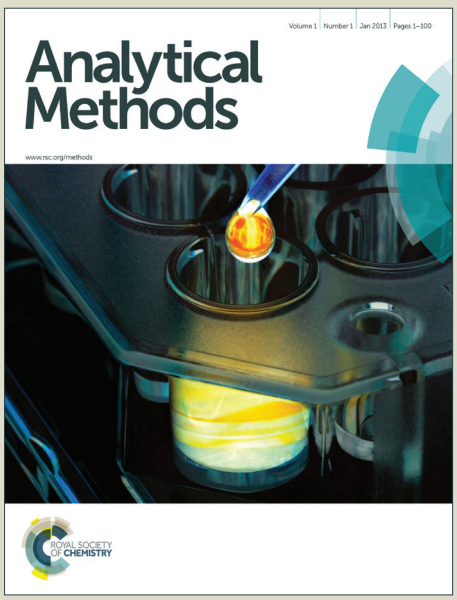
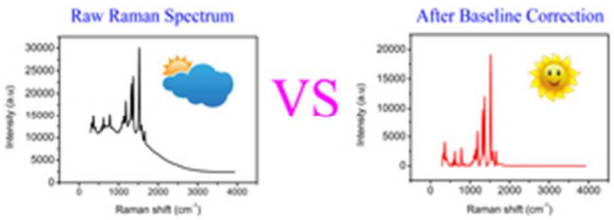
Analytical Methods

Accepted Manuscript



ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/methods



Graphical Abstract
27x9mm (300 x 300 DPI)

Baseline Correction for Raman Spectra Using Improved Asymmetric Least Squares

Shixuan He, Wei Zhang*, Lijuan Liu, Yu Huang, Jiming He, Wanyi Xie, Peng Wu and Chunlei Du

Key Laboratory of Multi-scale Manufacturing Technology, Chongqing Institute of Green and Intelligent Technology, Chinese Academy of Sciences, Chongqing, P. R. China, 400714

*Corresponding author, E-mail address: zhangwei@cigit.ac.cn

Abstract

Baseline shifts exist in many types of Raman spectrometers. Acquired spectra normally contain desired signal as well as undesirable elements such as background noise. In this paper, an improved asymmetric least squares (IAsLS) has been proposed for baseline correction of Raman spectra. The baseline correction algorithm is initiated by the raw spectrum baseline, and this baseline can be estimated by polynomial fitting method. For the simulated Raman spectra, the performance of proposed algorithm is evaluated and compared with that of asymmetric least squares (AsLS) and Jiang's method. The results showed that it is improved by sixteen fold and nine fold respectively. This proposed IAsLS method is successfully applied to practical Raman spectral data and the results in the paper indicate that the baseline of Raman spectra can be automatically subtracted.

Keywords: Baseline correction; Polynomial fitting; Improved asymmetric least squares; Raman spectroscopy

1. Introduction

Abbreviations: IAsLS, Improved asymmetric least squares; AsLS, Asymmetric least squares; RMSE, Root Mean Square Error; JAsLS, Jiang's asymmetric least squares baseline correction method

Raman spectroscopy is a rapid analytical technology and it provides detailed spectroscopic fingerprint information of target molecules. This technique is powerful and non-destructive, which requires only minimal sample preparation and it can realize the online analysis. Therefore, it is widely applied in food, materials, chemistry, biochemistry and other fields for qualitative or quantitative analytical purposes. However, in many types of Raman spectrometers, existences of unstable baselines are usually observed.[1-3] Obtained spectra often consist of desired signal as well as undesirable elements such as background noise from sample holder, instrument and sample themselves. The baseline may lead to serious problems if the data recorder reaches the detection limits during the practical operation. Moreover, the baseline correction is important since this step extracts the true Raman peak intensities, which is necessary for further numeric processing. Meanwhile, the unique strict requirements come with the different baseline correction methods.

Specifically, the baseline correction methods vary according to the precision acquired, patterns of losing and computing time obtained.[4-7] Therefore, an algorithm can be applied for the selection of a baseline correction method which is suitable for a given Raman spectra.

