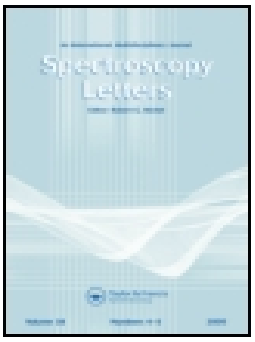

Taylor & Francis
Taylor & Francis Group

Spectroscopy Letters

An International Journal for Rapid Communication

ISSN: 0038-7010 (Print) 1532-2289 (Online) Journal homepage: https://www.tandfonline.com/loi/lstl20

Baseline correction for infrared spectra using adaptive smoothness parameter penalized least squares method

Feng Zhang, Xiaojun Tang, Angxin Tong, Bin Wang, Jingwei Wang, Yangyu Lv, Chunrui Tang & Jie Wang

To cite this article: Feng Zhang, Xiaojun Tang, Angxin Tong, Bin Wang, Jingwei Wang, Yangyu Lv, Chunrui Tang & Jie Wang (2020): Baseline correction for infrared spectra using adaptive smoothness parameter penalized least squares method, Spectroscopy Letters, DOI: 10.1080/00387010.2020.1730908

To link to this article: https://doi.org/10.1080/00387010.2020.1730908

Published online: 22 Feb 2020.

Baseline correction for infrared spectra using adaptive smoothness parameter penalized least squares method

Feng Zhang[a], Xiaojun Tang[a], Angxin Tong[a], Bin Wang[a], Jingwei Wang[a], Yangyu Lv[a], Chunrui Tang[b] and Jie Wang[b]

[a]State Key Laboratory of Electrical Insulation and Power Equipment, Xi'an Jiaotong University, Xi'an, China; [b]CCTEG Chongqing Engineering Co. Ltd, Chongqing, China

ABSTRACT

Baseline wander is a common problem in analysis with Fourier Transform Infrared Spectrometer (FTIR). And it is necessary to correct baseline drift for further quantitative and qualitative analysis. Several baseline correction algorithms based on penalized least squares have been proposed. However, these methods are usually used in noise-free or low-noise environments. In this paper, a novel algorithm named adaptive smoothness parameter penalized least squares was proposed. The smoothness parameters were set by user at first. Then, the smoothness parameter was updated iteratively according to the difference between the original spectrum and fitted baseline. When the iteration reaches the termination condition, the fitted baseline can be obtained. In the end of the paper, experimental results on simulated spectra and measured infrared spectra of methane were given. The simulated spectra results demonstrate that the proposed method has better performance than existing methods, especially when the spectra contain high noise. The results of infrared spectra confirm that the proposed method has good performance and can be applied to correct spectral baseline accurately.

ARTICLE HISTORY
Received 26 September 2019
Accepted 13 February 2020

KEYWORDS
Baseline correction; infrared spectra; penalized least squares

Introduction

Spectroscopic analysis technology such as Raman and infrared spectroscopy has been applied broadly in many fields such as environmental monitoring, power system fault diagnosis, coal mine disaster early warning, and so on, because of its rapid analysis speed and nondestructive since 1980s.[1–4] However, any spectra measured with a spectrometer usually contain unnecessary elements such as noise and background in addition to the desired signal itself.[5–6] Spectrometer is composed of numbers of precise optical components that are susceptible to environmental factors, such as light, temperature, humidity, vibration and so on during long-time continuous work.[7] These factors can easily lead to baseline drift and make it impossible to determine accurately peak height and peak area that were used to perform further analysis. Hence, it is necessary to correct the baseline of spectra before they are used to analyze gas qualitatively or quantitatively.

In recent years, researchers have proposed several baseline correction methods such as derivative method,[8,9] wavelet transform,[10–12] polynomial fitting,[9,13–15] automatic iterative moving average,[16] and penalized least squares. When the derivative method is used to correct the baseline of the spectra, the noise may be amplified easily. Once noise is amplified, the SNR (Signal-to-Noise Ratio) is reduced, and the spectra is distorted after the baseline correction is performed. Therefore, the spectra should be smoothed first in application. Wavelet transform assumes that baseline, noise and spectra signal exist in different frequency band, the baseline is a low-frequency signal, while the noise is high frequency signal. Therefore, high frequency noise and low-frequency baseline can be filtered by wavelet transform and reconstruction. However, when the absorption lines of the analyte are sparse, many peaks appear due to the limited resolution of the spectrometer, and the

CONTACT Xiaojun Tang xiaojun_tang@mail.xjtu.edu.cn State Key Laboratory of Electrical Insulation and Power Equipment, Xi'an Jiaotong University, Xi'an 710049, China.

absorption peak region is considered as a high-frequency signal. The wavelet transform method cannot distinguish the peak signal from the noise signal, which may result in the distortion of the corrected spectrum. Additionally, it is difficult to select the optimum wavelet basis, decomposition lever and threshold of wavelet coefficients. The polynomial fitting algorithm is simple to be performed and effective, but it is prone to over-fitting or under-fitting. And it needs to select appropriate fitting order. The principle of automatic iterative moving average[17] tends to over-estimate the baseline in peak regions, and it is not suitable to fit the baseline when the absorption spectra of every component of analyst overlap with each other heavily.

Baseline correction method based on penalized least squares was presented by taking into account the fidelity and smoothness of the fitted baseline. It is commonly used in various spectral preprocessing because of its fast speed and avoid the peak detection. Subsequently, several improved algorithms have been developed. Eilers designed asymmetric least squares method (AsLS)[18] to correct baseline of various spectra. For this method, weights are updated iteratively by estimating a baseline. If a signal is above a fitted baseline, it is considered as the absorption peak region and assigns small weight or sets weight zero to achieve automatic interpolation fitting. On the other hand, large weight was given when a signal is below a fitted baseline. Zhang et al. considered that AsLS algorithm need to optimize two parameters,λ and p to obtain a satisfactory baseline fitting results, and the asymmetry parameter p is all the same if the original signal is above a fitted baseline. In this aspect, an adaptive iteratively reweighted penalized least squares method (airPLS)[19] was proposed, which further improves the accuracy of baseline correction and reduces the calculation time. He and Zhang presented an improved asymmetric least squares (IAsLS)[20,21] method to correct baseline. Compared with airPLS method, its performance of predictability was further improved. But when a spectrum is contaminated with additive noise, both airPLS and IAsLS methods give under-fitting baseline.[6] In order to complete baseline correction in noisy environment, Park proposed an asymmetrically reweighted penalized least squares smoothing method (arPLS),[22] which give the weights by introducing a generalized logic function. This method can be applied in different noise environments. Although the arPLS baseline correction algorithm can obtain satisfactory results in the no peak regions, a boosted baseline will be given in the peak regions. According to our experiments, the value of smoothing parameter λ has a great influence on the performance of baseline correction. But for above four methods, the smoothing parameter remains unchanged when λ is assigned by user. Therefore, a baseline correction method of adaptive smoothness parameter penalized least squares (asPLS) is proposed in this paper by setting large λ in peak regions and a small λ in no peak regions.

