Asian Journal of Chemistry; Vol. 23, No. 12 (2011), 5229-5234

ASIAN JOURNAL OF CHEMISTRY

www.asianjournalofchemistry.co.in

Automated Background Subtraction Algorithm for Raman Spectra Based on Iterative Weighted Least Squares

H. RUAN and L.K. DAI*

State Key Laboratory of Industrial Control Technology, Zhejiang University, Hangzhou, P.R. China

*Corresponding author: Fax: +86 571 87952127; Tel: +86 571 87951894; E-mail: lkdai@iipc.zju.edu.cn

(*Received*: 15 October 2010; *Accepted*: 10 August 2011) AJC-10260

A major problem in Raman spectroscopy is that the spectrum is often suffered from intrinsic fluorescence which is orders of magnitude greater than the Raman signal. Background subtraction is essential for further analysis, particularly for quantitative analysis using multivariate calibration. In this paper, we propose a background removal algorithm which approximates the background by a polynomial and estimates the polynomial coefficients by iterative weighted least squares. The performance of the algorithm accompanied with two comparative methods are evaluated both on simulated and real Raman spectra. The results show that the proposed algorithm provides the best result using R^2 between the actual and extracted Raman peaks. It also improves the performance of background removal in quantitative Raman spectroscopy. Further more, the algorithm is least dependent on the choice of polynomial order.

Key Words: Raman spectroscopy, Background subtraction, Polynomial fitting, Iterative weighted least squares.

INTRODUCTION

Raman spectroscopy is a noninvasive analytical technique proven to be valuable in diverse applications. However, Raman spectra often suffer from the fluorescence interference which can be orders of magnitude more than the Raman peaks. This leads to a broad background under the Raman peaks. Distorted background results in incorrect integration or peak height determination which can be central to many spectral applications. Hence it is necessary to separate the Raman peaks from the background.

Both instrumental and computational methods have been proposed to perform background removal. Among instrumental methods, frequency-shifted excitation[1,2] and time gating[3] have been proven to be effective in rejecting fluorescence. But additional instrumental modification is required. Computational methods including first- and second-derivatives[4], frequency-domain filtering[5], wavelet transformation[6], optimization[7,8], polynomial fitting[9,10] and its improved versions[11,12] are substituted strategies with wide application. For the Raman scattering is intrinsically a series of Lorentzian peaks, most of the computational methods are based on the fact that it varies more quickly than the background does.

Since the polynomial fitting has the ability to approximate the background and meanwhile retain the contours of the Raman peaks, it has become a common background removal method. Simply fitting a polynomial curve to a raw Raman spectrum in a least-squares manner[9] does not provide accurate results because of the negative values in the extracted Raman spectra. To address the limitation, Lieber and Mahadevan-Jansen proposed the modified polynomial fitting (called ModPoly) method that fits the data in an iterative way and finally obtain the polynomial that beneath the raw spectrum[10]. The ModPoly method has now been widely used. Its main limitation is that, peak-free regions which are slightly higher than the fitted polynomial will be replaced by the fitted polynomial and thus generate incorrect result. This may lead to an upturn at the fringe of the extracted Raman spectra. The improved modified multi-polynomial fitting (called I-ModPoly) is a variation of the ModPoly method[11]. It suggests that points beneath the sum of the fitted value and a standard deviation will be incorporated in the next fitting procedure. It reduces the limitation of ModPoly, but inclines to yield Raman signal with negative value. Thus the further quantitative analysis is biased.

In this paper we present a new background subtraction algorithm that approximates the background by a polynomial and estimates the polynomial coefficients by iterative weighted least squares (IWLS). It is an extension of the ModPoly in methodology. Comparative study shows that this algorithm produces stable and reliable result in an automated way. It should be noted that this paper only dealing with the background subtraction. It is assumed that all the spectra applied by iterative weighted least squares have been denoised.

Background subtraction by iterative weighted least squares

Modeling: To formulate the background removal algorithm, we should define a model of Raman spectra. Given a set of points in $y = (y_1, ..., y_N)^T$ in y^N that denote an N-point spectrum, it can be considered as a combination of Raman peaks, background and noise,

$$y = s + b + n \quad (1)$$

where s denotes the ideal Raman spectrum, n denotes the physical noise and b denotes the background. Because the background is usually a broad fluorescence, it is sufficient to approximate it with a p^{th} order polynomial,

$$b = T\beta \quad (2)$$

where T and β represent the wavenumber Vandermonde matrix and the polynomial coefficient vector respectively, defined as

$$T = \begin{pmatrix} t_1^0 \cdots t_1^p \\ \vdots \quad \vdots \\ t_N^0 \cdots t_N^p \end{pmatrix}, \beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix} \quad (3)$$

To perform background removal, the estimation of β is required. Parameter estimation techniques consider a measurement equation of the form

$$y = T\beta + r \quad (4)$$

where r is an $N \times 1$ vector of residual.

