PHOTONIC SENSORS / Vol. 8, No. 4, 2018: 332–340

Improved Baseline Correction Method Based on Polynomial Fitting for Raman Spectroscopy

Haibing HU, Jing BAI, Guo XIA*, Wenda ZHANG, and Yan MA

Key Laboratory of Special Display Technology of the Ministry of Education, National Engineering Laboratory of Special Display Technology, National Key Laboratory of Advanced Display Technology, Academy of Photoelectric Technology, Hefei University of Technology, Hefei 230009, China

*Corresponding author: Guo XIA E-mail: gxia@zju.edu.cn

Abstract: Raman spectrum, as a kind of scattering spectrum, has been widely used in many fields because it can characterize the special properties of materials. However, Raman signal is so weak that the noise distorts the real signals seriously. Polynomial fitting has been proved to be the most convenient and simplest method for baseline correction. It is hard to choose the order of polynomial because it may be so high that Runge phenomenon appears or so low that inaccuracy fitting happens. This paper proposes an improved approach for baseline correction, namely the piecewise polynomial fitting (PPF). The spectral data are segmented, and then the proper orders are fitted, respectively. The iterative optimization method is used to eliminate discontinuities between piecewise points. The experimental results demonstrate that this approach improves the fitting accuracy.

Keywords: Raman spectrum; piecewise polynomial fitting; baseline correction; elimination of discontinuities

Citation: Haibing HU, Jing BAI, Guo XIA, Wenda ZHANG, and Yan MA, "Improved Baseline Correction Method Based on Polynomial Fitting for Raman Spectroscopy," *Photonic Sensors*, 2018, 8(4): 332–340.

1.Introduction

Raman spectrum analysis is widely carried out in various fields, such as chemistry and biology. After analyzing the scattering spectrum of the incident light, one can get the information of molecular vibration and rotation, and obtain the molecular structure characteristics. For example, from the Raman spectroscopy, one can analyze the structure of minerals, carry out archaeological studies [1], characterize the chemical properties of cells and intracellular fine molecules [2], detect the safety of food [3], and gauge geographic information [4]. However, Raman spectrum has only about 10^{-8} of the intensity of the original excitation signal [5]. The presences of shot noise and fluorescence background noise which consist of the baseline seriously affect the analysis of the Raman spectrum. Therefore, the baseline correction is crucial to restore the real spectrum. Instrumental improvement and mathematical calculation are two methods to reduce the baseline drift. Compared with the mathematical calculation method, the cost of the instrument improvement method is relatively high, which limits its application to some extent. Therefore, a mathematical calculation method is a better one.

As the background usually varies from sample to sample, several baseline correction methods have been proposed and improved. The commonly used

Received: 15 May 2018 / Revised: 2 August 2018

DOI: 10.1007/s13320-018-0512-y
Article type: Regular

methods include polynomial fitting [6–8], wavelet transform [9–11], and the least-square method [12–14]. Each of these methods has their own advantages when they are used in certain situations. As for the polynomial fitting method, it is the most popular and widely used method due to its efficiency and simplicity. Baeket *et al.* [8] sought the peaks by inspecting the smoothed derivative of a given spectrum. After clipping out the corresponding peak regions, they estimated the background by applying a modified linear interpolation. Feng *et al.* [15] removed the baseline by an iterative fitting process which discarded points whose intensities were above a threshold and fitted the remaining points into a straight line. Zhao *et al.* [7] took the signal noise distortion and the influence of large Raman peaks on fluorescence background fitting into account. Qin *et al.* [16] used a piecewise linear fitting to correct the baseline. They roughly defined the abscissa of the target points and directly found the minimum value among three points before and after the abscissa. Then, the baseline between the target points was fitted in turn. Sun *et al.* [17] refined and improved the method of Qin. They obtained the area of extreme value (target points) by the positive and negative of the derivative and located interval to get the minimum value. He *et al.* [18] used a genetic algorithm to filter the background points and used the cubic spline method to fit the background points obtained. The baseline correction was finally realized. Liu *et al.* [19] proposed a new cost function, whose property was that the cost increased when the fitting curve moved upward and away from the real baseline, thus improving the accuracy of the baseline.

In this paper, we present a piecewise polynomial fitting (PPF) method for the baseline correction. The paper is organized as follows. In Section 2, the theoretical model of PPF is presented. We obtain the segmentation points based on the right boundary of the Raman peak by deriving the spectral data. The appropriate points and order of each section are selected to perform polynomial fitting. At the same time, the number of data involved in the fitting may change with the process of iterative optimization in order to eliminate the discontinuity at the segmentation points. Section 3 presents the simulated data to illustrate the performances of the methods. We intuitively show the whole process of segmentation point selection and the discontinuity elimination. By applications to real Raman spectra, we demonstrate the validity of our algorithm in Section 4. Finally, some conclusions are drawn in Section 5.