In the following section, previous penalized least squares methods were briefly described at first. Then, the asPLS algorithm was introduced in detail. Next, we compared these baseline correction methods with simulated spectra and real infrared spectra. The experiments with simulated and infrared spectra confirmed that the proposed method can effectively handle diverse baseline types with different noise.

Methodology

The penalized least squares methods

Suppose $\mathbf{y}$ be the spectrum length of N, obtained by the spectrometer which is sampled at equal intervals. This assumption is applicable to most devices because signal is sampled by time, wavelength or frequency. Let $\mathbf{z}$ be the fitted vector. Then, $\mathbf{z}$ can be obtained by minimizing the following penalized least squares function:

$$F(\mathbf{z}) = (\mathbf{y}-\mathbf{z})^{\mathrm{T}}(\mathbf{y}-\mathbf{z}) + \lambda(\mathbf{D}\mathbf{z})^{\mathrm{T}}(\mathbf{D}\mathbf{z}) \quad (1)$$

where $\mathbf{D}$ is a second-order difference matrix, suppose N is 6, $\mathbf{D}$ can be expressed as follows:

$$\mathbf{D} = \begin{bmatrix} 1 & -2 & 1 & 0 & 0 & 0 \\ 0 & 1 & -2 & 1 & 0 & 0 \\ 0 & 0 & 1 & -2 & 1 & 0 \\ 0 & 0 & 0 & 1 & -2 & 1 \end{bmatrix} \quad (2)$$

The parameter λ is introduced to balance the fitness and smoothness. In order to get the minimum value of F. By calculating partial derivative of Eq. (1) for $\mathbf{z}$ and making it equal to 0, a linear equation can be obtained as follows:

$$\mathbf{z} = (\mathbf{I} + \lambda \mathbf{D}^{\mathrm{T}}\mathbf{D})^{-1}\mathbf{y} \tag{3}$$

where $\mathbf{I}$ is a unit matrix, in order to correct a baseline, a weight vector w is introduced. If a signal is above a fitted baseline, assigns small weight or sets weight zero to achieve automatic interpolation fitting. On the other hand, large weight was given when a signal is below a fitted baseline. For this case, the linear equation changes to the following function:

$$\mathbf{z} = (\mathbf{W} + \lambda \mathbf{D}^{\mathrm{T}}\mathbf{D})^{-1}\mathbf{W}\mathbf{y} \tag{4}$$

where $\mathbf{W} = \mathrm{diag}(\mathbf{w})$, $\mathbf{W}$ is a diagonal matrix with $\mathbf{w}$ on its diagonal. If the absorption peak regions are known beforehand, the weight of these areas can be set to zero. The system considers these regions as missing fragments. Then, these regions are interpolated automatically. In no peak regions, set w_i to 1, the original signal can be smoothed. But when applying this method to baseline correction, it is necessary to know the absorption peak regions in advance. To avoid detection of absorption peaks, AsLS and airPLS method were proposed. The weight vector $\mathbf{w}$ is adaptively obtained using an iterative method. At the beginning of the iteration, set $\mathbf{w}$ to have ones. For the AsLS method, update $\mathbf{w}$ as follows:

$$w_i = \begin{cases} p & y_i > z_i \\ 1-p & y_i \le z_i \end{cases} \tag{5}$$

where p is an asymmetry parameter, which is recommended to set between 0.001 and 0.1, the range of i is N. When the change of weight vector does not change any more or the number of iterations is reached, the iteration terminates. By adjusting parameters p and λ, a satisfactory baseline can be obtained. For the airPLS method, $\mathbf{w}$ can be obtained with the Eq. (6):

$$w_i = \begin{cases} 0, & z_i \le y_i \\ e^{t(y_i - z_i)/|\mathbf{d}^-|} & z_i > y_i \end{cases} \tag{6}$$

where vector $\mathbf{d}^-$ consists of negative elements of the differences between $\mathbf{y}$ and $\mathbf{z}$; t is the number of iterations. The iteration will stop either with the maximum number of iterations set beforehand or when the following terminative condition reached:

$$|\mathbf{d}| < 0.001 \times |\boldsymbol{y}| \tag{7}$$

In order to complete baseline correction in noisy environment, Park proposed an asymmetric reweighted penalized least squares (arPLS) algorithm, which assigns the weight vector $\mathbf{w}$ according to the follow function:

$$w_i = \begin{cases} \mathrm{logistic}(y_i - z_i, m_{\mathbf{d}^-}, \sigma_{\mathbf{d}^-}), & y_i \ge z_i \\ 1 & y_i < z_i \end{cases} \tag{8}$$

where $\mathbf{d}^-$ is the region where the original signal $\mathbf{y}$ is less than the fitted baseline $\mathbf{z}$. $m_{\mathbf{d}-}$ and $\sigma_{\mathbf{d}-}$ are the mean and the standard deviation of $\mathbf{d}^-$. The logistic function can be specified as follows:

$$\mathrm{logistic}(d_i, m_{\mathbf{d}^-}, \sigma_{\mathbf{d}^-}) = \frac{1}{1 + e^{2(d_i - (-m_{\mathbf{d}^-} + 2\sigma_{\mathbf{d}^-}))/\sigma_{\mathbf{d}^-}}} \tag{9}$$

The iteration will stop when the weight is no longer changes or the changes of weight is minimal.

The proposed method: asPLS

According to Park et al., AsLS and airPLS methods give a descending baseline when the spectral signal contains noise.[22] The arPLS method ensures that when the $\mathbf{d}$ is less than $m_{\mathbf{d}-}$, the weight values are almost the same and are close to one. In this way, a satisfactory baseline can be obtained in the no peak regions. However, in peak regions, the fitted baseline will be boosted. That is a natural consequence because the tail region of the absorption peak is low. And it is easy to be misjudged as the noise. If d_i is less than $-m_{\mathbf{d}-}$, w_i will close to one. As a result, the final baseline is overestimated in the peak regions. Moreover, when the original signal $\mathbf{y}$ contains different absorption peaks, in the region of small absorption peak, $\mathbf{d}$ is usually less than $m_{\mathbf{d}-}$, the small peak regions will be considered as a noise. This phenomenon is very common because there are many absorption peaks in an absorption spectrum, especially when the analyte is mixture. In such spectrum, the small peaks may be regarded as noise, which leads to the failure of following quantitative or qualitative analysis. Addressed the problems of the above methods, a baseline correction method of adaptive smoothness parameter penalty least square

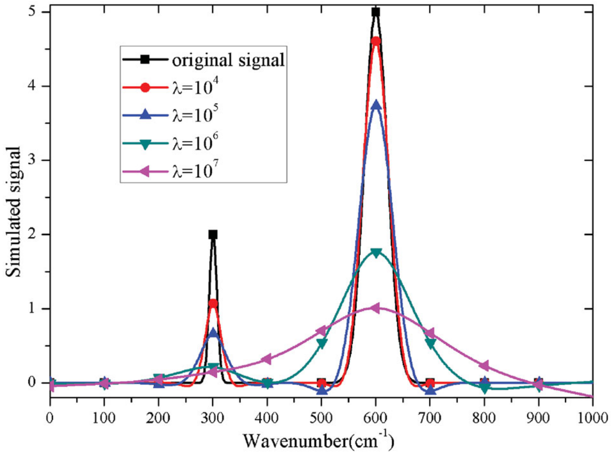

Figure 1. Simulated spectra smoothing results with different λ values. λ is smoothness parameter. The simulated baseline is $y = 0$. The figure reveals that with the increase of λ, the smoothing result curve is closer to the baseline in the peak regions, and in the no peak regions, the curve deviates further from the baseline.

