Automatic baseline correction of infrared spectra

Tiange Lan (兰天鸽), Yonghua Fang (方勇华), Wei Xiong (熊 伟), and Chao Kong (孔 超)

Remote Sensing Laboratory, Anhui Institute of Optics and Fine Mechanics, Chinese Academy of Sciences, Hefei 230031

Received April 25, 2007

A fast automatic algorithm is proposed for baseline correction of infrared (IR) spectral signals. It is devised based on iterative curve fitting where orthogonal polynomials are used. The algorithm can process both emission and absorption spectra automatically without human intervention. Orthogonal polynomials are used for curve fitting to reduce computation time. Both emission and absorption spectra are obtained and the results demonstrate the feasibility and practicability of this algorithm.

OCIS codes: 300.6340, 120.0280, 070.6020, 280.1120.

There is an important need for remote monitoring of air pollutants and greenhouse gases, as well as the remote detection and identification of toxic industrial compounds. Infrared (IR) spectroscopy is effective for the identification of these chemical species, since many of them show spectral features in the mid-IR spectral region. Feasibility studies on the use of passive Fourier transform infrared (FTIR) spectroscopy for remote monitoring of these compounds have shown that gases may be identified remotely with an optically modified commercial FTIR spectrometer when only a small (7 °C) temperature difference exists between the gas and background. However, it is the common problem that there is a baseline for the IR spectra collected, thus hampering the identification and quantitative analysis. It should be preprocessed from the original signal, so as to establish a stable, trustable qualitative or quantitative analysis model.

Schulze *et al.*[1] provided an extensive literature review and comparison of a wide range of preprocessing methods for the removal of typical baseline, without focusing on a particular instrumental method. Derivative method[2] is likely the best known baseline removal method. But it is often inadequate for proper removal of the baseline. The size of the spectral window used in the transformation must be selected carefully: a wide window can distort peaks, while a narrow window will amplify high frequency noise. Perhaps the biggest disadvantage of using derivative method is that the filtered spectra do not have the same appearance as the original spectra.

Another processing strategy is based on the frequency analysis of the spectra. It is believed that baseline is a low frequency signal, while the signal peaks will be middle-frequency features, and independent noise will be found at all frequencies. Baseline can be eliminated from the original spectral signal by treating the low frequency information in the frequency domain. Fourier transfer and wavelet transfer[3,4] are the general instruments of this method. However, both cannot give a theoretic description of the baseline information. It is difficult to clearly distinguish baseline information from others for Fourier transfer. And discouraging point about wavelet transforms is the wide range of parameters to be set (basis functions, transformation levels, and coefficients to remove), which can make the laborious process optimize.

Curve fitting is another well known method for baseline correction. Simply, a straight line or a fitted curve is taken as baseline, and then subtracted from the original spectrum. In fact, the straight line or the fitted curve often does not estimate the real baseline well. Recently, an iterative method based on curve fitting for automated estimation of baseline is proposed[2,5−7]. This method offers a promising approach to remove baseline effects in a simple, straightforward fashion.

In this paper, the iterative curve fitting method is amended for baseline correction of passive FTIR signals which are used for real-time monitoring of air pollutants. Because of the real-time application and the removal of baseline is but one step in a larger analysis scheme, including transforming interferograms to spectra, denoising, pattern recognition etc., the required time to correct a given spectrum should be as short as possible. Besides, the algorithm should process both emission and absorption spectra automatically without human intervention, because all these procedures are implemented on hardware. The improved method first identifies emission or absorption features of the spectrum at hand, and then adopts different processing strategies. Orthogonal polynomial is used for curve fitting to reduce computation time, because curve fitting with higher power needs much more time for computing matrix inversion and also leads to larger computational error.

This algorithm generates automatic threshold by curve fitting. First, the original signal is fitted following least square criterion. When the original signal has a complicated form, it can be predicted that the fitted result does not agree with the original signal well on the whole curve. At the sharp peak regions, the fitted result departs strongly from the original curve, and so it serves as the automatic thresholds. By substituting parts of the peaks for the automatic threshold, a truncated curve can be obtained, which is then used in further consecutive fitting. The iterative processes of fitting and truncating do not stop until a good estimate of the baseline is obtained. At the end, the baseline is subtracted from original signal to achieve the purpose of baseline correction.

