Optics and Laser Technology 96 (2017) 202–207

Contents lists available at ScienceDirect

Optics and Laser Technology

journal homepage: www.elsevier.com/locate/jolt

Full length article

A fully automated algorithm of baseline correction based on wavelet feature points and segment interpolation

Fang Qian, Yihui Wu*, Peng Hao

State Key Laboratory of Applied Optics, Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun 130033, China

ARTICLE INFO

Article history:
Received 24 November 2016
Received in revised form 13 April 2017
Accepted 15 May 2017

OCIS:
(300.0300) Spectroscopy
(300.6170) Spectra
(300.6450) Raman

Keywords:
Baseline correction
Spectral pre-processing
Wavelet transformation
Raman spectroscopy

ABSTRACT

Baseline correction is a very important part of pre-processing. Baseline in the spectrum signal can induce uneven amplitude shifts across different wavenumbers and lead to bad results. Therefore, these amplitude shifts should be compensated before further analysis. Many algorithms are used to remove baseline, however fully automated baseline correction is convenient in practical application. A fully automated algorithm based on wavelet feature points and segment interpolation (AWFPSI) is proposed. This algorithm finds feature points through continuous wavelet transformation and estimates baseline through segment interpolation. AWFPSI is compared with three commonly introduced fully automated and semi-automated algorithms, using simulated spectrum signal, visible spectrum signal and Raman spectrum signal. The results show that AWFPSI gives better accuracy and has the advantage of easy use.

1. Introduction

Spectroscopy provides detailed information and is routinely used in various application areas including biological processes and chemical analysis. The identification and quantification of the raw signals by spectral analysis methods are hindered because of the inherent artifacts such as noise and baseline. Therefore, to obtain meaningful information and deeper insight, baseline needed to be removed and the noise should be eliminated [1–6].

Baseline correction is a very important part of pre-processing. Various phenomenons like fluorescence, phosphorescence and black body radiation induce uneven amplitude shifts across different wavenumbers, manifesting itself as slowly varying curve called baseline. These amplitude shifts have to be compensated before proceeding with further analysis. There are many methods for baseline correction [7–12]. Manual baseline correction relies on user's experience, noise level and baseline characteristic. This kind of methods is not completely accurate. Automatic baseline correction can be broadly divided into fully automated [13–16] and semi-automated [17,18]. For fully automated baseline correction, the most commonly used method is polynomial fitting (PF) [4,19]. Selecting the appropriate polynomial order is extremely important. Higher order polynomial fitting may estimate some spectral information as background and can be affected by high frequency noise synchronously. Another automated method is small window moving average (SWMA) [16]. This is a moving window based method where at each point only three intensity values around this point are used for baseline estimation. It tends to have bias towards noise levels. Adaptive iteratively reweighted penalized least squares (AIRPLS) [17] is a recently introduced semi-automated method. The obvious disadvantage for this kind of method is the need to set parameters. Although default parameters are available for different signals, the accuracy and precision depend on further optimization.

Wavelet based method is widely used in chemometrics, pharmaceutics and bioinformatics, *etc.* [20]. Wavelet is localized both in time or space as well as frequency. In this study, an automated baseline correction method by means of continuous wavelet transformation and segment interpolation is proposed, called automated segment interpolation based on wavelet feature points (AWFPSI) algorithm.

2. Method

AWFPSI algorithm is developed based on the ideal of continuous wavelet transformation. The baseline is estimated using a simple linear interpolation. The algorithm is divided into five steps:

* Corresponding author.
E-mail address: qfmail@sina.cn (Y. Wu).

http://dx.doi.org/10.1016/j.optlastec.2017.05.021
0030-3992/© 2017 Published by Elsevier Ltd.