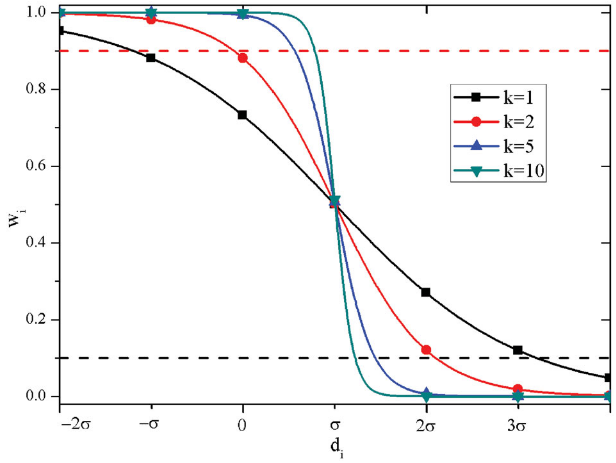

Figure 2. The relationship between asymmetric coefficient k and weight vector $\mathbf{w}$. With the increase of k, the curve becomes narrower. $\mathbf{d}$: the difference between the original signal $\mathbf{y}$ and the estimated baseline $\mathbf{z}$, $\mathbf{d} = \mathbf{y} - \mathbf{z}$. $\mathbf{d}^-$: the consist of negative elements of the $\mathbf{d}$; σ: the standard deviation of $\mathbf{d}^-$; k: asymmetric coefficient; $\mathbf{w}$: weight vector; subscript i is the ith element of the vector.

method (asPLS) was proposed. The core idea of asPLS algorithm is to set large λ in peak regions, and set a small λ in no peak regions. The asPLS algorithm can be described as follows.

Firstly, Gauss function is used to generate two absorption peaks, and different λ is chosen to observe the relationship between λ and $\mathbf{z}$ according to Eq. (3). That is shown in Fig. 1. The real baseline is $\mathbf{y} = 0$, from Fig. 1, it can be clearly seen that in the peak regions, with the increase of λ, the fitted curve $\mathbf{z}$ is closer to the baseline, and in the no peak regions, the curve deviates further from the baseline. Based on this fact, a coefficient vector α was introduced, and the length of α is N. At the initial iteration process, each element in the vector α was set to one. Then, according to following equation to obtain a new α.

$$\alpha_i = \frac{\mathrm{abs}(y_i - z_i)}{\max(abs(\mathbf{y} - \mathbf{z}))} \tag{10}$$

where abs($y_i - z_i$) is the absolute value of the subtraction between spectrum signal $\mathbf{y}$ and fitted vector $\mathbf{z}$ at the ith spectrum signal, max(abs($\mathbf{y} - \mathbf{z}$)) means the maximum value of subtraction between spectrum signal $\mathbf{y}$ and fitted vector $\mathbf{z}$. From Eq. (10), each element in α is determined by the difference between $\mathbf{y}$ and $\mathbf{z}$. And at the peak position regions, α_i can obtain a larger value, and at the no peak regions, the value of α_i is very small and close to zero. By multiplying $\boldsymbol{\alpha}$ and λ, the new smoothness parameter λ_i at each wavenumber position can be obtained. Thus, the Eq. (4) changes to the following function:

$$\mathbf{z} = (\mathbf{W} + \lambda\alpha\mathbf{D}^{\mathrm{T}}\mathbf{D})^{-1}\mathbf{W}\mathbf{y} \tag{11}$$

This ensures that the value of λ_i in the peak regions are greater than that in the no peak regions, so that the fitted baseline $\mathbf{z}$ is close to the baseline in all regions.

Secondly, $\mathbf{w}$ is initialized to have ones. And in the process of iteration, the weight vector $\mathbf{w}$ can be obtained by using the following expression:

$$w_i = \frac{1}{1 + e^{k(d_i - \sigma_{\mathbf{d}^-})/\sigma_{\mathbf{d}^-}}} \tag{12}$$

where $\mathbf{d}$ is the difference between the original signal $\mathbf{y}$ and the fitted baseline signal $\mathbf{z}$, $\mathbf{d}^-$ consists of negative elements of the $\mathbf{d}$, and $\sigma_{\mathbf{d}-}$ is the standard deviation of $\mathbf{d}^-$. k is asymmetric coefficient, the relationship between k and $\mathbf{w}$ is depicted in Fig. 2.

As you see in Fig. 2, with the increase of k, the curve becomes narrower. In the extreme case, k is infinite, the Eq. (12) be a reverse step function. According to PauTa criterion, if a signal is in the 3σ from the estimated noise mean,[22] we can consider this area as an absorption peak region. At this point, the w_i should be set to close zero. The minimum k can be solved according to the following formula:

$$1-\int_{3\sigma}^{\inf}\frac{1}{1+e^{k(x-\sigma)/\sigma}}\mathrm{dx}\geq 0.9973 \qquad (13)$$

According to Eq. (13), when $\sigma = 2$, the minimum k is 1.3949, because the value of σ is usually less than two, so the value of k can be set to two. Finally, the iteration stops when $|\mathbf{w}^i - \mathbf{w}^{i-1}|/|\mathbf{w}^{i-1}|$ is less than a threshold or it reaches the number of iterations. The proposed method of asPLS can be described more briefly by the flow chart shown in Fig. 3.

The proposed asPLS method is easy to implement in MATLAB environment. Supposing $\mathbf{y}$ is an absorbance spectrum measured by an infrared spectrometer and contains N wavenumbers, baseline correction can be finished quickly by setting the values of smoothing parameters λ and iteration termination parameters ε. The parameter λ is recommended to set between 10^4 and 10^8, and the ε is set to 10^{-4}. The following code describes the implementation of the asPLS method.

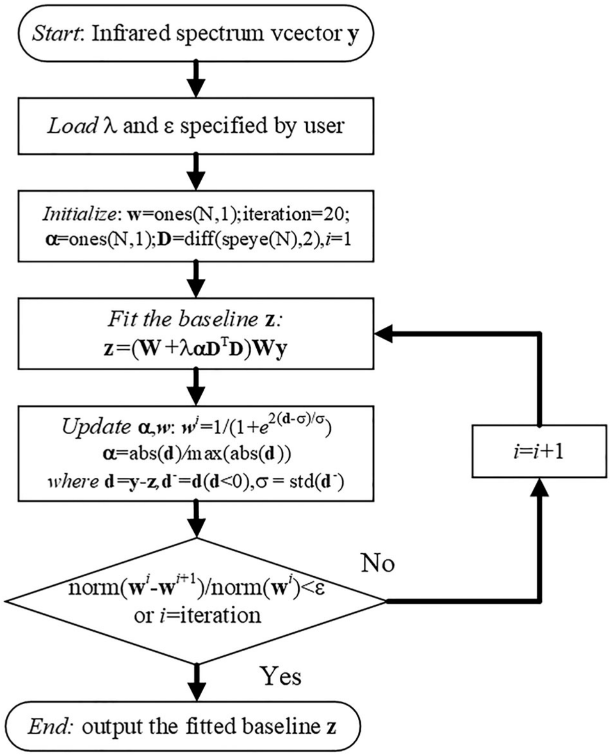
Start: Infrared spectrum vcector y
Load λ and ε specified by user
Initialize: w=ones(N,1);iteration=20;
α=ones(N,1);D=diff(speye(N),2),i=1
Fit the baseline z:
z=(W+λαD^TD)Wy
Update α,w: w^i=1/(1+e^{2(d-σ)/σ})
α=abs(d)/max(abs(d))
where d=y-z,d^-=d(d<0),σ = std(d^-)
i=i+1
norm(w^i-w^{i+1})/norm(w^i)<ε
or i=iteration
No
Yes
End: output the fitted baseline z