As far as IR signal concerned, there are two classes of spectra dependent on the relative temperature of the measured target and the background behind the target. Infrared detection is the response of the thermal radiation. If the temperature of target is higher than that of background, the IR signal obtained is an emission spectrum. On the contrast, if the temperature of target is lower than that of background, the IR signal obtained is an absorption spectrum. The algorithm must process two kinds of spectra automatically. Emission spectrum has feature peaks upwards while absorption spectrum has feature peaks downwards. The truncation criterion of upwards peaks is different from that of downwards peaks. In the first case, the points larger than the threshold are substituted while the points smaller than the threshold are changed when the peaks are downwards. So, when loading an original signal, algorithm should judge the direction of the peaks first.

Based on the discussion above, an algorithm for baseline correction is proposed as follows. Step 0: Load the original spectral signal and identify the direction of feature peaks; Step 1: Get a fitted curve by fitting the original signal curve with the least square criterion; Step 2: The fitted curve serves as automatic threshold and getting a truncated curve by truncating the original signal with different truncation criteria with respect to emission or absorption spectrum; Step 3: Get a newly fitted curve by fitting the truncated curve; if the fitted curve superposes with the last one, go to step 4; otherwise, go to step 2; Step 4: Take the fitted curve as the estimated baseline, and it is subtracted from the original signal, stop.

The termination of the iterative process is based on the following consideration: as long as the spectrum contains peaks or components of the peaks, the fitted curve lies between the real baseline and the peaks at the peak regions, and then some points are truncated by the automatic threshold. Newly fitted curve by fitting the truncated curve is different from the last one. It means that the peak components are removed step by step and the fitted curve approaches to the real baseline step by step. When the truncated signal includes few components of the peaks, newly fitted curve almost equals the former. Then the iteration is over. The fitted curve which contains no peak components is the baseline. Generally, 3 to 5 iterations are enough for an estimation process.

The principle of curve fitting is introducing a new function $f(x)$ to approximate the original signal $y(x)$, which can be expressed as

$$\begin{aligned} y(x) = c_0T_0(x) + c_1T_1(x) + c_2T_2(x) + \cdots \\ + c_nT_n(x) + e(x), \end{aligned} \tag{1}$$

where c_0, c_1, c_2, $\cdots$, c_n are the coefficients to be determined. $T_i(x)$ is function of x. n is the number of the basis function, and $e(x)$ is fitting error. To a spectrum with m data points, x_1, x_2, $\cdots$, x_m are the wavelength values, Eq. (1) can be written in matrix form as

$$\begin{bmatrix} y(x_1) \\ y(x_2) \\ y(x_3) \\ \cdots \\ y(x_m) \end{bmatrix} = \begin{bmatrix} T_0(x_1) & T_1(x_1) & T_2(x_1) & \cdots & T_n(x_1) \\ T_0(x_2) & T_1(x_2) & T_2(x_2) & \cdots & T_n(x_2) \\ T_0(x_3) & T_1(x_3) & T_2(x_3) & \cdots & T_n(x_3) \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ T_0(x_m) & T_1(x_m) & T_2(x_m) & \cdots & T_n(x_m) \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \\ c_2 \\ \cdots \\ c_n \end{bmatrix} + \begin{bmatrix} e(x_1) \\ e(x_2) \\ e(x_3) \\ \cdots \\ e(x_m) \end{bmatrix}, \tag{2}$$

it can be written in the following concise matrix form,

$$\mathbf{y} = \mathbf{Tc} + \mathbf{e}. \tag{3}$$

The vector $\mathbf{c}$ of fitting coefficients can be calculated by

$$\mathbf{c} = (\mathbf{T}^{\mathrm{t}}\mathbf{T})^{-1}\mathbf{T}^{\mathrm{t}}\mathbf{y}, \tag{4}$$

which follows the least square method by making the sum square of $\mathbf{e}$ minimum. Then, the fitting function is

$$f(x) = T(T^{\mathrm{t}}T)^{-1}T^{\mathrm{t}}y, \tag{5}$$

where superscripts t and -1 denote the matrix transpose and matrix inversion, respectively. Mexican Hat wavelet and polynomial have been taken as T in Eq. (5)[5-7]. However, when n is large enough (i.e., $n \geq 7$), the calculation of matrix inversion $(T^{\mathrm{t}}T)^{-1}$ leads to large computational error and needs much more time for computation, which is unacceptable in the case of real time application. If the orthogonal polynomial is used for curve fitting, this problem can be resolved easily. The basic property of the orthogonal polynomials is

$$T_{\mathrm{r}}(x)^{\mathrm{t}}T_{\mathrm{s}}(x) = 0, \quad r \neq s. \tag{6}$$

Because of this property, orthogonal polynomial yields diagonal normal matrices $T^{\mathrm{t}}T$. Consequently, matrix inversion $(T^{\mathrm{t}}T)^{-1}$ can be calculated easily and precisely.

