Baseline correction method based on doubly reweighted penalized least squares

Article *in* Applied Optics · May 2019

DOI: 10.1364/AO.58.003913

4 authors, including:

Degang Xu
Central South University
33 PUBLICATIONS **334** CITATIONS
SEE PROFILE

Yaoyi Cai
Hunan Normal University
11 PUBLICATIONS **52** CITATIONS
SEE PROFILE

Chunhua Yang
Central South University
539 PUBLICATIONS **6,985** CITATIONS
SEE PROFILE

Analytical Methods

PAPER

View Article Online
View Journal | View Issue

Cite this: *Anal. Methods*, 2018, **10**, 3525

Baseline correction for Raman spectra using penalized spline smoothing based on vector transformation†

Yaoyi Cai,[ab] Chunhua Yang,[a] Degang Xu *[a] and Weihua Gui[a]

Baseline drift always negatively affects the qualitative or quantitative analytical results of Raman spectroscopy for many types of Raman spectrometers. Several baseline correction methods have been applied for processing Raman spectra. The parameters of these methods, however, are usually set through a complex and time-consuming process. More importantly, existing methods cannot normally provide a good estimation of the dramatically changing baselines. In this paper, we propose a penalized spline smoothing method based on vector transformation (VTPspline) for baseline estimation. The proposed baseline correction method is initiated by the raw spectrum baseline which is set as the original Raman spectrum. Meanwhile a vector ***v***, the number of elements of which is equal to that of the spectral wavenumbers is randomly generated, and all elements are set as 1 to indicate that the corresponding Raman spectral wavenumbers are considered to be background points. Vector transformation is used to transform vector ***v*** into a new sequence to make random spectral wavenumbers turn into the suspected characteristic peak channels. Based on the penalized spline smoothing algorithm and iterative process, the points that belong to the spectral background region are automatically and gradually preserved. The performance tests using both simulated and experimental Raman spectral data demonstrate that the proposed method outperforms the other existing methods for baseline estimation, such as adaptive iteratively reweighted penalized least squares and improved asymmetric least squares methods. The results also indicate that the proposed VTPspline method can handle complex and severely drifting baselines well, while maintaining the original characteristic Raman features.

Received 23rd April 2018
Accepted 17th June 2018

DOI: 10.1039/c8ay00914g

rsc.li/methods

1. Introduction

Raman spectroscopy is an efficient analytical tool and it provides detailed spectroscopic fingerprint information on target molecules.[1–3] Information on the oscillation of functional groups and structure of molecules has also been applied for both the qualitative and quantitative analysis of materials. As a non-destructive and non-contact technique, it is suitable for the analysis of water-rich samples and requires no sample preparation, which supports on-line quantitative analysis and process monitoring in many areas. Currently, Raman spectroscopy has been widely applied in the food industry, chemical processes, geological prospecting, biochemistry and other fields.[4–6] However, the obtained spectrum basically consists of Raman scattering and a varying background signal, which mainly originate from fluorescence. In the measurement process of many types of Raman spectrometers, unstable baselines are always observed. The baseline effects may lead to problems by blurring useful data and have an impact on the precision of the quantitative analysis. Moreover, baseline correction is an important step for extracting true Raman signals from raw spectra. Several methods have been used to address the fluorescence background of Raman spectra. Mathematical pre-processing is widely used to eliminate the baseline drifts of Raman spectra due to its low experimental demands.

Early analysts manually fit a curve as the baseline by piecewise linear approximation.[7] However, the accuracy is, highly dependent on the experience of the analysts and piecewise linear approximation is not convenient for use in the case of a complex baseline. Therefore, attention has been drawn by scientific researchers to baseline correction for extracting the true spectral peak intensities. Since the baselines among samples are usually different, a large number of baseline correction methods have been proposed for eliminating the varying background, such as polynomial fitting, spline fitting, wavelet transform, asymmetric least squares, genetic algorithm based cubic spline smoothing and combinations of these methods.[8–14]

[a]*College of Information Science and Engineering, Central South University, Changsha 410083, China. E-mail: dgxu@csu.edu.cn*

[b]*College of Engineering and Design, Hunan Normal University, Changsha, Hunan, 410083, China*

† Electronic supplementary information (ESI) available. See DOI: 10.1039/c8ay00914g

Manual or automatic polynomial fitting algorithms are the most widely used to estimate linear and curved baselines.[15] However, a baseline with strong nonlinearity and overlapped peaks is hardly fitted well by these methods. Wavelet transform as a frequency domain transformation method is also introduced to correct the varying background.[16–18] With respect to the real spectral signal, due to the difficulty of determining the appropriate wavelet function and threshold, it is hard to obtain a satisfactory baseline estimation effect. In addition, the associated optimization is complex to implement. The spline fitting method requires the appropriate selection of the interpolation node. This operation is usually dependent on the researcher's experience and greatly affected by noise.[19,20] The methods based on the Whittaker smoother and penalized least squares are also applied for baseline correction. Eilers *et al.* first obtained an effective baseline estimator which combined a smoother with asymmetric least squares (AsLS).[21] To improve the accuracy of the AsLS baseline correction method, some new methods have been proposed, such as adaptive iteratively reweighted penalized least squares (airPLS)[22] and improved asymmetric least squares (IAsLS).[23] These methods can be easily applied and do not require prior information about peak detection. However, the performances of these baseline correction methods depend on the selection of key parameters and the selection process is decided by experts' experience and spectral types. More importantly, these methods cannot handle spectra with strong and complex fluorescence background in real Raman spectral pre-processing.

Generally, the existing baseline correction algorithms for Raman spectra involve complicated and time-consuming processes to select appropriate key parameters. There are some problems existing in the analysis of samples, such as weak signals, shot noise and strong fluorescence interference. These methods may not eliminate the strong fluorescence background and smooth shot noise well. In this respect, a new baseline correction approach, which is a penalized spline smoothing method based on vector transformation (VTPspline), is developed for Raman spectra in this research. Combining the penalized spline smoothing algorithm with a vector transformation strategy, the points that belong to the spectral background region are automatically and gradually preserved.[24] Finally, the performance and effectiveness of the proposed method is tested with both simulated and real spectra.

