarea to obtain a low predictive error.

3. Results and discussion

The FT-MIR spectra of 181 human blood samples were showed in the full-scanning spectral range of 600-4000 cm^{-1} (see **Fig. 1**). There are severe affects from the spectral responses of water molecular at the absorption peaks of about 690 cm^{-1}, 1650 cm^{-1} and 3280 cm^{-1} respectively. The fingerprint region for human blood HGB was selected as 900-1600 cm^{-1}, which does not contain any of the water absorption peaks, because it was seen from **Fig. 1** that different samples had obvious discrepant absorbance in this fingerprint region. According to the spectral continuity, the full-scanning spectral data and the fingerprint data were used for modelling, respectively.

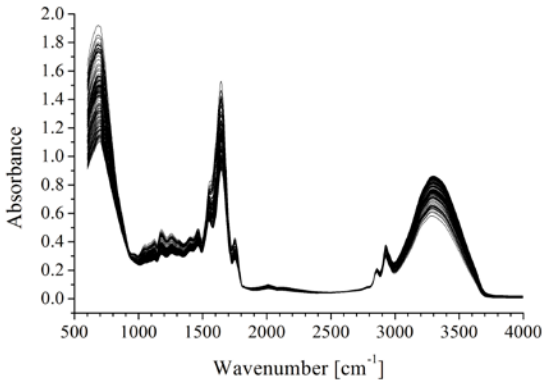


Fig.1 FT-MIR spectra of 181 human blood samples

3.1 Grid-search LSSVR models for the raw data

Grid-search LSSVR models were established for all 30 divisions of calibration set and validation set. It is worth noting that the two parameters of γ and σ^2 represent the regularization extension and the kernel width when using the RBF kernel. A grid-search for screening these two parameters is particularly necessary for a smooth subarea to obtain a low predictive error. We have γ changed from 10 to 300 with a step of 10, and σ changed consecutively from 1 to 20 (i.e. $\sigma^2=1^2, 2^2 \ldots 20^2$). The grid-search LSSVR models were established for each division in the calibration-validation process. According to the minimum RMSEV, the optimal parameters and predictive effects were selected in the full-scanning region and the fingerprint region, respectively. The modelling parameters (γ and σ) of the optimal LSSVR models, as well as the predictive results (RMSEV and R_V), were obtained. **Table 1** listed the parameters and the predictive results for each of the 30 divisions.

We obtained from **Table 1** that γ and σ of the optimal LSSVR models for each division was smaller than the corresponding maximum values of their changing ranges, which meant the tuning range of (γ, σ) was appropriate. However, the predictive effects (RMSEV and R_V) of the optimal LSSVR models had serious frustration for the different divisions, and also, the modelling parameters (γ and σ) were totally different. It meant that the separate optimization for each division cannot find the stable parameters, and we cannot catch the robust outperformed model that is stable and optimal simultaneously for all 30 divisions. Under this situation, it is necessary to further calculate the $RMSEV_{Ave}$ and $R_{V,Ave}$ for each combination of modelling parameters, and used as the optimizing indicator for the selection of a stable calibration model.

In searching for the minimum values of $RMSEV_{Ave}$ and $R_{V,Ave}$, the optimal LSSVR modelling parameters were selected as γ=230 and σ=12 for the full-scannnig region, and the corresponding $RMSEV_{Ave}$ and $R_{V,Ave}$ were found as 8.63 (g/L) and 0.885, respectively. While for the fingerprint region, the optimal LSSVR modelling parameters were selected as γ=120 and σ=9 and the corresponding $RMSEV_{Ave}$ and $R_{V,Ave}$ were found as 8.14 (g/L) and 0.896, respectively. These predictive effects for HGB of human blood was obviously acceptable in comparison with the results by using NIR spectrometry.[11]

Table 1 The predictive results and the modelling parameters for each of the 30 divisions, corresponding to the optimal grid-search LSSVR models in the full-scanning range and the fingerprint region