The equations used to generate the orthogonal polynomials are

$$\begin{aligned} T_0(x) &= 1, \\ T_1(x) &= x - \bar{x}, \\ &\vdots \\ T_n(x) &= x^n + k_{n,n-1}T_{n-1}(x) + k_{n,n-2}T_{n-2}(x) + \cdots \\ &\quad + k_{n,0}T_0(x), \end{aligned} \tag{7}$$

$$\begin{aligned} k_{n,j} &= -\sum_{i=1}^{m} x_i^n T_j(x_i) \Big/ \sum_{i=1}^{m} T_j^2(x_i), \\ j &= n-1, n-2, \cdots, 1, 0, \end{aligned} \tag{8}$$

where $\bar{x}$ is the mean of x.

So far, an algorithm is proposed for baseline correction of IR spectral signal by iterative orthogonal polynomial curve fitting with automatic threshold. It is tested with several passive FTIR signals which are measured with spectrometer designed by us. Interferograms are transformed into spectra and background spectra are subtracted before baseline correction. The range is from 1300 to 700 cm^{-1} and the resolution is about 2 cm^{-1}. The only parameter in this algorithm is the power of the orthogonal polynomial, which should be set appropriately. If the power is too large, the useful parts of the signal may be fitted into baseline and then subtracted from the original signal. Oppositely, if a smaller power is used, baseline can not be estimated correctly. Figure 1 shows an original signal (solid line) and the baseline (dashed line) estimated with the algorithm proposed in this paper. The power of the orthogonal polynomial is 8. Obviously, there are two big peaks near 780 and 940 cm^{-1} after baseline correction, which are shown in Fig. 2.

Figure 3 shows the same original signal as that of Fig. 1, and the estimated baseline is also shown with the dashed line. Power of the orthogonal polynomial is 16. It can be seen that almost all the original signals are fitted into baseline. The signal after baseline correction is shown in Fig. 4, which has no obvious feature peak. The two results are different because powers of the orthogonal polynomial used in each estimate are different. So, the power of the orthogonal polynomial should be set appropriately based on the feature peaks and the real baseline of the signal will be processed. Luckily, the selection of the power is not so difficult because the spectra are clear and the power is robust (e.g., selecting 8 or 7 as the power has no evident influence on the result of baseline correction which serves as the preprocessing step of pattern recognition in this work).

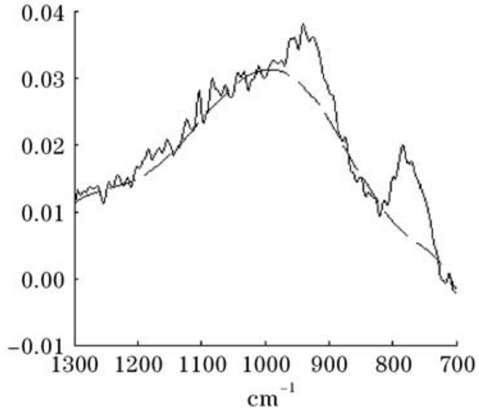

Fig. 1. Original signal (solid line) and estimated baseline (dashed line). The power for the polynomial is 8.

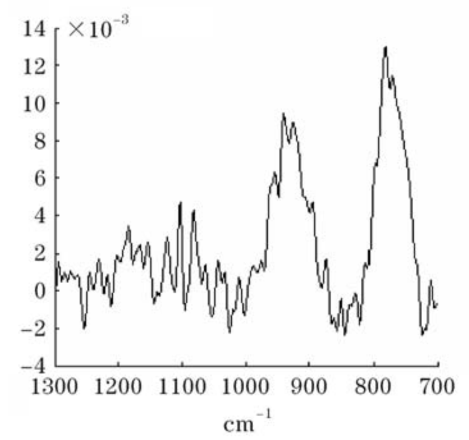

Fig. 2. Corrected spectral signal of Fig. 1.