2. Theory

2.1 Penalized spline smoothing

In this paper, the baseline for Raman spectra is modelled by a penalized spline smoothing method, where B-splines basis functions and difference penalties are combined to achieve tune smoothness. A set of points $\{x_i, y_i\}_{i=1}^{n}$ that belong to the background in the non-Raman characteristic peak region are obtained, where vector x_i indicates the spectral wavenumber and vector y_i indicates the intensity of the Raman spectrum. The objective is to estimate the baseline with these points. Usually, the baseline is modelled in the form of

$$y_i = f(x_i) + \varepsilon_i, \tag{1}$$

where $f(\cdot)$ is an unknown smoothing function for baseline estimation and ε_i is an error vector. In the proposed penalized spline smoothing method, $f(x_i)$ is represented by a linear combination of B-splines as:

$$f(x_i) = \sum_{j=1}^{m} B_j(x_i)\beta_j, \tag{2}$$

where $B_j(x_i)$ is the value at x_i of the j-th cubic B-spline with m equidistant knots and the coefficients $\beta = (\beta_1, \ldots, \beta_m)^{\mathrm{T}}$ of the B-spline basis function determine the scaling of the function f. Considering the regression of m points belonging to the background in a set of n B-splines basis function, we can estimate β by minimizing the least squares objective function

$$S = \sum_{i=1}^{n}\left[y_i - \sum_{i=1}^{m} B_j(x_i)\beta_j\right]^2 = \|\mathbf{Y} - \mathbf{B}\boldsymbol{\beta}\|^2. \tag{3}$$

where $\mathrm{Y} = (y_1, \ldots, y_n)'$, $\beta = (\beta_1, \ldots, \beta_n)'$, and $\|\cdot\|$ is the Euclid norm.

Unfortunately, the fitted curve shows more variation than the slowly varying baseline. To avoid over-fitting and obtain a smoother baseline, the objective function (3) is extended with a penalized regression term on the differences of β:

$$\begin{aligned} S &= \sum_{i=1}^{n}\left[y_i - \sum_{i=1}^{m} B_j(x_i)\beta_j\right]^2 + \lambda\sum_{j=3}^{m}\left(\Delta^2\beta_j\right)^2 \\ &= \|\mathbf{y} - \mathbf{B}\boldsymbol{\beta}\|^2 + p\|\mathbf{D}_2\boldsymbol{\beta}\|^2, \end{aligned} \tag{4}$$

where Δ^2 is a difference operator of the order 2 and it works on coefficients β as $\Delta^2\beta_j = (\beta_{\mathrm{j}} - \beta_{\mathrm{j-1}}) - (\beta_{\mathrm{j-1}} - \beta_{\mathrm{j-2}})$, while the positive parameter λ controls the degree of smoothness. The m $\times$ n matrix $\mathbf{D}_2$ is given by

$$\mathbf{D}_2 = \begin{bmatrix} 1 & -2 & 1 & \cdots & 0 & 0 \\ 0 & 1 & -2 & 1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots \\ 0 & 0 & \cdots & 1 & -2 & 1 \end{bmatrix}. \tag{5}$$

Then, the coefficient vector $\boldsymbol{\beta}$ can be estimated using

$$\hat{\boldsymbol{\beta}} = \underset{\beta}{\operatorname{argmin}}\left\{\|\mathbf{Y} - \mathrm{B}\boldsymbol{\beta}\|^2 + \lambda\|\mathbf{D}_2\boldsymbol{\beta}\|^2\right\}. \tag{6}$$

By computing the vector of partial derivatives and equating it to zero, the optimal B-spline coefficients are operated by

$$\hat{\boldsymbol{\beta}} = (\mathbf{B}'\mathbf{B} + \lambda\mathbf{D}'_2\mathbf{D}_2)^{-1}\mathbf{B}'\mathbf{y}. \tag{7}$$

here, it should be noted that the positive parameter λ determines the smoothness of the fitted baseline. The shape of the final baseline is mostly concerned with the smoothing parameter λ and less influenced by the richness of knots. The applications with the same simulated data are smoothed by penalized spline smoothing for two values of λ and the results are illustrated in Fig. 1.

2.2 Penalized spline smoothing based on vector transformation (VTPspline)

The main procedures of the VTPspline method are outlined in this section. At every step of the proposed method, the details are described as outlined below. Meanwhile, the flowchart of the proposed method is shown in Fig. 2.

(1) Step 1: coding scheme and initialization of parameters.

The vector $\boldsymbol{v}$ with the quantity of elements that is equal to the Raman spectral wavenumbers is randomly generated with binary code. The element (basic information) in vector $\boldsymbol{v}$ is coded as either 1 or 0, with 1 indicating that the corresponding Raman spectral wavenumber is considered as a background point in the non-Raman characteristic peak region, and 0 implying that the corresponding Raman spectral wavenumber is regarded as a part of the characteristic peaks. The parameters are then initialized, such as the proportion of elements to be flipped (flip rate), maximum iteration, and smoothing parameter of penalized spline smoothing. Meanwhile, all elements in the vector $\boldsymbol{v}$ are set to 1 and the original Raman spectrum is assumed to be the initial value of the spectral residual in the first iteration process. In the VTPspline method, the flip rate is ten percent and the maximum iteration is set as 100.

(2) Step 2: penalized spline smoothing.

The estimated baseline is fitted by the penalized spline smoothing method according to the suspected background points in the vector $\boldsymbol{v}$. In addition, the penalized parameter for spline smoothing is initialized to keep the estimated baseline smooth (5 for the VTPspline baseline correction method). The number of knots in the penalized spline smoothing algorithm for baseline correction is selected as 60. In fact, the fitted baseline obtained in the first iteration is the original Raman spectrum. The reason is that all elements in the vector $\boldsymbol{v}$ on which the fit is based are 1, that is, all the corresponding spectral wavenumbers belong to the background region.