Iterative procedure: The IWLS is based on weighted least-squares which estimate β by minimizing

$$(y - T\beta)^T W(y - T\beta) \quad (5)$$

where the weight matrix is defined as

$$W = \text{diag}(w_1, w_2, ... w_N); w_i \geq 0, i = 1,2\cdots N \quad (6)$$

For a given weight matrix W, the solution for eqn. 5 is given by

$$\beta = (T^T WT)^{-1} T^T Wy \quad (7)$$

To avoid the influence of Raman peaks, an iterative weighted least-squares procedure was used. We initial the weight W with an identity matrix and solve eqn. 7 for this weight. Then we use the solution β to define a new weight and repeat this process. The definition of a new weight is the core of iterative weighted least squares.

Definition of the new weight: The new weight for iteration is defined by a weight function. Since we assume that the spectrum has been denoised, it can be considered as a combination of Raman peaks and background. The design of weight function should remain the background yet exclude Raman peaks. Based on Lieber and Mahadevan-Jansen's method[10], we propose a new algorithm which incorporates a function of the coarseness of the residual to define the new weight.

The residual is obtained by $r = y - T\beta$, where β is the solution of a given weight. A moving window standard deviation of the first derivative of the residual is adopted to represent its coarseness. Specifically, given a series of numbers $r = (r_1, ..., r_N)^T$ that represent the residual, we firstly calculate its first derivative $d = (d_1, ..., d_N)^T$ by numerical differentiation method. For a fixed window with $2M + 1$ points, the coarseness of the residual is defined as

$$c_i = \sqrt{\frac{1}{2M}\sum_{j=-M}^{M}(d_{i+j} - \bar{d}^{(i)})^2}, \quad i = 1\cdots N \quad (8)$$

where $\bar{d}^{(i)}$ is the mean value for $(d_{i-M}, \cdots, d_{i+M})$, M is a user defined parameter that controls the width of the moving window. Finally, the coarseness $c = (c_1, \cdots, c_N)^T$ is normalized to range between 0 and 1.

For a raw spectrum shown in Fig. 1(a), it was fitted by a 5^{th} order polynomial. The residual and its 1^{st} derivative are shown in Fig. 1(b) and (c). Fig. 1(d) shows the coarseness of the residual where the parameter M is set to 5.

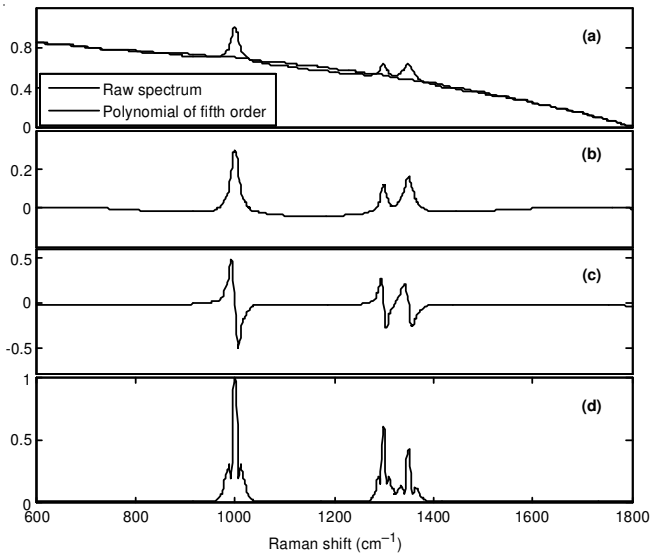


Fig. 1. (a) The raw spectrum, (b) the residual of the spectrum, (c) the first derivative of the residual, (d) the coarseness of the residual

After we have obtained the coarseness of the residual, the new weight $W \in ?^{n\times n}$ is defined in eqn. 6 and

$$w_i = \varphi(r_i) \times \psi(c_i) \quad (9)$$

where r is the residual, c is the coarseness of the residual.

The function $\varphi(\cdot)$ calculates the weight according to the residual, which can be defined as

$$\varphi(x) = \begin{cases} 1 & \text{if } x < 0 \\ \varepsilon & \text{otherwise} \end{cases} \quad (10)$$

where $\varepsilon > 0$ and close to 0.

The function $\psi(\cdot)$ calculates the weight according to the coarseness of the residual.

$$\psi(x) = \begin{cases} 1 & \text{if } x < \eta \\ \varepsilon & \text{otherwise} \end{cases} \quad (11)$$

where η is the threshold.