Division	Full-scanning region				Fingerprint region			
	γ	σ	RMSEV	Rv	γ	σ	RMSEV	Rv
1	250	7	8.98	0.849	170	8	8.92	0.856
2	50	8	9.14	0.841	90	11	9.08	0.865
3	80	10	9.69	0.810	180	11	9.62	0.827
4	130	13	8.42	0.906	170	11	7.69	0.927
5	190	8	8.59	0.883	130	11	8.53	0.882
6	180	16	9.70	0.783	120	9	9.63	0.786
7	50	10	8.23	0.886	130	7	8.63	0.893
8	190	10	8.87	0.845	220	8	8.80	0.830
9	220	11	8.47	0.912	160	10	8.50	0.897
10	140	10	9.15	0.871	240	13	9.09	0.844
11	180	8	9.89	0.760	180	8	9.82	0.782
12	240	6	9.94	0.781	230	12	9.87	0.764
13	290	12	9.68	0.798	80	8	9.61	0.812
14	140	7	8.64	0.894	190	8	8.44	0.896
15	140	14	8.94	0.829	190	6	8.87	0.845
16	250	10	9.02	0.864	140	10	8.96	0.860
17	90	10	8.78	0.856	60	7	8.71	0.903
18	170	9	8.40	0.874	240	8	8.58	0.873
19	220	11	8.61	0.902	270	10	8.09	0.898
20	60	10	9.65	0.825	190	10	9.58	0.800
21	220	7	9.84	0.768	250	12	9.77	0.771
22	290	11	8.65	0.886	220	14	8.59	0.901
23	140	11	7.99	0.909	140	10	8.55	0.887
24	230	8	8.56	0.889	230	8	8.41	0.876
25	170	9	8.55	0.880	90	14	8.49	0.885
26	120	6	9.35	0.785	250	6	9.28	0.784
27	80	12	8.69	0.899	120	10	8.32	0.896
28	60	10	9.99	0.773	250	7	9.92	0.777
29	130	9	9.37	0.796	180	9	9.30	0.797
30	130	13	8.57	0.877	290	16	8.51	0.878

3.2 Optimization of combined pre-processing mode with grid-search LSSVR model

We applied respectively the MSC, SNV, SGS, MSC-SGS and SNV-SGS pre-processing modes to the raw spectral data, and re-establish the grid-search LSSVR models. MSC and SNV are two fundamental methods; they processed the spectral data obeying their own algorithms. The SGS pre-processing method had 3 parameters for tuning; we set OD changed as 1, 2, 3, 4, DP changed as 2, 3, 4, 5, NSP changed from 5 to 75 (odd numbers only). In LSSVR model optimization, we still have γ changed from 10 to 300 with a step of 10 and σ changed consecutively from 1 to 20. The combined optimization of pre-processing mode and modelling parameters were performed on the full-scanning range data and on the fingerprint region data, respectively. To get the stable pre-processed improved models, we calculated the

$RMSEV_{Ave}$ and $R_{V,Ave}$ of all the 30 divisions for the combined optimization of grid-search LSSVR model with different pre-processing mode. The optimal models were selected based on goal of the minimum $RMSEV_{Ave}$.

Table 2. The predictive results and the corresponding parameters of the optimal LSSVR models for different pre-processing mode, for the full-scanning data and the fingerprint data

	Pre-processing	SGS parameter (OD,DP,NSP)	LSSVR (γ, σ)	$RMSEV_{Ave}$	$R_{V,Ave}$
Full-scanning region	—	—	(230, 12)	8.63	0.885
	MSC	—	(220, 8)	8.32	0.899
	SNV	—	(130, 14)	8.21	0.912
	SGS	(0, 2, 49)	(150, 9)	8.17	0.924
	MSC-SGS	(1, 4, 47)	(130, 13)	7.83	0.927
	SNV-SGS	(2, 4, 35)	(210, 11)	7.68	0.931
Fingerprint region	—	—	(120, 9)	8.14	0.896
	MSC	—	(260, 10)	8.05	0.918
	SNV	—	(120, 12)	7.88	0.922
	SGS	(1, 2, 31)	(190, 16)	7.61	0.932
	MSC-SGS	(2, 5, 55)	(170, 14)	7.33	0.937
	SNV-SGS	(1, 3, 45)	(110, 12)	6.98	0.941

We observed the prospective modelling results (see **Table 2**). It could be seen from **Table 2** that, for one thing, the modelling performance was quite improved by pre-processing for both the full-scanning data and the fingerprint data. The combined pre-processing modes lead to better pre-processed effects than the simple separate modes, and the optimized pre-processing mode was SNV-SGS. For another, the predictive results on the fingerprint region were obviously better than on the full-scanning region.