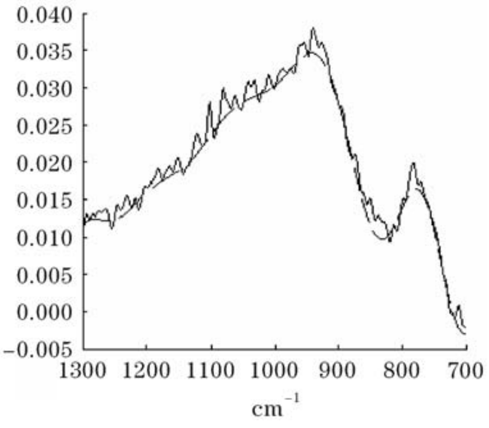

Fig. 3. Same original signal as Fig. 1 (solid line) and estimated baseline (dashed line). The power for the polynomial is 16.

Figure 5 shows a representative IR spectral signal in the range of $1300 - 700$ cm^{-1}. It is obvious that there is a baseline. The baseline must be corrected before a further analysis is operated. The dashed line is the fitted curve of the original signal by orthogonal polynomial least square curve fitting. The power of the orthogonal polynomial is 8. The value of the fitted result serves as the automatic threshold. By truncating the parts of the peaks that is larger than the automatic threshold, a truncated curve can be obtained, which is shown in Fig. 6. Then the truncated curve is used in further consecutive fitting. The iterative process of fitting and truncating does not stop until a good estimate of the baseline is obtained. Figure 7 shows the final estimated baseline. At the end, the baseline is subtracted from original signal to achieve the purpose of baseline correction. Figure 8 shows the

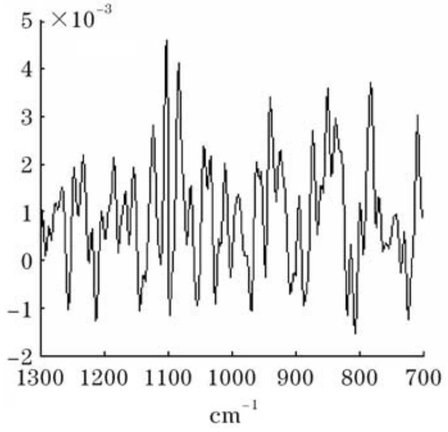

Fig. 4. Corrected spectral signal of Fig. 3.

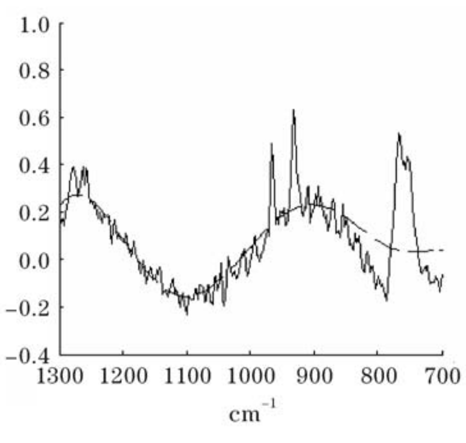

Fig. 5. Original signal and the first fitted curve (dashed line).

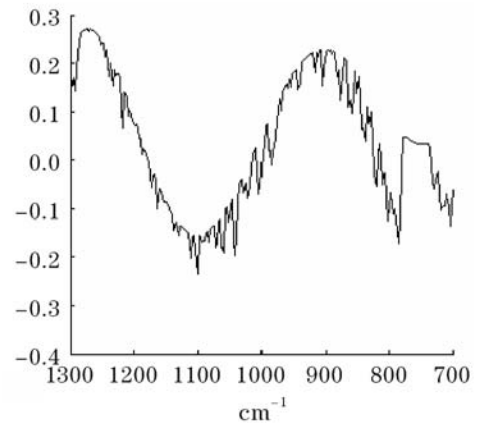

Fig. 6. Truncated curve of Fig. 5.

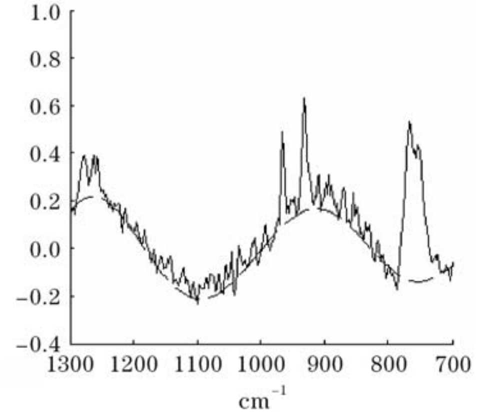

Fig. 7. Final estimated baseline (dashed line).