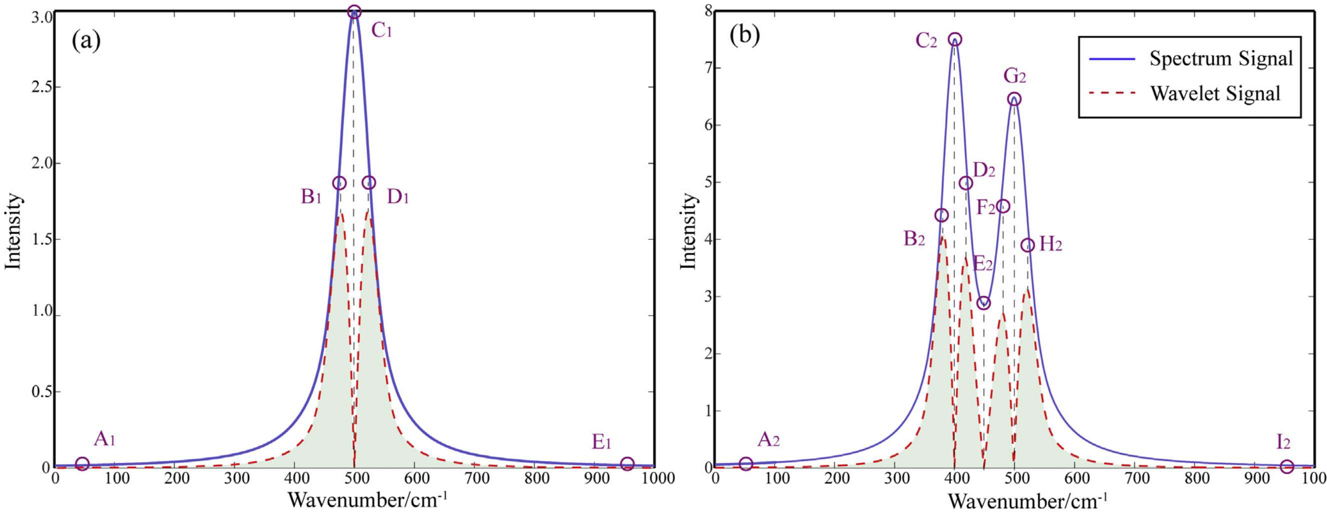

Fig. 1. Simulated spectrum signal. (a) Single peak and (b) overlap peak.

Table 1
The correspondence between wavelet extreme points and the spectral feature points.

Peak style	Single peak	Overlap peak	Spectrum signal	Wavelet signal	Spectral first derivative
Starting point	A1	A2	Minimum	Minimum	×
Left inflection point	B1	B2, F2	Rising edge	Maximum	×
Peak	C1	C2, G2	Maximum	Minimum	∨
Right inflection point	D1	D2, H2	Trailing edge	Maximum	×
Valley	None	E2	Minimum	Minimum	∨
Ending point	E1	I2	Minimum	Minimum	×

Table 2
Simulated pure spectrum signal.

Peak position (cm^{-1})	645	1090	1535	2130	2200	2270
Peak height	3	6	9	6	6	6
FWHM	25	30	35	40	45	50

Step1: pre-processing

Pre-processing is required to eliminate effects of unwanted signals. It is necessary to minimize the noise of the raw signals to increase the accuracy of the feature recognition. Savitzky-Golay (SG) filter [14] is a commonly used smoothing method. The SG filter is a moving window based on local polynomial fitting procedure. In this study, the size of moving window is three and polynomial order is zero. Three-point zero-order Savitzky-Golay filter can reduce the noise through the minimum window and retain all important spectral bands.

Step 2: continuous wavelet transformation

The continuous wavelet transformation of signal $s(t)$ at scale a ($a \in \mathrm{R}$) and translational value b ($b{\in}\mathrm{R}$) can be expressed by the following integral:

$$s_w(a,b) = \frac{1}{|a|}\int_{-\infty}^{+\infty} s(t)\psi\left(\frac{t-b}{a}\right)dt \tag{1}$$

where $\psi(t)$ is a continuous function in both the time domain and the frequency domain called mother wavelet. When the signals are decomposed with continuous wavelet transformation, we can extract some local extreme values from wavelet coefficients. These extreme values correspond to the feature points in the spectrum signal, including peak, valley, starting point, ending point and inflection point. There are many wavelet families available in the literature such as Harris, Daubechies, Biorthgonal, Ceiflets and Symlets and different wavelet families have different mother wavelets. A series of functions which can be obtained by the scale and translation of the mother wavelets are wavelet basic functions.