Termination criteria: Since the algorithm excludes regions with Raman peaks, convergence is an important issue. We use the root-mean-square of the difference between $\beta^{(n)}$ and $\beta^{(n+1)}$ as the termination criteria, where $\beta^{(n+1)}$ and $\beta^{(n)}$ represent the solution of step n and $(n + 1)$ respectively. Once this value is small enough to reach a threshold, the iteration procedure is terminated. Empirically, the iterative procedure reliably converges in about 10 iterations. At the point, $\beta^{(n+1)}$ is considered as an optimized estimation of the background. Subtracting the estimated background from the raw spectrum yields a true Raman signal.

Summary of the algorithm: The computational procedure used to subtract background for Raman spectra is summarized as follows:

Step-1: Given a raw spectrum described by $y = (y_1, \cdots, y_N)^T$, set the polynomial order p, form the affinity matrix T according to eqn. 3, initialize the weight matrix $W \in ?^{n\times n}$ with an identity matrix.

Step-2: Repeat the following iterative process. (1) Estimate the polynomial coefficient vector β by eqn. 7. (2) Calculate the root-mean-square of the difference between β and the result at the previous iteration. If the value is small enough to reach a threshold, then go to Step-3. (3) Calculate the residual and its coarseness by eqn. 8 and set the new weight W by eqn. 9.

Step-3: Obtain the estimated background by $\beta = T\beta$, the extracted Raman spectrum is given by $y - b$.

Simulations: In this section, we present a comparative study of the IWLS, ModPoly[10] and I-Mod-Poly[11]. A set of synthetic spectra with known Raman and background contributions is used to evaluate the three algorithms. Each spectrum consists of a Raman signal and a background. They are generated with varying signal-to-background ratios (SBR). The SBR is defined as the maximum peak height above background divided by the difference between the maximum and minimum background points[13].

The Raman signal is generated as a series of Lorentzian peaks,

$$L(\omega) = \sum_{i=1}^{N} \frac{2A_{0i}}{\pi} \frac{\omega_{Li}}{4(\omega-\omega_{0i})^2 + \omega_{Li}^2} \quad (12)$$

where ω represents the Raman shift, ω_{0i} is the position of the peak center, A_{0i} is the total area under the curve and ω_{Li} is the bandwidth of the peak at full-width at half maximum (FWHM). The Lorentzian function parameters used for creating the synthetic spectra are listed in Table-1. Fig. 2 shows the synthetic Raman peaks.

TABLE-1
LORENTZIAN FUNCTION PARAMETERS USED FOR CREATING SYNTHETIC SPECTRUM

ω_0	ω_L	A_0
730	13	12
763	20	9
890	15	15
930	12	8
1100	8	20
1260	9	11
1280	9	9
1310	15	20
1480	10	67
1658	11	30

The Gaussian and Sigmoid backgrounds are used in this study. Both of them are non-polynomial. The Gaussian background is defined as

$$G(\omega) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{(\omega-\mu_g)^2}{2\sigma_g^2}} \quad (13)$$

where $\sigma_g = 400$, $\mu_g = 1000$.

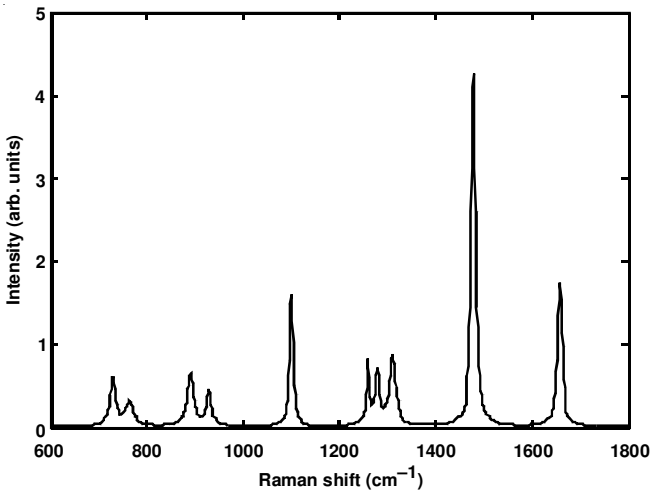


Fig. 2. Synthetic Raman peaks created using a series of Lorentzian peaks

The Sigmoid background is defined as:

$$S(\omega) = \frac{1}{1+e^{\frac{\omega-\mu_s}{\sigma_s}}} \quad (14)$$

where $\sigma_s = 100$, $\mu_s = 1200$.

The two types of backgrounds are shown in Fig. 3. They are then multiplied by a factor to synthesize the spectra at specified SBR. The synthesized spectrum can be written as:

$$y(\omega) = L(\omega) + \delta_1 G(\omega) + \delta_2 S(\omega) \quad (15)$$

where δ_1 and δ_2 are determined by the SBR of spectrum.