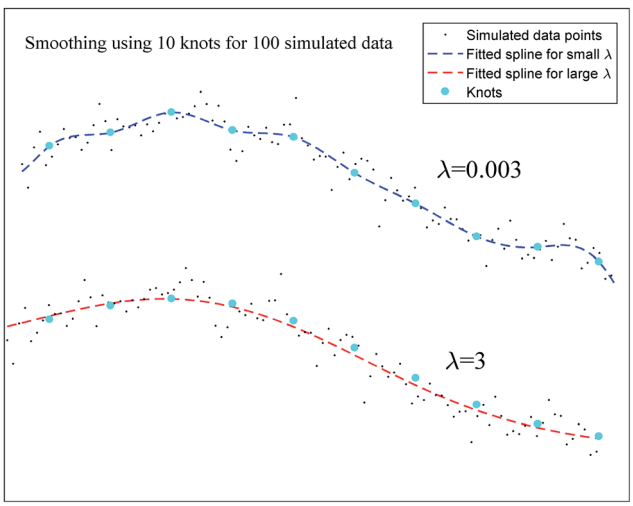

Fig. 1 Illustration of penalized spline smoothing for small λ (upper part of the figure) and large λ (lower part of the figure).

(3) Step 3: termination of the iterative process.

The root mean square error (RMSE) between the estimated curve baseline and spectral residual and the number of suspected background points are used to terminate the iterative process. The detailed course of computing is defined as follows

$$S^t = \sqrt{\sum_i (yb_i - yr_i)^2 \Big/ n} \tag{8}$$

$$m^t = \sum_i v_i{}^t \tag{9}$$

$$d^t = S^t \times m^t/(S^t + m^t) \tag{10}$$

where S^t is the RMSE between the estimated baseline **yb** and spectral residual **yr** in the t iteration step, m^t is the number of suspected background points in the t iteration step, and n is the number of Raman spectral wavenumbers. Vector $\boldsymbol{v}$ involves the binary elements used to distinguish the Raman spectral wavenumbers that belong to the background or characteristic peak region, while d^t is the value to evaluate the fitted baseline in the t iteration step.

The termination criterion is defined by:

$$|d^t - d^{t-1}| < 10^{-4},\ t \geq 2 \tag{11}$$

When the eqn (11) is met, the algorithm is terminated, and the optimal baseline can be obtained. If the condition to terminate the method or the maximum permission iteration is dissatisfied, the method will return to Step 4.

(4) Step 4: vector transformation operation.

In this step, a vector transformation operation has been used to transform vector $\boldsymbol{v}$ into a new sequence in the current iteration. We define the flip operation as a special vector transformation operation. The flip operation is used to randomly change the value of some elements in vector $\boldsymbol{v}$.

The flip operation is defined as follows:

$$v_{l_i} = 0,\ i = 1, 2, \ldots, \beta, \tag{12}$$

where

$$l_i = [\mathrm{rand} \cdot L] \ \text{ and } \ \beta = p_f \sum_i v_i. \tag{13}$$

here v_l is the l-th element in vector $\boldsymbol{v}$, and rand is used to generate a random value between 0 and 1. L is the number of all elements in the vector $\boldsymbol{v}$. The function $[\cdot]$ is used to round a number to the nearest integer, and β is the total number of elements that are treated by the flip operation. p_f is the flip rate which has been initially set as ten percent. After the flip operation, random elements in the vector $\boldsymbol{v}$ represent that the corresponding Raman channels are suspected as background points that have been set as zero. Through the flip operation, random spectral wavenumbers are transformed to the ones that are suspected as characteristic peak channels.

(5) Step 5: updating the spectral residual and vector $\boldsymbol{v}$.

This step is the most critical procedure of the VTPspline baseline correction method for obtaining the estimated

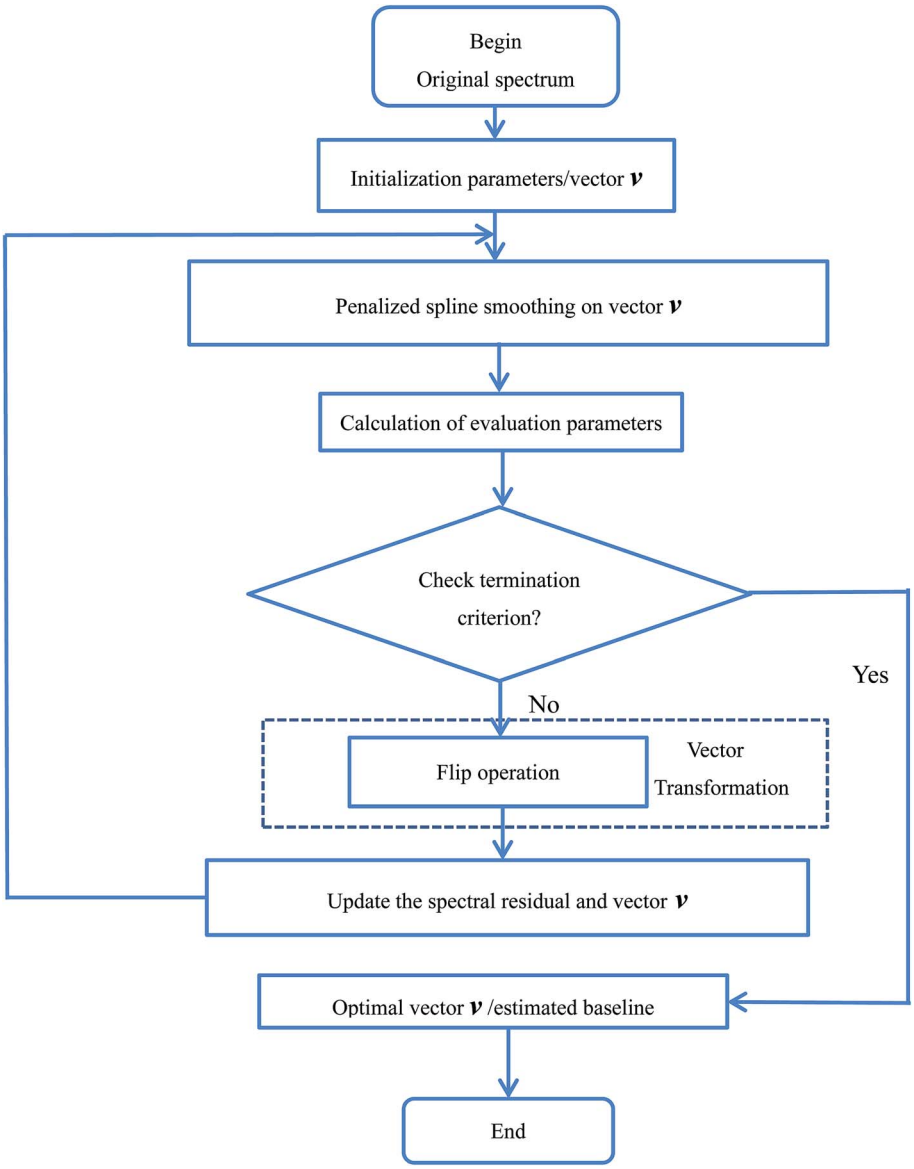

Fig. 2 The flowchart of the proposed VTPspline baseline correction method.