Using the wavelet basic functions, frequency-like information from the spectrum signals can be extracted. In order to extract the extreme values, the wavelet basis function should have first order vanishing moment, such as Harris, Daubechies and Biorthgonal. If the wavelet basis function is odd symmetry, the corresponding wavelet filter coefficient has linear phase. This characteristic makes the extreme value points unbiased in any scale. But Daubechies does not have symmetry properties. Harris and Bior1.1 have bad localization properties. Considering the exact reconstruction, Bior1.3 is the wavelet basis function in this algorithm.

Simulated spectrum signal with a single peak and its wavelet transformation signal are shown in Fig. 1(a). Simulated spectrum signal with overlap peaks and the wavelet transformation signal are shown in Fig. 1(b).

The wavelet extreme points corresponding to the feature points in the spectrum signal are shown in Fig. 1(a) and (b).

Step 3: identifying the feature points in the spectrum signal

Table 1 shows the peak, valley, starting point and ending point are minimum values in the wavelet curve graph. If the first derivative of the point goes through zero, "∨" is set, otherwise "×" is set.

The first derivative of the peak and valley in the spectrum signal go through zero point. By comparing with these characteristics, we can identify the different feature points.

Step 4: linear segment interpolation

In this step, linear segment interpolation between the starting point x_k and ending point x_{k+n} are computed. A smoothed signal is estimated from the original spectrum signal as baseline.

Starting with an array of intensities $y = [y_k, y_{k+1}, \ldots, y_{k+n}]$ at set interval $x = [x_k, x_{k+1}, \ldots, x_{k+n}]$, the interpolation polynomial is calculated as follows:

$$L_k(x) = y_k l_k(x) - y_{k+n} l_{k+n}(x) \tag{2}$$

where y_k and y_{k+n} are linear interpolation factor, $L_k(x)$ is the primary function.

$$\begin{aligned} l_k(x) &= \tfrac{x - x_{k+n}}{x_k - x_{k+n}} \\ l_{k+n}(x) &= \tfrac{x - x_k}{x_{k+n} - x_k} \end{aligned} \tag{3}$$

The interpolation result array is $Y_k = [L_k(x_k), L_k(x_{k+1}), \ldots, L_k(x_{k+n})]$.

Step 5: baseline correction

Every segment interpolation is calculated. The baseline can be updated as follows:

$$baseline = [Y_1, Y_2, \ldots, Y_k, Y_{k+1}, \ldots, Y_N] \tag{4}$$

The residual signal between the original spectrum signal and the estimated baseline is the correct spectrum signal.

3. Experimental result and analysis

Both simulated and experimental data are used to evaluate the performance of the AWFPSI algorithm. All data are compared with three other baseline correction algorithms: PF, SWMA and AIRPLS.

3.1. Simulated signals for comparison

Mathematically, the simulated spectrum signal $s(t)$ can be expressed as follows:

$$s(t) = x(t) + b(t) + n(t) \tag{5}$$

where $s(t)$ is the simulated spectrum signal, $x(t)$ is the pure spectrum signal, $b(t)$ is the baseline, $n(t)$ is the noise.

Three single peaks and three overlap peaks comprise the simulated spectrum signal $s(t)$: each peak varied in intensity. The parameters are shown in Table 2. FWHM is full width at half maximum.

The baseline includes five different forms: linear, Gaussian, exponential, sigmoidal and polynomial function, shown in Fig. 2.

Fig. 3 shows the simulated signals with pure spectrum signal, noise and baseline.

The difference between the ideal baseline and estimated baseline is calculated. Root mean square error (RMSE) is used to verify the availability.

$$RMSE = \sqrt{\sum\nolimits_{n=1}^{L} [BL_{ideal}(n) - BL_{est}(n)]^2 \Big/ L} \tag{6}$$

where L is the spectral length, BL_{ideal} is the ideal baseline, BL_{est} is the estimated baseline.

For each baseline, RMSE is calculated in both low and high SNR (signal to noise ratio) for all algorithms. A range of SNR factors are multiplied by the Gaussian white noise created to evaluated the ability of the algorithm to perform baseline correction in different SNR levels. The factors from the set {0.1, 0.2, 0.4, 0.6, 0.8, 1.0} and {2, 4, 6, 8, 10} are used for low and high SNR levels respectively. Fig. 4 shows the RMSE of the different algorithms for the various baseline and SNR used.

