Digital Signal Processing 40 (2015) 53–62

Contents lists available at ScienceDirect

Digital Signal Processing

www.elsevier.com/locate/dsp

# A robust baseline elimination method based on community information

Yanling Wu, Qingwei Gao *, Yuanyuan Zhang

*School of Electrical Engineering and Automation, Anhui University, China*

ARTICLE INFO

*Article history:*
Available online 4 March 2015

*Keywords:*
Baseline correction
Genetic programming
Robust estimation
Community information

ABSTRACT

Baseline correction is an important pre-processing technique used to separate true spectra from interference effects or remove baseline effects. In this paper, an adaptive iteratively reweighted genetic programming based on excellent community information (GPEXI) is proposed to model baselines from spectra. Excellent community information which is abstracted from the present excellent community includes an automatic common threshold, normal global and local slope information. Significant peaks can be firstly detected by an automatic common threshold. Then based on the characteristic that a baseline varies slowly with respect to wavelength, normal global and local slope information are used to further confirm whether a point is in peak regions. Moreover the slope information is also used to determine the range of baseline curve fluctuation in peak regions. The proposed algorithm is more robust for different kinds of baselines and its curvature and slope can be automatically adjusted without prior knowledge. Experimental results in both simulated data and real data demonstrate the effectiveness of the algorithm.



## 1. Introduction

Fourier transform infrared spectroscopy can be a valuable tool for measuring many chemical and physical properties of materials. However, it is a severe problem that spectra generally consist of peaks and noise superimposed on a baseline. Usually these baselines can be either flat, linear, curved or a combination of all three. Compared with peaks, their main character is that they vary much more slowly than the peaks do. The worst is that baselines vary greatly from spectrum to spectrum, even in similar samples. Thus, it is hard to eliminate them and this situation hampers the interpretation of spectra, which makes the removal of baseline drift necessary.

Baseline elimination for spectral data has been studied intensively and several methods have already been presented. These methods can be divided into two categories: manual and automatic techniques. In the manual method [1], the baseline is constructed by using linear, polynomial, or spline functions fitted on the no signal (baseline) points selected by users. If the points are correctly selected, the construction would produce satisfactory results. Obviously, this technique is subjective, time-consuming, and poorly reproducible [2].

In contrast, automatic baseline correction is called for and more widely employed. Among these methods, the wavelet transform has become a useful tool in background removal [3,4]. However, inappropriate wavelet and resolution level selection are detrimental to baseline estimation. Fourier transform method which can generate the frequency components from the original spectrum is used to make a discrimination among baseline (low frequency), signal (mid frequency), and noise (high frequency) components. Then these frequency components can be filtered by a band-pass or high-pass filter to eliminate unwanted spectral components. But the filter parameters are difficult to set for separating the baseline from the signal effectively [5]. In general, these approaches are based on a hypothesis that the background can be well separated (in the transformed domain) from the real signal. The derivative method [6] uses first derivatives or second derivatives to remove constant off-sets or linear baselines from the spectra. But the threshold, which determines how many peaks are selected from the smoothed differentiated spectrum, is difficult to set.

Recently, baseline correction algorithms with asymmetric least squares smoothing are proposed [7–10]. The Whittaker smoother described by Eilers is used and only two parameters related to the rigid of the fitted curve and the noise level need to be tuned. But how to set the parameters is not always an easy task.

An iterative method based on polynomial curve fitting for automated estimation of baseline is proposed [11–16]. These algorithms generate automatic threshold to distinguish the baseline

* Corresponding author.
*E-mail address:* qingweigao@ahu.edu.cn (Q. Gao).

http://dx.doi.org/10.1016/j.dsp.2015.02.015
1051-2004/

from peaks by a fitted curve. Linear programming is used for baseline correction [17], the polynomial order is selected based on a criterion instead of the user's experience, but the criterion can be used only when these baseline correction processes with different polynomial orders have been completed and only be used in comparing results of these processes. These methods offer a promising approach to removing baseline effects in a simple, straightforward fashion. However, their performance depends on the two parameters predefined by the users. The parameters include the polynomial order and the threshold which is related to the noise level and other characters about the spectrum. Therefore, the accuracy of the estimation still depends on the user's prior knowledge.