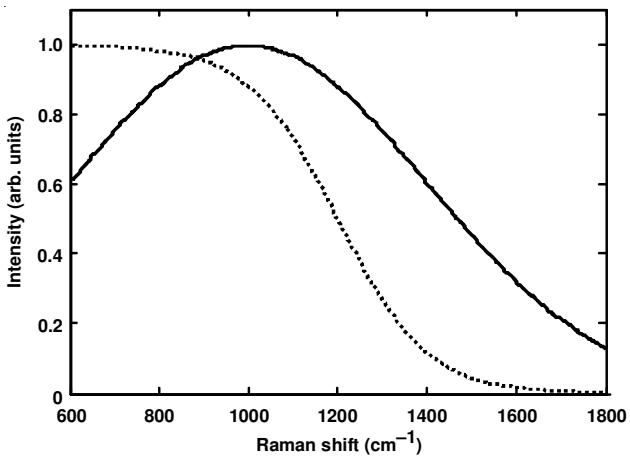


Fig. 3. Two types of backgrounds, solid line for Gaussian background, dotted line for sigmoid background

The spectra are submitted to ModPoly, I-Mod-Poly and the IWLS algorithm proposed in this paper. The coefficient of determination (R^2) between the actual and extracted Raman peaks is used. All the three fitting methods involve a user defined parameter p that specifies the fitting polynomial order. To investigate the influence of this parameter, two spectra with identical SBR (SBR = 0.1) but two distinct backgrounds (Gaussian and sigmoid) are used in the study. The two spectra are shown in Fig. 4. We choose the polynomial order p ranging from 5 to 12. For each choice, the ModPoly, I-Mod-Poly and the IWLS fitting algorithms are applied respectively, a series of results are shown in Figs. 5 and 6. The IWLS parameters are empirically chosen as $\varepsilon = 0.01$ and $\eta = 5$ %.

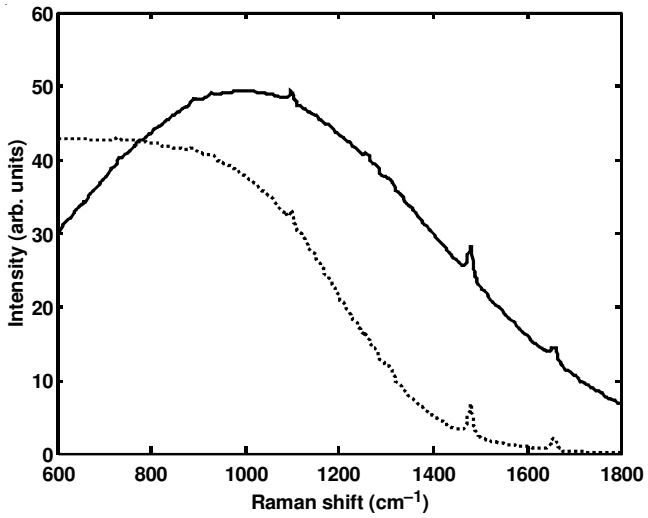


Fig. 4. Synthetic spectra with Gaussian background (solid line) and sigmoid background (dotted line)

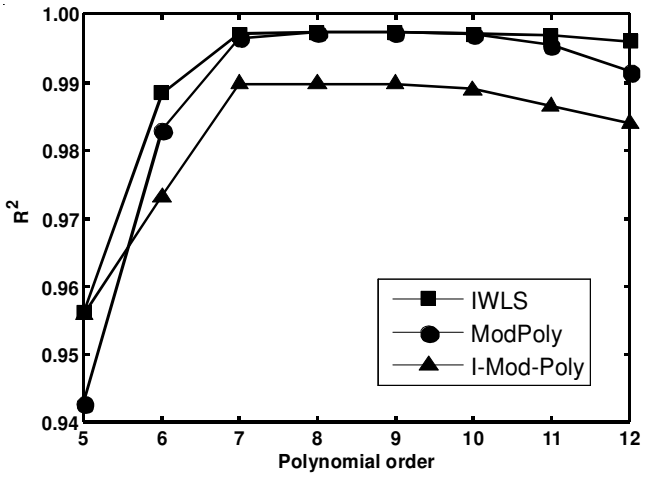


Fig. 5. A comparison of the three algorithms for the spectrum with Gaussian background

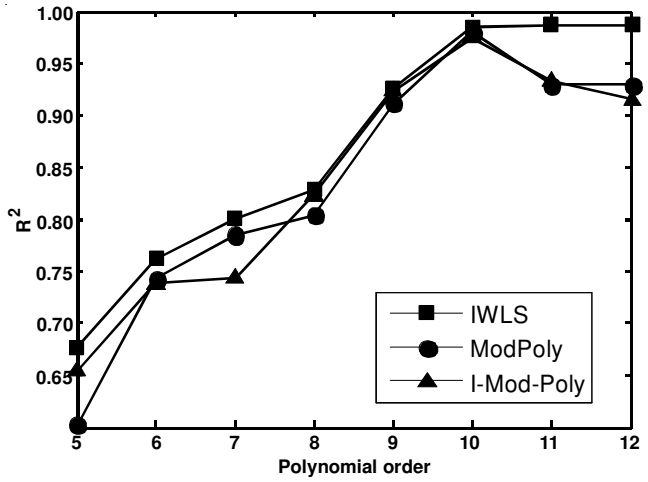


Fig. 6. A comparison of the three algorithms for the spectrum with sigmoid background