A large number of baseline correction methods are utilized by the research scientists, such as differencing and filtering, interpolation fitting, manual or automatic polynomial, using an asymmetric function [8, 9] and the combination of methods which are mentioned above.[10] Currently, the baseline correction methods used are semi-manual, subjective, time consuming and lack of repeatability. The linear and constant drift can be eliminated by using the differencing method. [11] However, the differencing method may amplify the high frequency noise in Raman spectrum as well. Secondly, the interpolation fitting method requires the automatic selection of the interpolation node and this is dependent on the artificial experience.[12, 13]The baseline fitting is not the same in different interpolation functions. For the case of wavelet transform, [14, 15] it is important to choose an optimal decomposition way, appropriate wavelet and the threshold values. These values can be employed to distinguish the high frequency noise, low frequency baseline and middle frequency signals. Both manual and automatic polynomial fitting method [16]require the user to manually identify the 'non-Raman' locations, determine the order of polynomial, then the baseline curve is formed by fitting these locations. The asymmetric least squares (AsLS) [17]combine a smoother with asymmetric weighting of deviations from the trend of smooth to form an effective baseline estimation method. However, the limitation of this algorithm is that only the smoothness constraint with the second derivative is considered. In practical, the method requires the baseline fitting the raw data well, and the first derivative is very close. Therefore, based on the AsLS, Jiang proposed an asymmetric least squares which considered the first derivative constraint term and the background values as the initial baseline as well (it is described in reference [18] and abbreviated as JAsLS).

An extensive literature review has been provided by Eilers and Schulze *et al.*, [19, 20]which summarized the preprocessing methods for the typical spectral backgrounds, without focusing on any particular instrumental method. Based on the advantages of

AsLS and polynomial fitting, an improved asymmetric least squares (IAsLS) is proposed in this paper. And then, this hypothesized method is employed to remove the background noise of practical Raman spectral.

2. Theory

2.1 The asymmetric least squares (AsLS)

Based on the Whittaker smoother, [20] the asymmetric least squares (AsLS) is proposed for background removal by Eilers.[19] A given vector $y = \{y_1, y_2, ..., y_i\}$ is defined as i observed frequency domain spectral intensities. And the smoothing series $z = \{z_1, z_2, ..., z_i\}$ are faithful to y. Then, the penalized least squares function is minimized:

$$F = \sum_i (y_i - z_i)^2 + \lambda \sum_i (\Delta^2 z_i)^2 \quad (1)$$

with $\Delta^2 z_i = (z_i - z_{i-1}) - (z_{i-1} - z_{i-2}) = z_i - 2z_{i-1} + z_{i-2}$, $i \in [1, 2, 3, ..., m]$. The parameter λ is introduced to tune the balance between the smoothness and fitness. Finally, a vector w is defined as weights of fitness and the minimized function is introduced as follows:

$$F = \sum_i w_i (y_i - z_i)^2 + \lambda \sum_i (\Delta^2 z_i)^2 \quad (2)$$

The minimization of equation (2) can lead to the following equations:

$$(W + \lambda D^T D) z = Wy \quad (3)$$

with $W = diag(w)$, D is second order differential matrix: $Dz = \Delta^2 z$.

Generally, a lighter smoothing is capable of removing the noise, otherwise, the stronger smoothing will eliminated true signal. In order to estimate the true background, much more attention should be paid to the deviations in positive direction for baseline correction. However, the weights of both negative and positive residuals $y - z$ are the same while using the Whittaker smoother. Therefore, a key parameter of asymmetric least squares for baseline correction, $p (0 < p < 1)$, is introduced and computed as follows: $w_i = p$ if $y_i > z_i$ and $w_i = 1 - p$ otherwise.

2.2 The improved asymmetric least squares (IAsLS)

The limitation of AsLS is that it only considers the smoothness constraint with the second derivative. In practical, on the premise of the baseline smoothness, the baseline correction method requires the baseline is well fitted, and the first derivatives for baseline data are close to each other. Therefore, the penalized least squares function of the improved asymmetric least squares proposed in this paper is minimized and shown as follows:

$$F = \sum_i (y_i - z_i)^2 + \lambda_1 \sum_i (\Delta(y_i - z_i))^2 + \lambda \sum_i (\Delta^2 z_i)^2 \quad (4)$$

where $\Delta(y_i - z_i) = (y_i - z_i) - (y_{i-1} - z_{i-1}) = y_i - y_{i-1} - (z_i - z_{i-1})$,