If there is some slope or curvature information about the baseline, the parameter which is related to the rigid of the fitted curve and the threshold would be easier to set and these baseline correcting methods should have more chances to present satisfied results.

Usually there isn't any information about baseline before a baseline correction process, but more and more knowledge about the baseline can be obtained with the deepening of the process. In this paper, this knowledge is used in the adaptive iterative baseline correction process to help automatically define the rigid of the fitted curve. Reweighted genetic programming based on excellent community information (GPEXI) is proposed to recognize and model baseline automatically. Here excellent community information includes common automatic threshold, global slope, local slope, and curvature information which are obtained from these present common baseline areas determined by excellent community selected from the current population of GP. The proposed method uses an automatic threshold defined by excellent community information instead of one curve to discriminate baseline areas and peaks. The order of polynomial is automatically determined during the learning process without prior knowledge of spectra. By this way, an iteratively procedure is executed to gradually approximate a complex baseline.

In Section 2, some useful and important preliminary ideas are discussed. The proposed reweighted genetic programming based on excellent community information (GPEXI) are given in Section 3. These methods about how to extract excellent community information from each generation and how to use this information are also given in this section. Section 4 presents some simulated data which are used to illustrate the performances of the proposed method. The effectiveness of the method is also demonstrated through applications on experimental spectra. Finally, some conclusions are given in Section 5.

## 2. Preliminaries

### 2.1. Problem modeling

Assume that the I-point spectrum is $\{(x_1, y(x_1)), \cdots (x_i, y(x_i)), \cdots, (x_I, y(x_I))\}$. It can be modeled as $y(x_i) = b(x_i) + e(x_i)$, $1 \le i \le I$ where: $x_i$ is a wavelength value. $y = (y(x_1), y(x_2), \cdots, y(x_I))$ is a $I$ point positive peak spectrum. $b = (b(x_1), b(x_2), \cdots, b(x_I))$ denotes the baseline itself. $e = (e(x_1), e(x_2), \cdots, e(x_I))$ denotes the residual, peaks, and physical noise. The baseline can be modeled as

$$b(x_i) = f(x_i, a). \tag{1}$$

Here, $f(\cdot)$ and $a$ are functions and parameters. Baseline should have the following properties: 1) being smooth, but 2) also being faithful to $y$ [18].

### 2.2. Polynomial curve fitting algorithms

In these methods, a baseline $b$ can be modeled as a p order polynomial function. It can be written as $b' = Xa$. Here,

$$X = \begin{pmatrix} x_1^0 & x_1^1 & \cdots & x_1^p \\ x_2^0 & x_2^1 & \cdots & x_2^p \\ \vdots & \vdots & & \vdots \\ x_I^0 & x_I^1 & \cdots & x_I^p \end{pmatrix}, \qquad a = \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_p \end{pmatrix}$$

is the polynomial coefficients. Try to minimize the criterion $J(a)$, then find the coefficients $a$ of the polynomial curve.

$$J(a) = \sum_{i=1}^{I} \varphi\big(y(x_i) - X_i a\big) \tag{2}$$

$X_i$ represents the $i$th row of $X$. The cost function $\varphi$ has a critical influence on the criterion. The two following cost functions are used and $a$ is estimated by an iterative technique [12,19].