baseline. In this step, the estimated baseline which was obtained in Step 2, is updated by comparing it with the spectral residual. All parts of the spectral residual that are larger than the estimated baseline are found and all points on these parts are updated by the corresponding points on the estimated baseline. Meanwhile, the corresponding elements are updated with 0. Then go back to Step 2.

(6) Step 6: obtaining the estimated baseline.

Once the process is terminated, the last iterated vector is obtained, and the estimated baseline is fitted with the optimal vector $\boldsymbol{v}$. In the iteration process, the spectral residual could converge to the real baseline and the optimal vector $\boldsymbol{v}$ can reach the actual distribution of spectral wavenumbers in the background and characteristic peak regions gradually. Theoretically, the estimated baseline could cover the actual spectral background while the VTPspline baseline correction process is terminated.

2.3 Programming and implementation

All programs are implemented in Matlab 2016b on a LENOVO notebook computer with a CPU of 2.50 GHz and RAM of 8GB. The existing airPLS and IAsLS methods are programmed with the code in previous papers.[22,23]

3. Applications

In general, baseline drifts of Raman spectra will be caused by Rayleigh scattering, instrument influence, fluorescence scattering *etc.* In some samples, the fluorescence background is quite strong that it overlaps with the characteristic peaks of the Raman spectrum. Meanwhile, the baseline shifts could be different from spectrum to spectrum for the same samples. There are two typical types of baselines: slightly drifting baseline and severely drifting baseline.

In this section, simulated and real Raman spectra are tested to show the performance of the proposed baseline correction method. Simulated spectra consist of various types of curve baselines, analytical spectral signals and random noises. They can be mathematically described by the following formula:

$$y(t) = s(t) + b(t) + n(t), \quad (16)$$

where $y(t)$ denotes the resulting simulated data, which is the sum of $s(t)$ the pure spectral data, $b(t)$ the simulated baseline data and $n(t)$ the random noise data.

The performance of the proposed VTPspline baseline correction method is discussed in this subsection for the two simulated Raman spectra y_1 and y_2. Broader Gaussian peaks are treated as curved baselines and several narrower Gaussian peaks with different intensities are simulated as the Raman features. The simulated pure Raman spectral and curved baselines are given as eqn (17–19):

$$s_1(t) = 2e^{-(t-60)^2/200} + 8e^{-(t-100)^2/200} + 5e^{-(t-240)^2/200} + 3e^{-(t-350)^2/200} + 9e^{-(t-370)^2/300} + 10e^{-(t-500)^2/250} + 7e^{-(t-550)^2/250} + 4e^{-(t-600)^2/400} + 3e^{-(t-840)^2/200} + 4e^{-(t-880)^2/450}, \quad (17)$$

$$b_1(t) = 10^{-3}t + 6e^{-(t-200)^2/10^6/2} + 10e^{-(t-600)^2/10^4/12}, \quad (18)$$

$$b_2(t) = 10^{-3}t + 15e^{-(t-300)^2/10^6/2} + 30e^{-(t-800)^2/10^4/12}, \quad (19)$$

where $t = 1, 2, \ldots, 1000$. The simulated Raman spectra y_1 and y_2 are superposed by $s_1 + b_1$ and $s_1 + b_2$. As is shown in y_1 and y_2, two broader Gaussian peaks b_1 and b_2 are treated as the flatly and strongly changed curve backgrounds respectively, and narrower Gaussian peaks in s_1 are treated as the characteristic Raman peaks.

To investigate the performance of the proposed VTPspline baseline correction method, the root mean squares error (RMSE) is calculated for simulated spectra as follows:

$$\mathrm{RMSE} = \sqrt{\sum_i (ys_i - yf_i)^2 / N} \quad i = 1, 2, 3, \ldots, N, \quad (20)$$

where ys is the real baseline of the simulated spectra, yf is the estimated baseline in the last iteration, and N is the number of Raman shifts.

Fig. 3 demonstrates the process of baseline correction on the simulated spectrum y_1. As shown in Fig. 3, the initial value of the spectral residual is the original spectrum. By the vector transformation, the estimated baseline is updated according to the suspected background points in the vector $\boldsymbol{v}$. As shown in oval mark 1, when the estimated baseline is less than the spectral residual at the region of characteristic Raman peaks, the new spectral residual is replaced with the estimated baseline. This means that the corresponding elements, which actually belong to the region of characteristic Raman peaks, have been updated to 0. Meanwhile, as shown in oval mark 2, when the estimated baseline is less than the real baseline at the suspected background region, the spectral residual remains unchanged. In the VTPspline method, the vector transformation and iterative methods are used to automatically and gradually preserve the points that belong to the real background region. Then, these suspected background points are fitted with the penalized spline smoothing algorithm and the estimated baseline is obtained.

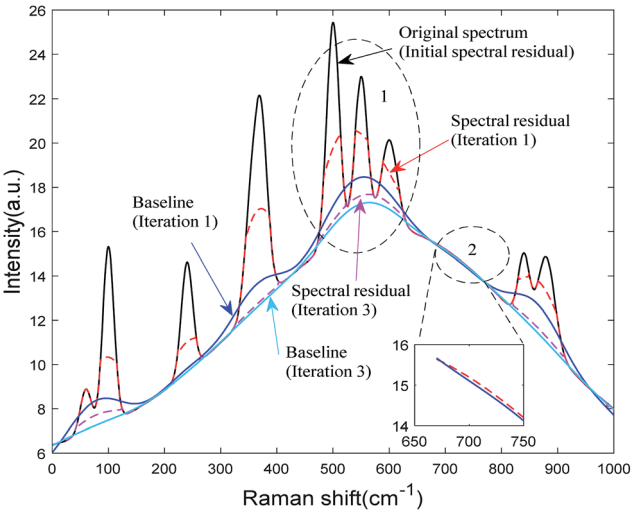

Fig. 3 The process of baseline estimation in the VTPspline method.