The best model was established on the fingerprint region, the optimized LSSVR parameters (γ, σ) were (110, 12), matched the SNV-SGS pre-processing mode with the SGS parameters of 1st-OD, 3rd-DP and 45-NSP. The corresponding output $RMSEV_{Ave}$ and $R_{V,Ave}$ were 6.98 (g/L) and 0.941, respectively. It revealed that the pre-processed LSSVR model predictive effect remarkably outperformed the ones obtained by raw LSSVR modelling. Hereafter we denote this optimal model as SNV-SGS-LSSVR model.

Further, we investigated the running procedure of SGS and studied the optimization of its own parameters. As OD and NSP are the core parameters in SGS algorithm, we sketched the optimal $RMSEV_{Ave}$ curve of the LSSVR model with SNV-SGS pre-processing. **Fig. 2** showed each optimal $RMSEV_{Ave}$ corresponds to each value of NSP for each OD. As for comparison, the minimum value of $RMSEV_{Ave}$ of the non-pre-processed LSSVR modelling was also drawn (as a straight line) in **Fig. 2**. It was indicated in **Fig. 2** that the predictive effect in 0th and 1st order derivative could be better than the raw LSSVR model if a certain NSP selected. Much improved predictive results can be achieved in 1st order derivative, using 15 smoothing points or more. Some acceptable predictive results can be obtained by using a larger number of smoothing points when using 2nd, 3rd, 4th order derivative. The best SGS pre-processing parameters were 1st-OD, 45-NSP, and the corresponding DP was 3rd. We successively provided the sketch of the smoothed spectra by SNV-SGS, using the best optimal SGS parameters (see **Fig. 3**).

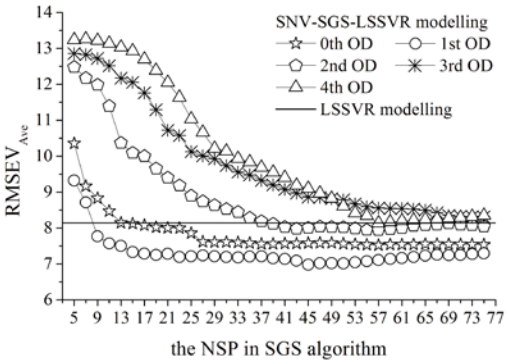


Fig.2 $RMSEV_{Ave}$ values corresponding to for each NSP and OD of SG smoother

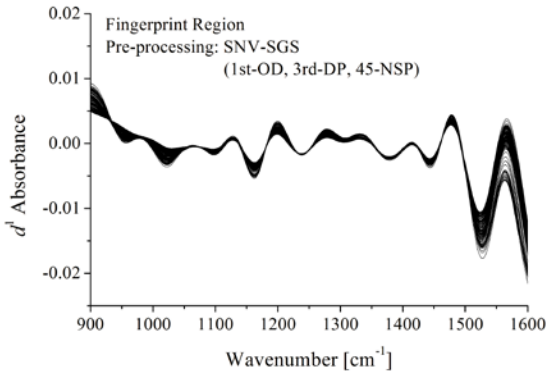


Fig.3 The pre-processed spectra in the fingerprint region by combination of SNV and SGS (1st derivative)

Moreover, we investigated how the LSSVR parameters (γ and σ) influence the predictive effect based on the optimal pre-processing mode (SNV-SGS). For the fingerprint data, the $RMSEV_{Ave}$ corresponding to each γ and each σ for the optimal SNV-SGS-LSSVR model can be also seen in **Fig. 4** (sub-figure (a) for γ and sub-figure (b) for σ). Obviously, the $RMSEV_{Ave}$ had a slowly minimum value when γ was around 110, but a relatively sharp trough at the point of σ valuing 12.