2. Methods

Defining a model of the spectrum will allow us to better access the useful information. In the process of baseline processing, the intensities of the Raman spectra of N points are expressed as $y=(y_1,\cdots,y_N)$ and $y[x]=s[x]+b[x]$, where N is the number of measured spectral data, $y[x]$ is the measured spectrum, $s[x]$ is the analytical spectrum, and $b[x]$ is the background spectrum.

The background is modeled as a piecewise polynomial function, which can be written as

$$b=\begin{cases} b_1+\alpha_1 x+\alpha_2 x^2+\cdots+\alpha_{p_1}x^{p_1}, & x_1\le x\le \mathrm{seg}_1 \\ b_2+\beta_1 x+\beta_2 x^2+\cdots+\beta_{p_2}x^{p_2}, & \mathrm{seg}_1< x\le \mathrm{seg}_2 \\ \vdots \quad , & \vdots \\ b_n+\gamma_1 x+\gamma_2 x^2+\cdots+\gamma_{p_n}x^{p_n}, & \mathrm{seg}_n< x\le x_n \end{cases} \tag{1}$$

where x_1 and x_n are the Raman shifts corresponding to the first point and the last point, seg_1 to seg_n are the Raman shifts of segmentation points, and p_1 to p_n represent the highest order of polynomial in each fitting, respectively.

2.1 Selection of segmentation points

In order to conduct piecewise polynomial fitting of the baseline, the fitting points involved should contain as much background information as possible, and an isolated peak is not allowed for baseline fitting. Here, we choose the derivative method to

detect the peaks. For the discrete data, the derivative is approximated with the difference as follows:

$$\frac{dy(x)}{dx} \approx y[k]-y[k-1]. \quad (2)$$

The maximum value of the peak can be found with the sign change of the derivative from the positive to the negative. The boundary of the peak can be determined with adjacent zero positions of the derivative. Then, the segmentation points are determined according to the boundary of the peaks. If the spectrum is more complicated, you can manually select the segmentation point according to the actual situation.

Compared with the characteristic peaks in a Raman spectrum, the variation of the baseline is moderate. Therefore, the change in the derivative value is relatively small where there are no characteristic peaks. Then, the derivative's difference of each peak's right n (n increases from one in increment of one) points is calculated. If the difference is less than a given threshold, it is assumed that the right side of the peak is the background information. Then, the point of the right boundary is determined as a segmentation point. The flow chart for selection of segmentation points is shown in Fig. 1.

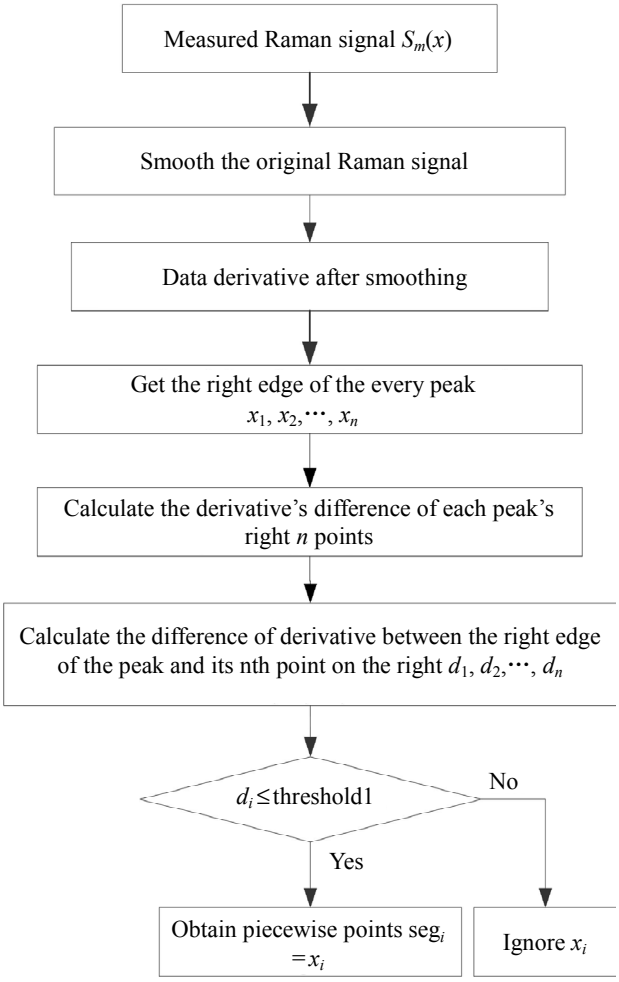


Fig. 1 Flow chart for selection of segmentation points.