The simulated Raman spectra that are treated by the VTPspline method are shown in Fig. 4 with the baseline corrected spectra. The suspected backgrounds which are preserved by the VTPspline method have been denoted as cyan asterisks. Meanwhile the estimated baselines are plotted with broken red lines. It is obvious that the estimated baselines almost coincide with the real baselines, especially for the simulated spectrum with a strong background in Fig. 4(b). After the background elimination, the corrected Raman spectra are plotted with blue lines. The result shows that the drifting baselines of the simulated spectra overlapped with peaks and high fluorescence backgrounds can be automatically corrected by the VTPspline method.

The VTPspline, airPLS and IAsLS baseline correction methods are applied to the simulated spectra y_1 and y_2 for comparing their performances. The optimized parameters for airPLS and IAsLS are obtained by the minimization of the RMSEs of the baseline corrected spectra. The estimated baselines and performance indices of VTPspline, airPLS and IAsLS are shown in Fig. S1† and Table 1. It is found that the baselines of the original Raman spectra are over-fitted by the airPLS method at 50–150 cm^{-1} (baseline b_1) and 500–650 cm^{-1} (baseline b_2). Obviously, the intensities of Raman characteristic peaks are weakened in these areas. In addition, the IAsLS method did not fit the baselines well at 480–650 cm^{-1} (baseline b_1), 500–650 cm^{-1} (baseline b_1) and 700–950 cm^{-1} (baseline b_2). In contrast, based on the proposed VTPspline method, the baselines in simulated spectra y_1 and y_2 are fitted well throughout the entire Raman spectrum.

In addition, as shown in Table 1, the RMSE of the proposed method is the smallest among these three methods. The reason is that it is harder for the IAsLS and airPLS methods to achieve a balance between fitting the characteristic peaks and obtaining

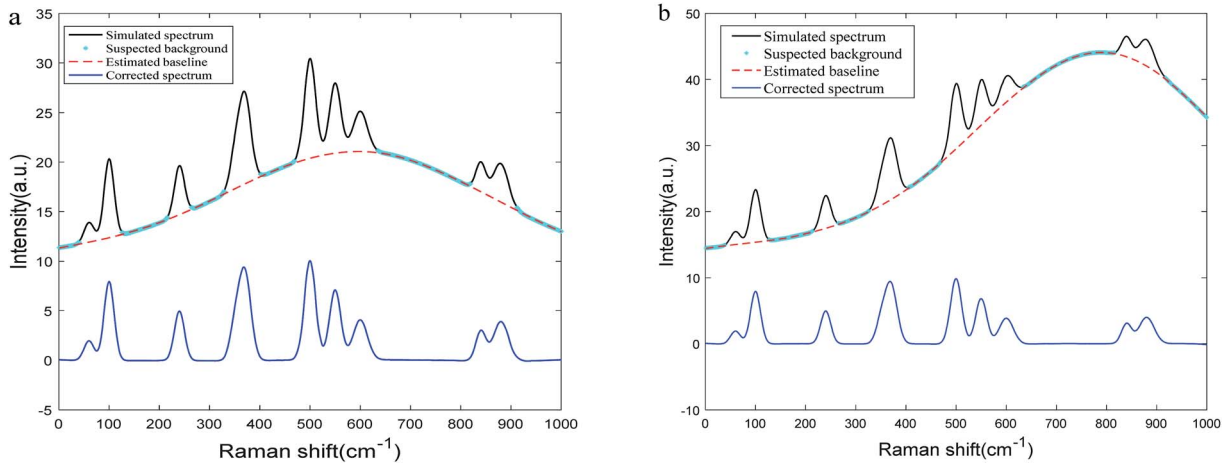

Fig. 4 The estimated baselines and corrected Raman spectra for simulated spectra y_1 (a) and y_2 (b) respectively.

smoothing baselines for the small curvature radius of curved backgrounds. The iteration time of VTPspline is longer compared with those of airPLS and IAsLS. However, it is less than one second, which can meet the requirement of online baseline correction of Raman spectra. More importantly, the RMSEs of the proposed VTPspline method are the least, only 0.037 (baseline b_1) and 0.068 (baseline b_2). Additionally, the Raman spectra with a strong fluorescence background (baseline b_2) could be better treated by the proposed VTPspline method than the IAsLS and airPLS methods. The corrected Raman characteristic peaks are not weakened by the proposed baseline correction process. Therefore, the validity of the VTPspline method is approved.

Furthermore, the proposed VTPspline baseline correction method was tested for sensitivity to random noise. Low and high Gaussian-distributed pseudorandom noise was added in the simulated spectra. The SNR (signal to noise ratio) of the low noise spectrum was set to 35 dB and that of the high noise spectrum was set to 25 dB. The SNR is measured as follows:

$$\mathrm{SNR} = 10\log_{10}(P_s/P_n), \quad (21)$$

where P_s and P_n are the measured energies of the spectrum and noise. The simulation results for evaluating the performance of the proposed VTPspline method in low and high noise spectra are shown in the ESI (Fig. S2–S5†). The simulation shows that the RMSE of the VTPspline method is much less than those of the other methods for the Raman spectra in low and high noise, which means that the baselines are more accurately estimated by the VTPspline method. It is important to note that the performance of the VTPspline method for spectrum y_2, which is composed of multiple overlapping peaks and a strong fluorescence background, is far better than those of the other two methods.

In the proposed VTPspline baseline correction method, the smoothing parameter λ and the number of equidistant knots k affect the performance of the method. Therefore, guidance for the choice of suitable values for the parameters λ and k should be given. Using the simulated spectra y_1 and y_2, the parameters λ and k are varied by a certain degree and the RMSE was calculated to obtain an optimal range of the parameters. The simulation processes of choosing parameters λ and k are also shown in the ESI (Fig. S6–S9†). The results show that, to obtain the optimal performance of the proposed VTPspline method, the smoothing parameter λ should be set in the range of 2.5–8, and the value of k could be set as one twentieth of the total wavenumbers.