The scatter plots for RMSE using various SNR of each algorithm are shown in Fig. 4(a)–(e). The RMSE declines as the SNR increases. AWFPSI tends to work better for Gaussian, exponential and sigmoidal baseline correction. Fig. 4(f) shows AWFPSI has lower average error and higher stability.

For different kinds of baseline, the mean RMSE of all four algorithms is calculated, shown in Table 3.

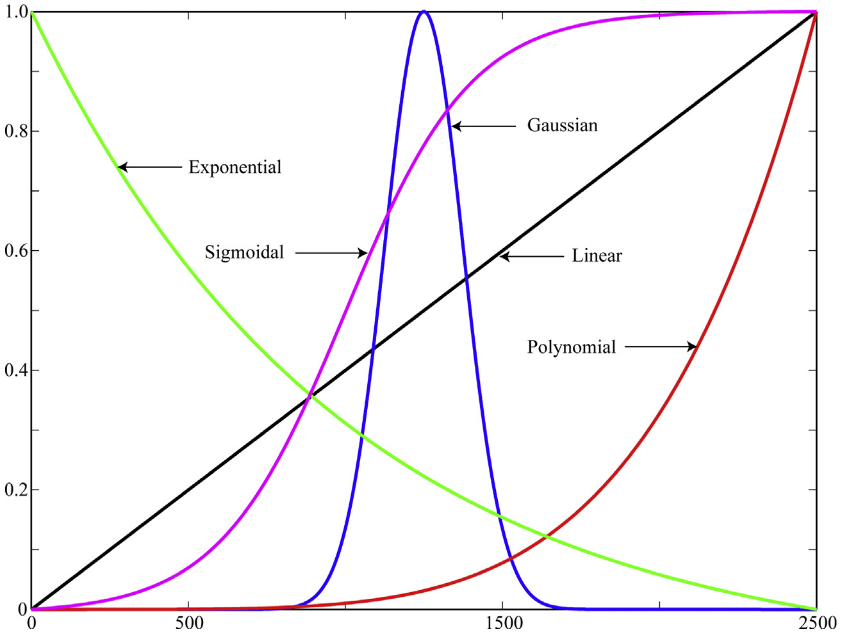

Fig. 2. Five kinds of baselines.