Figure 3. The flowchart of baseline correction by proposed algorithm. The proposed algorithm obtained the baseline by updating smoothing parameter and weight vector adaptively, setting the threshold and max iteration times.

Experiments

In this section, the performance of the above four baseline correction methods was evaluated by simulated spectra data and infrared spectra, respectively. The simulated spectra data contains four different baselines, and each baseline contains two different levels of noise. The infrared spectral are 7 different concentrations of methane and ethane scanned by infrared spectrometer. All the experiments were performed in MATLAB 9.1.0(R2016b) (MathWorks, MA, USA).

Simulated data

The simulated spectra data are composed of pure signal, baseline and noise, [23] which can be expressed as the following equation:

$$y(v) = s(v) + b(v) + n(v) \qquad (14)$$

where y(v) denotes the simulated spectra data, $s(v)$ represents the pure spectral data, $b(v)$ stand for the simulated baseline and $n(v)$ displays the random noise. The spectra obtained by FTIR is actually a convolution of the real spectrum and the window function, so the shape of the peak is similar to that of Gauss function. Therefore, the pure spectral signal can be simulated by several Gauss peaks. The Gauss function is expressed as follows.

$$s(v) = He^{-\left(\frac{v-b}{\sigma}\right)^2} \qquad (15)$$

where H is the height of the peak, b denotes the position of the peak, and σ is related to the full width at half maxima (FWHM).[24] In this experiment, the length of the simulated spectra is 1300 and the number of absorption peaks is 8. The Gauss peak parameters are shown in Table 1.

Table 1. The Gauss peak parameters used for constructing simulated spectral.

Peak number	H	b	σ
1	2	100	20
2	1	200	20
3	2	400	40
4	1	500	30
5	4	800	50
6	0.5	1000	15
7	1	1100	20
8	1.5	1200	20

H is the height of the peak, b denotes the position of the peak, and σ is related to the full width at half maxima. The pure spectral signal can be simulated using H, b and σ.

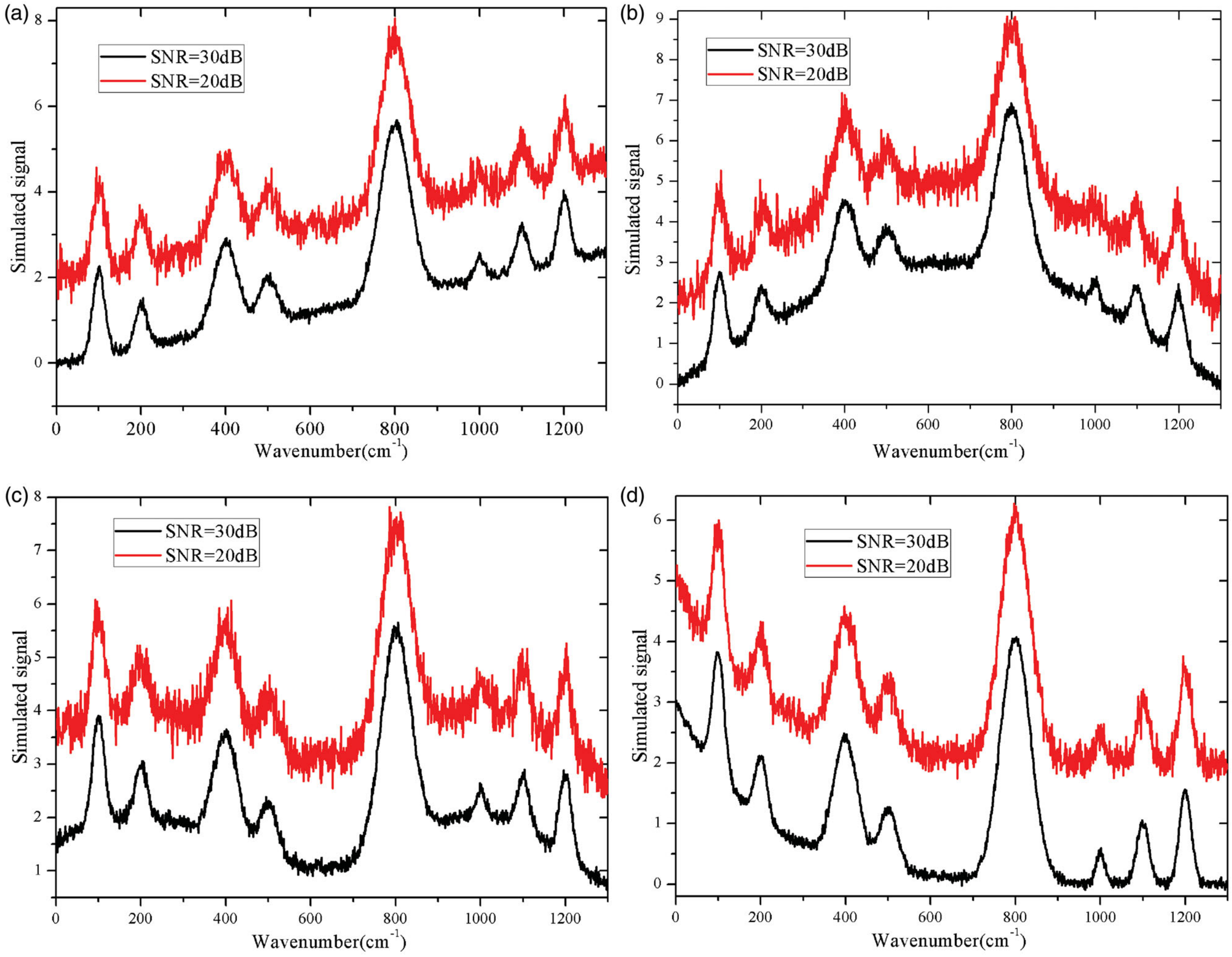

Figure 4. Simulated spectra with different types of baselines in high noise and low noise. (a) Linear baseline; (b) Sine baseline; (c) Gaussian baseline; (d) Exponential baseline. The simulated spectra are composed of analytical signal, baseline and noise. In the view of easy distinction, simulated spectra with high noise simulated spectra were moved up for 2. SNR-signal-noise ratio.