Furthermore, to verify the reliability of the nonlinear LSSVR modelling results, we have the optimal stable SNV-SG-LSSVR model compared with the linear models established by the most commonly-used PLS method. Similarly, PLS models were established for all 30 divisions of calibration set and validation set by tuning of the number of latent valuables. The optimal combined pre-processing mode (SNV-SGS) was also applied to the modelling process. The values of $RMSEV_{Ave}$ and $R_{V,Ave}$ were calculated for each latent valuable. According to the minimum $RMSEV_{Ave}$, the optimal stable PLS model were selected with 5 latent valuables, and the corresponding $RMSEV_{Ave}$ and $R_{V,Ave}$

were found as 7.94 (g/L) and 0.907, respectively. Obviously, the nonlinear LSSVR modelling, with the optimal pre-processing mode (SNV-SG), reached a slightly better predictive effect than the linear PLS using the full-scanning data, and output an obviously improved result when using the fingerprint data. We concluded that the combined optimization of LSSVR modelling and SNV-SGS pre-processing mode can improve the predictive ability of FT-MIR spectrometry.

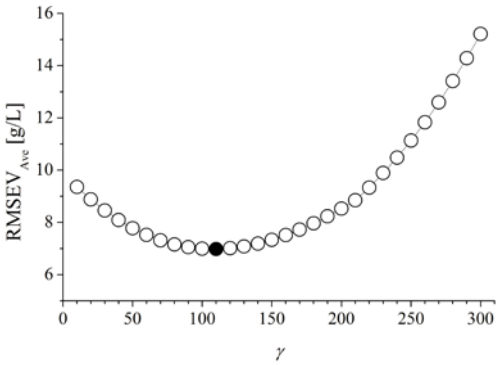


Fig.4 sub-figure (a)

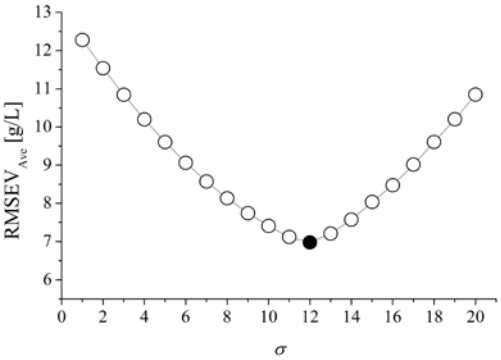


Fig.4 sub-figure (b)

Fig.4 $RMSEV_{Ave}$ corresponding to the optimized LSSVR models (sub-figure (a) distributes the minimum of $RMSEV_{Ave}$ for each γ, and sub-figure (b) distributes the minimum of $RMSEV_{Ave}$ each value of σ)

3.3 Evaluation for the optimized SNV-SG-LSSVR model

The 46 test human blood samples, excluded in the calibration-validation process, were used to evaluate the optimized SNV-SG-LSSVR model. The HGB values of test samples were predicted by using the optimizationally designated pre-processing mode with the satisfactory parameters of SGS and LSSVR algorithms.

The FT-MIR predicted values of HGB for 46 test samples were obtained close to the measured chemical values, with the RMSEP and R_P were 0.792 (g/L) and 0.908. The results demonstrated a root mean square error no more than 6% of the mean chemical value and a correlation coefficient higher than 0.9. This prediction effect was much satisfactory for the randomly selected test samples, as the optimization of SNV-SG-LSSVR model was a combined tuning of LSSVR modelling parameters and the optimal pre-processing parameters. The correlation relationship between the FT-MIR predicted values and the measured chemical values was showed in **Fig 5**.

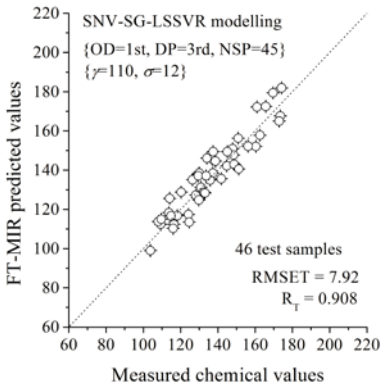


Fig.5 The correlation relationship between the FT-MIR predicted values and the measured chemical values for the test samples

4 Conclusions

In this paper, enhanced chemometric methods were applied to the FT-MIR spectroscopic quantitative determination of HGB in human blood. We constructed the framework for optimizing the separate and combined use of the spectral pre-processing methods of MSC, SNV and SGS, in which the SGS parameters were discussed tunable in a certain designated range. The performances of different pre-processing modes were evaluated in combination of LSSVR modelling, and the gird-search technique was proposed to use for the LSSVR parameter optimization. These analytical methods were executed on the FT-MIR fingerprint region of human blood HGB, with comparison to the full-scanning region.