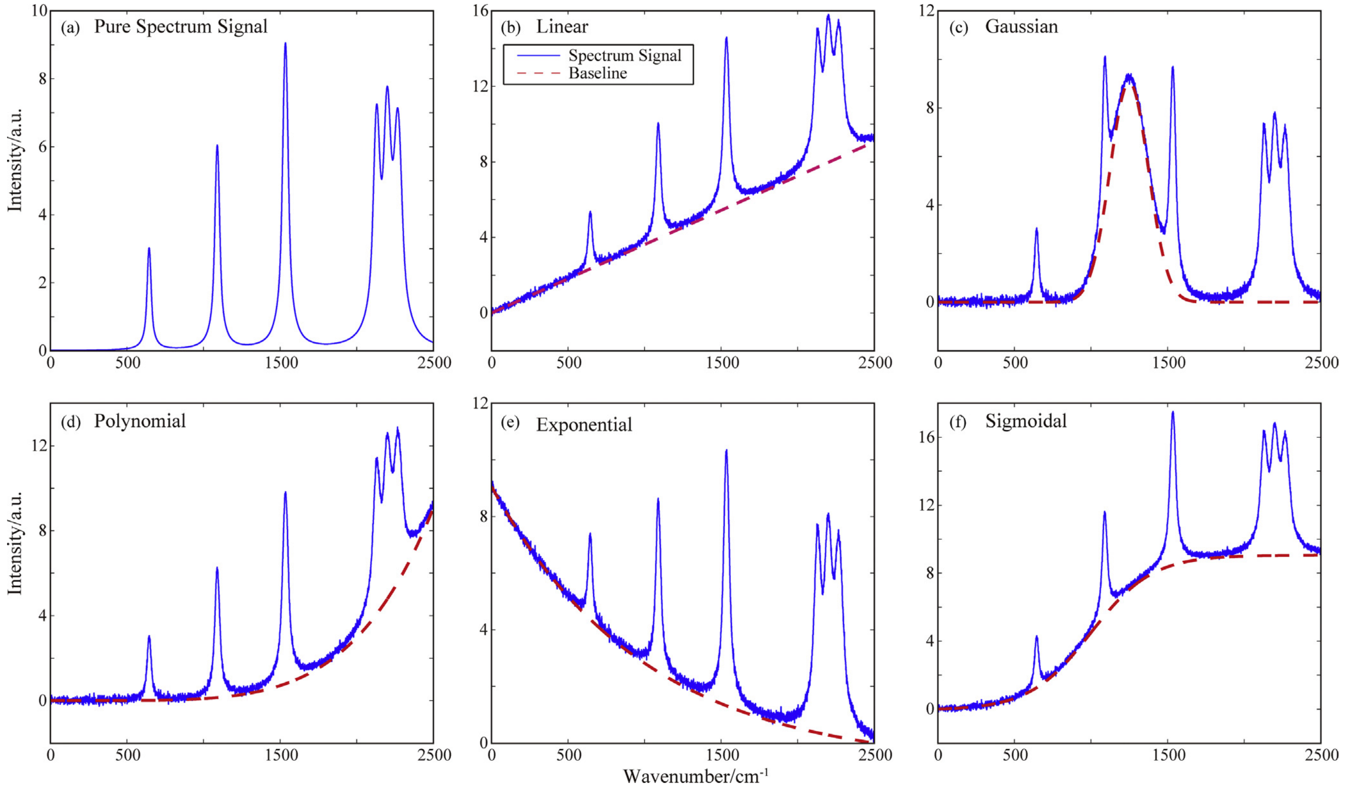

Fig. 3. Simulated data. (a) Pure spectrum signal; (b) pure spectrum signal with linear baseline and low noise; (c) pure spectrum signal with Gaussian baseline and low noise; (d) pure spectrum signal with polynomial baseline and low noise; (e) pure spectrum signal with exponential baseline and low noise; (f) pure spectrum signal with sigmoidal baseline and low noise.

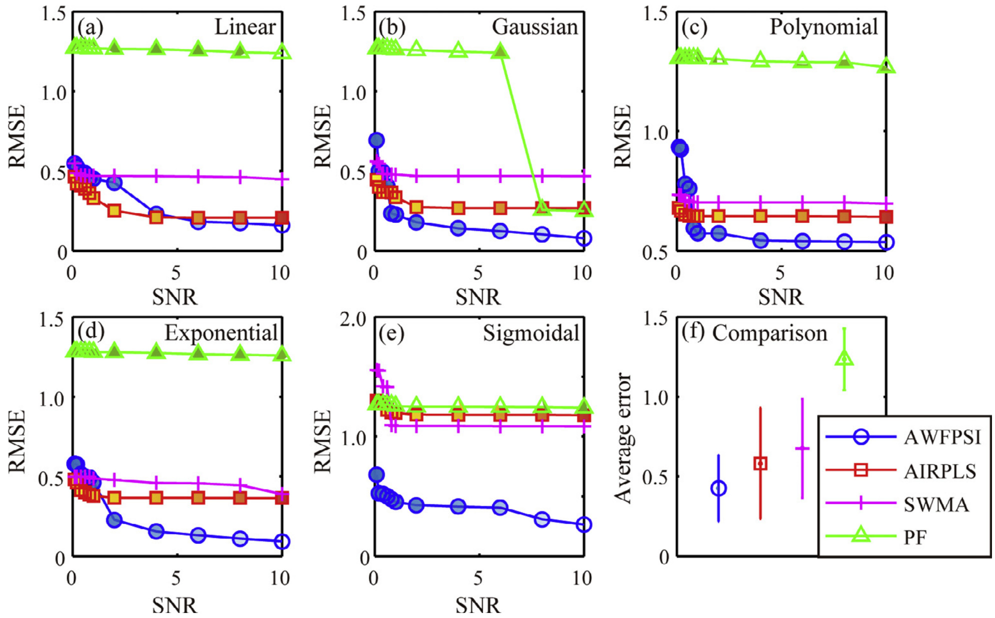

Fig. 4. Error curve of PF, AIRPLS, SWMA and AWFPSI using different SNR signals. (a) Signals with linear baseline; (b) signals with Gaussian baseline; (c) signals with polynomial baseline; (d) signals with exponential baseline; (e) signals with sigmoidal baseline; (f) box plot of comparison.