2.2 Polynomial fitting method in each segment

For each section of the background, the curve is fitted with Zhao's method (I-ModPoly) [7], which takes the noise effect into account and avoids the mismatch peaks caused by noise.

Firstly, the polynomial fitting $P(x)$ of the original spectral $O(x)$ in the segmented region is performed. Then, the residual $R(x)$ and standard deviation DEV of the fitting data and the original data are calculated according to (3) and (4), respectively.

$$R(x) = O(x) - P(x) \quad (3)$$

$$DEV = \sqrt{\frac{(R_1-\bar{R})^2 + (R_2-\bar{R})^2 + \cdots + (R_n-\bar{R})^2}{n}}. \quad (4)$$

If it is the first iteration and the original spectral value is greater than the sum of the polynomial fitting value and the standard deviation, it proves that the peak exists, and the peak removal is needed. After removing the peak and re-fitting the spectral data, we can obtain the processed spectrum. Then, we calculate the standard deviation of the processed spectrum and add it to each point of the processed spectrum. The sum is compared with the contrastive spectrum (the contrastive spectrum is the original spectrum initially). If the value is larger than that of the contrastive spectrum, the contrastive spectrum doesn't need to be changed. Otherwise, the contrastive spectrum is changed into the sum. After several iterations, we can fit out the spectral background. If $\left|(DEV_i - DEV_{i-1}) / DEV_i\right| < 0.05$, which indicates that additional iterations

cannot significantly improve the fitting, the iteration stops.

2.3 Elimination of discontinuities at the segmentation points

Because the orders and fitting data involved in each piecewise polynomial fitting section are different, the polynomial curves are not the same. Discontinuities occur at the endpoints (i.e., segment points) of the polynomial curve, which may introduce additional peaks or increase the background of some areas. We overcome the discontinuities by the following iterative optimization methods:

(1) Fit the baseline by selecting the spectral data between two initially selected segment points.

(2) Calculate the difference between two adjacent fitting curves at the segmented points in the corresponding spectral shift according to (5) as

$$\mathrm{dif}_i = \left| P_{i+1}(\ \mathrm{seg\_1}_i) - P_i(\mathrm{seg}_i) \right| \tag{5}$$

where $P_i(x)$ represents the ith polynomial, and $\mathrm{seg\_1}_i$ is the next Raman shift of the segmentation point seg_i.

(3) Set an acceptable value dif_0 as the threshold. If $\mathrm{dif}_i \le \mathrm{dif}_0$, the result is the final fitting value. If $\mathrm{dif}_i > \mathrm{dif}_0$, extend m points on basis of the left and right segmented points on both sides at each iteration. Then, perform piecewise polynomial fitting again.

(4) Repeat Steps 2 and 3 until the background is obtained.

Through the iterative optimization method, the discontinuity of segmentation points is eliminated.

After fitting the background polynomial curve, the analytical spectrum $s[x]$ is obtained by subtracting fitting background $b[x]$ from the original spectral $y[x]$. Piecewise polynomial fitting can solve the problem that one polynomial curve cannot effectively fit the background, and finally, the required analytical spectra are obtained. A detailed diagram of the PPF algorithm is shown in Fig. 2.