$\Delta^2 z_i = (z_i - z_{i-1}) - (z_{i-1} - z_{i-2}) = z_i - 2z_{i-1} + z_{i-2}$. The first two part in F measure the fit to the signal and fit to the first derivative of the signal respectively, however, the rest part is a penalty on non-smooth behavior of z. The balance between the three parts is tuned by the parameters λ_1 and λ. Finally, a vector w of weights and minimizes is introduced as follows:

$$F = \sum_i [w_i (y_i - z_i)]^2 + \lambda_1 \sum_i (\Delta(y_i - z_i))^2 + \lambda \sum_i (\Delta^2 z_i)^2 \quad (5)$$

The minimization of equation (5) can lead to the following equations:

$$(W^T W + \lambda_1 D_1^T D_1 + \lambda D^T D) z = (W^T W + \lambda_1 D_1^T D_1) y \quad (6)$$

with $W = diag(w)$, and the weight coefficient w is defined the same as AsLS.

$D_1 = \Delta(y - z)$ and $Dz = \Delta^2 z$ are first and second order differential matrix respectively.

Because of the mutual interaction of weights and smooth curve, the solution for the equations seems complicated. However, it can be transformed into two easy computations in iterative application. At the same time, the estimation values of initial baseline background for the raw spectrum are very important for the improved asymmetric least squares algorithms. Compared with background values estimated using in JAsLS method, the initial baseline background is estimated by polynomial fitting method in this paper. The flow chart of IAsLS algorithm is shown in Figure 1.

Initial baseline of the raw spectrum y is fitted by second order polynomials, say $z^{(0)}$. According to $z^{(0)}$, it is easy to compute new weights, say $w^{(0)}$. Based on these weights, a new estimate of z is obtained by solving equation (6). And then, we repeat these steps until the weights stay without changing anymore.

Eilers[19] reported that, for initial baseline estimation, the algorithm above can achieve convergence in about 5 to 10 iterations. However, it may not be able to estimate the signal background completely. Hence, the proposed algorithm is working iteratively. The residual of baseline reflected in the spectral background. Therefore, the spectrum is updated by the residual spectrum in the next iteration, and then repeats the estimation of the baseline with polynomial fitting. While the residual of baseline is almost unchanged in two iterations continuously, then the algorithm terminated. Using this method, the algorithm convergence can be obtained as the iterations always go downhill the gradient direction. And it is described by simulated spectra in details in Figure 3.

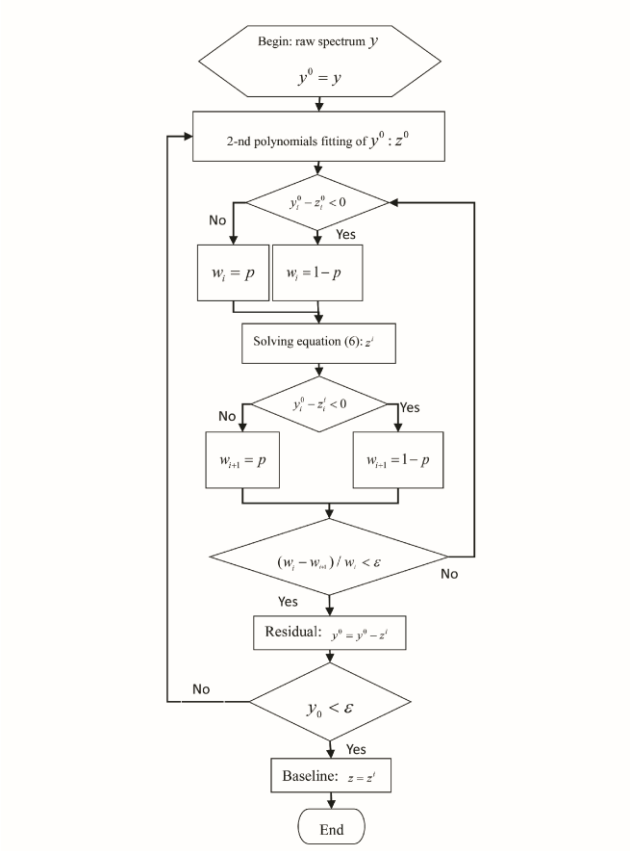


Figure 1 The flow chart of IAsLS algorithm

2.3 Programming and Software

All programs are written using Matlab R2011b and run under Windows 7 on a personal computer (RAM 4G, CPU 2.83GHz).The real Raman spectra were obtained from an Ocean Optic Peak Seeker Pro Raman spectrometer 25 equipped with a TE-cooled CCD detector, a 785 nm radiation from a He-Ne laser with a power of 300 mW for sample excitation.