The spectrum baseline is prone to linear and non-linear drift. Therefore, we simulated four different types of baselines to verify the performance of the above baseline correction methods: (1) a linear baseline; (2) a sine baseline; (3) a Guassian baseline; (4) an exponential baseline. Four baselines can be described by the following formula:

$$\begin{cases} b_1 = -0.01 + 0.002x \\ b_2 = 3 \times \sin\left(x\dfrac{\pi}{1300}\right) \\ b_3 = 2e^{-\left(\frac{x-200}{400}\right)^2} + 2e^{-\left(\frac{x-1000}{300}\right)^2} \\ b_4 = 3e^{-x/200} \end{cases} \tag{16}$$

Adding the Gauss peak to the baseline, the spectra signal without noise is obtained. Noise exists in in absorbance spectrum inherently. In order to improve the accuracy and detection limit of instrumental technique, the baseline correction method of spectra with different noise should be considered. Here, awgn function in MATLAB is used to generate noise signals with SNR of 20 and 30, respectively. The simulated spectra with different types of baselines in high noise and low noise can be obtained, as shown in Fig. 4. In the view of easy distinction, simulated spectra with high noise were moved up for two.

As we know the real baseline which is given as Eq. (16), we can compare the performance of baseline correction methods using RMSE (root mean square error). RMSE is calculated as follows:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}\left[b(i) - \hat{b}(i)\right]^2}{N}} \tag{17}$$

where b is the real baseline, and $\hat{b}$ is the estimated baseline. To evaluate the performance of

Table 2. The RMSE value obtained by four baseline correction methods.

		Baseline type			
noise	Method	Linear	Sine	Gaussian	Exponential
Low noise	airPLS	0.0706	0.1170	0.0972	0.0713
	IAsLS	0.0246	0.0435	0.0815	0.0840
	arPLS	0.0244	0.0249	0.0380	0.0309
	asPLS	0.0119	0.0177	0.0174	0.0275
High noise	airPLS	0.3328	0.4397	0.3637	0.2563
	IAsLS	0.1772	0.2670	0.2140	0.1248
	arPLS	0.1131	0.1129	0.1256	0.0778
	asPLS	0.0290	0.0528	0.0585	0.0490

RMSE: the root mean square error; airPLS: adaptive iteratively reweighted penalized least squares; IAsLS: improved asymmetric least squares; arPLS: asymmetrically reweighted penalized least squares smoothing; asPLS: adaptive smoothness parameter penalized least squares, the proposed baseline correction method. Under the same noise, the smaller the RMSE, the better performance of the algorithm for baseline correction. The results showed that the smaller RMSE values can be obtained by the asPLS algorithm compared to the others, whether in low noise or high noise.

the proposed method, the airPLS, IAsLS and arPLS methods are used for comparison. All the above baseline correction methods have one or more parameters to be adjusted. For IAsLS method, we choose the optimum parameters according to reference.[20] For the other three methods, λ is changed from 10^2 to 10^8 as λ is recommended to vary in the log scale,[18] we can find the optimal parameter λ by linear search method. The RMSE value of four baseline correction methods under optimal parameters are listed in Table 2.

From Table 2, among the four baseline methods, airPLS method obtains the worst RMSE values. The RMSE value of IAsLS method is close to that of arPLS when the baseline type is linear in low noise. It also can be found that the RMSE values obtained by above baseline correction methods in high noise were larger than twice that obtained in low noise condition. Therefore, we believe that noise have a great impact on baseline correction. Moreover, it is obviously found that the smaller RMSE values can be obtained by the asPLS algorithm compared to the others, whether in low noise or high noise. Hence, we can get a conclusion that asPLS method obtains the better baseline estimation result compared to airPLS, IAsLS, arPLS, especially when the spectra contain high noise.

Since the RMSE values obtained by the four baseline correction methods are small in low noise environment, it is not easy to evaluate the performance of four baseline correction methods from the figure. In order to clearly distinguish the estimated baselines obtained by the above methods, only the estimated baselines in high noise environment were given, as shown in Fig. 5.

From Fig. 5, one can find that the estimated baseline by asPLS method almost overlaps the real baseline and is proved to have the best performance among the four baseline correction methods. On the contrary, the estimated baseline obtained with IAsLS and arPLS are relative worse, while the airPLS method obtains the worst result. The estimated baseline obtained by airPLS and IAsLS methods trend to pass through the lowest point of the spectrum, resulting in different degrees of under-fitting in the all regions. The arPLS method gives an overestimated baseline in peak regions, the reason is that the tail of the peak regions is considered noise and the weight value is set to one.

Experimental infrared spectra

In this paper, the type of Fourier transform spectrometer used was the Spectrum Two produced by Perkin Elmer, Waltham, United States. The optical path is 10 cm. The spectral resolution was set to 1 cm^{-1}. The spectral range was set to 400–4000 cm^{-1}. The apodization function was selected as Norton–Beer medium, and each spectrum is scanned 8 times. Thus, there are 3601 spectral lines for each spectrum. The spectra of methane and ethane are shown in Fig. 6.

It can be seen from Fig. 6 that the baseline drift becomes more serious as the concentration decreases. This is because the SNR of spectrum is calculated by signal and noise. When the concentration decreases, the pure spectrum signal becomes smaller, while the noise changes little, resulting in the reduction of the SNR. Thus, the baseline drift is easy to occur for a slight change of environmental factors.

The baseline corrected spectra of methane by four methods are shown in Fig. 7. As you see in this figure, IAsLS and airPLS methods get a boosted baseline corrected spectrum, arPLS get a declining baseline corrected spectrum in the peak regions. The baseline is well estimated by asPLS method, the corrected spectrum almost overlaps to zero in no peak regions, it can obviously be