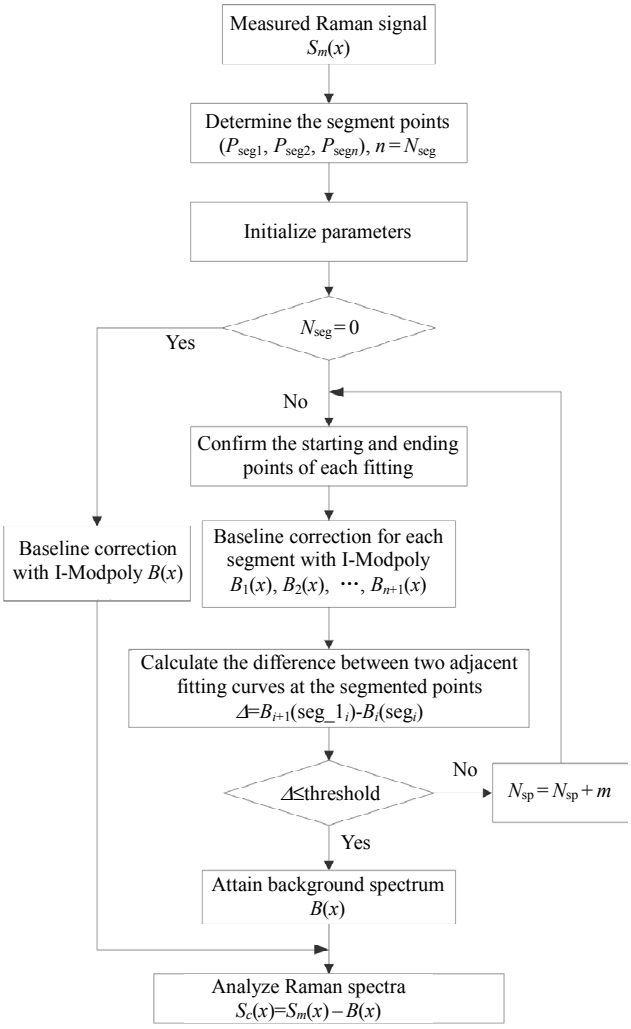


Fig. 2 A detailed diagram of the PPF algorithm.

2.4 Noise analysis

Noise usually occurs in the real spectrum. So, the threshold dif_0 is determined by noise to ensure that the difference between the values of the two adjacent curves at the segment points is within the noise range. The measured spectrum noise can be classified as four types which are independent from each other, i.e. readout noise, photoelectric noise, dark noise, and fixed pattern noise. The total noise can be described as

$$N^2_{e^-} = R^2_{e^-} + N^2_{Se^-} + N^2_{De^-} + N^2_{FPNe^-} \tag{6}$$

where $N^2_{e^-}$ is the total noise, $R^2_{e^-}$ is the readout noise, $N^2_{De^-}$ is the dark noise, $N^2_{Se^-}$ is the photoelectron noise, and $N^2_{FPNe^-}$ is the fixed pattern noise.

3. Simulation

We simulate a Raman spectrum whose signals contain a series of Lorentzian peaks on a null baseline with appropriate locations, heights, and widths.

$$y_R = \sum_{i=1}^{N} \frac{2A_{i0}}{\pi} \frac{w_{0i}}{4(r - r_{0i})^2 + {w_{0i}}^2} \tag{7}$$

where w_{0i} is the bandwidth of the peak at the full width at half-maximum (FWHM), r_{0i} is the position of the peak, and A_{i0} is the total area under the curve from the baseline. The Raman spectrum is composed of 7 peaks with intensities and positions. The parameters used in the simulation are listed in Table 1.

Table 1 Parameters of Lorentzian function used for the synthetic Raman spectrum.

Number	r_0	w_0	A_0
1	200	6	40
2	230	6	80
3	380	20	160
4	610	20	180
5	675	10	110
6	710	8	20
7	850	20	160

For the modeling of the Raman spectrum, the simulated spectrum is produced by mixing the baseline with several simulated Lorentzian peaks, which is shown in Fig. 3.

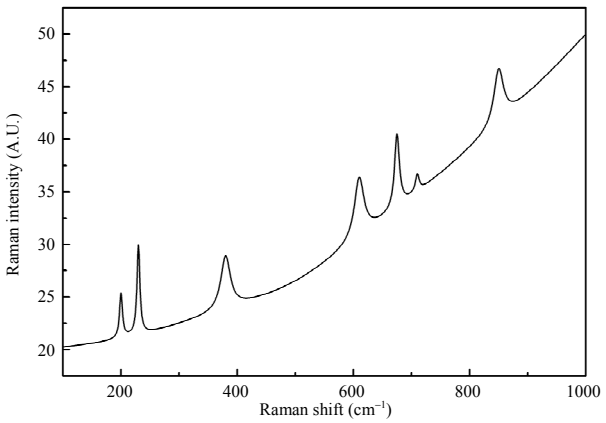


Fig. 3 Mathematical examples of simulated Raman spectrum superimposed on the baseline.

We can obtain segmentation points by the following steps. Start from the right boundary of the peak and find the *n*th point as "S". The value of *n* is determined by the specific spectrum, which we choose is 40. If the difference of derivative between the point S and the right boundary of the peak is less than a certain threshold, the point S is not on the characteristic peaks, and its intensity is only the background value, which can be used as segment points. Otherwise, the point S is on the next characteristic peak.

After calculating the derivative of the synthesized Raman spectrum as shown in Fig. 3, the derivatives are shown in Fig. 4. After a short calculation, the right boundary of 7 peaks can be obtained, which are denoted as A – G. Points A, D, and E cannot be used as segmentation points, because they are on the next characteristic peak. Besides, the last point G cannot be used as a segmentation point as well because there is no peak behind the last point.