Asymmetric Huber function

$$\varphi\big(y(x_i) - X_i a\big) = \begin{cases} (y(x_i) - X_i a)^2 & \text{if } y(x_i) - X_i a < \Delta \\ 2\Delta(y(x_i) - X_i a) - \Delta^2 & \text{otherwise} \end{cases} \tag{3}$$

Asymmetric truncated quadratic

$$\varphi\big(y(x_i) - X_i a\big) = \begin{cases} (y(x_i) - X_i a)^2 & \text{if } y(x_i) - X_i a < \Delta \\ \Delta^2 & \text{otherwise} \end{cases} \tag{4}$$

Other polynomial curve fitting algorithm [11,13–16] use an iterative algorithm to estimate the baseline on a spectrum in which peaks are eliminated. The signal areas (peaks) are redefined at each iteration by a estimated baseline. Each point is set equal to the estimated baseline if the corresponding spectrum intensity is higher than the estimated baseline, otherwise it is set equal to the spectrum [14]. This method is equivalent to minimize an asymmetrical truncated quadratic when $\Delta = 0$.

The threshold $\Delta$ and the order of a polynomial function provide a threshold to discriminate baseline areas from signal areas. The two parameters need to be predefined by a user.

### 2.3. Genetic programming

GP was proposed by J.R. Koza in 1990's. It starts with an initial population created randomly and produces offspring by performing genetic operators in the current population. The iterative process continues until the stopping criterion is satisfied after several generations [20]. Then the best individual is gotten. GP has been widely used in many fields [21,22]. The steps of GP is:

(1) A random population of size $N$ is created. Each individual is represented as a tree structure. The individuals in the initial population are generated by recursively generating a rooted point-labeled tree.
(2) Calculate the fitness of each individual in the current population.
(3) Use the selection, recombination, and mutation operators on the current population and generate a offspring population.
(4) Back to step 2 until the final conditions are satisfied. The best individual ever encountered during the run is the solution to the problem.

## 3. The baseline correction algorithm based on community information

Here, genetic programming provides multiple estimated baseline curves with different smoothness and recognizes baseline areas by community information which is abstracted from all these estimated curves. The characteristic of GP, which is a population based optimization technique, is used to improve the accuracy of

the baseline region recognition. Then GP models the baseline without pre-specifying the structure of it.

Considering that a baseline generally varies much slower than a signal, so slope and curvature information can be used to distinguish signal free regions (baseline area) from signal areas (peaks). With the help of this information, reweighted genetic programming based on excellent community information (GPEXI) is proposed to eliminate baselines in spectra.

First several best fitted curves (individuals in the current population) are used to identify peak regions from baseline regions. Because several fitted curves are used, the results should be more credible than those results obtained by only one fitted curve. Then, in order to avoid mistaking baseline regions as peaks, slope and curvature information are used to retest whether regions belong to peaks or baseline regions.

### 3.1. Noise removal

As mentioned before, the baseline shows very broad spectral features and its derivative changes very slowly too. In the proposed method, derivative information of spectra is used to identify whether a point belongs to peaks, but noise degrades effectiveness of the method. So noise removal is needed. A Savitzky–Gloay filter [23,24] is used as a generalized moving average filter and is denoted as $\bar{y}(x_i) = \text{smoothing}(y(x_i), N_1)$, $N_1$ is the number of neighboring data points on the either side of $y(x_i)$. $\bar{y}(x_i)$ is only used to get slope (derivative) information in baseline regions, so the parameter $N_1$ is robust which can be demonstrated in subsequent experiments.

### 3.2. Excellent community and community information

In order to utilize the characteristic information of the baseline to prevent erroneous identification of baseline areas and peak areas, excellent community and community information are combined with GP. The characteristic information includes global slope information, local slope information, and the curvature information of a baseline. Generally, this information is produced based on baseline areas of a spectrum. But unfortunately no information about baseline can be used before a correction processes generally. With the deepening of the correction process, more information can be obtained. The proposed method try to use and update these information during all the process based on current excellent community. Thus they are called community information.

Excellent community consists of the several optimal individuals selected from the current population and is updated during all the evolutionary process. Here an individual represents an estimated fitted curve. Excellent community is defined as: Sort all individuals in the current population in an increasing order based on their fitness. The first 20% in the queue are chosen to make up the excellent community.