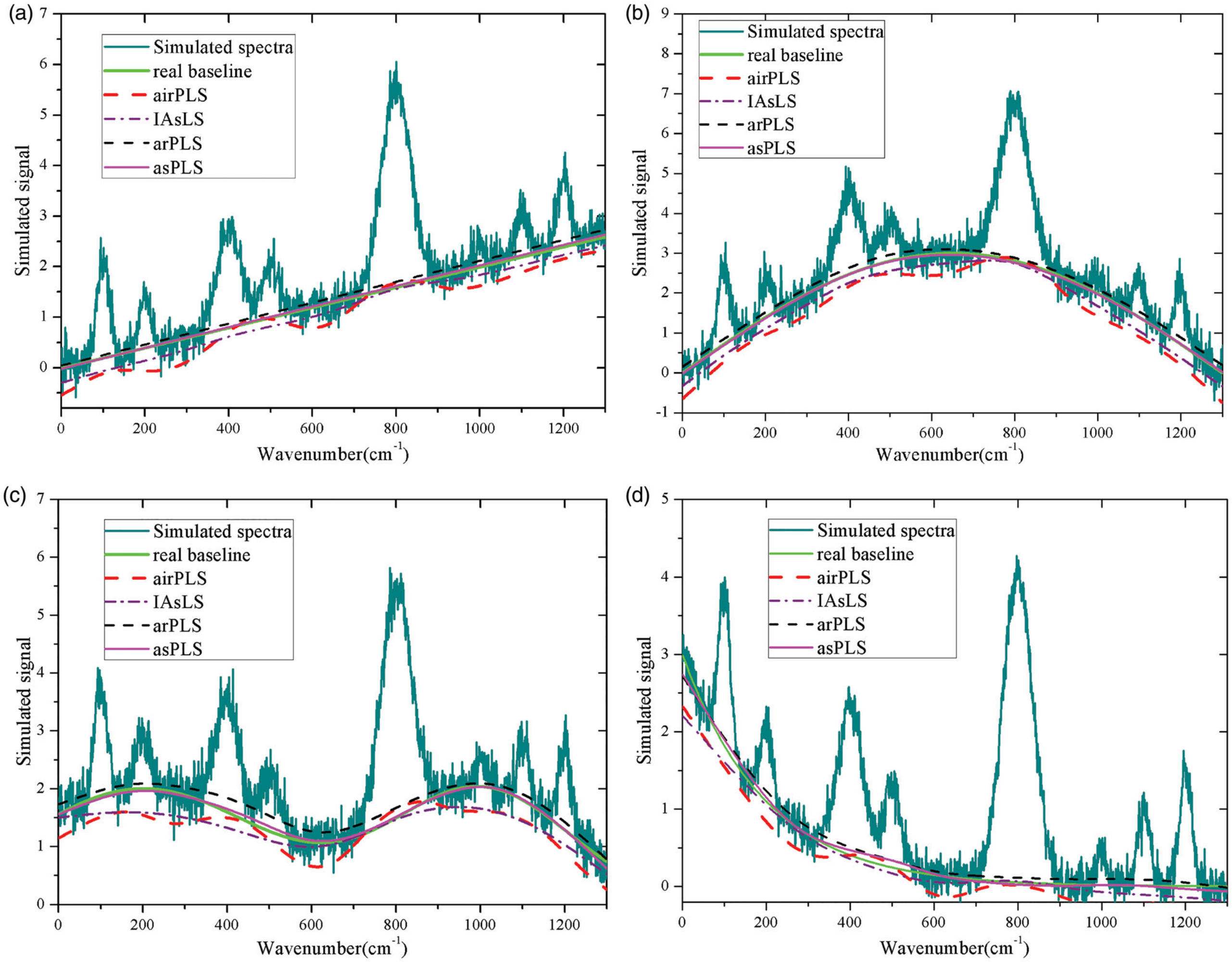

Figure 5. The estimated baselines with four baseline correction methods in high noise (SNR = 20dB). (a) Linear baseline; (b) Sine baseline; (c) Gaussian baseline; (d) Exponential baseline. It shows that the estimated baseline by asPLS method almost overlaps the real baseline and get the best performance among the four baseline correction methods, airPLS and IAsLS methods get underfitting baselines in the all regions, arPLS method give overestimated baselines in peak regions. airPLS: adaptive iteratively reweighted penalized least squares; IAsLS: improved asymmetric least squares; arPLS: asymmetrically reweighted penalized least squares smoothing; asPLS: adaptive smoothness parameter penalized least squares.

found that the baselines corrected by asPLS methods are more precise than others. When the concentration of methane is 10 ppm, the performance of above baseline correction methods has been improved. It is also proved that noise has a great influence on baseline correction.

The baseline corrected spectra of seven different concentrations of methane and ethane using the asPLS method are shown in Fig. 8. Comparing with Fig. 6, it is evident that the baseline corrected spectra successfully eliminate baseline drift. This phenomenon is more obvious when gas concentration is low, and the baseline corrected spectra look quite similar. It could be concluded that the proposed method can correct the baseline with different SNR.

In order to observe the effect of the asPLS baseline correction method more clearly, we give the relationship between concentration and absorbance at the main absorption peak of methane, as shown in Fig. 9. According to Lambert's law, the relationship between concentration and absorbance is linear at lower concentration. But from Fig. 9a, they are not shown a linear relationship, however, after baseline correction using asPLS method, they show a linear relationship. This also proves that the proposed method can be applied to infrared spectrum.

Partial least squares (PLS) can effectively reduce the dimension and eliminate the collinear relationship between variables with simple operation and high accuracy. In the paper, we use the

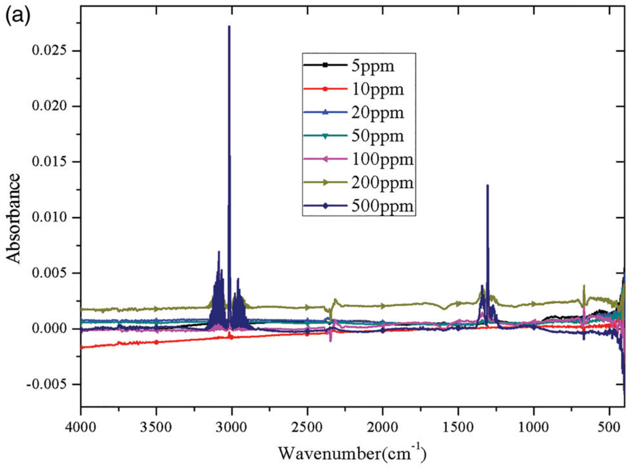

Figure 6. Original spectra of seven different concentrations before baseline correction. (a) methane; (b) ethane. The spectra range was 4000–400 cm^{-1}, resolution set 1 cm^{-1}, scanned by Spectrum Two Fourier transform spectrometer. It can be found that baseline drift becomes more serious as the concentration decreases. ppm: parts per million.

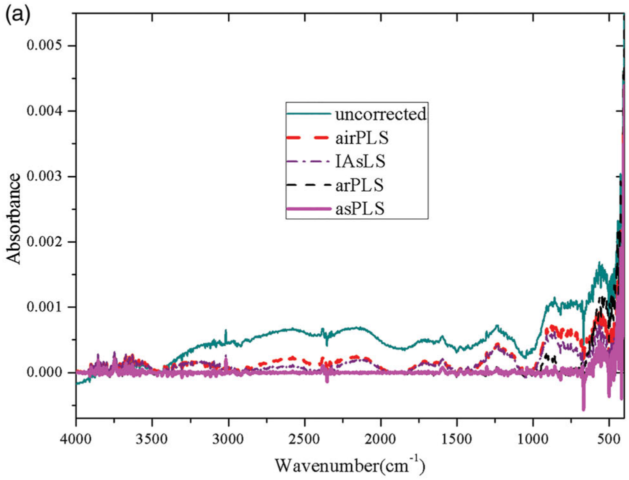

Figure 7. Baseline corrected spectra of methane by four methods. (a) the concentration of methane is 5 ppm; (b) the concentration of methane is 10 ppm. It can obviously find that the baseline corrected by asPLS methods are more precise than others, especially when the concentration is 5 ppm. airPLS: adaptive iteratively reweighted penalized least squares; IAsLS: improved asymmetric least squares; arPLS: asymmetrically reweighted penalized least squares smoothing; asPLS: adaptive smoothness parameter penalized least squares. ppm: parts per million.

methane and ethane corrected spectra by the above four methods to build PLS models, the performance of the four methods are evaluated by correlation coefficient of cross validation (Q^2) and Root mean square error of cross validation (RMSECV). The Q^2 and RMSECV were listed in Table 3. It is apparent that the values of Q^2 and RMSECV obtained by the proposed method were evidently better than airPLS, IAsLS and arPLS. Compared with the PLS model without baseline correction, the performance of the PLS model after baseline correction is improved, which shows that baseline correction can improve the performance of the model's accuracy.

