Joint Baseline-Correction and Denoising for Raman Spectra

Hai Liu,[a,*] Zhaoli Zhang,[a,*] Sanya Liu,[a,*] Luxin Yan,[b] Tingting Liu,[a] Tianxu Zhang[b]

[a] *Central China Normal University, National Engineering Research Center for E-Learning, Wuhan, Hubei 430079, China*
[b] *Huazhong University of Science and Technology, School of Automation, Wuhan, Hubei 430074, China*

Laser instruments often suffer from the problem of baseline drift and random noise, which greatly degrade spectral quality. In this article, we propose a variation model that combines baseline correction and denoising. First, to guide the baseline estimation, morphological operations are adopted to extract the characteristics of the degraded spectrum. Second, to suppress noise in both the spectrum and baseline, Tikhonov regularization is introduced. Moreover, we describe an efficient optimization scheme that alternates between the latent spectrum estimation and the baseline correction until convergence. The major novel aspect of the proposed algorithms is the estimation of a smooth spectrum and removal of the baseline simultaneously. Results of a comparison with state-of-the-art methods demonstrate that the proposed method outperforms them in both qualitative and quantitative assessments.

Index Headings: **Raman spectroscopy; Optical data processing; Baseline correction; Denoising; Morphological operation; Regularization.**

INTRODUCTION

Baseline and random noise are common degradation problems in Raman spectra. Baseline degradation could be the result of instrument fluctuations or spurious background-signal influence; for example, the Raman spectra observed from glass often reaches a maximum around 500 cm^{-1}. The causes of random noise include experimental conditions and detector errors. The baseline and random noise may severely interfere with further processing,[1–4] complicate quantification,[3,5,6] or hinder the presentation and visualization of relevant data.[7] To overcome those problems, many baseline-correction and spectral-denoising methods have been proposed to separately improve the spectral quality. Formally, the degrading process can be described as:

$$\mathbf{g} = \mathbf{f} + \mathbf{b} + \mathbf{n} \quad (1)$$

where $\mathbf{f}$ is the ideal version of the degraded spectrum $\mathbf{g}$, $\mathbf{b}$ is the baseline, and $\mathbf{n}$ is the random noise. To perform further qualitative or quantitative analysis, we need to correct the baseline using a reasonable method. The methods found in the literature can be classified into two major groups: experimental elimination (EE) methods and mathematical estimation (ME) methods.

Received 14 October 2014; accepted 10 March 2015.
* Authors to whom correspondence should be sent. E-mail: hailiu0204@gmail.com, zl.zhang@mail.ccnu.edu.cn, lsy5918@mail.ccnu.edu.cn
DOI: 10.1366/14-07760

For the EE methods, baselines are measured using instrument techniques and eliminated from the observed spectra. O'Grady et al.[8] reported the employment of pseudo-second derivatives of subtracted Raman spectra to suppress the baseline. Osticioli et al.[9] used shift excitation difference spectroscopy and subtracted shifted Raman spectroscopy methods to reject the baseline caused by fluorescence in spectra of painting materials. More recently, excitation with specific wavelengths,[10] standard addition-based derivative spectra,[11] single-shot interferometric approach,[12] and the gas correlation method[13] have been common choices. However, these algorithms can work only for certain types of spectra, which limits the extent of their applications.

For the ME methods, baselines are automatically estimated using mathematical algorithms, and then, latent spectra are derived through subtracting them from the degraded spectra. Gerow and Rutan modeled the baseline from the abstract spectra obtained using factor analysis and an adaptive Kalman filter.[14] Perez-Pueyo et al.[15] proposed an impressive baseline-correction method based on an adaptive least squares threshold and morphology operations. Zhang et al.[16] applied the continuous wavelet transform to correct baseline drift. Peng et al.[17] proposed an effective algorithm for multiple-spectra baseline correction based on asymmetric least squares smoothing. Green et al.[18] proposed a novel method to remove low-frequency baseline variations and noise by simply omitting the approximation wavelet coefficients for further analysis. Schulze et al.[19] proposed a model-free full-baseline-removal method to correct Raman spectra. Recently, iterative methods based on curve fitting for the automatic estimation of baselines have been proposed.[20–25] These methods offer a promising approach to removing baseline effects in a simple, straightforward fashion. Nevertheless, although the estimated baseline is smooth and accurate most of the time for the methods mentioned so far, the latent spectrum seems rather noisy and the strong noise level in the degraded spectrum may lead to an unreasonable estimation of the baseline.

To overcome the disadvantages of these methods, we propose an adaptive method to estimate the baseline and denoise the degraded spectrum simultaneously. The major novel aspect of this work is the coupling of baseline correction and spectral denoising in a unified variation model that can remove the baseline as well as suppress the noise effectively. In addition, we describe an efficient optimization scheme that alternates between the latent spectrum estimation and baseline estimation until convergence. The present study is related to recent

0003-7028/15/6909-1013/0

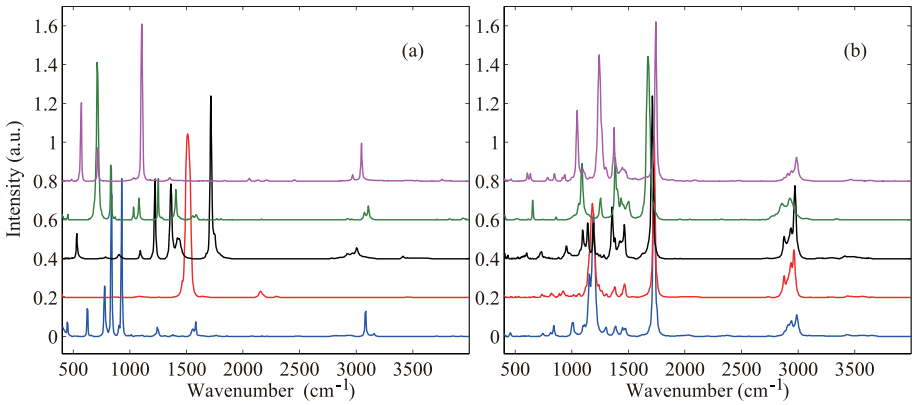

FIG. 1. Characteristics analysis of ground truth spectra. The spectra are sparse, and most positions approach zero intensity. (**a**) Ground truth spectral data from 4000 to 400 cm^{-1}. (**b**) Ground-truth spectroscopic data.

approaches in that it employs some type of regularization scheme.[26,27] It is an attempt to incorporate both baseline estimation and denoising problems into a unified framework, which has not been considered in earlier studies.

SPECTRAL CURVE CHARACTERISTICS ANALYSIS

Briefly, two key findings are revealed by the analysis. Figure 1 shows the true (ground truth) spectra from the spectral data (only 10 of them are illustrated here). In the remainder of this section, we provide a detailed analysis leading to these findings, followed by an illustration of the mathematical expressions describing these findings.

Key Finding 1. The first key finding is that the spectral baseline has the same characteristics as the degraded spectrum. The ground truth spectra are sparse (Fig. 1); that is, they have a low intensity (approach zero) in most positions. The degraded spectra have roughly the same characteristics as the spectral baseline.