3. Applications

Normally, there are many types of background origin including Rayleigh scattering, long, short Stokes shift fluorescence, sample holder and instrument effects in Raman spectra. The fluorescence effects exist as three smooth features which are as to a function of Raman shifts. These features include offset, linear baselines and exponential curves**.** Sometimes, the fluorescence background is so strong that it overlaps with the peaks of Raman. Furthermore, removing the fluorescence background with high intensity will lead to the lower signal-to-noise ratio, because the Raman peaks could probably be weakened as well. The baseline shifts could be different from spectrum to spectrum, even for the same samples. Therefore, inconsistent baselines may lead to increasing of complexity and decreasing of applicability of mathematical model for qualitative and quantitative analysis.

To obtain a better performance of baseline correction algorithm, we computed the optimal values for parameters (p , λ_1 and λ). With the true baseline, say b , as a reference. The Root Mean Square Error (*RMSE*) is minimized using the parameters

p , λ_1 and λ which are varied on a fine grid as follows:

$$RMSE = \sqrt{\sum_i (b_i - b_i^0)^2 / n} \quad (7)$$

As we don't know b in reality, but it can provide us a value for the optimal performance case, and it is worthwhile for us to find optimal parameters for algorithms if $RMSE$ can be made small.

The simulation was carried out in order to imitate real spectral data sets that contain various backgrounds and Raman signals. A broader Gaussian peak was treated as curved background, a linear function was treated as sloping background, and a narrower Gaussian peak was treated as the spectra of interest. The simulated Raman spectrum consisted of 1024 channels. The amplitudes of simulate spectrum were 0.25, 0.1, 0.12, 0.25, 0.1, 0.13. The slope of the linear part of the baseline was set to 0.00001. The amplitudes of curves background were 0.15, 0.20 (baseline 1) and 0.15, 0.30 (baseline 2) respectively. The two simulated spectra which include representative features such as overlapping peaks and high fluorescence background respectively are shown in Figure 2.

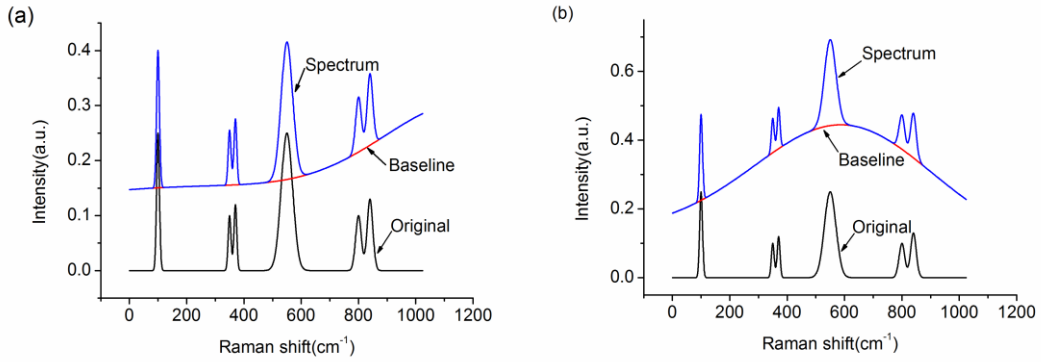


Figure 2 The simulated spectra with complex baseline background
(a) baseline 1 (b) baseline 2