As shown in Figs. 5 and 6, the optimal polynomial order varies according to the background shape. For the spectrum with Gaussian background, all the three methods achieve the optimum performance at p = 12. But for the spectrum with sigmoid background, the optimal polynomial order is 10. Increasing the order leads to degradation of performance for ModPoly and I-Mod-Poly. On the contrary, IWLS performs consistently as polynomial order increase. Using high order polynomial in IWLS algorithm accurately retains the Raman spectral component. Since the shape of background varies in practice, choosing an order that best approximates the background become problematic to ModPoly and I-Mod-Poly. In comparison with them, IWLS is much less dependent on the polynomial orders. The extracted spectra by the three methods with optimal order and high order (p = 12) are shown in Figs. 7 and 8.

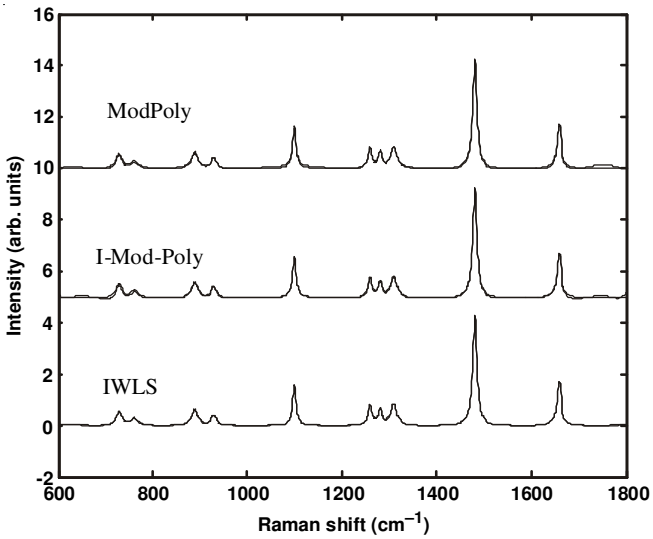


Fig. 7. Extracted spectra from the synthetic spectra with Gaussian background. The solid line and dotted line are for 7th order and 12th order polynomial extracted spectra respectively

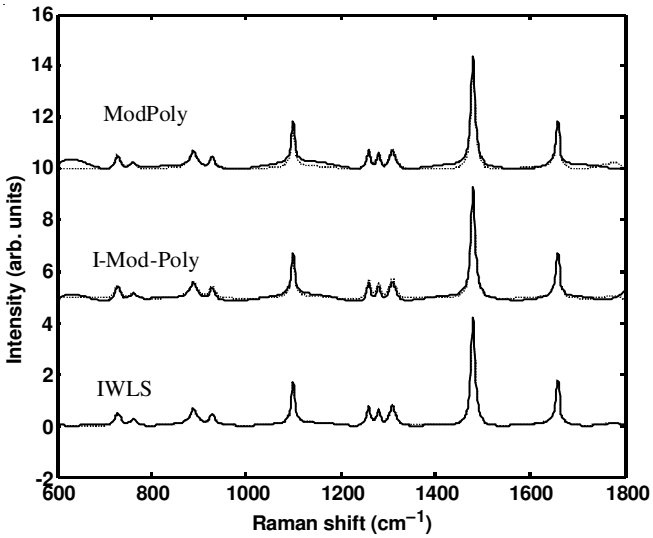


Fig. 8. Extracted spectra from the synthetic spectra of sigmoid background. The solid line and dotted line are for 10th order and 12th order polynomial extracted spectra respectively

A second comparative study is performed using a series of synthetic spectra at different SBR (0.001, 0.01, 0.1, 1 and 10). According to the results obtained, we use 7th and 10th polynomial to approximate the Gaussian and Sigmoid background respectively. The results are shown in Tables 2 and 3. For high

TABLE-2
R^2 FOR SIMULATION OF SPECTRA WITH GAUSSIAN BACKGROUND, HIGHLIGHTED IS THE ALGORITHM WITH HIGHEST R^2 VALUE FOR A SPECIFIC SBR

	SBR				
	0.001	0.01	0.1	1	10
ModPoly	0.5627	0.9812	0.9987	0.9980	0.9980
I-Mod-Poly	0.7250	0.9792	0.9912	0.9916	0.9916
IWLS	0.7129	0.9844	0.9978	0.9981	0.9981