Key Finding 2. The second key finding is that the baseline and spectral data are relatively smooth lines, just like the smoothed ground truth spectra shown in Fig. 1. Figure 2 presents the Raman spectrum of amidated pectin (C.P. Kelco) from 3600 to 200 cm^{-1} at 1 cm^{-1} resolution. The Raman spectrum (black line) and baseline (red line) of amidated pectin are shown in Fig. 2d. This baseline is measured by the instrument. Baselines can be removed using experimental methods that employ instrument measurement techniques to

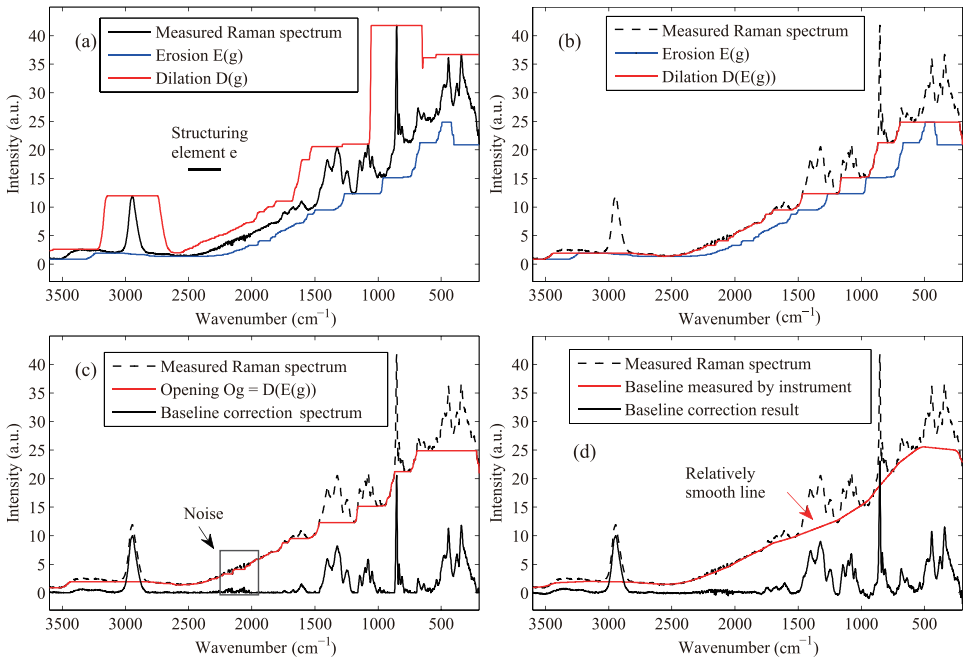

FIG. 2. Spectrum of amidated pectin, 3600–200 cm^{-1} at 1 cm^{-1} resolution, illustrating morphological operations. (**a**) Raman spectrum (black line), and erosion and dilation operation. (**b**) Opening operation result (red line), which is first processed by the erosion operation (blue line) and then the dilation operation (red line). (**c**) Baseline-correction result (black line), produced by subtracting the opening operation (red line) from the Raman spectrum (black dotted line). (**d**) Actual baseline is measured by instrument (red line). The background-estimation result (red line in (c)) is similar to the measured baseline (red line in (d)).

estimate the baseline. The erosion correction (blue line) and dilation correction (red line) are shown in Fig. 2a. The correction result of erosion and dilation is shown in Fig. 2b (red line). The corrected spectrum (black solid line) is achieved by subtracting the baseline **b** (red line in Fig. 2c) from the measured spectrum **g** (black dotted line). We observe that the baseline is a smooth line. For the corrected spectrum (black solid line in Fig. 2c), there is residual noise at the black rectangle (Fig. 2c), which should be suppressed. The roughness of the spectral signals **f** and **b** can be written as the sum of the squares of the differences, denoted R:

$$R(\mathbf{f}) = \sum_{i=0} (\mathbf{f}_i - \mathbf{f}_{i+1})^2 \tag{2}$$

where i denotes the ith spectral point in Raman spectrum. Equation 2 is often called Tikhonov regularization (TR) in spectral processing. We adopt TR to suppress noise in both the latent spectrum and the estimated baseline. Given these observations, we can extract the characteristics of the degraded spectrum using morphological operations and then use those characteristics for the baseline estimation.

Morphological operations are the foundation of the morphological image processing, which consists of a set of operators that transform images. It is often based on simple dilation and erosion operations. Perez-Pueyo et al.[15] introduced morphology-based automated baseline removal for the Raman spectra of artistic pigments. The structuring element e is defined as a line-structuring element, a window centered at $\mathbf{g}_i$ with r as half the window width. This is shown in Fig. 2a. The erosion (**E**) of the spectral line **g** is defined using:

$$\mathbf{E}(\mathbf{g}_i) = \min(\mathbf{g}_{i+j}), \quad j = -r, \ \ldots, \ r \tag{3}$$

where r denotes the neighborhood region around the current point $\mathbf{g}_i$, and j denotes the index. The dilation (**D**) of the spectral line **g** is given by:

$$\mathbf{D}(\mathbf{g}_i) = \max(\mathbf{g}_{i+j}), \quad j = -r, \ \ldots, \ r \tag{4}$$

The opening of **g**, namely **Og**, is processed by the erosion operator **E**, followed by the dilation operator **D**:

$$\mathbf{Og} = \mathbf{D}(\mathbf{Eg}) \tag{5}$$

The goal of the opening operation in Eq. 5 is to remove the peaks whose widths are less than or equal to the structuring element size. Given a suitable structuring element, the opening of a spectrum line will remove all bands narrower than the structuring element size. With the help of mathematic morphology, the characteristics of the baseline can then be roughly estimated. We find that the **Og** curve (red line in Fig. 2c) approaches the baseline (red line in Fig. 2d).

PROPOSED VARIATION MODEL

The purpose of this study is to use mathematical morphology for the automatic removal of the baseline from Raman spectra, which is mainly caused by the Raman spectrograph during the spectral measurement. Formally, given the baseline-drifted and noisy spectrum **g**, we expect the latent spectrum and baseline to be estimated as:

$$\left\{\hat{f}, \hat{b}\right\} = \arg\min_{f,b} \Phi(\mathbf{f} + \mathbf{b} - \mathbf{g}) + \beta R_{\mathbf{f}}(\mathbf{f}) + \gamma R_{\mathbf{b}}(\mathbf{b}) \tag{6}$$

where $\hat{f}$ is the latent spectrum estimation and $\hat{b}$ is the baseline estimation.

According to this analysis, both the baseline and latent spectrum are smooth, and the baseline has the same characteristics as the degraded spectrum. These significant observations motivate us to:

1. Keep the characteristics of the baseline **b** as that of the degraded spectrum **g**.
2. Penalize any roughness of the latent spectrum and baseline.

Thus, we translate motivation 1 into the following data-fidelity term:

$$\Phi(\mathbf{f} + \mathbf{b} - \mathbf{g}) = \frac{1}{2}\int (\mathbf{f}_i + \mathbf{b}_i - \mathbf{g}_i)^2 d\nu + \alpha \int (\mathbf{b}_i - \mathbf{Og}_i)^2 d\nu \tag{7}$$

where ν denotes the wavenumber, and α is the regularization parameter that balances the two terms. The first term is the conventional reconstruction constraint. The aim of the second term is to keep the characteristics of the baseline **b** similar to the degraded spectrum **g**. The morphology-based line **Og** is adopted to describe the characteristics of the degraded spectrum g in our model.

Motivation 2 leads to the regularization terms:

$$R_{\mathbf{f}}(\mathbf{f}) = \int |\nabla \mathbf{f}_i|^2 d\nu \tag{8}$$

$$R_{\mathbf{b}}(\mathbf{b}) = \int |\nabla \mathbf{b}_i|^2 d\nu \tag{9}$$

where $\nabla \mathbf{f}_i = (\mathbf{f}_{i+1} - \mathbf{f}_i)$ and $\nabla \mathbf{b}_i = (\mathbf{b}_{i+1} - \mathbf{b}_i)$.

Combining the terms in Eqs. 7–9, we construct a minimization problem with the following energy function:

$$E(\mathbf{f}, \mathbf{b}) = \frac{1}{2}\int (\mathbf{f}_i + \mathbf{b}_i - \mathbf{g}_i)^2 d\nu + \alpha \int (\mathbf{b}_i - \mathbf{Og}_i)^2 d\nu + \beta \int |\nabla \mathbf{b}_i|^2 d\nu + \gamma \int |\nabla \mathbf{f}_i|^2 d\nu \tag{10}$$

where α, β, and γ denote the regularization parameters. We call our proposed method the joint baseline-correction and denoising (JBCD) method. Note that the proposed model in Eq. 10 turns into a pure baseline-correction (PBC) method if the fourth term, $\gamma|\nabla \mathbf{f}_i|^2 d\nu$, is removed.