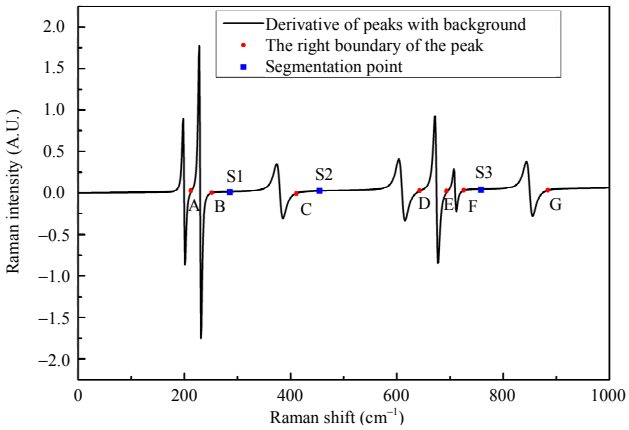


Fig. 4 Derivative of the synthetic Raman spectrum.

Respectively, starting from the points B, C, and F, the next 40th points are the segmentation points, recorded as S1, S2, and S3. Spectral data are successfully segmented.

The threshold at selection of segmentation points is determined by the derivative of the spectrum. The dash dot shadow area in Fig. 5 is enlarged, and the

effect diagram is shown in the lower right corner. The threshold is set between two dash lines, which needs to be set manually according to different spectra.

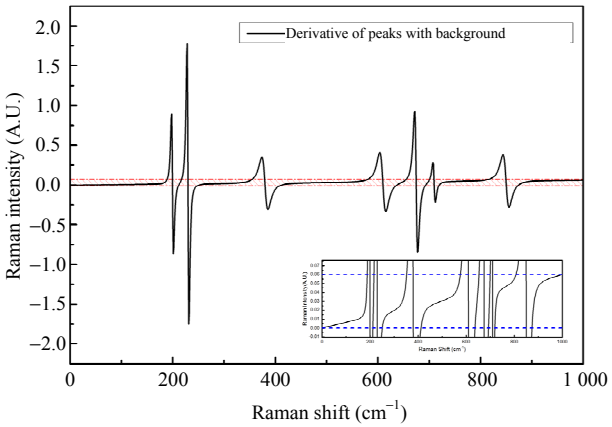


Fig. 5 Derivative of the synthetic Raman spectrum.

In the fitting of each segment, we select the data between the two segments point merely. The discontinuities at the segmentation points are caused, which is shown in Fig. 6(a). Then, the iterative optimization method is used, and the final analytical spectrum is shown in Fig. 6(b), whose dif_0 is set to a minimum of 0.005 because there is no noise.

Then, we add noise to the simulated Raman spectrum. According to Huang's theory [21], the readout noise and dark noise both belong to the white noise whose mathematic models are a simple Gaussian function, and the value of photoelectron noise is proportional to the root-mean-square of the signal intensity. Figure 7 shows that under noisy condition, the PPF method can still fit the background well. By comparison, we find the accuracy of PPF is higher than that of the method of I-ModPoly and piecewise linear fitting (PLF).

All programs are written using Matlab R2016b and run under Windows 7 on a personal computer (RAM 8G, CPU 3.20 GHz). The I-ModPoly and PLF programs are written with the description of previous reports [7, 17].

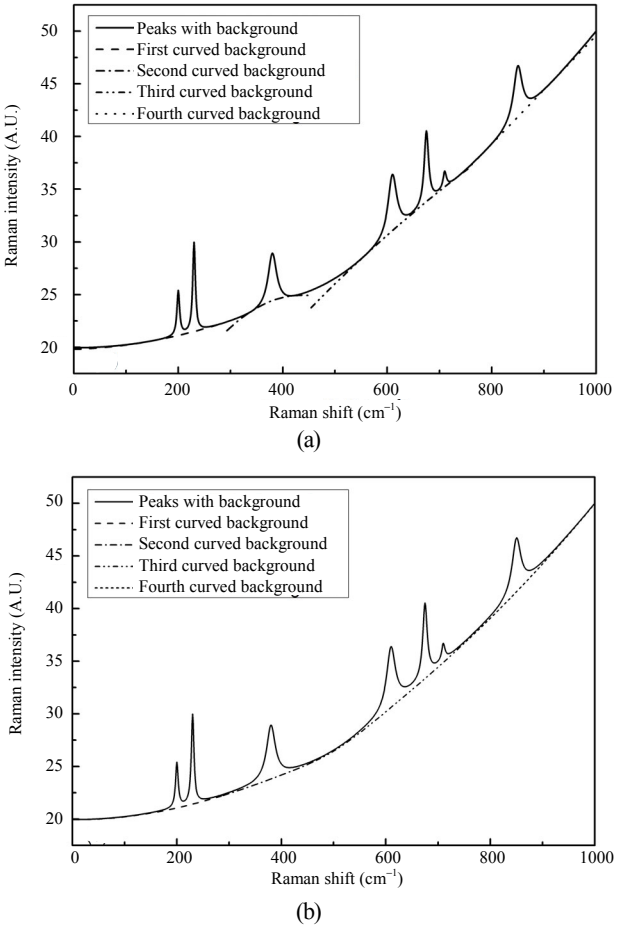


Fig. 6 Baseline correction for synthetic Raman spectrum: (a) final fitted background with discontinuity and (b) final fitted background without discontinuity.