TABLE-3
R^2 FOR SIMULATION OF SPECTRA WITH SIGMOID BACKGROUND, HIGHLIGHTED IS THE ALGORITHM WITH HIGHEST R^2 VALUE FOR A SPECIFIC SBR

	SBR				
	0.001	0.01	0.1	1	10
ModPoly	0.6030	0.8042	0.9841	0.9960	0.9975
I-Mod-Poly	0.7273	0.8287	0.9735	0.9885	0.9908
IWLS	0.7106	0.8289	0.9865	0.9971	0.9977

SBR spectra, IWLS and ModPoly are better then I-Mod-Poly. Both produce result that $R^2 > 0.995$. For low SBR spectra, I-Mod-Poly and IWLS produce almost identical results, which are much better then ModPoly.

EXPERIMENTAL

In quantitative Raman spectroscopy, particular in multivariate calibration, the performance of background subtraction is critical to the final result. A slight distortion of the Raman peaks will yield considerable estimation errors. In this section, IWLS is applied to perform background subtraction for Raman spectra of gasoline. The extracted spectra are then used to determine the contents of methyl tertiary butyl ether (MTBE) in gasoline. Partial least squares (PLS) and cross-validation using leave-one-out (LOO) approach are applied to evaluate the regression model and background subtraction algorithm.

Forty gasoline samples with known methyl tertiary butyl ether contents are prepared. The intensity of the background varies according to the concentration of fluorescent substance in gasoline. Raman spectra are recorded using a fiber-probe-based Raman system. In the system, a 500 mW, 785 nm diode laser (B&W Tek, USA) is used for sample excitation. The laser is coupled to a 200 μm core diameter excitation fiber contained in an optical probe (InPhotonics, USA). The Raman signal is collected by the probe and delivered to a spectrometer (QE65000, Ocean Optics, USA) by a 400 μm core diameter fiber. The spectrometer is configured with a 25 μm slit, a 1200 lines/mm grating and a deep-depletion back-illuminated, thermoelectrically cooled CCD detector of 1024 × 58 pixel count. All the spectra are recorded with an integration time of 15 s. Then the Savitzky-Golay filter with a half-width of 5 pixels is applied for denoising.

RESULTS AND DISCUSSION

Fig. 9 displays the Raman spectra of the prepared gasoline samples. The overall signal changes obviously in several cases due to the fluorescence. Ideal background subtraction method is desired because even a slight distortion of Raman signal has a significant effect on quantitative analysis of a test sample. Fig. 10 shows the background removal for a typical spectrum using a 5th-, 6th- and 7th-order polynomial. The ModPoly method produces desirable result using a 5th-order polynomial. But obvious artifacts at upper spectral boundary region (1700-2000 cm^{-1}) are observed when using a 7th-order polynomial. The results of I-Mod-Poly contain several regions with negative value, which is non-physical and consequently not suitable for quantitative analysis. For the IWLS algorithm, artifact peaks at the spectral boundary regions are not observed. Using different polynomial orders, the extracted Raman spectra are desirable and almost identical. The algorithm is the least dependent on the choice of polynomial order.

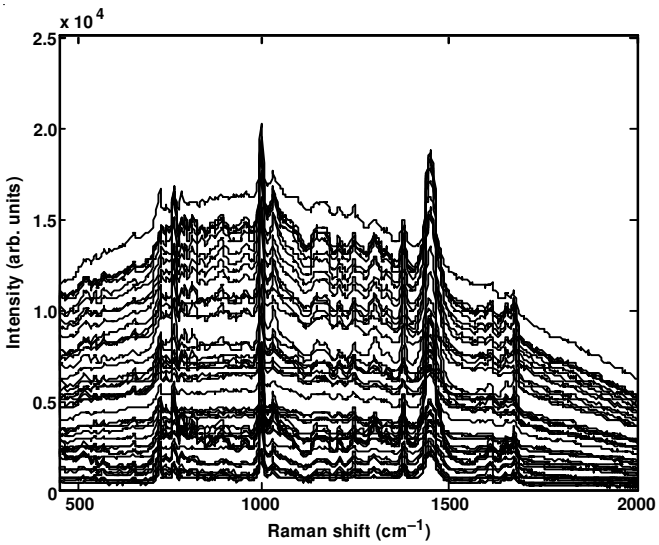
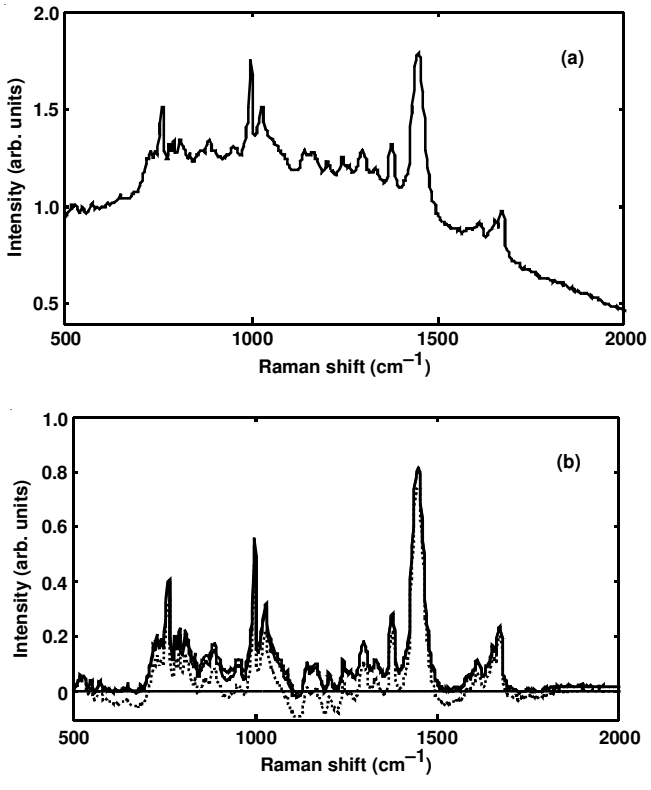