OPTIMIZATION AND PARAMETERS DETERMINATION

Optimization Method. There are two unknowns of **f** and **b** to be estimated. The most commonly used approach is an alternative iteration process.[28] This iterative method starts with an initial baseline $\mathbf{b}_0 = \mathbf{Og}$. First, for a fixed **b**, the energy functional is minimized

with respect to **f**; thus, the denoised spectrum is estimated. Then, for a fixed **f**, a new estimation for **b** is obtained. This procedure continues to repeat, **f** and **b** being updated in a cyclic fashion, until convergence is reached. The alternative iteration scheme can be described in two steps.

Step f. Given the baseline **b**, compute the latent spectrum **f**:

$$\hat{f} = \arg\min_{f} \left[\frac{1}{2} \int (\mathbf{f}_i + \mathbf{b}_i - \mathbf{g}_i)^2 d\nu + \gamma \int |\nabla \mathbf{f}_i|^2 d\nu \right] \quad (11)$$

We then employ the conjugate-gradient optimization method to solve this energy function minimization problem. The gradient can be obtained using:

$$\nabla E(\mathbf{f}) = \int (\mathbf{f}_i + \mathbf{b}_i - \mathbf{g}_i + 2\gamma \Delta \mathbf{f}_i) d\nu \quad (12)$$

where $\Delta \mathbf{f}_i = (\mathbf{f}_{i+1} + \mathbf{f}_{i-1} - 2\mathbf{f}_i)$. The iteration is terminated when $||\mathbf{f}^{n+1} - \mathbf{f}^n||/||\mathbf{f}^n||$ is sufficiently small with respect to a threshold T_1. The detailed optimization process has been omitted here; interested readers can refer to Nocedal and Wright[29] for more details.

Step b. Given the latent spectrum **f**, compute the baseline **b**. We drop the constant term with respect to **f** in the function in the model in Eq. 10 and obtain:

$$\hat{b} = \arg\min \left[\frac{1}{2} \int (\mathbf{f}_i + \mathbf{b}_i - \mathbf{g}_i)^2 d\nu + \alpha \int (b_i - \mathbf{Og}_i)^2 d\nu + \beta |\Delta \mathbf{b}_i|^2 d\nu \right] \quad (13)$$

Then the gradient with respect to **b** is given by:

$$\nabla E(\mathbf{b}) = \int [\mathbf{f}_i + \mathbf{b}_i - \mathbf{g}_i + 2\alpha(b_i - \mathbf{Og}_i) + 2\beta \Delta \mathbf{b}_i] d\nu \quad (14)$$

where $\mathbf{Og}_i$ is a constant vector. The conjugate-gradient optimization method is also employed to solve for the baseline **b**. The solution steps are similar to those of the **f** step.

We declare that convergence has been reached when, for more than two consecutive iterations, both the baseline and latent spectra change less than the threshold values: $||\mathbf{f}^{n+1} - \mathbf{f}^n||/||\mathbf{f}^n|| < T_1$, $||\mathbf{b}^{n+1} - \mathbf{b}^n||/||\mathbf{b}^n|| < T_2$, where T_1 and T_2 are predetermined coefficients.

Parameters Determination. To make the model in Eq. 10 work, the window size of the structuring element should first be determined. If the window size is too small, the peak shapes will change. In contrast, if the size is too large, some information about the characteristics will be omitted. To avoid the influence of the size of the structuring element, we compute the window size adaptively. In our previous study,[30] we obtained the relationship between full width half-maximum (FWHM) and the steep distance (SD), $\mathrm{FWHM} = \sqrt{2\ln 2} \times \mathrm{SD}$, for a Gaussian-shape band of any height and width. Thus, each band width can be computed adaptively using this formula. In this study, we set the window size equal to r = FWHM/2. This extracts the characteristics of the degraded spectrum successfully.

The regularization parameter α tunes up the first and second data-fidelity terms in Eq. 10, and the parameters β and γ are used to balance the data-fidelity and regularization terms (the third and fourth terms), respectively. As α is increased, the recovered baseline approaches the characteristics of the degraded spectrum **Og** more precisely. Increasing the parameter β leads to a smoother baseline but at the expense of a lower fidelity to **Og**. In addition, increasing the parameter γ leads to a smoother spectrum, but some small peaks will be removed with the noise as well. Some approaches have been developed to determine this parameter automatically, such as the discrepancy principle[31] and generalized cross-validation.[32] However, the universal applicability of these approaches for different types of spectra and regularization problems has not been effectively validated. Therefore, in this study, we have determined these parameters heuristically.

Furthermore, two parameters in our algorithm, β and γ in Eq. 10, are adjustable. They correspond to the probability parameters in Eqs. 12 and 14 for the spectrum and baseline smoothness, and their values are adapted from their initial values over the iterations of the optimization. We set $\beta = 1.2\beta$ and $\gamma = \gamma/1.1$. Then, after each iteration of optimization, the values of β and γ are updated, which can reduce the influence of the spectral smoothing and increase that of the spectrum likelihood (similarity to the original spectrum).

The JBCD method was implemented as a visual program in Matlab 2012a (The Mathworks), and it is available upon request. The algorithm described in this section can be summarized as follows.

Algorithm 1: JBCD algorithm for Raman spectra
Input degraded spectrum **g**

1: Select α, β, and γ
2: Initialize $\mathbf{f}^0 = \mathbf{g}$, $\mathbf{b}^0 = \mathbf{Og}$
3: WHILE $\{||\mathbf{f}^{n+1} - \mathbf{f}^n||/||\mathbf{f}^n|| > T_1, ||\mathbf{b}^{n+1} - \mathbf{b}^n||/||\mathbf{b}^n|| > T_2\}$
(i) Fix $\mathbf{b}^{n+1} = \mathbf{b}^n$; solve $\mathbf{f}^{n+1} = \arg\min_f E(\mathbf{f}, \mathbf{b}^n)$ using the conjugate-gradient method. Updated $\gamma = \gamma/1.1$.
(ii) Fix $\mathbf{f}^{n+1} = \mathbf{f}^n$; solve $\mathbf{b}^{n+1} = \arg\min_b E(\mathbf{f}^{n+1}, \mathbf{b})$ using the conjugate-gradient method. Updated $\beta = 1.2\beta$.

END
Output latent spectrum **f** and baseline **b**

In this study, T_1 and T_2 are small positive constants between 10^{-7} and 10^{-5}.

EXPERIMENTS AND DISCUSSION

In this section, we test the proposed estimation method using both simulated and real spectra. To investigate performance of the proposed method, we compared it with two state-of-the-art methods: the adaptive least square (ALS) method [33] and morphology-based method.[15] We also compare it to the PBC method. To quantitatively assess the corrected spectra, we employed two metrics, the root mean square error (RMSE) at peak height and the total spectrum. The