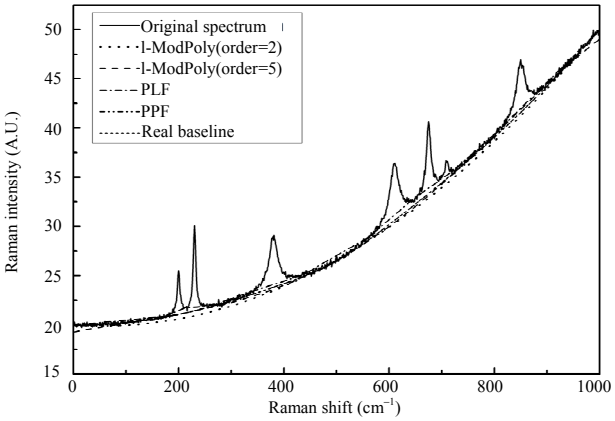


Fig. 7 Baseline correction for simulated Raman spectra with noise.

4. Experimental results and discussion

In order to demonstrate the utility of the proposed baseline correction algorithm, we obtain

real Raman spectral data on two minerals from the database about RRUFF™ Project [22]; the minerals contain different impurities and therefore different baselines. The minerals used are magnetite (R080025) and topaz (R060024).

Figure 8 shows the baseline correction for magnetite (780 nm), whose baseline has multiple curvatures. Figure 8(a) shows the original spectra and the background, and Fig. 8(b) shows the final pure Raman spectra fitted by three different methods with the appropriate order. For better comparison, we conduct the fifth-order fitting and the seventh-order fitting for the I-ModPoly method. For the fifth-order polynomial fitting, the spectra within the range of 150 cm^{-1} – 950 cm^{-1} are fitted well. But the fitting effect is poor in the range of 950 cm^{-1} – 1300 cm^{-1}. For the seventh-order polynomial fitting, the background curve at the beginning of the spectrum will be greatly bent, resulting in additional peaks, and the intensity of the peaks are also reduced. The background can be well fitted at the end region of the spectrum. Besides, other orders will lead to a worse fitting. Meanwhile, we also carry out the PLF. The fitting effect is better than that of the I-ModPoly. But if the width of the peak is large and the background below is a curve, this method is not very good. However, for the PPF method, the different orders can be used in different areas because of the segmentation spectrum data. The overall background fitting effect is perfect.

Figure 9 shows the baseline correction for topaz (532 nm), whose baseline curvature varies greatly. Figure 9(a) shows the original spectra and the background, and Fig. 9(b) shows the final pure Raman spectra fitted by three different methods with the appropriate order. From the original spectrum, we can find that the fitting order of background spectra within the range of 210 cm^{-1} – 450 cm^{-1} is different from other areas obviously. Therefore, if the third-order polynomial fitting is selected, the background spectra within the range of 210 cm^{-1} – 450 cm^{-1} cannot be fitted. If the ninth-order polynomial is used, the other background areas will have oscillation. Moreover, because PLF uses the derivation to find the extreme and selects the neighborhood minimum, leading to the problem of overlay peaks within the range of 250 cm^{-1} – 310 cm^{-1}. The PPF method uses the fifth-order in the range of 210 cm^{-1} – 450 cm^{-1} and the lower order in other parts. The method can overcome the existing problems and achieve good results.