Fig. 9. Raw spectra of the gasoline samples



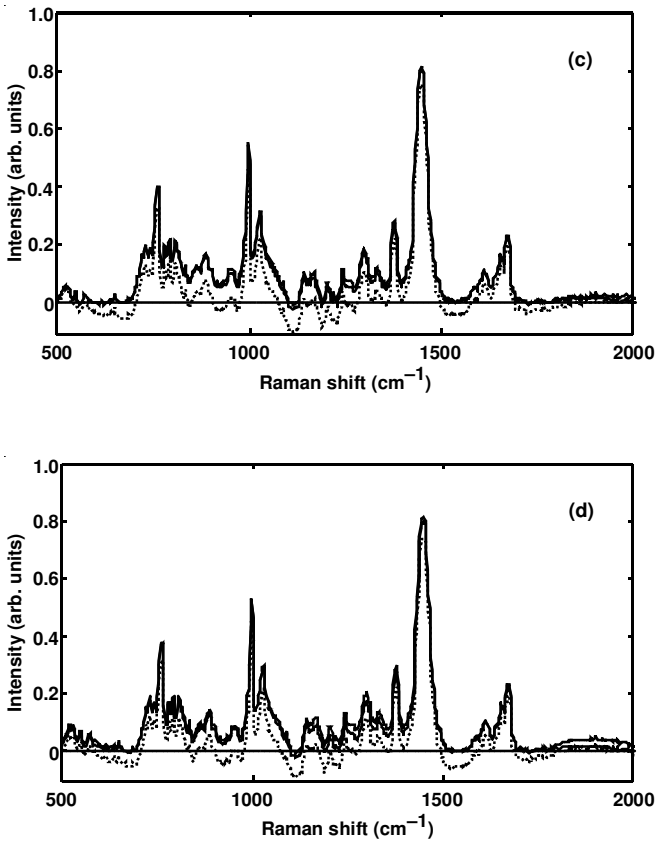


Fig. 10. (a) Raman spectrum of a sample. (b), (c) and (d) background subtraction with 5th-, 6th- and 7th-order polynomial respectively. For each figure, solid bold line, solid thin line and dashed line are for IWLS, ModPoly and I-Mod-Poly respectively

All of the extracted spectra by IWLS with a 6th-order polynomial are shown in Fig. 11. Partial least square and leave-one-out cross-validation is applied to evaluate the regression model and background removal methods. The PLS factors are set to 4 according to Cooper's work[14]. Fig. 12 shows the evaluation results of the models which are built by the extracted spectra using the three background removal methods. The standard error of cross-validation (SECV) is 3.3, 8.6 and 19.7 % for IWLS, ModPoly and I-Mod-Poly respectively. The IWLS algorithm gives a satisfactory calibration with low SECV.