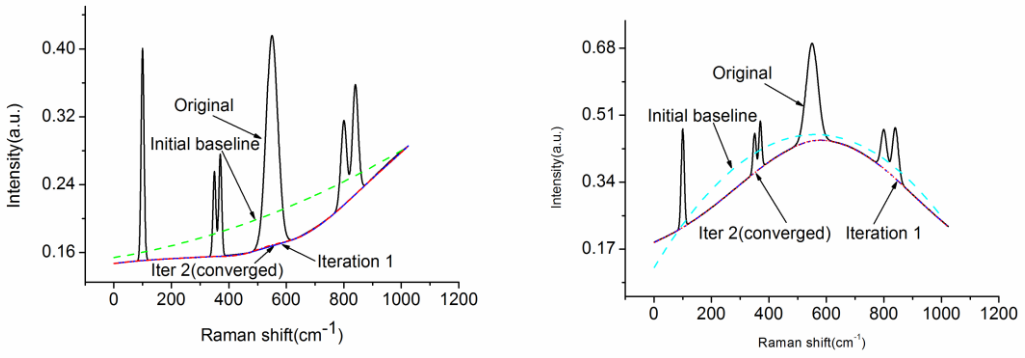


Figure 3 Iteratively estimated baseline by IAsLS of simulated Raman spectra

First, the convergence of IAsLS baseline correction algorithm is shown in Figure 3.The two iterations and the final results of the IAsLS algorithm on the simulated Raman spectra are demonstrated, and convergence occurs after 2 iterations. The initial estimated baselines, which are fitted by the second order polynomials, are strongly

improved by the new iterations. Therefore, the convergence of IAsLS baseline correction method is verified.

And then, considering the range of the parameters in reference, [18] by choosing proper parameters ($\lambda = 10^4$, $\lambda_1 = 10^{-2}$, $p = 0.001$), the performance of IAsLS baseline correction is compared with AsLS and JAsLS, and the results are shown in Figure 4. It is found that the background of raw spectrum is over-fitted by the AsLS algorithms and the intensity of spectrum is weakened as well (500-600 cm^{-1} in baseline 1 and baseline 2). In addition, the JAsLS method cannot fits background well at 800-1000cm^{-1} (baseline 1) and 0-100 cm^{-1} (baseline 2), 500-700 cm^{-1} (baseline 2), however, it fits similarly to the IAsLS method at 100-800cm^{-1} (baseline 1) and 100-500 cm^{-1} (baseline 2), 700-800cm^{-1} (baseline 2). Finally, in summary, using the proposed IAsLS method, the baseline background can be fitted well at whole Raman spectrum range.

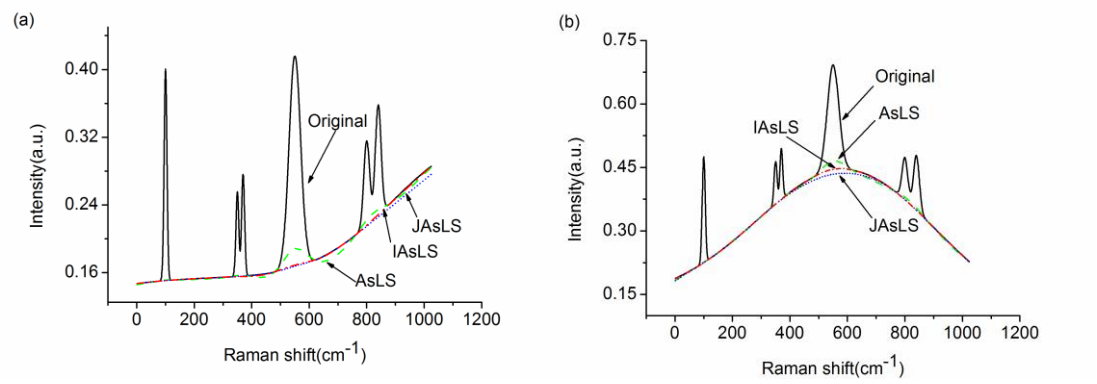


Figure 4 The performance of baseline correction based on AsLS, JAsLS and IAsLS. (a) baseline 1 (b) baseline 2

Table 1 The performance of different baseline correction algorithms

Algorithms	Parameters	Baseline 1			Baseline 2		
		Time(s)	Iterations	RMSE	Time(s)	Iterations	RMSE
AsLS	$\lambda = 10^4$	0.0468	——	0.0064	0.0156	——	0.0064
JAsLS	$\lambda_1 = 10^{-2}$	0.1092	3	0.0032	0.0156	2	0.0038
IAsLS	$p = 0.001$	0.2652	2	0.0006	0.2964	2	0.0004

In addition, compared with three algorithms in Table 1, the computation time of AsLS is equal or shorter, but its *RMSE* is the largest. And, the computation time of JAsLS and IAsLS are equal or longer compared with AsLS. However, they are all less than one second that it doesn't affect the offline or online baseline correction of Raman spectra. The *RMSEs* of the proposed IAsLS algorithm are the least, only 0.0006 (baseline 1) and 0.0004 (baseline 2) respectively. They are reduced by five fold and eleven fold comparing with JAsLS and AsLS respectively in simulated Raman spectrum (baseline 1), nine fold and sixteen fold comparing with JAsLS and AsLS respectively in simulated Raman spectrum (baseline 2). At the same time, iterations of

the IAsLS method are less than JAsLS (baseline 1). The simulation shows that, with proper parameters, the obtained results are quite near to the true baseline. Therefore, the validity of the IAsLS method is approved.