Table 3 shows AWFPSI gives better accuracy than other two automated baseline correction algorithms, PF and SWAM. Although AWFPSI is shown to have poorer accuracy than AIRPLS in linear and polynomial baseline correction, the performance of AWFPSI and AIRPLS is comparable. AWFPSI has higher average accuracy than AIRPLS and it does not need to optimize parameters.

3.2. Experimental signals for comparison

Experimental signals from visible spectrum signal and Raman spectrum signal are used to show the applicability of the AWFPSI algorithm in actual databases. Information about the experimental condition is shown in Table 4.

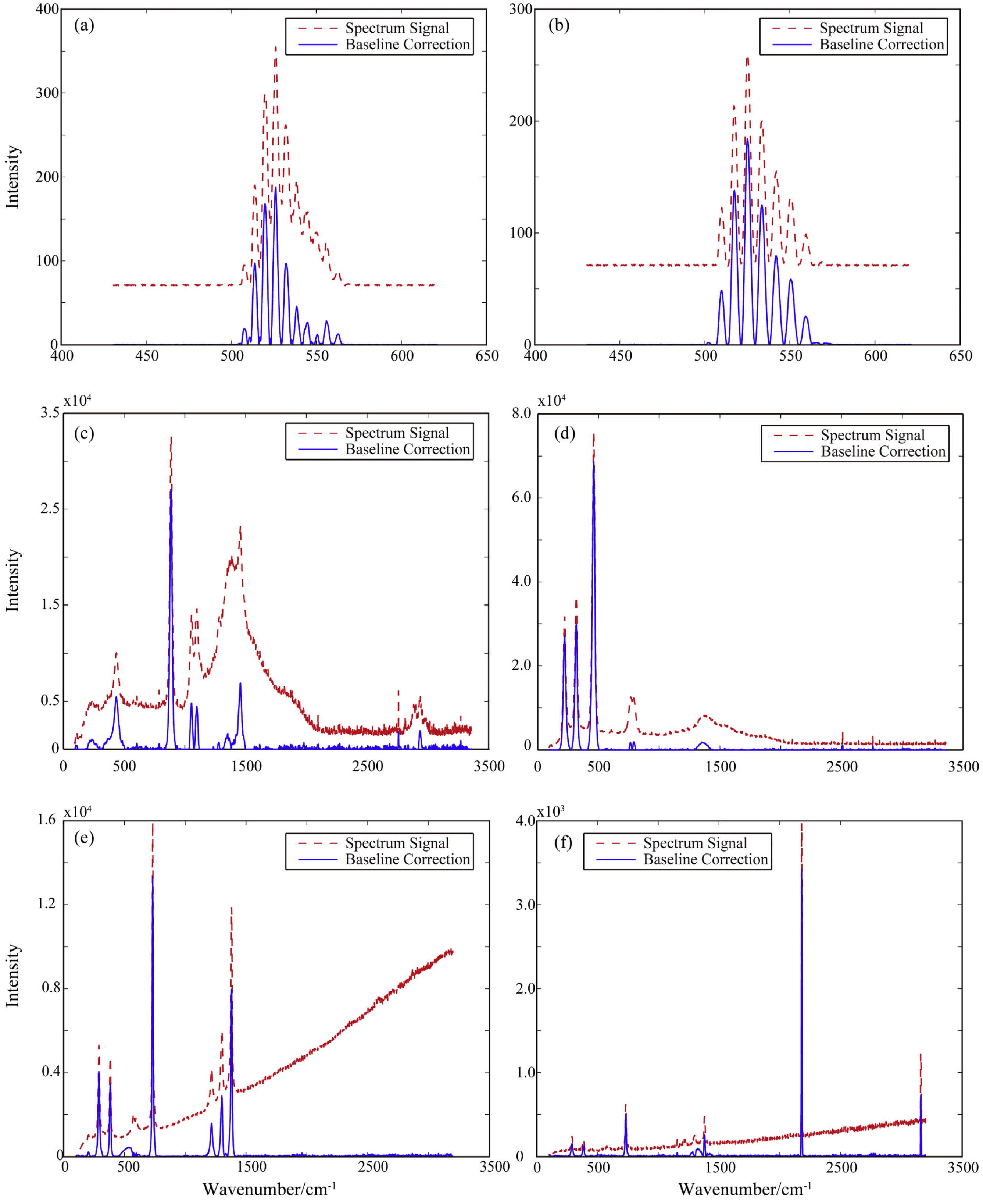

Fig. 5. Baseline correction. (a) 7% glycerin and water; (b) 28% glycerin and water; (c) alcohol; (d) carbon tetrachloride; (e) polytetrafluoroethylene; (f) methanol.

The compound of glycerin and water in different concentrations is measured based on microfiber coupler using a 532 nm LED as optical source by grating spectrometer. The different chemical substances, such as alcohol and carbon tetrachloride, are measured using a laser of 785 nm wavelength for excitation by Raman spectrometer. Polytetrafluoroethylene and methanol are measured using a laser of 488 nm wavelength for excitation by confocal microscope Raman spectrometer. The results of baseline correction are shown in Fig. 5.

For visible spectrum signal and Raman spectrum signal, comparison of the four algorithms is done by calculating the reduction

Table 3
Comparison of RMSE for all the algorithms.

Baseline style	PF	AIRPLS	SWMA	AWFPSI
Linear	1.2629	0.3141	0.4751	0.3759
Gaussian	1.0733	0.3290	0.4881	0.2883
Polynomial	1.2986	0.6501	0.7076	0.6631
Exponential	1.2755	0.3953	0.4729	0.3489
Sigmoidal	1.2562	1.2178	1.2330	0.4520

Table 4
Experimental condition used to evaluate AWFPSI algorithm.

Spectral type	Description
Visible spectrum	Grating spectrometer Optical source: 532 nm LED Resolution: 1 nm Integration time: 2 ms–2 s
Raman spectrum	Raman spectrometer Optical source: 785 nm laser Resolution: 3 nm Integration time:100 ms–10 s Confocal microscope Raman spectrometer Optical source: 488 nm laser Resolution: 1 nm Integration time:10 s

Table 5
Comparison of reduction of the convex hull area of the four algorithms using grating spectrometer.