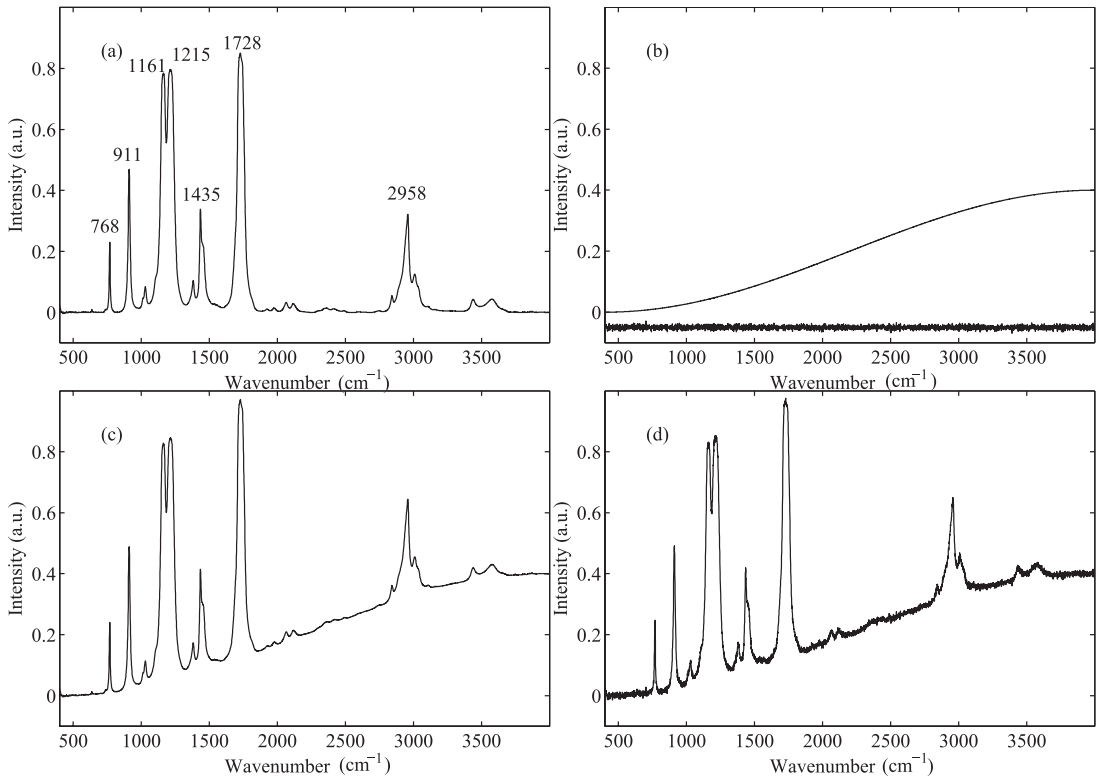

FIG. 3. Simulation of the degraded spectrum. (**a**) Infrared spectrum of methyl formate ($C_2H_4O_2$). (**b**) Baseline (sine curve) and random Gaussian noise. (**c**) Spectrum coupled with baseline curve (noise-free). (**d**) Spectrum contaminated by Gaussian noise (2%).

smaller the RMSE value is, the better the spectrum quality.

Simulated Experiments. Following Eq. 1, we simulated the degraded spectra data on the basis of the experimental infrared (IR) spectra. Figure 3 presents this simulation using the IR spectrum of methyl formate ($C_2H_4O_2$) from 400 to 4000 cm^{-1}. The original spectrum (Fig. 3a) consists of peaks with several heights (listed in Table I). The curved baseline $\mathbf{b}(\nu)$ is a sine curve (shown in Fig. 3b). To investigate the robustness to noise of the proposed methods, we added white Gaussian noise to the degraded spectrum with different noise levels (1 and 2% of the spectral intensity). The degraded spectrum without and with Gaussian noise is illustrated in Figs. 3c and 3d, respectively. The degraded spectra become highly distorted with the peak heights becoming higher, such as for peaks at 1728 and 2958 cm^{-1}. In Fig. 3d, it is difficult to extract the peak positions at 1728, 1215, and 1161 cm^{-1} because the heavy noise drifts the peak positions (the positions of the highest point).

First, we applied the ALS and JBCD methods to the noise-free degraded spectrum. Figure 4 shows the estimated baselines (Figs. 4a and 4b) and the corrected spectra (Figs. 4c and 4d) for the noise-free degraded spectrum in Fig. 3c. Both of the baseline-corrected spectra are clearly resolved. When we used the JBCD method, the baseline converged after 75 iterations. The estimated baseline appears to have good similarity to the ground truth baseline (Fig. 3b). Furthermore, for the spectra corrected using the ALS and JBCD methods, we investigated the height distortions. In the original spectrum (Fig. 3a), the seven peaks at 768, 911, 1162, 1215, 1434, 1729, and 2958 cm^{-1} were taken as references. Table I lists these peak-height distortions between the corrected (Figs. 4c and 4d) and ground truth (Fig. 3a) spectra. After we preprocessed them using ALS and JBCD, all the peak heights were closer

TABLE I. Peak heights comparison estimated by the ALS and JBCD methods for the noise-free spectrum.

	Peak positions (cm^{-1})[a]							
	768	911	1162	1215	1434	1729	2958	RMSE
Ground truth spectrum	0.230	0.469	0.785	0.797	0.338	0.850	0.321	—
Morphology-based estimation	−0.010	+0.011	+0.023	+0.027	+0.019	+0.015	−0.010	0.0465
ALS estimation	−0.009	+0.015	+0.020	+0.021	+0.020	+0.011	−0.004	0.0410
JBCD estimation	−0.007	+0.014	+0.017	+0.018	+0.019	+0.009	+0.001	0.0361

[a] "+" or "−" indicates larger or smaller distortions than the true peak height, respectively.

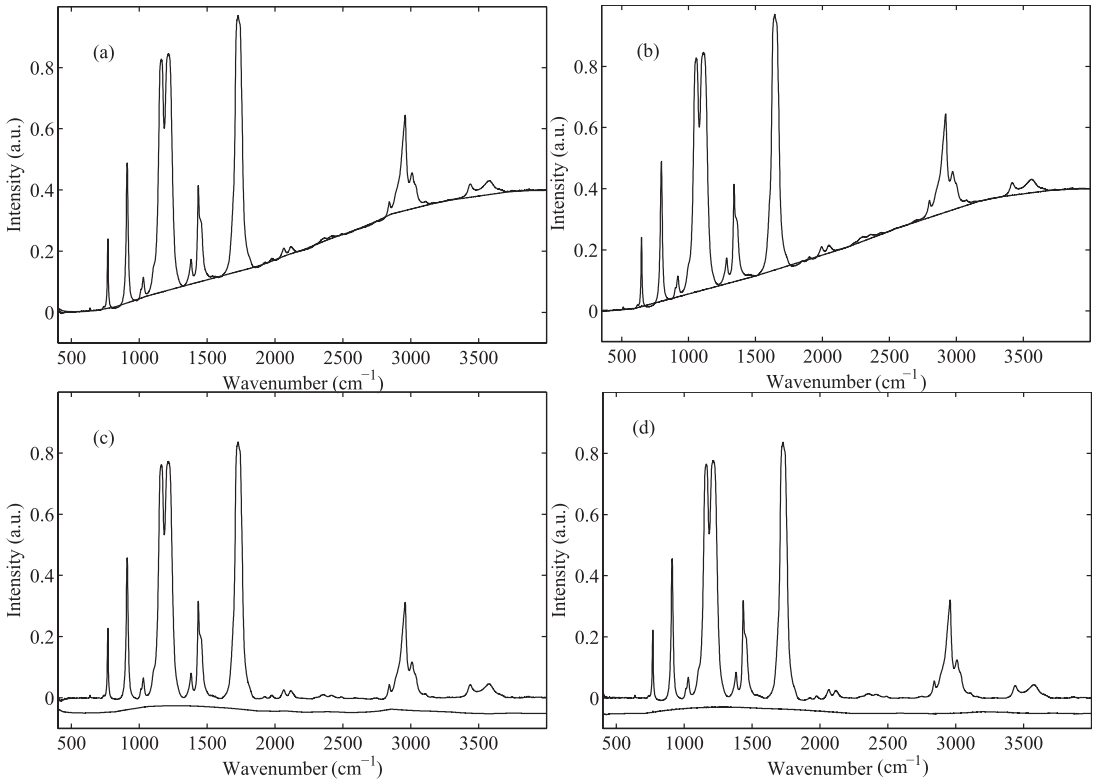

FIG. 4. Correction methods applied to degraded, noise-free spectrum in Fig. 3c. (**a**) Baseline correction. using the ALS method. (**b**) Baseline correction using JBCD method. (**c**) Latent spectrum correction using ALS method. (**d**) Latent spectrum correction using JBCD method.

to the expected height, which means that both methods can remove the baseline from spectra well. We calculated the RMSEs at the seven peaks and found that the RMSE values obtained using our method are smaller than those obtained using the ALS method. Therefore, it seems that the JBCD method produces a more accurate spectrum with more details than the ALS method.