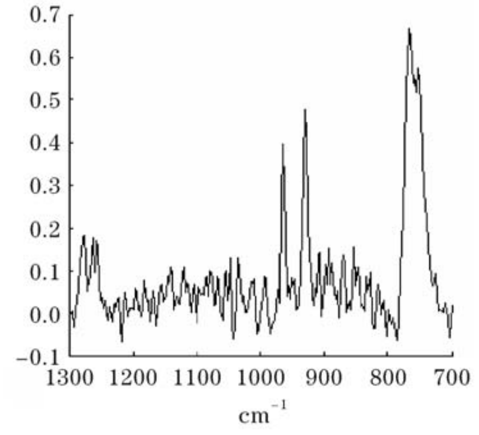

Fig. 8. Corrected spectral signal of Fig. 5.

corrected spectral signal.

Figure 9 shows another type of IR spectral signal. It is an absorption spectrum and the feature peaks are downwards. The truncation criterion of downwards peaks is different from that in Fig. 5. The parts of peaks smaller than the fitted result which serves as automatic threshold are truncated. Dashed line is the final estimated baseline. The power for the orthogonal polynomial is also 8. And the corrected spectral signal is shown in Fig. 10.

In this paper, the algorithm for baseline correction of IR spectral signal is devised based on the characteristics of IR spectrum and the need of real time application.

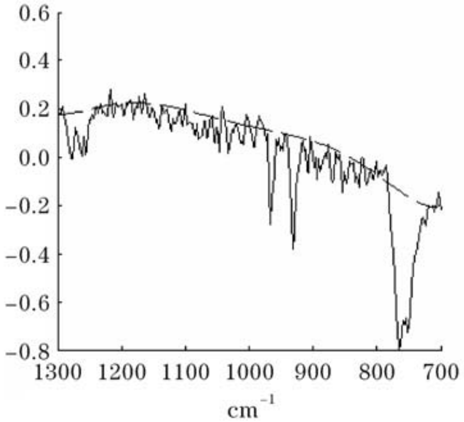

Fig. 9. Original signal and final estimated baseline (dashed line). The power for the polynomial is 8.

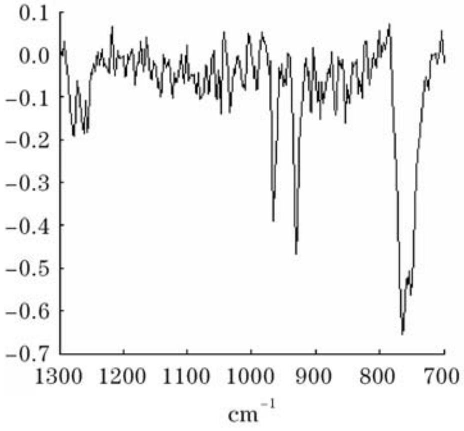

Fig. 10. Corrected spectral signal of Fig. 9.

Essentially, it can be expressed as iterative fitting with automatic threshold. The advantages of this method are that it is a robust method with only one parameter and can process both emission and absorption spectra automatically without human intervention. Orthogonal polynomials are used for curve fitting to reduce computation time. Both emission and absorption spectra are processed and the results demonstrate the feasibility and practicability.

T. Lan's e-mail address is tingerlan@163.com.

References

1. G. Schulze, A. Jirasek, M. M. L. Yu, A. Lim, R. F. B. Turner, and M. W. Blades, Appl. Spectrosc. **59,** 545 (2005).
2. M. N. Leger and A. G. Ryder, Appl. Spectrosc. **60,** 182 (2006).
3. B. K. Alsberg, A. M. Woodward, and D. B. Kell, Chemometrics and Intelligent Laboratory Systems **37,** 215 (1997).
4. C. Kong, Y. Fang, T. Lan, W. Xiong, D. Dong, and D. Li, Opt. Precision Eng. (in Chinese) **14,** 1094 (2006).
5. C. A. Lieber and A. Mahadevan-Jansen, Appl. Spectrosc. **57,** 1363 (2003).
6. Y. Wang and J. Y. Mo, Chemical J. Internet **5,** (2) 16 (2003).
7. F. Gan, G. Ruan, and J. Mo, Chemometrics and Intelligent Laboratory Systems **82,** 59 (2006).