Algorithm	Percentage reduction in area of convex hull (%)
PF	0.32%
AIRPLS	89.08%
SWMA	19.41%
AWFPSI	78.86%

Table 6
Comparison of reduction of the convex hull area of the four algorithms using Raman spectrometer.

Algorithm	Percentage reduction in area of convex hull (%)
PF	54.69%
AIRPLS	64.76%
SWMA	58.51%
AWFPSI	66.21%

in area of convex hull of the principal components analysis (PCA) plots. This is because the compactness and separation in principal components pattern space would improve clustering and classification results.

Table 5 shows AWFPSI is ranked second in the comparison and it is 10.22% behind the AIRPLS. However, the performance of AWFPSI is much better than the other two algorithms. In this paper, visible spectrum database is small and more studies should be done to confirm whether AWFPSI is suitable for visible spectrum baseline correction or not.

Table 6 shows the performance of the four algorithms in Raman spectrum baseline correction is comparable. AWFPSI is ranked first in the comparison.

4. Conclusion

As mentioned previously, baseline correction is a very important part of spectral preprocessing. A fully automated segment interpolation based on wavelet transformation (AWFPSI) algorithm is proposed in this study. Both simulated and experimental data are used to evaluate and compare the performance of the AWFPSI algorithm. AWFPSI is compared with two fully automated algorithms, namely PF, SWMA and a semi-automated algorithm, namely AIRPLS. The simulated signals use five different baseline. The RMSE between the ideal baseline and estimated baseline is calculated. The results show that AWFPSI ranks first in terms of the Gaussian, sigmoidal and exponential baseline correction. AWFPSI ranks second in terms of the linear and polynomial baseline correction, and is only behind AIRPLS. AWFPSI is a fully automated baseline correction algorithm whereas AIRPLS requires to optimize its parameters which is not desirable to process large datasets. The reduction in area of convex hull of the PCA plots is calculated. The much better performance of AWFPSI compared to the other three algorithms indicates that AWFPSI is more suitable for baseline correction of Raman spectrum. Thus, AWFPSI is a potentially useful baseline correction algorithm and has advantage of ease of use.