Above all, the effectiveness of the proposed VTPspline method has proven it to be an excellent way to correct simulated Raman spectra with overlapping peaks, high fluorescence backgrounds and different level noises. Moreover, the Raman spectra treated by the VTPspline method are better than those treated by the airPLS and IAsLS methods.

For further investigating the performance of the proposed VTPspline baseline correction method, the Raman spectra of two materials were used for the experiments. The materials for measuring the Raman spectra were aegirine and tricyclazole. The normal Raman spectra of aegirine and tricyclazole were

Table 1 Comparison of airPLS, IAsLS and VTPspline baseline correction methods for simulated Raman spectra

	Simulated spectrum y_1			Simulated spectrum y_2		
Methods	Parameters	Time	RMSE	Parameters	Time	RMSE
airPLS	$\lambda = 10^5$, $p = 0.05$	0.007	0.389	$\lambda = 10^5$, $p = 0.05$	0.007	0.424
IAsLS	$\lambda_1 = 10^5$, $\lambda_2 = 10^{-2}$, $p = 0.05$	0.18	0.073	$\lambda_1 = 10^5$, $\lambda_2 = 10^{-2}$, $p = 0.05$	0.22	0.597
VTPspline	$P_f = 0.1$, $\lambda = 5$	0.5	0.037	$P_f = 0.1$, $\lambda = 5$	0.48	0.068

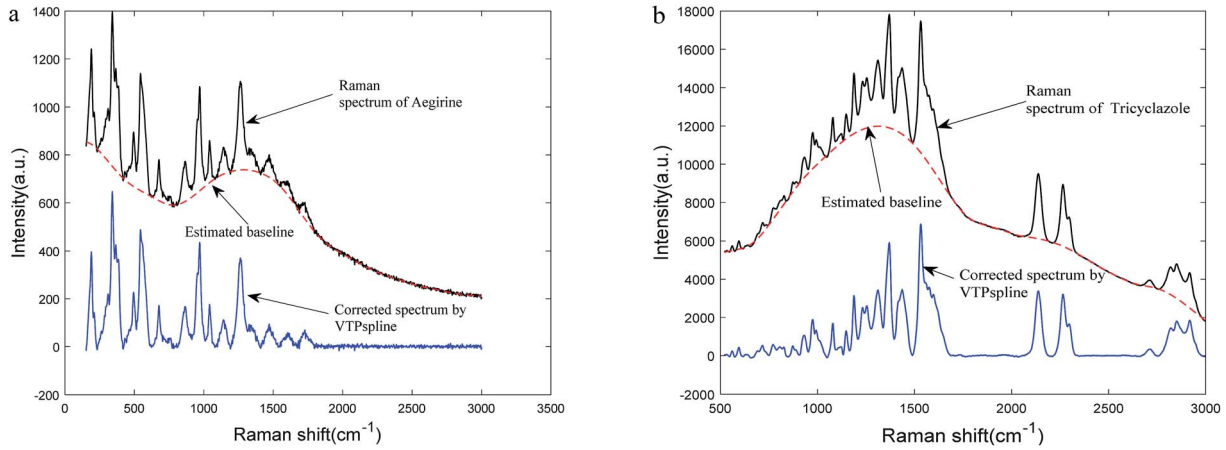

Fig. 5 Original Raman spectra and estimated baselines of aegirine and tricyclazole by the proposed VTPspline method ($\lambda = 2.5$, $k = 60$).

acquired using a Renishaw inVia-reflex micro-Raman spectrometer equipped with a 785 nm-diode laser. The spectra were collected *via* a static scan in the wavenumber region of 100–4000 cm^{-1}. The collection time for each Raman spectrum was 10 seconds, and three accumulations were used for each sample. The proposed baseline correction method was used to estimate the baselines of the experimental Raman spectra. The parameters of each method are determined by the simulation process. The results of the baseline estimations for Raman spectra are shown in Fig. 5. As shown in Fig. 5, the original