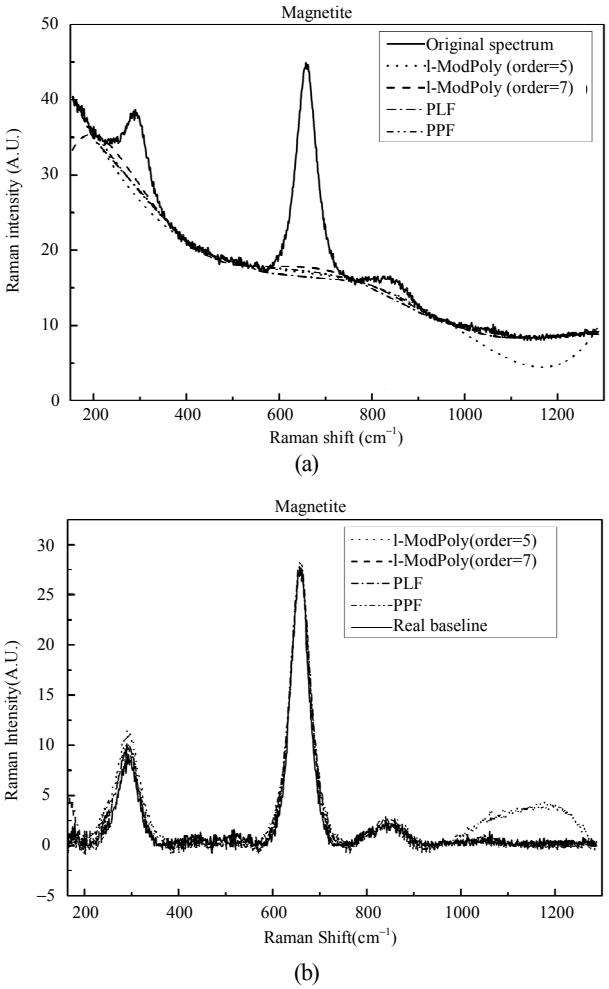
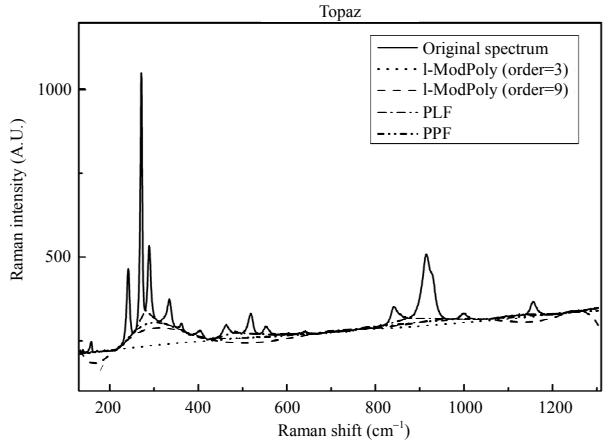
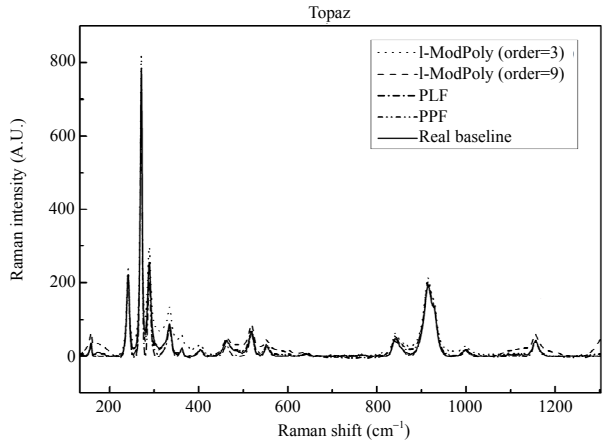


Fig. 8 Baseline correction for magnetite: (a) the original spectra and the background and (b) final pure Raman spectra fitted by the three different methods in the appropriate order.



(a)



(b)

Fig. 9 Baseline correction for topaz: (a) the original spectra and the background and (b) final pure Raman spectra fitted by the three different methods in the appropriate order.

5. Conclusions

In this paper, we propose a novel algorithm for the baseline correction of the Raman spectrum. This baseline correction method overcomes the problem that the background cannot be fitted perfectly by one polynomial curve, especially for the Raman spectrum whose curvature of background varies greatly. In the analysis and simulation, we establish a theoretical model of the PPF, put forward an idea of segmentation, and solve the discontinuity problem in the segmentation process. Compared with the experimental results of the I-ModPoly and PLF method, the results of the PPF method show a high accuracy and validity.

Acknowledgment

This work was supported by the Key Research and Development Program of Anhui Province (Grant No. 1804d08020310).

References

[1] M. Bouchard and D. C. Smith, "Catalogue of 45 reference Raman spectra of minerals concerning research in art history or archaeology, especially on corroded metals and coloured glass," *Spectrochimica Acta Part A Molecular & Biomolecular Spectroscopy*, 2003, 59(10): 2247–2266.

[2] J. Chan, S. Fore, S. W. Hogiu, and T. Huser, "Raman spectroscopy and microscopy of individual cells and cellular components," *Laser & Photonics Reviews*, 2010, 2(5): 325–349.

[3] S. X. He, X. H. Liu, W. Zhang, W. Y. Xie, H. Zhang, W. L. Fu, *et al.*, "Discrimination of the coptis chinensis, geographic origins with surface enhanced Raman scattering spectroscopy," *Chemometrics & Intelligent Laboratory Systems*, 2015, 146: 472–477.