Acknowledgements

This work was funded by the 863 Program (2015AA042402). This work was supported by Jilin Postdoctoral Science Foundation.

References

[1] Rekha Gautam, Sandeep Vanga, Freek Ariese, et al., Review of multidimensional data processing approaches for Raman and infrared spectroscopy, EPJ Tech. Instrum. 2 (8) (2015) 1–10.
[2] T. Bocklitz, A. Walter, K. Hartmann, et al., How to pre-process Raman spectra for reliable and stable models?, Anal Chim. Acta 704 (2011) 47–56.
[3] P.M. Ramos, I. Ruisanchez, Noise and background removal in Raman spectra of ancient pigments using wavelet transform, J. Raman Spectrosc. 36 (2005) 848–856.
[4] R. Gautam, A. Samuel, S. Sil, et al., Raman and infrared imaging: applications and advancements, Curr. Sci. 108 (2015) 341–356.
[5] M. Diem, A. Mazur, K. Lenau, et al., Molecular pathology via IR and Raman spectral imaging, J. Biophoton. 6 (2013) 855–886.
[6] B. Singh, R. Gautam, S. Kumar, et al., Application of vibrational microspectroscopy to biology and medicine, Curr. Sci. 102 (2012) 232–244.
[7] M.S. Friedrichs, A model-free algorithm for the removal of baseline artifacts, J. Biomol. NMR 5 (2) (1995) 147–153.
[8] A.F. Ruckstuhl, M.P. Jacobson, R.W. Field, et al., Baseline subtraction using robust local regression estimation, J. Quant. Spectrosc. Radiat. Transf. 68 (2) (2001) 179–193.
[9] C.A. Lieber, A. Mahadevan-Jansen, Automated method for subtraction of fluorescence from biological Raman spectra, Appl. Spectrosc. 57 (11) (2003) 1363–1367.
[10] G. Schulze, A. Jirasek, M.M.L. Yu, et al., Investigation of selected baseline removal techniques as candidates for automated implementation, Appl. Spectrosc. 59 (5) (2005) 545–574.
[11] M. Morháč, An algorithm for determination of peak regions and baseline elimination in spectroscopic data, Nucl. Instrum. Methods Phys. Res., Sect. A 600 (2) (2009) 478–487.
[12] Z.M. Zhang, S. Chen, Y.Z. Liang, et al., An intelligent background-correction algorithm for highly fluorescent samples in Raman spectroscopy, J. Raman Spectrosc. 41 (6) (2010) 659–669.
[13] Bhaskaran David Prakash, Yap Chun Wei, A fully automated iterative moving averaging (AIMA) technique for baseline correction, The Analyst 136 (2011) 3130–3135.
[14] H.G. Schulze, R.B. Foist, A. Ivanov, et al., Fully automated high-performance signal-to-noise ratio enhancement based on an iterative three-point zero-order Savitzky-Golay filter, Appl. Spectrosc. 62 (10) (2008) 1160–1166.
[15] H.G. Schulze, R.B. Foist, K. Okuda, et al., A model-free, fully automated baseline-removal method for Raman spectra, Appl. Spectrosc. 65 (1) (2011) 75–84.
[16] H.G. Schulze, R.B. Foist, K. Okuda, et al., A small-Window moving average-based fully automated baseline estimation method for Raman spectra, Appl. Spectrosc. 66 (7) (2012) 757–764.
[17] Z.M. Zhang, S. Chen, Y.Z. Liang, Baseline correction using adaptive iteratively reweighted penalized least squares, Analyst 135 (5) (2010) 1138–1146.
[18] G. Li, Removing background of Raman spectrum based on wavelet transform, Fut. Comput. Commun. (2009) 198–200.
[19] F. Gan, G. Ruan, J. Mo, Baseline correction by improved iterative polynomial fitting with automatic threshold, Chemomet. Intell. Lab. Syst. 82 (1) (2006) 59–65.
[20] G. Schulze, A. Jirasek, M.L. Lu, et al., Investigation of selected baseline removal techniques as candidates for automated implementation, Appl. Spectrosc. 59 (2005) 545–574.