Next, we tested the three methods on the noisy spectrum with noise levels of 1 and 2% (Fig. 3d). Figure 5 presents the baseline-corrected results using the ALS and JBCD methods with a noise level of 2%. For the ALS method, the estimated spectrum appears rather noisy, especially at a high noise level (see Fig. 5c). On the whole, the JBCD method suppresses the noise better than the ALS method. In addition, because the ALS method does not work well when there are high levels of noise, it cannot accurately distinguish noise peaks and small signal peaks. The baseline cannot be estimated accurately, especially at wide peaks, such as 2958 cm^{-1}. After we preprocessed the spectra using the morphology-based,[15] ALS, and JBCD methods, all the peak heights were closer to the expected height, which means that all three methods can remove the baseline well. We also tested the JBCD method without the fourth term of Eq. 10 (the PBC method).

The JBCD method achieved almost the best results in both wide and narrow peaks. That is, the JBCD method can correct the baseline more accurately than the ALS and morphology-based methods do while suppressing noise effectively and preserving details. Moreover, the fourth term allows the proposed method to work well for noise suppression.

The RMSEs between the ground truth and corrected spectra using each baseline-correction method are listed in Table II. Every result reported is an average of five trials for the different levels of random noise. The mean and standard deviation of the RMSE are denoted by R-RMSE and S-RMSE. As we can see, the proposed algorithm consistently provides the best RMSE values for the various levels of Gaussian noise. The simulated experiment results show that the proposed method has significantly enhanced baseline-correction performance as well as being able to suppressing random noise.

Application to Experimental Spectra. Three more challenging real examples are shown in Figs. 6–8, all of which contain baseline drift and random noise. The Raman spectra of hyoscine-N-butylbromide (HNBB) (Fig. 6), Semtex (Fig. 7), and high methoxy pectin (Fig. 8) were measured using a dispersive spectrometer using a 785 nm laser (Thorlabs). Those spectra were obtained during an investigation of hyperspectral imaging of nanomaterials with structural features that are significantly smaller than the laser-spot size (500 nm). The thickness of the polymeric shell of the ultrafine fibers was on the order of 10 nm. Because of these factors, a high level of noise exists at 2200 cm^{-1}.

The blue lines in Figs. 6a–6c, 7, and 8 represent the degraded spectra, the red lines denote the baselines

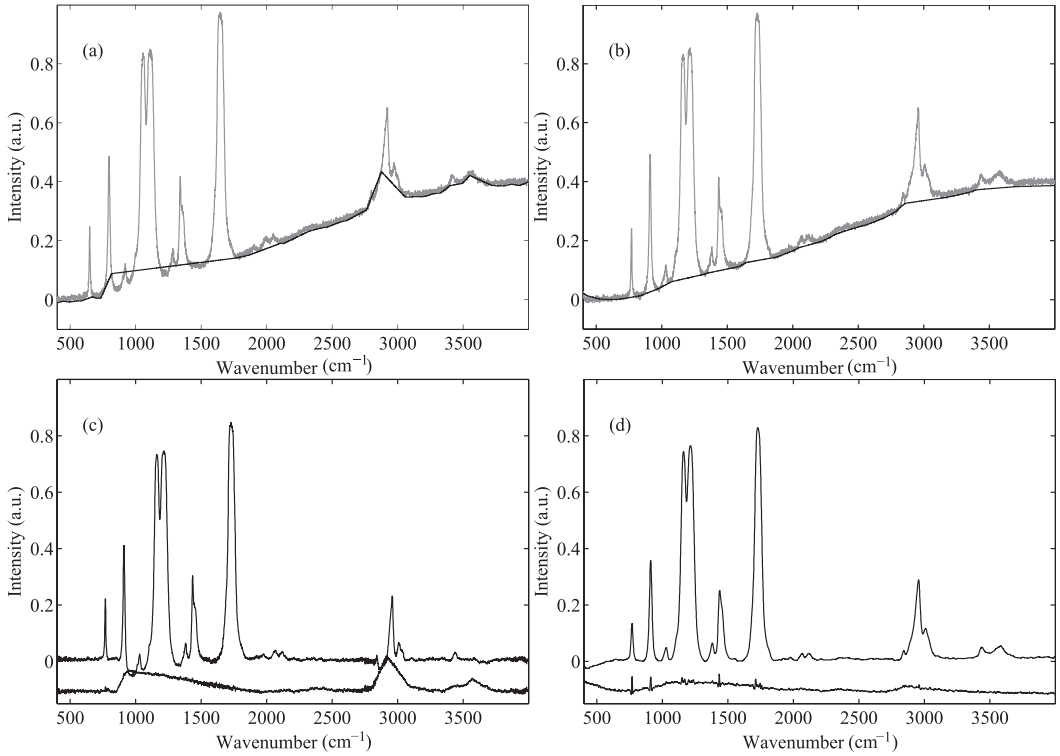

FIG. 5. Correction and estimation methods applied to the degraded noisy spectrum in Fig. 3d. Gray line represents the noisy spectrum with noise level 2%. (**a**) Baseline estimation using ALS method. (**b**) Baseline estimation using JBCD method. (**c**) Latent spectrum estimation using ALS method. (**d**) Latent spectrum estimation using JBCD method.

estimated by the different methods, and the green lines are the baseline-corrected results. The polynomial approach is the most popular method for subtracting the baseline from a spectrum in Raman spectra. This method consists of finding the polynomial that best fits the baseline.

Figure 6 shows the raw spectrum (blue), estimated baseline (red), and recovered (green) Raman spectrum of HNBB. This spectrum is a fairly simple test for the baseline subtraction procedure because of the high signal-to-noise ratio, as well as the gradually varying baseline. Figure 6a shows the baseline correction result using the fifth-order polynomial fitting method, Fig. 6b shows the results for the ALS method, and Fig. 6c shows the results for the JBCD method. Figures 6d and 6e show close-up views of the corrected spectra in Figs. 6b and 6c, respectively, from 2000 to 2500 cm^{-1}. As shown in Fig. 6b, ALS, the state-of-art method, can correct the baseline effectively. As shown in Figs. 6c and 6e, JBCD, our proposed method, can estimate the baseline accurately and also suppress the noise effectively. Because this region (2200 cm^{-1}) is enlarged in Fig. 6e, it looks like a broad band. In fact, the intensity of this band is about 1% of the original spectrum. Thus, this band can be ignored when we compare it to the original spectrum.

TABLE II. Comparison of RMSE values of the original and corrected spectra.

Noise level	Index[a]	Morphology-based	ALS	PBC	JBCD
Noise-free	M-RMSE	0.4031	0.4537	0.4802	0.3312
	S-RMSE	—	—	—	—
Noise level 1%	M-RMSE	0.5427	0.6423	0.6894	0.3923
	S-RMSE	0.0281	0.0451	0.0542	0.0213
Noise level 2%	M-RMSE	0.7142	0.8341	0.8516	0.4522
	S-RMSE	0.0310	0.0564	0.0671	0.0276

[a] M-RMSE, mean of the RMSEs; S-RMSE, standard deviation of the RMSEs.