In practical, we only have the experimental data which can guide us in choosing of suitable values for parameters p, λ_1 and λ. Simultaneously, it is difficult for us to find a fail-safe cross-validation algorithm yet. Therefore, we report here of our experience with relatively ad-hoc computations. These can assist us to set approximately optimal parameters. Using the true baseline b, the simulated spectrum (baseline 1) as a reference, we varied p, λ_1 and λ on a fine grid and the *RMSE* is calculated afterwards to obtain an appropriate range of parameters.

For each pair of parameters (p, λ_1 and λ), a baseline was estimated and the *RMSE* of the proposed method was computed. Simulation results for choosing of suitable values for parameters p, λ_1 and λ are shown in electronic supplementary information (Figure S1-Figure S4). The simulation shows that, in order to obtain the optimal performance, the value of p should be set to less than 0.1, and the range of λ value to 10^2 to 10^6, and the range of λ_1 value to less than 10^{-4}.

Now, we demonstrate the two applications of the IAsLS baseline correction procedure. The examples given are of different complexity (overlapping peaks and high fluorescence background) and which illustrate the performance and limitations of the algorithm.

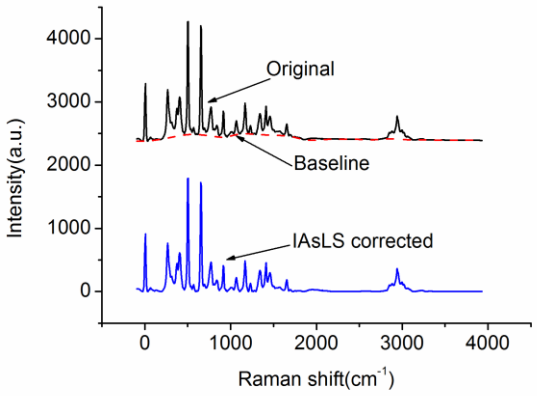


Figure 5 Original spectrum and the estimated baseline of dimethoate solution by IAsLS baseline correction method ($\lambda = 10^2$, $\lambda_1 = 10^{-5}$, $p = 0.001$)

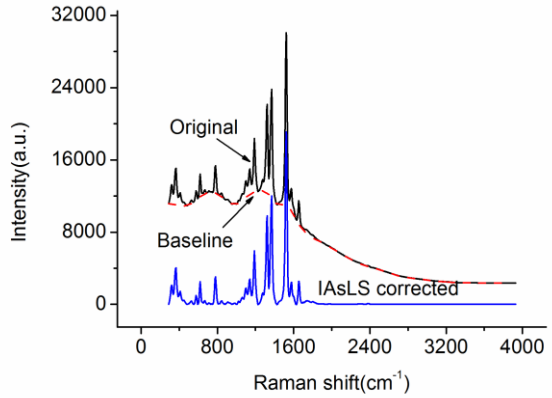


Figure 6 Original spectrum and the estimated baseline of solid Rhodamine6G by IAsLS baseline correction method ($\lambda = 10^2, \lambda_1 = 10^{-5}, \ p = 0.01$)

Figure 5 shows that the IAsLS baseline correction result of spectrum for dimethoate solution, while $\lambda = 10^2$, $\lambda_1 = 10^{-5}$, $p = 0.001$. The raw spectrum consisting of 1024 Raman shift units was obtained from the Raman spectrometer. The spectrum exhibits a linear baseline which can be attributed to the glass substrate. The spectrum for dimethoate solution with much overlapping peaks can be corrected well by IAsLS compared with other baseline correction methods (Figure S5). Then, the IAsLS baseline correction result of spectrum for solid Rhodamine6G are shown in Figure 6, while $\lambda = 10^2$, $\lambda_1 = 10^{-5}$, $p = 0.01$. The spectrum exhibits a baseline disturbance that can be attributed to the fluorescence background of sample. The results show that IAsLS is able to remove the fluorescence signal better comparing with the other two methods, but not over-fitted. Above all, both of these irrelevant spectral disturbances can be removed well by the IAsLS baseline correction.