Finally, it is worth mentioning that RMSE cannot be applied to select the best parameters because we do not know the real baseline. For asPLS used in this paper, λ is set to 10^7. As you see in Figs. 8 and 9, the drifting baselines are successfully corrected by the asPLS method, we can get a conclusion that asPLS is robust to λ.

Conclusion

In this work, the asPLS baseline correction method based on penalized least squares was proposed and used to handle diverse baseline types under different noise. This method updates the

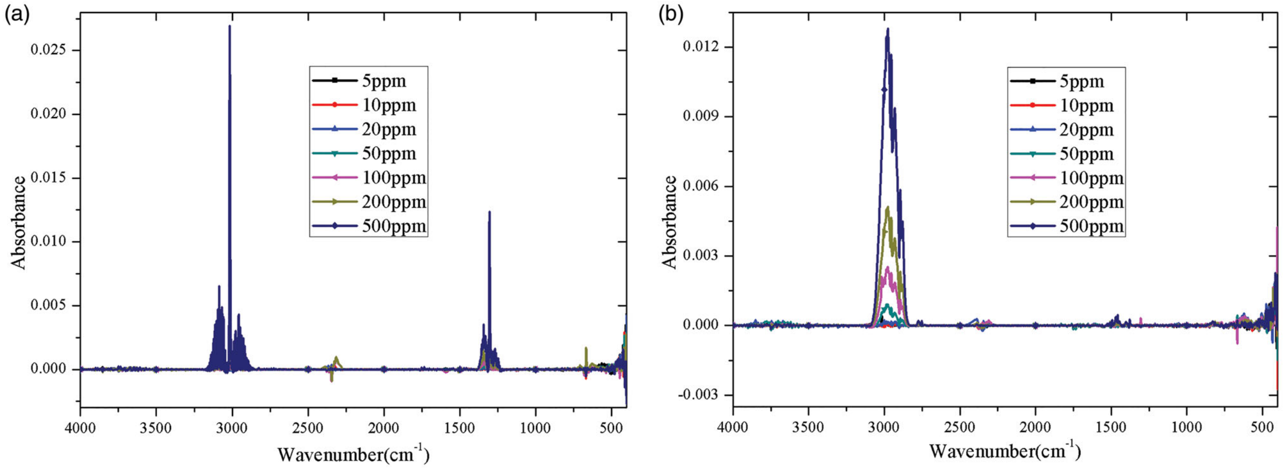

Figure 8. Baseline corrected spectra of seven different concentrations were obtained by the proposed baseline correction method. (a) methane; (b) ethane. In the graphs, the baseline corrected spectra look quite similar, and the baseline obtained by proposed method were almost equal to zero, it is shown that the proposed baseline correction method successfully eliminate baseline drift. ppm: parts per million.

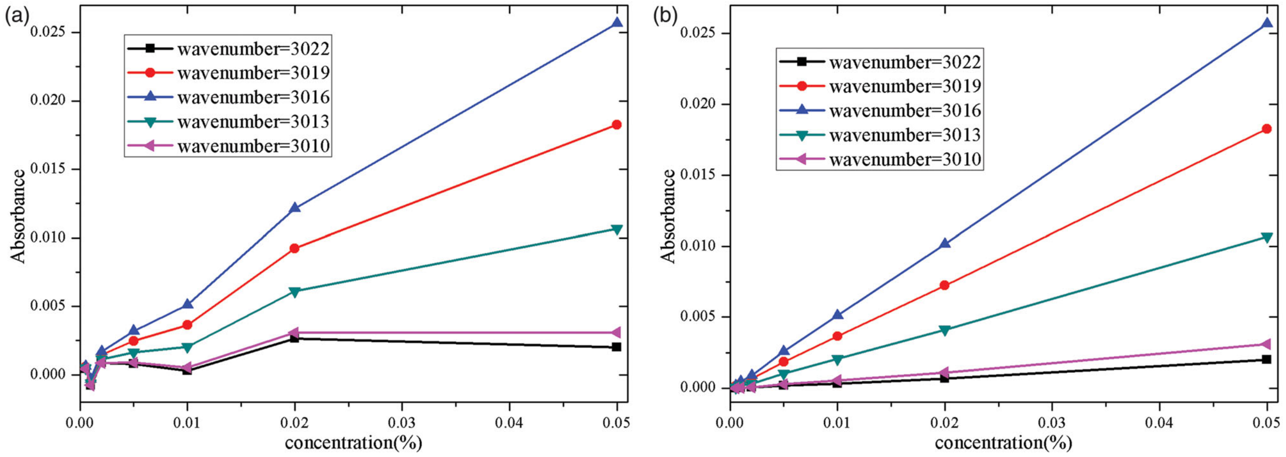

Figure 9. The relationship between concentration and absorbance at the main absorption peak of methane. (a) methane spectra before baseline corrected; (b) methane spectra after baseline corrected. Form Fig. 9a, concentration and absorbance do not show a linear relationship, however, form Fig. 9b, after baseline correction using proposed method, they show a linear relationship. It is proved that the proposed method can be applied to infrared spectrum.

smoothness parameter λ adaptively according to the difference between the original signal and the fitted baseline signal ($\mathbf{d}$) in the iterative process. Thus, d_i in the peak regions is larger than that in no peak regions, which ensured that we can obtain a larger λ_i in the peak regions and a smaller λ_i in the no peak regions.

The experimental results of the simulated spectra confirm that the asPLS method obtains a better performance than airPLS, IAsLS and arPLS methods in both high and low noise environments. Moreover, the asPLS method is more effective in dealing with different types of baselines. The baseline correction results of the experimental infrared spectra also demonstrate that the asPLS can handle various kinds of baseline with different SNR, which could improve the performance of quantitative and qualitative analysis models significantly. These consistent results show that the proposed baseline correction method is accurate and can handle various types of baseline drift. Therefore, we hope that asPLS method will replace the existing baseline correction methods in the future. Although this method only performed on the infrared spectra in this paper, we believe that it can also be used in Raman spectra and chromatograms.

Table 3. Comparison of prediction performances of methane and ethane.

	Methane		Ethane	
Methods	Q^2	RMSECV	Q^2	RMSECV
PLS	0.7190	0.0088	0.9863	0.0019
airPLS	0.9404	0.0040	0.9985	0.00064
IAsLS	0.9860	0.0020	0.9992	0.00046
arPLS	0.9888	0.0017	0.9993	0.00044
asPLS	0.9962	0.0010	0.9995	0.00039

Q^2: correlation coefficient of cross validation; RMSECV: mean square error of cross validation; None: without use baseline correction method; airPLS: adaptive iteratively reweighted penalized least squares; IAsLS: improved asymmetric least squares; arPLS: asymmetrically reweighted penalized least squares smoothing; asPLS: adaptive smoothness parameter penalized least squares. An excellent model should have high relative coefficients values and low mean square error of cross validation. The results confirmed that the larger Q^2 and smaller RMSECV values can be obtained by the asPLS algorithm compared to the others, whether in low noise or high noise.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

The authors are grateful for financial support by National Great Program on Scientific Instrument and Equipment Research of China [2016YFF0102805].