Community information is automatically abstracted based on the common information provided by the present excellent community. This information includes common automatic threshold, global slope, local slope and curvature information which are abstract from common preliminary baseline areas. These information will be used to further confirm whether the preliminary results derived from common automatic threshold are true or false.

### 3.3. Identify preliminary peaks

Each individual curve generates an automatic threshold to discriminate peaks from baseline regions. A region is identified as a peak region by an individual if the original spectra have a value higher than the corresponding value in the individual fitted curve.

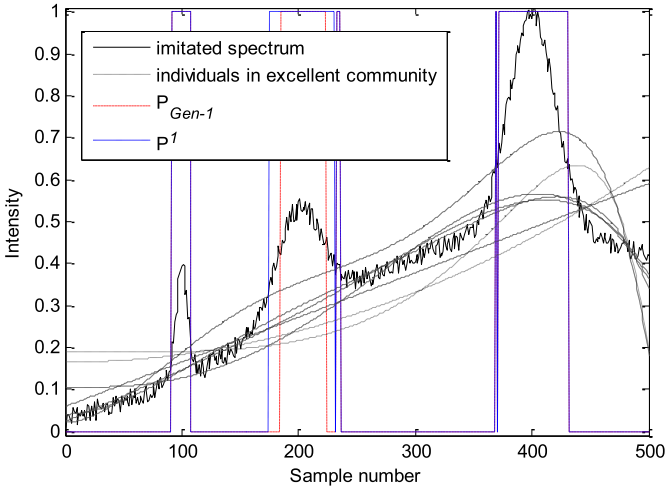


**Fig. 1.** Identify preliminary peaks.

Preliminary baseline and peak information is derived from common automatic threshold which are abstracted from the present excellent community.

(1) Suppose, the current generation is *Gen*. $P_{Gen-1}(x_i)$ contains the peak distribution information obtained from the last generation.

$$P_{Gen-1}(x_i) = \begin{cases} 0 & \text{if } x_i \in \text{baseline regions} \\ 1 & \text{if } x_i \in \text{peak regions} \end{cases} \tag{5}$$

(2) $p_j(x_i)$ is the peak distribution identified by an individual $\hat{b}_j$ $(1 \le j \le J)$ in the excellent community.

$$p_j(x_i) = \begin{cases} 0 & \text{if } y(x_i) - \hat{b}_j(x_i) \le 0 \\ 1 & \text{if } y(x_i) - \hat{b}_j(x_i) > 0 \end{cases}. \quad \text{If } p_j(x_i) = 1,$$

a wavelength value $x_i$ is identified to be belonging to a peak area by $\hat{b}_j$.

(3) A wavelength point $x_i$ is initially identified as belonging to peaks if $P^1(x_i) = 1$. $P^1(x_i) = P_{Gen-1}(x_i) \vee (p_1(x_i) \wedge p_2(x_i) \wedge \cdots \wedge p_J(x_i))$ contains the preliminary peak distribution information obtained from the current excellent community and the last generation. Here $\wedge$ is an and operator, $\vee$ is a or operator.

(4) In order to avoid spikes in the baseline areas (single or few points belonging to peaks) [9] which are caused by noise usually, $P^1(x_i)$ is modified by this way: $P^1(x_i) = 1$ requires not only that the point $x_i$ satisfies $P^1(x_i) = 1$ but also that two of its neighbors do.

Fig. 1 shows the current preliminary peak distribution information. The area about 440 is initially identified as a peak area based $P_{Gen-1}$ although this area is a baseline area based the information from the current excellent community.

Because the effects of peak areas will be rarely considered in the subsequent baseline correction process, the determination of peak areas requires more attention. Global slope, local slope and curvature information will be used to further determine whether a point is really in a peak region.