To avoid serious modelling frustration, we have the division of calibration and validation sets performed 30 times, and presented the modelling results using the averaging value of the 30 divisions. The model establishing results showed that the models on fingerprint data are obviously better than those on the full-scanning range. The best pre-processing mode was SNV-SGS, with the SGS parameters of 1st-OD, 3rd-DP and 45-NSP. The combined optimized LSSVR parameters (γ, σ) were selected as (110, 12) in the calibration-validation process. This selected model provided the predictive $RMSEV_{Ave}$ of 6.98 (g/L) and R_V of 0.941. The predictive effect was obviously better than that obtained by LSSVR models without pre-processing, and in comparison, was also better than that of PLS modelling. It was theoretically verified the feasibility of SNV-SG-LSSVR methodology performing on the FT-MIR analysis of HGB concentration of human blood samples.

The optimized SNV-SG-LSSVR model was further evaluated by the randomly selected 46 test samples excluded in the calibration-validation process. The evaluation model was also established on the FT-MIR fingerprint region. The predicted values of HGB were obtained close to the measured chemical values, with the RMSET and R_T were 0.792 (g/L) and 0.908 respectively. The results demonstrated a root mean square error no more than 6% of the mean chemical value and a correlation

coefficient higher than 0.9. The results showed that the combined optimization of LSSVR modelling with the best pre-processing mode obviously improved the predictive ability of FT-MIR spectroscopic analysis of HGB in human blood.

Acknowledgment

This work was supported by Natural Scientific Foundation of China (11226219) and Natural Scientific Foundation of Guangxi (2013GXNSFBA019004, 2014GXNSFBA118023).

Notes and references

College of Science, Guilin University of Technology, 12 Jian'gan Road, Guilin 541004, China. Tel: +86 773 5896179;
**Corresponding Authors. E-mail: huazhouchen@163.com (H.Z. Chen); tanggq@glut.edu.cn (G.Q. Tang)*
DOI: 10.1039/b000000x/