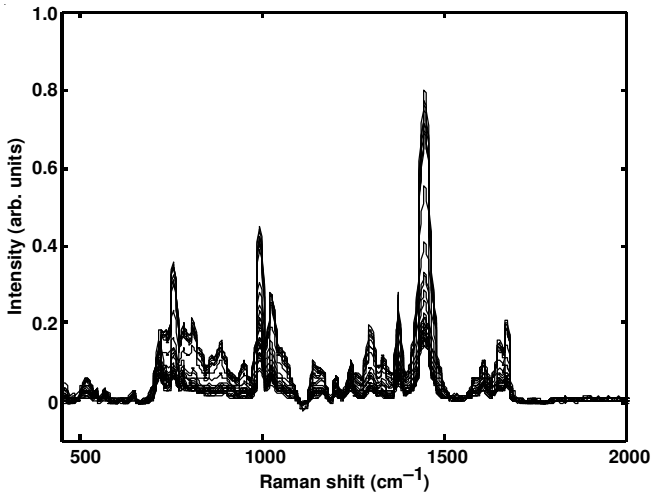


Fig. 11. Extracted Raman spectra of all gasoline samples using IWLS

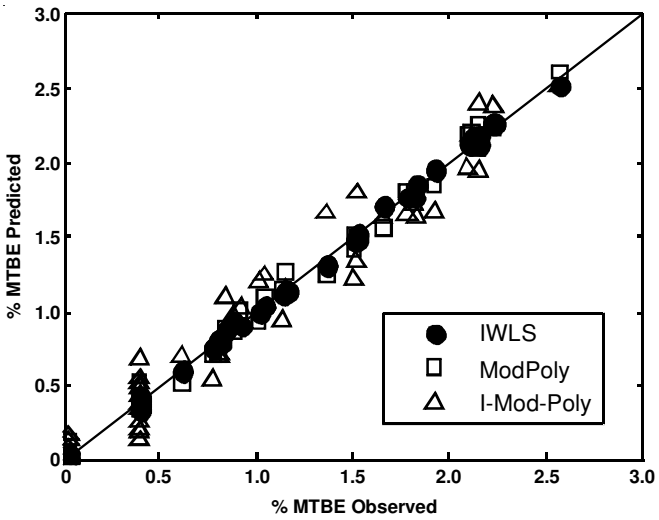


Fig. 12. PLS calibration plot of observed vs predicted concentration values of MTBE

Conclusion

This paper presents an automated background subtraction algorithm for Raman spectra. The algorithm approximates the background by polynomial and estimates the polynomial coefficients by iterative weighted least-squares (IWLS). Simulation comparison study shows that the IWLS provide the best results. Further more, IWLS is the least dependent on the choice of polynomial order. To prove the improvement brought by this algorithm, we have analyzed the methyl tertiary butyl ether contents of gasoline samples. Partial least sqaure calibration of the extracted Raman spectra by IWLS gave the lowest SECV. The experiment results demonstrated that the IWLS is a stable and reliable automated background subtraction algorithm for quantitative Raman spectroscopy analysis.

ACKNOWLEDGEMENTS

This research work was supported by the National High Technology Research and Development Program ("863"Program) of China, Grant No. 2009AA04Z123.

REFERENCES

1. P. Shreve, N.J. Cherepy and R.A. Mathies, *Appl. Spectrosc.*, **46**, 707 (1992).
2. S.T. McCain, R.M. Willett and D.J. Brady, *Opt. Express.*, **16**, 10975 (2008).
3. P. Matousek, M. Towrie, A. Stanley and A.W. Parker, *Appl. Spectrosc.*, **53**, 1485 (1999).
4. A. O'Grady, A.C. Dennis, D. Denvir, J.J. McGarvey and S.E.J. Bell, *Anal. Chem.*, **73**, 2058 (2001).
5. P.A. Mosier-Boss, S.H. Lieberman and R. Newbery, *Appl. Spectrosc.*, **49**, 630 (1995).
6. Z.M. Zhang, S. Chen, Y.Z. Liang, Z.X. Liu, Q.M. Zhang, L.X. Ding, F. Ye and H. Zhou, *J. Raman Spectrosc.*, **41**, 659 (2010).
7. V. Mazet, C. Carteret, D. Brie, J. Idier and B. Humbert, *Chemometr. Intell. Lab.*, **76**, 121 (2005).
8. B.D. Beier and A.J. Berger. *Analyst*, **134**, 1198 (2009).
9. J.F. Brennan, Y. Wang, R.R. Dasari and M.S. Feld, *Appl. Spectrosc.*, **51**, 201 (1997).
10. C.A. Lieber and A. Mahadevan-Jansen, *Appl. Spectrosc.*, **57**, 1363 (2003).
11. J.H. Zhao, H. Lui, D.I. McLean and H. Zeng, *Appl. Spectrosc.*, **61**, 1225 (2007).
12. A. Cao, A.K. Pandya, G.K. Serhatkulu, R.E. Weber, H.B. Dai, J.S. Thakur, V.M. Naik, R. Naik, G.W. Auner, R. Rabah and D.C. Freeman, *J. Raman Spectrosc.*, **38**, 1199 (2007).
13. G. Schulze, A. Jirasek, M.M.L. Yu, A. Lim, R.F.B. Turner and M.W. Blades, *Appl. Spectrosc.*, **59**, 545 (2005).
14. J.B. Cooper, K.L. Wise, W.T. Welch, R.R. Bledsoe and M.B. Sumner, *Appl. Spectrosc.*, **50**, 917 (1996).