References

[1] Schmitt, J.; Flemming, H. FTIR-Spectroscopy in Microbial and Material Analysis. *International Biodeterioration and Biodegradation* **1998**, *41*, 1–11. DOI: 10.1016/S0964-8305(98)80002-4.

[2] He, S.; Xie, W.; Zhang, W.; Zhang, L.; Wang, Y.; Liu, X.; Liu, Y.; Du, C. Multivariate Qualitative Analysis of Banned Additives in Food Safety Using Surface Enhanced Raman Scattering Spectroscopy. *Spectrochimica Acta, Part A* **2015**, *137*, 1092–1099. DOI: 10.1016/j.saa.2014.08.134.

[3] Cai, Y.; Yang, C.; Xu, D.; Gui, W. Baseline Correction for Raman Spectra Using Penalized Spline Smoothing Based on Vector Transformation. *Analytical Methods* **2018**, *10*, 3525–3533. DOI: 10.1039/C8AY00914G.

[4] He, S.; Fang, S.; Liu, X.; Zhang, W.; Xie, W.; Zhang, H.; Wei, D.; Fu, W.; Pei, D. Investigation of a Genetic Algorithm Based Cubic Spline Smoothing for Baseline Correction of Raman Spectra. *Chemometrics and Intelligent Laboratory Systems* **2016**, *152*, 1–9. DOI: 10.1016/j.chemolab.2016.01.005.

[5] Schulze, G.; Jirasek, A.; Yu, M. M. L.; Lim, A.; Turner, R. F. B.; Blades, M. W. Investigation of Selected Baseline Removal Techniques as Candidates for Automated Implementation. *Applied Spectroscopy* **2005**, *59*, 545–574. DOI: 10.1366/0003702053945985.

[6] Xu, D.; Liu, S.; Cai, Y.; Yang, C. Baseline Correction Method Based on Doubly Reweighted Penalized Least Squares. *Applied Optics* **2019**, *58*, 3913–3920. DOI: 10.1364/AO.58.003913.

[7] Zhao, A.; Tang, X.; Li, W.; Zhang, Z.; Liu, J. The Piecewise Two Points Auto-Linear Correlated Correction Method for Fourier Transform Infrared Baseline Wander. *Spectroscopy Letters* **2015**, *48*, 274–279. DOI: 10.1080/00387010.2013.874530.

[8] Ni, Z.; Hu, C.; Feng, F. Progress and Effect of Spectral Data Pretreatment in NIR Analytical Technique. *Chinese Journal of Pharmaceutical Analysis* **2008**, *28*, 824–829.

[9] Leger, M. N.; Ryder, A. G. Comparison of Derivative Preprocessing and Automated Polynomial Baseline Correction Method for Classification and Quantification of Narcotics in Solid Mixtures. *Applied Spectroscopy* **2006**, *60*, 182–193. DOI: 10.1366/000370206776023304.

[10] Bertinetto, C. G.; Vuorinen, T. Automatic Baseline Recognition for the Correction of Large Sets of Spectra Using Continuous Wavelet Transform and Iterative Fitting. *Applied Spectroscopy* **2014**, *68*, 155–164. DOI: 10.1366/13-07018.

[11] Qian, F.; Wu, Y.; Hao, P. A Fully Automated Algorithm of Baseline Correction Based on Wavelet Feature Points and Segment Interpolation. *Optics and Laser Technology* **2017**, *96*, 202–207. DOI: 10.1016/j.optlastec.2017.05.021.

[12] Shao, L.; Griffiths, P. R. Automatic Baseline Correction by Wavelet Transform for Quantitative Open-Path Fourier Transform Infrared Spectroscopy. *Environmental Science and Technology.* **2007**, *41*, 7054–7059. DOI: 10.1021/es062188d.

[13] Gan, F.; Ruan, G.; Mo, J. Baseline Correction by Improved Iterative Polynomial Fitting with Automatic Threshold. *Chemometrics and Intelligent Laboratory Systems* **2006**, *82*, 59–65. DOI: 10.1016/j.chemolab.2005.08.009.

[14] Lieber, C. A.; Mahadevan-Jansen, A. Automated Method for Subtraction of Fluorescence from Biological Raman Spectra. *Applied Spectroscopy.* **2003**, *57*, 1363–1367. DOI: 10.1366/000370203322554518.

[15] Lan, T.; Fang, Y.; Xiong, W.; Kong, C. Automatic Baseline Correction of Infrared Spectra. *Chinese Optics Letters* **2007**, *5*, 613–616.

[16] Prakash, B. D.; Wei, Y. C. A Fully Automated Iterative Moving Averaging (AIMA) Technique for Baseline Correction. *Analyst* **2011**, *136*, 3130–3135. DOI: 10.1039/c0an00778a.

[17] Johnsen, L. G.; Skov, T.; Houlberg, U.; Bro, R. An Automated Method for Baseline Correction, Peak Finding and Peak Grouping in Chromatographic

Data. *Analyst* **2013**, *138*, 3502–3511. DOI: 10.1039/c3an36276k.

[18] Eilers, P. H. C. A Perfect Smoother. *Analytical Chemistry*. **2003**, *75*, 3631–3636. DOI: 10.1021/ac034173t.

[19] Zhang, Z. M.; Chen, S.; Liang, Y. Z. Baseline Correction Using Adaptive Iteratively Reweighted Penalized Least Squares. *Analyst* **2010**, *135*, 1138–1146. DOI: 10.1039/b922045c.

[20] He, S.; Zhang, W.; Liu, L.; Huang, Y.; He, J.; Xie, W.; Wu, P.; Du, C. Baseline Correction for Raman Spectra Using an Improved Asymmetric Least Squares Method. *Analytical Methods* **2014**, *6*, 4402–4407. DOI: 10.1039/C4AY00068D.

[21] He, S.; Liu, X.; Zhang, W.; Xie, W.; Zhang, H.; Fu, W.; Liu, H.; Liu, X.; Xu, Y.; Yang, D.; Gao, Y. Discrimination of the Coptis Chinensis Geographic Origins with Surface Enhanced Raman Scattering Spectroscopy. *Chemometrics and Intelligent Laboratory Systems* **2015**, *146*, 472–477. DOI: 10.1016/j.chemolab.2015.07.002.

[22] Baek, S. J.; Park, A.; Ahn, Y. J.; Choo, J. Baseline Correction Using Asymmetrically Reweighted Penalized Least Squares Smoothing. *Analyst* **2015**, *140*, 250–257. DOI: 10.1039/C4AN01061B.

[23] Liland, K. H.; Almøy, T.; Mevik, B. Optimal Choice of Baseline Correction for Multivariate Calibration of Spectra. *Applied Spectroscopy* **2010**, *64*, 1007–1016. DOI: 10.1366/000370210792434350.

[24] Chen, Y.; Dai, L. EXPRESS: An Automated Baseline Correction Method Based on Iterative Morphological Operations. *Applied Spectroscopy* **2018**, *72*, 731–739. DOI: 10.1177/0003702817752371.