The baseline-corrected Raman spectrum of Semtex is displayed Fig. 7. This spectrum poses a greater challenge in that the relatively weak Raman spectrum is superimposed on an intense polynomial-like baseline. In spite of this, the JBCD method (Fig. 7c) succeeded in correcting the baseline to give a more reasonable result than that of the ALS method (Fig. 7b) and polynomial fitting (Fig. 7a). The green lines show the expanded plots of the final corrected spectrum. As we can see, the ALS algorithm is not robust to noise and introduces some negative points (such as 1250 cm^{-1}) in the recovered result (Fig. 7b), whereas the JBCD method corrects the degraded spectrum to positive bands (Fig. 7c). Furthermore, from 1800 to 2400 cm^{-1}, the corrected baseline in Fig. 7c is smoother than the one in Fig. 7b. We conclude that the JBCD method can remove the baseline in a reasonable way in addition to suppressing the noise well.

Figure 8 shows the Raman spectrum of high methoxy pectin from 400 to 3600 cm^{-1} (blue lines). The ALS method (Fig. 8b) estimates the baseline accurately for most bands. On the whole, the latent spectrum produced using the JBCD method (Fig. 8c) appears smoother than those in Figs. 8a and 8b. That is, unlike the other two methods, the JBCD method suppresses noise well from 2000 to 2500 cm^{-1}. Furthermore, using

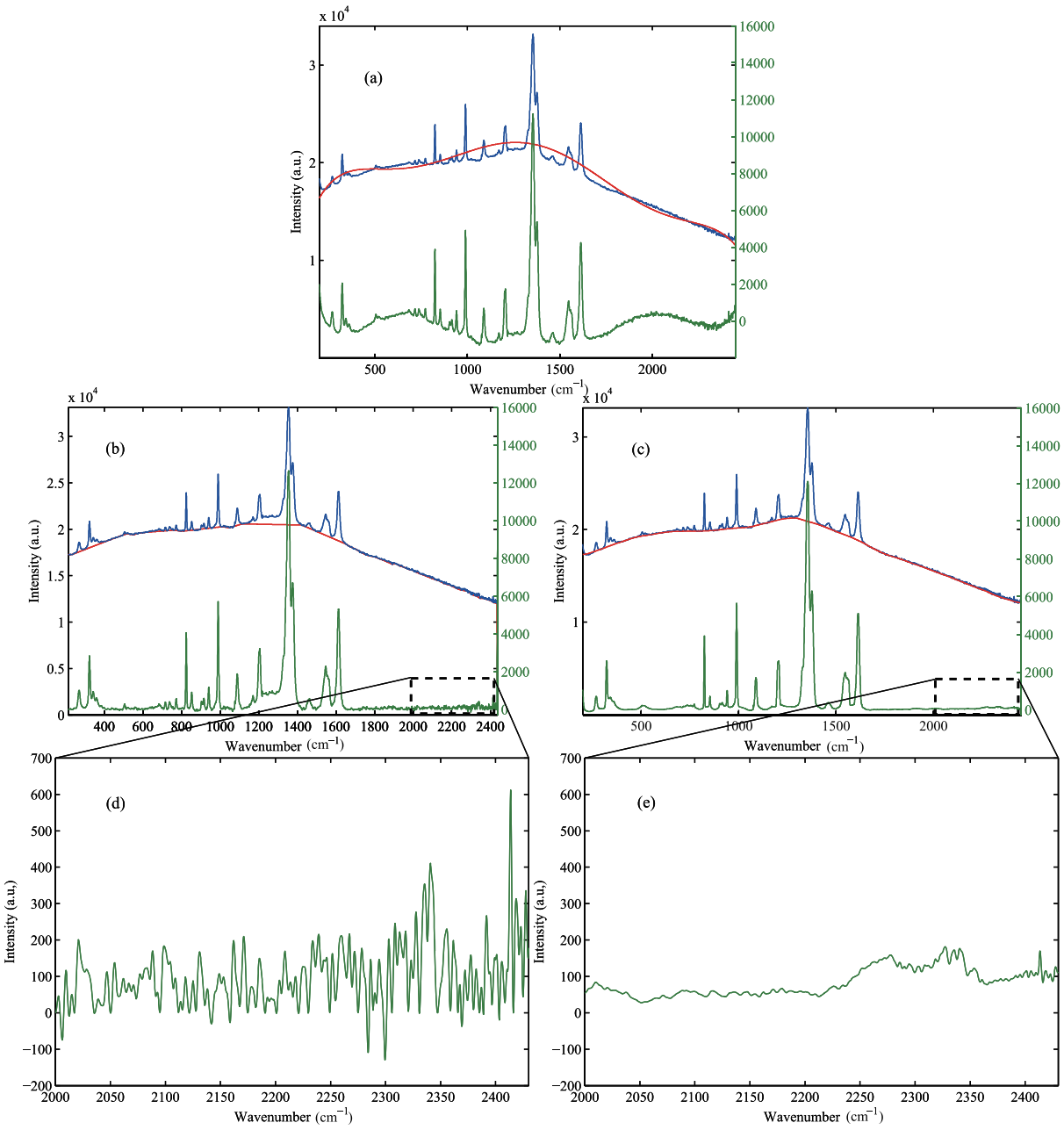

FIG. 6. Baseline correction of Raman spectrum of HNBB. (**a**) Using fifth-order polynomial fitting method. (**b**) Using ALS method. (**c**) Using JBCD method. (**d**) Residual noise in (b) (close up from 2000 to 2500 cm^{-1}). (**e**) Residual noise in (**c**) (close up from 2000 to 2500 cm^{-1}). In (**a**)–(**c**), blue lines, degraded spectrum; red lines, estimated baselines; green lines, baseline-corrected results. The raw data were provided by Professor Weakley.

the proposed JBCD method, the baseline profile we obtain is much more reasonable. In Fig. 8b, for the signal from 700 to 1200 cm^{-1}, which has a high noise level, part of the latent spectrum is interpreted as the baseline by the ALS method, leading to an unreasonable result. The computing time for these examples was between 12 and 25 s.

Finally, to demonstrate the effectiveness of the JBCD method in chemometric analyses, we introduced principal component analysis (PCA) to process the results. We first used the PBC, ALS, and JBCD methods on the same real spectra. Then we processed their results using PCA. The first 10 principal components (PCs) and their corresponding normalized eigenvalues are plotted in the Fig. 9. We can see that the JBCD algorithm can better distinguish the PCs from the minor components. In other words, the results using the JBCD method are more suitable for extracting the spectral features of and identifying unknown chemical mixtures.

CONCLUSION

In this article, a unified variation model that couples baseline-correction with spectral denoising and demonstrate its applications on real Raman spectra and IR spectra was proposed. By combining morphological operation and Tikhonov regularization, the proposed