[4] S. X. He, W. Y. Xie, W. Zhang, L. Q. Zhang, Y. X. Wang, X. L. Liu, *et al.*, "Multivariate qualitative analysis of banned additives in food safety using surface enhanced Raman scattering spectroscopy," *Spectrochimica Acta Part A Molecular & Biomolecular Spectroscopy*, 2015, 137: 1092–1099.

[5] R. M. Jarvis, A. Brooker, and R. Goodacre, "Surface-enhanced Raman spectroscopy for bacterial discrimination utilizing a scanning electron microscope with a Raman spectroscopy interface," *Analytical Chemistry*, 2004, 76(17): 5198–5202.

[6] C. A. Lieber and A Mahadevanjansen, "Automated method for subtraction of fluorescence from biological Raman spectra," *Applied Spectroscopy*, 2003, 57(11): 1363–1367.

[7] J. Zhao, H. Lui, D. I. Mclean, and H. Zeng, "Automated auto fluorescence background subtraction algorithm for biomedical Raman spectroscopy," *Applied Spectroscopy*, 2007, 61(11): 1225–1232.

[8] S. J. Baek, A. Park, J. Kim, A. Shen, and J. Hu, "A simple background elimination method for Raman spectra," *Chemometrics & Intelligent Laboratory*

Systems, 2009, 98(1): 24–30.

[9] C. G. Bertinetto and T. Vuorinen, "Automatic baseline recognition for the correction of large sets of spectra using continuous wavelet transform and iterative fitting," *Applied Spectroscopy*, 2014, 68(2): 155-1–155-11.

[10] C. M. Galloway, R. E. Le, and P. G. Etchegoin, "An iterative algorithm for background removal in spectroscopy by wavelet transforms," *Applied Spectroscopy*, 2009, 63(12): 1370–1376.

[11] Y. G. Hu, T. Jiang, A. Shen, W. Li, X. P. Wang, and J. M. Hu, "A background elimination method based on wavelet transform for Raman spectra," *Chemometrics & Intelligent Laboratory Systems*, 2007, 85(1): 94–101.

[12] S. J. Baek, A. Park, Y. J. Ahn, and J. Choo, "Baseline correction using asymmetrically reweighted penalized least squares smoothing," *Analyst*, 2015, 140(1): 250–257.

[13] Z. M. Zhang, S. Chen, and Y. Z. Liang, "Baseline correction using adaptive iteratively reweighted penalized least squares," *Analyst*, 2010, 135(5): 1138–1146.

[14] S. X. He, W. Zhang, L. J. Liu, Y. Huang, J. M. He, W. Y. Xie, *et al.*, "Baseline correction for Raman spectra using an improved asymmetric least squares method," *Analytical Methods*, 2014, 6(12): 4402–4407.

[15] X. Feng, Z. Zhu, and M. Shen, "The method of baseline drift correction of Raman spectrum based on polynomial fitting," *Computers & Applied Chemistry*, 2009, 26(6): 759–762.

[16] Z. J. Qin, Z. H. Tao, J. X. Liu, and G. W. Wang, "Baseline correction of Raman spectrum based on piecewise linear fitting," *Spectroscopy & Spectral Analysis*, 2013, 33(2): 383–386.

[17] K. Sun, H. Su, Z. X. Yao, and P. X. Huang, "Baseline correction for Raman spectra based on piecewise linear fitting," *Spectroscopy*, 2014, 29(2): 54–61.

[18] S. X. He, S. X. Fang, X. H. Liu, W. Zhang, W. Y. Xie, H. Zhang, *et al.*, "Investigation of a genetic algorithm based cubic spline smoothing for baseline correction of Raman spectra," *Chemometrics & Intelligent Laboratory Systems*, 2016, 152: 1–9.

[19] J. T. Liu, J. Y. Sun, X. Z. Huang, G. J. Li, and B. Q. Liu, "Goldindec: a novel algorithm for Raman spectrum baseline correction," *Applied Spectroscopy*, 2015, 69(7): 834–842.

[20] Y. C. Sun, C. Huang, G. Xia, S. Q. Jin, and H. B. Lu, "Accurate wavelength calibration method for compact CCD spectrometer," *Journal of the Optical Society of America A Optics Image Science & Vision*, 2017, 34(4): 498–505.

[21] C. Huang, G. Xia, S. Jin, M. Hu, S. Wu, and J. Xing, "Denoising analysis of compact CCD-based spectrometer," *Optik*, 2017, 157: 693–706.

[22] The RRUFF™ Project [database]: http://rruff.info/about/about_general.php.