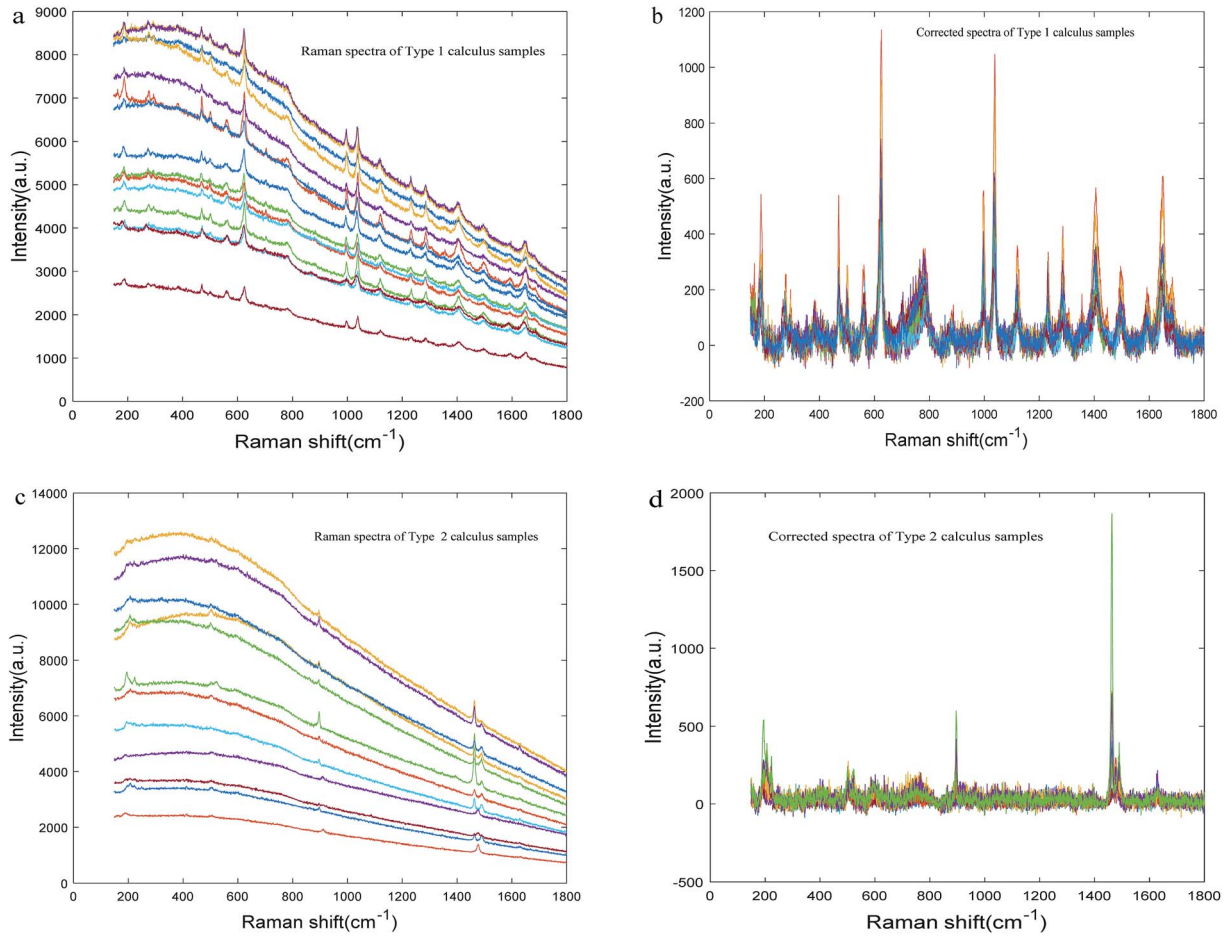

Fig. 6 Baseline correction results of Raman spectra of type 1 and 2 calculus samples. (a) and (c) are the original Raman spectra, while (b) and (d) are the corrected Raman spectra.

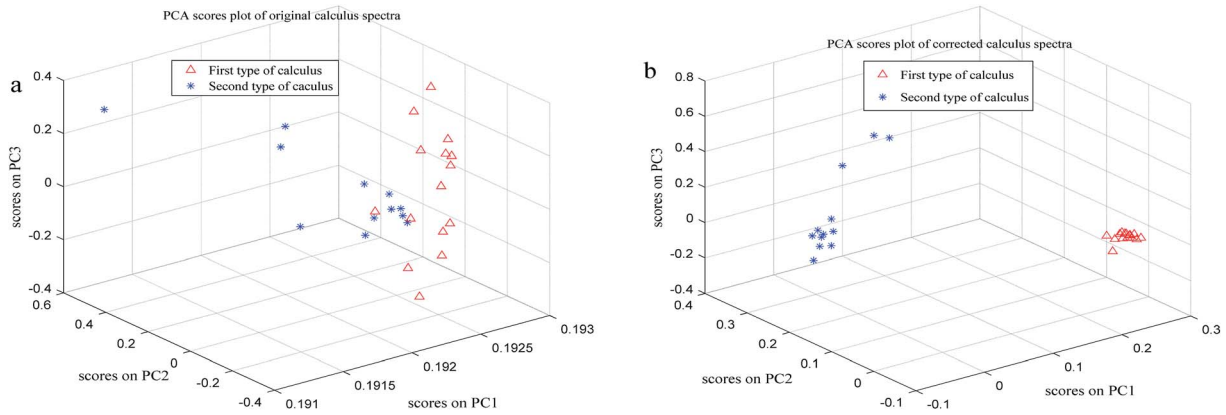

Fig. 7 Plots of the first three PCA scores of the original and corrected calculus spectra. (a) Principal components of the PCA scores of the original calculus spectra without baseline correction. (b) Principal components of the PCA scores of the corrected calculus spectra.

spectra for these actual materials exhibit multiple overlapping peaks and high backgrounds in the Raman characteristic peak regions.

The experimental results obviously show that the proposed VTPspline method can precisely estimate the baseline throughout the full wavenumber range. Then, the Raman spectra of aegirine and tricyclazole were corrected by the above three baseline correction methods, and the results are shown in Fig. S10.† In comparison with the IAsLS and airPLS baseline correction methods, the severely drifting baselines in the Raman spectra can be better estimated by the proposed VTPspline method, as shown by the black dotted ellipse areas in Fig. S10(a) and (b).†

Furthermore, the proposed VTPspline method was also used to handle the Raman spectra of complex mineral samples which were complex mixtures of hundreds of organic and inorganic constituents. The raw ore samples were mainly composed of quartz, sericite, stibnite, sphalerite, pyrite, calcite and so on. The Raman spectra of these samples were measured with an Ocean Optics confocal micro-Raman spectrometer, excited with a frequency-doubled Nd:YAG laser (785 nm, P_{max} of 300 mW). The original and corrected Raman spectra, which were treated by the proposed VTPspline method, are shown in Fig. S11 and S12.† Actually, the proposed VTPspline baseline correction method was used to eliminate the strong fluorescence background for extracting the characteristic Raman peaks of most types of minerals that were included in the raw ore samples. The corrected Raman spectra could be used to identify and determine the mineral composition in the raw ore samples.

In the field of medicine, the proposed VTPspline method was applied to Raman spectra of two types of calculus with high fluorescence backgrounds (see Fig. 6). The first type of calculus consists of uricite and calcium oxalate monohydrate. The second type of calculus consists of calcium oxalate monohydrate, calcium oxalate dihydrate and carbonate apatite. As shown in Fig. 6(b) and (d), all baselines of 27 calculus samples were successfully corrected, and the characteristic peaks were retained. The features of the Raman spectra of calculus were extracted by principal component analysis (PCA), and the classification result of the proposed VTPspline method was investigated. PCA was performed on two matrices, which consisted of the original and corrected spectra. The number of principal components was set as 3, and the classification results are shown in Fig. 7. As shown in Fig. 7(a), the first and second types of calculus samples were mixed in the principal component spaces, which means that the characteristic peaks in the Raman spectra of the two types of calculus samples are partially masked by the strong fluorescence background. Then, PCA was also performed with the corrected Raman spectra, which were treated by the proposed VTPspline method to correct severely drifting baselines. Fig. 7(b) shows the three-dimensional scatter-plots of the three principal components. It is clear that the classification results based on the corrected spectra were significantly improved, which was obviously attributed to the VTPspline method.