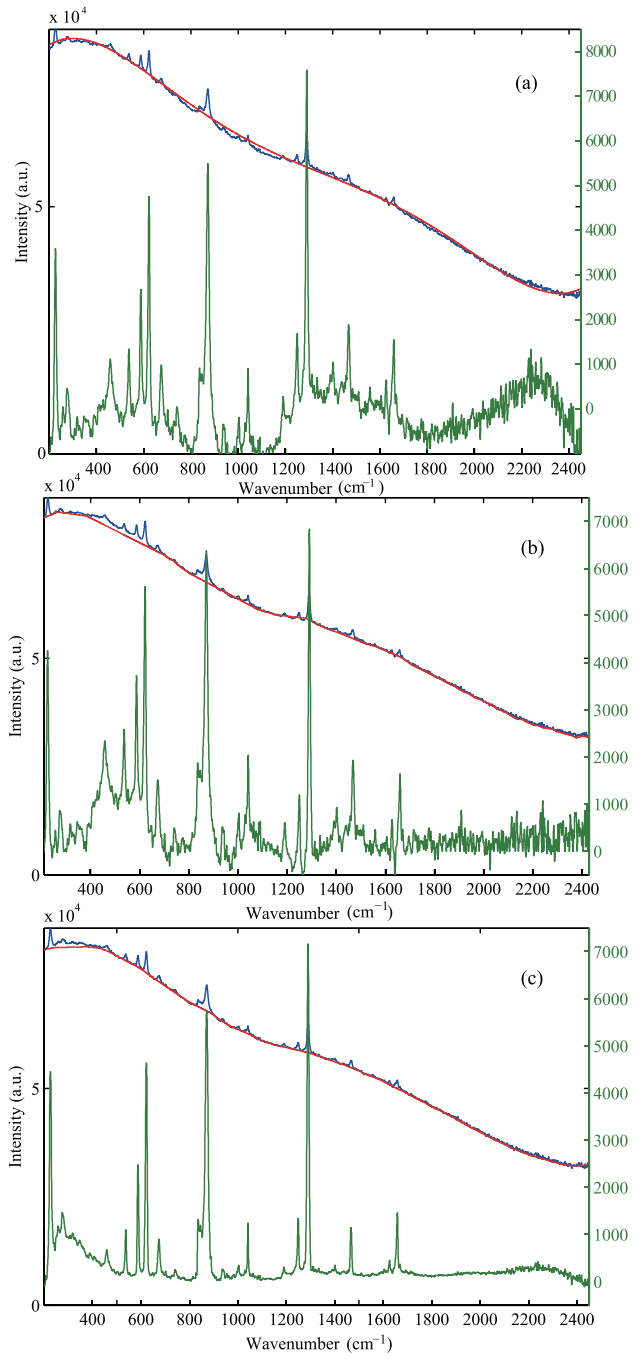

FIG. 7. Baseline correction of the Raman spectrum of Semtex. (**a**) Using a fifth-order polynomial fitting method. (**b**) Using the ALS method. (**c**) Using the JBCD method. Blue lines, degraded spectrum; red lines, estimated baselines; green lines, baseline-corrected results.

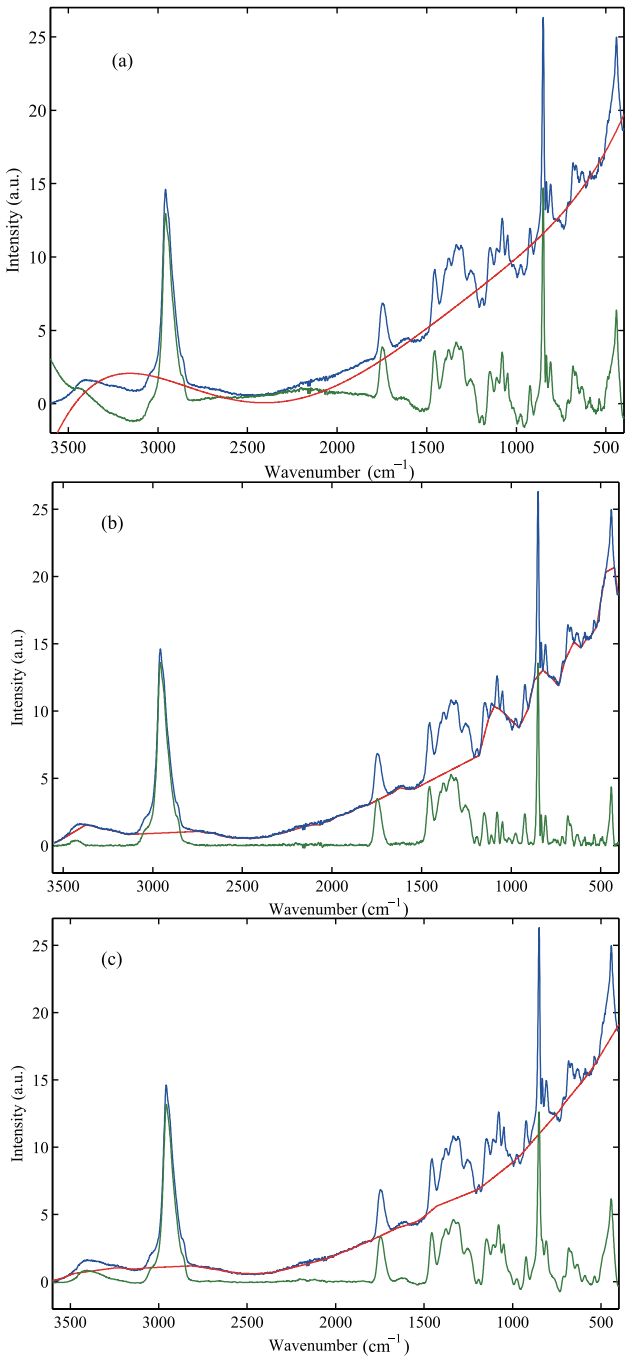

FIG. 8. Baseline correction of Raman spectrum of high methoxy pectin. (**a**) Using fifth-order polynomial fitting method. (**b**) Using LS method. (**c**) Using JBCD method. Blue lines, degraded spectrum; red lines, estimated baselines; green lines, baseline-corrected results.

model could estimate well the latent spectra and baseline simultaneously. Both visual inspection and quantitative evaluation show that the proposed method performs quite well in different cases. The Raman spectra processed using the JBCD method are more suitable for extracting the spectral features of and identifying unknown chemical mixtures. Although only IR and Raman spectra have been investigated in this article, the method can be applied to other analytical instrument signals, such as chromatograms and fluorescence spectra. Baseline-corrected spectra are commonly used for building a clustering, classification, or regression model, which we will examine in future work.

ACKNOWLEDGMENTS

The authors thank the editor and anonymous reviewers for their valuable suggestions. This research was partially funded by the National Social Science Fund of China (14BGL131), the National Natural Science Foundation of China under Grant No. 61505064, 60902060, the Project of the Program for National Key Technology Research and Development Program (2013BAH72B01, 2013BAH18F02, 2015BAH33F02), the Self-Determined Research Funds of CCNU from

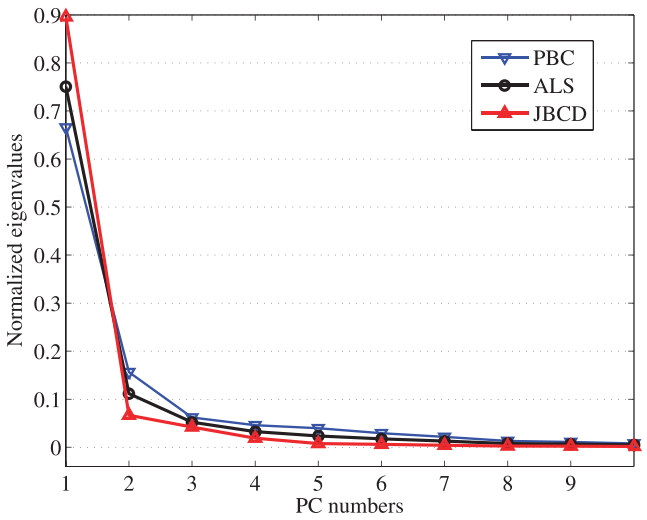

FIG. 9. First 10 PCs and their corresponding normalized eigenvalues using the PBC, ALS, and JBCD methods.

the Colleges' Basic Research and Operation of MOE (CCNU15A05009 and CCNU15A05010), Scientific R&D Project of the State Education Ministry and China Mobile (MCM20121061), National Social Science Fund of China (14BGL131), New PhD Researcher Award from the Chinese Ministry of Education, New Century Excellent Talents in University (NCET-11-0654). The authors thank Professor Andrew T. Weakley, Professor Peter R. Griffiths, and C.P. Kelco for providing the raw Raman spectra.