1 A. Shaw, K.Z. Liu, A. Man, T.C. Dembinski and H.H. Mantsch, *Clinical Chemistry*, 2002, **48**, 499-506.
2 C. Petibois, V. Rigalleau, A.M. Melin, A. Perromat, G. Cazorla, H. Gin and G. Deleris, *Clinical Chemistry*, 1999, **45**, 1530-1535.
3 S. Gunasekaran, T.S. Renuga Devi and P.S. Sakthivel, *Asian Journal of Clinical Cardiology*, 2007, **10**, 19-29.
4 T. Fujii and Y. Miyahara, *Applied Spectroscopy*, 1998, **52**, 128-133.
5 D.D. Purkayastha and V. Madhurima, *Journal of Molecular Liquids*, 2013, **187**, 54-57.
6 Y. Furutani and H. Kandori, *Biochimica et Biophysica Acta*, 2014, **1837**, 598-605.
7 N.A. Ngo Thi and D. Naumann, *Analytical and Bioanalytical Chemistry*, 2007, **387**, 1769-1777.
8 N. Nicolaou, Y. Xu and R. Goodacre, *Analytical Chemistry*, 2011, **83**, 5681-5687.
9 Z.M. Chuah, R. Paramesran, K. Thambiratnam and S.C. Poh, *Chemometrics and Intelligent Laboratory Systems*, 2010, **104**, 347-351.
10 W.J. McAuley, M.D. Lad, K.T. Mader, P. Santos, J. Tetteh, S.G. Kazarian, J. Hadgraft and M.E. Lane, *European Journal of Pharmaceutics and Biopharmaceutics*, 2010, **74**, 413-419.
11. I.V. Nagy, K.J. Kaffka, J.M. Jako, E. Gonczol and G. Domjan, *Clinica Chimica Acta*, 1997, **264**, 117-125.
12. Y. C. Shen, A. G. Davies, E. H. Linfield, P. F. Taday, D. D. Arnone, and T. S. Elsey, *Journal of Biological Physics*, 2003, **29**, 129–133.
13. G. Hosafci, O. Klein, G. Oremek, and W. Mantele, *Analytical and Bioanalytical Chemistry*, 2007, **387**, 1815–1822.
14. M. Brandstetter, T. Sumalowitsch, A. Genner, A. Posch, C. Herwig, A. Drolz, V. Fuhrmann, T. Perkmann, and B. Lendl, *Analyst*, 2013, **138**, 4022–4028.
15 A.L.Q. Baddini, L.E.R. Cunha, A.M.C. Oliveira and R.J. Cassella, *Analytical Biochemistry*, 2010, **397**, 175-180.
16. R. A. Shaw, C. Rigatto, M. Reslerova, S. L. Ying, A. Man, B. Schattka, C. F. Battrell, J. Matthewson, and C. Mansfield, *Analyst*, 2009, **134**, 1224–1231.
17 Y. Lee, S. Lee, J.Y. In, S.H. Chung and J.H. Yon, *Journal of Korean Medical Science*, 2008, **23**, 674-677.
18 D. Perez-Guaita, J. Ventura-Gayete, C. Pérez-Rambla, M. Sancho-Andreu, S. Garrigues, and M. de la Guardia, *Analytical and Bioanalytical Chemistry*, 2012, **404**, 649–56.
19 M. Daszykowski, M.S. Wrobel, H. Czarnik-Matusewicz and B. Walczak, *Analyst*, 2008, **113**, 1523-1531.
20 J.W. Jin, Z.P. Chen, L.M. Li, R. Steponavicius, S.N. Thennadil, J. Yang and R.Q. Yu, *Analytical Chemistry*, 2012, **84**, 320-326.
21 S.P. Magalhaes da Silva, A.M. da Costa Lopes, L.B. Roseiroa and R. Bogel-Łukasik, *RSC Advances*, 2013, **3**, 16040-16050
22 M. Daszykowski, M.S. Wrobel, H. Czarnik-Matusewicz, and et al. *Analyst*, 2008, **113**, 1523–1531.
23 D. Perez-Guaita, J. Kuligowski, G. Quintás, S. Garrigues, and M. de la Guardia, *Talanta*, 2013, **107**, 376–385.
24 H.Z. Chen, G.Q. Tang, Q.Q. Song and W. Ai, *Analytical Letters*, 2013, **46**, 2060-2074.
25 I. Barman, C.R. Kong, G.P. Singh, R.R. Dasari and M.S. Feld, *Analytical Chemistry*, 2010, **82**, 6104-6114.
26 I. Barman, N.C. Dingari, G.P. Singh, J.S. Soares, R.R. Dasari and J.M. Smulko, *Analytical Chemistry*, 2012, **84**, 8149-8156.
27 J.A.K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor and J. Vandewaller, *Least Squares Support Vector Machines*, World Scientific Publishing, 2002.
28 A. Rinnan, F. van den Berg and S.B. Engelsen, *Trends in Analytical Chemistry*, 2009, **28**, 1201–1222
29 P. Geladi, D. Macdougall and H. Martens, *Applied Spectroscopy*, 1985, **39**, 491–500.
30 T. Fearn, C. Riccioli, A. Garrido-Varo and J.E. Guerrero-Ginel, *Chemometrics and Intelligent Laboratory Systems*, 2009, **96**, 22–26.
31 A. Savitzky and M.J.E. Golay, *Analytical Chemistry*, 1964, **36**, 1627-1637.
32 H.H. Madden, *Analytical Chemistry*, 1978, **50**, 1383-1386.
33 H. Chen, T. Pan, J. Chen and Q. Lu, *Chemometrics and Intelligent Laboratory Systems*, 2011, **107**, 139-146.
34 I.A. Vasilieva, *Journal of Quantitative Spectroscopy & Radiative Transfer*, 2006, **101**, 159-165.
35 T. Pan, Z.H. Chen, J.M. Chen and Z.Y. Liu, *Analytical Methods*, 2012, **4**, 1046-1052.
36 H.Z. Chen, Q.Q. Song, G.Q. Tang and L.L. Xu, *Journal of Cereal Science*, 2014, **60**, 595-601.
37 H. Chen, W. Ai, Q. Feng, Z. Jia and Q. Song, *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 2014, **118**, 752-759.
38 N. Cristianini and J. Shawe-Taylor, *An introduction to support vector machines and other kernel-based learning methods*, Cambridge University Press, New York, 2000.
39 T.T Zou, Y. Dou, H. Mi, J.Y. Zou and Y.L. Ren, *Analytical Biochemistry*, 2006, **355**, 1-7.