In the food industry, the proposed VTPspline method was applied to the Raman spectra of peanut oil for detecting and quantifying peanut oil adulteration with other types of oil by combining Raman spectroscopy and chemometrics. The spectra of 45 peanut oil samples with soybean, rapeseed and sunflower seed oils (3%, 5%, 8%, 10%, 12%, 15%, 20%, 25%, 30%, 40%, 50%, 60%, 70%, 80% and 90%) were collected using a portable Raman spectroscopy system. As shown in Table S1, Fig. S13 and S14,† the proposed VTPspline can eliminate the strong fluorescence background in the Raman spectra and improve the prediction accuracy of the calibration model for the quantitative analysis of adulterants in peanut oil.

In summary, the proposed VTPspline method was demonstrated to be an excellent way to estimate the multiple overlapping Raman characteristic peaks with slightly and severely drifting baselines.

4. Conclusions and discussion

We propose a new method for baseline estimation and validate it with both simulated and experimental Raman spectra. The proposed method provides an adaptive and effective way to estimate various drifting baselines with multiple overlapping

peaks. Meanwhile, the selection strategies for the appropriate parameters λ and k were determined through comparison of fitting effects and analysis of the baselines of simulated spectra using different parameter values. When the appropriate parameters were found, the experimental results of the simulated spectra demonstrated that the VTPspline method provided better performance for baseline estimation than the IAsLS and airPLS methods. The baseline correction results of the experimental Raman spectra also showed that the VTPspline method could handle various types of backgrounds for real Raman spectra. Compared with the popular baseline correction method, which is based on an asymmetric least squares smoothing approach, the proposed VTPspline method avoids the complex and time-consuming process of choosing the parameters to determine the amount of asymmetry. Meanwhile, based on the penalized splines smoothing approach, the proposed baseline correction method is not sensitive to noise. Based on these reasons, the proposed VTPspline method can be applied to pre-process real Raman spectra.

Matlab toolbox is designed to estimate the baseline using the proposed VTPspline method. In the future, we would provide the R-package.

Conflicts of interest

The authors declare no competing financial interest.

Acknowledgements

This work was financially supported in part by the National Natural Science Foundation of China under Grant 61473319 and 61533021, the Australia-China Science and Research Fund 2016YFE0101300, the Foundation for Innovative Research Groups of the National Natural Science Foundation of China under Grant 61621062, and the Innovation-driven Plan in Central South University.

References

1 S. Mazurek and R. Szostak, *Vib. Spectrosc.*, 2016, **83**, 1–7.
2 Z. Tan, T. T. Lou, Z. X. Huang, J. Zong, K. X. Xu, Q. F. Li and D. Chen, *J. Agric. Food Chem.*, 2017, **65**, 6274–6281.
3 M. S. Bergholt, M. B. Albro and M. M. Stevens, *Biomaterials*, 2017, **140**, 128–137.
4 X. Han, Z. X. Huang, X. D. Chen, Q. F. Li, K. X. Xu and D. Chen, *Fuel*, 2017, **207**, 146–153.
5 S. X. He, X. H. Liu, W. Zhang, W. Y. Xie, H. Zhang, W. L. Fu, H. Liu, X. L. Liu, Y. J. Xu, D. J. Yang and Y. M. Gao, *Chemom. Intell. Lab. Syst.*, 2015, **146**, 472–477.
6 T. Yaseen, D. W. Sun and J. H. Cheng, *Trends Food Sci. Technol.*, 2017, **62**, 177–189.
7 M. N. Leger and A. G. Ryder, *Appl. Spectrosc.*, 2006, **60**, 182–193.
8 F. Gan, G. Ruan and J. Mo, *Chemom. Intell. Lab. Syst.*, 2006, **82**, 59–65.
9 X. Wang, X. Fan, Y. Xu, X. Wang, H. He and Y. Zuo, *Meas. Sci. Technol.*, 2015, **26**, 1–7.
10 L. Shao and P. R. Griffiths, *Environ. Sci. Technol.*, 2007, **41**, 7054–7059.
11 J. Peng, S. Peng, A. Jiang, J. Wei, C. Li and J. Tan, *Anal. Chim. Acta*, 2010, **683**, 63–68.
12 T. Lan, Y. Fang, W. Xiong and C. Kong, *Chin. Opt. Lett.*, 2007, **5**, 613–616.
13 S. X. He, S. X. Fang, X. H. Liu, W. Zhang, W. Y. Xie, H. Zhang, D. P. Wei, W. L. Fu and D. S. Pei, *Chemom. Intell. Lab. Syst.*, 2016, **152**, 1–9.
14 S. X. He, W. Y. Xie, W. Zhang, L. Q. Zhang, Y. X. Wang, X. L. Liu, Y. L. Liu and C. L. Du, *Spectrochim. Acta, Part A*, 2015, **137**, 1092–1099.
15 M. Mecozzi, *APCBEE Proc.*, 2014, **10**, 2–6.
16 C. G. Bertinetto and T. Vuorinen, *Appl. Spectrosc.*, 2014, **68**, 155–164.
17 H. Shin, M. P. Sampat, J. M. Koomen and M. K. Markey, *OMICS: J. Integr. Biol.*, 2010, **14**, 283.
18 Ł. Górski, F. Ciepiela, M. Jakubowska and W. W. Kubiak, *Electroanalysis*, 2011, **23**, 2658–2667.
19 J. J. deRooi and P. H. C. Eilers, *Chemom. Intell. Lab. Syst.*, 2012, **117**, 56–60.
20 K. H. Liland, E.-O. Rukke, E. F. Olsen and T. Isaksson, *Chemom. Intell. Lab. Syst.*, 2011, **109**, 51–56.
21 P. H. C. Eilers and H. F. M. Boelens, *Leiden University Medical Centre Report*, 2005.
22 Z. M. Zhang, S. Chen and Y. Z. Liang, *Analyst*, 2010, **135**, 1138–1146.
23 S. X. He, W. Zhang, L. J. Liu, Y. Huang, J. M. He, W. Y. Xie, P. Wu and C. L. Du, *Anal. Methods*, 2014, **6**, 4403–4407.
24 P. H. C. Eilers and B. D. Marx, *Statistical Science*, 1996, **11**, 89–121.