4. Conclusion

In this paper, we present an iterative method to estimate the backgrounds of Raman spectra. The simulation results show that the background of raw spectrum is over-fitted by the AsLS algorithms and the intensity of spectrum is weakened as well. In addition, the JAsLS method cannot fits background well at 800-1000cm^{-1} and 0-100 cm^{-1} (baseline 2),500-700 cm^{-1}(baseline 2) on simulated spectra respectively. However, it fits similarly to the IAsLS method at 100-800cm^{-1}(baseline 1) and 100-500 cm^{-1} (baseline 2), 700-800cm^{-1}(baseline 2). Finally, using the IAsLS method, the baseline background can be fitted well over the whole Raman spectrum range. When the appropriate parameters were given, the simulation results indicate that the performance of IAsLS algorithm is improved by nine fold and sixteen fold comparing with that of JAsLS and AsLS respectively for the simulated Raman spectra. The actual Raman spectra experiments show that Raman peaks are eliminated automatically and only the baseline is subtracted. Furthermore, the position and shape of peak can be maintained as original form with this proposed method.

Acknowledgements

This work was supported by research funding from Chongqing Key Scientific and Technological Program Project of China (cstc2012ggC70002), National Natural Science Foundation of China (Grant No.21207165), The CAS action-plan for West Development (Grant No.KZCX2-XB3-14), West Light Foundation of Chinese Academy of Sciences, The Bureau of Chongqing Human Resource and Social Security (Grant No. 2012004).

Reference

[1] M. N. Leger, A. G. Ryder, Appl. Spectrosc., 2006, 60, 182-193.
[2] A. F. Ruckstuhl, M. P. Jacobson, Field R W and J. A. Dodd, J. Quant. Spectrosc. Radiat. Transfer, 2001, 68, 179-193.
[3] H. F. Boelens, R. J. Dijkstra, P. H. Eilers, F. Fitzpatrick and J. A. Westerhuis, J. Chromatogr. A, 2004, 1057, 21-30.
[4] K. H. Liland, T. Almøy, B. H. Mevik, Appl. Spectrosc., 2010, 64, 1007-1016.
[5] C. Rowlands, S. Elliott , J. Raman Spectrosc., 2011, 42, 363-369.
[6] Ł. Komsta, Chromatographia, 2011, 73, 721-731.
[7] A T Weakley, P R Griffiths and D E Aston, Appl. Spectrosc., 2012, 66, 519-529.
[8] V. Mazet, C. Carteret, D. Brie, J. Idier and B. Humbert, Intell. Lab. Syst., 2005, 76, 121-133.
[9] Z. M. Zhang, S. Chen and Y. Z. Liang, Analyst, 2010, 135, 1138-1146.
[10] C. V. Di Anibal, L. F. Marsal, M. P. Callao and I. Ruis ánchez, Spectrochim. Acta, Part A, 2012, 87,135-141.
[11] H. G. Schulze, R. B. Foist, K. Okuda, A. Ivanov and R. F. B. Turner, Appl. Spectrosc., 2012, 66, 757-764.
[12] J. J. deRooi, P. H. C. Eilers, Chemom. Intell. Lab. Syst., 2012, 117, 56-60.
[13] K. H. Liland, E.-O. Rukke, E. F. Olsen, T. Isaksson, Chemom. Intell. Lab. Syst., 2011, 109, 51-56.
[14] Y. G. Hu, T. Jiang, A. G. Shen, W. Li, X. P. Wang and J. M. Hu , Chemom. Intell. Lab. Syst., 2007, 85, 94-101.
[15] G. Schulze, A. Jirasek, M. M. Yu, A. Lim, R. F. Turner and M. W. Blades, Appl. Spectrosc., 2005, 59, 545-574.
[16] F. Xinwei, Z. Zhongliang, S. Mengjie and C. Peisheng, Comput. Appl. Chem., 2009, 6, 020.
[17] P. H. C. Eilers, Anal. Chem., 2004, 76, 404-411.
[18] J. An, P. Jiangtao, X. Qiwei, W. Jiping and P. Silong, Comput. Appl. Chem., 2012, 29, 537-540.
[19] P. H. C. Eilers, H. F. M. Boelens, Baseline correction with asymmetric least squares smoothing, Leiden University Medical Centre report, 2005.
[20] P. H. C. Eilers, Anal. Chem., 2003, 75, 3631-3636.