1. A.L. Gallego, A.R. Guesalaga, E. Bordeu, A.S. González. "Rapid Measurement of Phenolics Compounds in Red Wine Using Raman Spectroscopy". IEEE T. Instrum. Meas. 2011. 60(2): 507-512.
2. H. Liu, T. Zhang, L. Yan, H. Fang, Y. Chang. "A MAP-Based Algorithm for Spectroscopic Semi-Blind Deconvolution". Analyst. 2012. 137(16): 3862-3873.
3. D. Chang, C.D. Banack, S.L. Shah. "Robust Baseline Correction Algorithm for Signal Dense NMR Spectra". J. Magn. Reson. 2007. 187(2): 288-292.
4. H. Liu, S. Liu, Z. Zhang, J. Sun, J. Shu. "Adaptive Total Variation-Based Spectral Deconvolution with the Split Bregman Method". Appl. Opt. 2014. 53(35): 8240-8248.
5. J. Zhao, H. Lui, D.I. McLean, H. Zeng. "Automated Autofluorescence Background Subtraction Algorithm for Biomedical Raman Spectroscopy". Appl. Spectrosc. 2007. 61(11): 1225-1232.
6. H. Liu, Z. Zhang, S. Liu, T. Liu, L. Yan, T. Zhang. "Richardson-Lucy Blind Deconvolution of Spectroscopic Data with Wavelet Regularization". Appl. Opt. 2015. 54(7): 1770-1775.
7. Y.V. Karpievitch, E.G. Hill, A.J. Smolka, J.S. Morris, K.R. Coombes, K.A. Baggerly, J.S. Almeida. "PrepMS: TOF MS Data Graphical Preprocessing Tool". Bioinformatics. 2007. 23(2): 264-265.
8. A. O'Grady, A.C. Dennis, D. Denvir, J.J. McGarvey, S.E.J. Bell. "Quantitative Raman Spectroscopy of Highly Fluorescent Samples Using Pseudosecond Derivatives and Multivariate Analysis". Anal. Chem. 2001. 73(9): 2058-2065.
9. I. Osticioli, A. Zoppi, E.M. Castellucci. "Shift-Excitation Raman Difference Spectroscopy Difference Deconvolution Method for the Luminescence Background Rejection from Raman Spectra of Solid Samples". Appl. Spectrosc. 2007. 61(8): 839-844.
10. B. Auguié, A. Reigue, E.C. Le Ru, P.G. Etchegoin. "Tiny Peaks vs Mega Backgrounds: A General Spectroscopic Method with Applications in Resonant Raman Scattering and Atmospheric Absorptions". Anal. Chem. 2012. 84(18): 7938-7945.
11. N. Li, X.-Y. Li, Z.-X. Zou, L.-R. Lin, Y.-Q. Li. "A Novel Baseline-Correction Method for Standard Addition Based Derivative Spectra and Its Application to Quantitative Analysis of Benzo(a)pyrene in Vegetable Oil Samples". Analyst. 2011. 136(13): 2802-2810.
12. Y.J. Lee, M.T. Cicerone. "Single-Shot Interferometric Approach to Background Free Broadband Coherent Anti-Stokes Raman Scattering Spectroscopy". Opt. Express. 2009. 17(1): 123-135.
13. X. Lou, G. Somesfalean, S. Svanberg, Z. Zhang, S. Wu. "Detection of Elemental Mercury by Multimode Diode Laser Correlation Spectroscopy". Opt. Express. 2012. 20(5): 4927-4938.
14. D.D. Gerow, S.C. Rutan. "Background Correction for Fluorescence Detection in Thin-Layer Chromatography Using Factor Analysis and the Adaptive Kalman Filter". Anal. Chem. 1988. 60(9): 847-852.
15. R. Perez-Pueyo, M.J. Soneira, S. Ruiz-Moreno. "Morphology-Based Automated Baseline Removal for Raman Spectra of Artistic Pigments". Appl. Spectrosc. 2010. 64(6): 595-600.
16. Z.-M. Zhang, S. Chen, Y.-Z. Liang, Z.-X. Liu, Q.-M. Zhang, L.-X. Ding, F. Ye, H. Zhou. "An Intelligent Background-Correction Algorithm for Highly Fluorescent Samples in Raman Spectroscopy". J. Raman Spectrosc. 2010. 41(6): 659-669.
17. J. Peng, S. Peng, A. Jiang, J. Wei, C. Li, J. Tan. "Asymmetric Least Squares for Multiple Spectra Baseline Correction". Anal. Chim. Acta. 2010. 683(1): 63-68.
18. G.C. Green, A.D.C. Chan, R.A. Goubran, B.S. Luo, M. Lin. "A Rapid and Reliable Method of Discriminating between Listeria Species Based on Raman Spectroscopy". In: 2008 IEEE International Instrumentation and Measurement Technology Conference (I^2MTC 2008) Proceedings. Piscataway, NJ: IEEE, 2008. Pp. 513-517.
19. H.G. Schulze, R.B. Foist, K. Okuda, A. Ivanov, R.F.B. Turner. "A Model-Free, Fully Automated Baseline-Removal Method for Raman Spectra". Appl. Spectrosc. 2011. 65(1): 75-84.
20. H. Liu, Z. Zhang, J. Sun, S. Liu. "Blind Spectral Deconvolution Algorithm for Raman Spectrum with Poisson Noise". Photon. Res. 2014. 2(6): 168-171.
21. M.N. Leger, A.G. Ryder. "Comparison of Derivative Preprocessing and Automated Polynomial Baseline Correction Method for Classification and Quantification of Narcotics in Solid Mixtures". Appl. Spectrosc. 2006. 60(2): 182-193.
22. F. Gan, G. Ruan, J. Mo. "Baseline Correction by Improved Iterative Polynomial Fitting with Automatic Threshold". Chemom. Intell. Lab. Syst. 2006. 82(1–2): 59-65.
23. C.A. Lieber, A. Mahadevan-Jansen. "Automated Method for Subtraction of Fluorescence from Biological Raman Spectra". Appl. Spectrosc. 2003. 57(11): 1363-1367.
24. P.J. Cadusch, M.M. Hlaing, S.A. Wade, S.L. McArthur, P.R. Stoddart. "Improved Methods for Fluorescence Background Subtraction from Raman Spectra". J. Raman Spectrosc. 2013. 44(2): 1587-1595.
25. H. Liu, M. Zhou, Z. Zhang, J. Shu, T. Liu, T. Zhang. "Multi-Order Blind Deconvolution Algorithm with Adaptive Tikhonov Regularization for Infrared Spectroscopic Data". Infrared Phys. Technol. 2015. 71: 63-69.
26. P.H.C. Eilers, H.F.M. Boelens. "Baseline Correction with Asymmetric Least Squares Smoothing". 2005. http://zanran_storage.s3.amazonaws.com/www.science.uva.nl/ContentPages/443199618.pdf [accessed 12 Jun 2015].
27. Z.-M. Zhang, S. Chen, Y.-Z. Liang. "Baseline Correction Using Adaptive Iteratively Reweighted Penalized Least Squares". Analyst. 2010. 135(5): 1138-1146.
28. Y. Yu-Li, M. Kaveh. "A Regularization Approach to Joint Blur Identification and Image Restoration". IEEE Trans. Image Process. 1996. 5(3): 416-428.
29. J. Nocedal, S.J. Wright. Numerical Optimization. New York: Springer, 2006.
30. H. Liu, L. Yan, Y. Chang, H. Fang, T. Zhang. "Spectral Deconvolution and Feature Extraction with Robust Adaptive Tikhonov Regularization". IEEE T. Instrum. Meas. 2013. 62(2): 315-327.
31. H.W. Engl, M. Hanke, A. Neubauer. "Continuous Regularization Methods". In, Mathematics and Its Applications: Regularization of Inverse Problems. Dordrecht: Kluwer Academic, 1996. Pp. 83-88.
32. H. Liao, M.K. Ng. "Blind Deconvolution Using Generalized Cross-Validation Approach to Regularization Parameter Estimation". IEEE Trans. Image Process. 2011. 20(3): 670-680.
33. A.T. Weakley, P.R. Griffiths, D.E. Aston. "Automatic Baseline Subtraction of Vibrational Spectra Using Minima Identification and Discrimination via Adaptive, Least-Squares Thresholding". Appl. Spectrosc. 2012. 66(5): 519-529.