### 3.4. Identify peak areas by global slope information

Global slope information should be obtained from the whole baseline of a spectrum. Slopes of a baseline in peak regions can be represented by the slopes of their neighboring baseline areas, because baselines vary much slower than peaks do and these peak areas can't provide any effective baseline information generally.

$$\mathrm{d}b(x_i) = \begin{cases} \mathrm{d}\bar{y}(x_i) & \text{if } P^1(x_i) = 0 \\ \mathrm{d}Line(Peak_m) & \text{if } P^1(x_i) = 1 \text{ and } kstest2 = 0 \\ \mathrm{d}(B_m) & \text{if } P^1(x_i) = 1 \text{ and} \\ & L(x_i, B_m) \le 0.5(L(B_m) + L(B_{m+1})) \\ \mathrm{d}(B_{m+1}) & \text{if } P^1(x_i) = 1 \text{ and} \\ & L(x_i, B_{m+1}) < 0.5(L(B_m) + L(B_{m+1})) \end{cases} \quad (6)$$

Here, $\mathrm{d}b(x_i)$ is the slope of a estimated baseline at $x_i$. $\mathrm{d}\bar{y}(x_i) = \bar{y}(x_{i+1}) - \bar{y}(x_i)$. $Peak_m$ is a continuous peak area which locates between its neighbor baseline areas $B_m$ and $B_{m+1}$. $kstest2 = 0$ means that these absolute derivatives of the spectrum in $Peak_m$ and these absolute derivatives in $B_m$ and $B_{m+1}$ are drawn from the same underlying data based on Kolmogorov–Smirnov (K–S) test with the significance level 5%. If they are drawn from the same underlying data, $Peak_m$ may be a broad peak and too much parts of it are mistaken as some parts of $B_m$ and $B_{m+1}$. So slope information obtained from $B_m$ and $B_{m+1}$ contains wrong information from $Peak_m$ and use this slope information to represent the slopes in $Peak_m$ is improper. $Line(Peak_m)$ is the slope of the straight line which connects both sides of $Peak_m$. $L(B_m)$ and $L(B_{m+1})$ represent the number of points in $B_m$ and $B_{m+1}$ respectively. $L(x_i, B_m)$ and $L(x_i, B_{m+1})$ are distances from $x_i$ to $B_m$ and $B_{m+1}$ respectively. $\mathrm{d}(B_m)$ is a value generated from a normal distribution with mean $m(B_m)$ and standard deviation $\sigma(B_m)$. $m(B_m)$ and $\sigma(B_m)$ are mean and standard deviation of slops in $B_m$. $\mathrm{d}(B_{m+1})$ are generated by the same way used in $\mathrm{d}(B_m)$, but $B_{m+1}$ is used instead of $B_m$.

In order to avoid the influence of spikes, the neighbor baseline areas $B_m$ and $B_{m+1}$ on both sides of $Peak_m$ are defined as: if a wavelength point $x_i$ belongs to $B_m$, $x_i$ must satisfy that $\min(x_{Peak_m,left} - L(Peak_m), x_{Peak_m,left} - L_{bl}) \le x_i \le x_{Peak_m,left} - 1$ and $P^1(x_i) = 0$, $x_{Peak_m,left}$ is the left most position of $Peak_m$. $L_{bl}$ is the length of a continuous baseline area which is adjacent to the left side of the peak. $B_{m+1}$ is defined by the way used in defining $B_m$, but the right most position of the peak and a continuous baseline area located at the right side are used instead of $x_{Peak_m,left}$ and $L_{bl}$.

With the deepening of the correction process, the area of a peak becoming clearer and some peaks become very broad, the information from these peak regions can't be ignored. At this condition, slopes of baseline in peak regions are represented by the slope of a straight line which connects both sides of the peak and the value of $\mathrm{d}b(x_i)$ in these peak regions should be modified. A peak is considered to be very broad if the length of a peak is larger than the sum of number of points in its neighbor baseline areas. Here, the peak and its neighbor baseline areas are defined by the way used in defining $Peak_m$, $B_m$ and $B_{m+1}$, but $P_{Gen-1}$ is used here instead of $P^1$.

By this way, the estimated baseline slope information $\mathrm{d}b(x_i)$ on the whole spectrum can be obtained. If $\mathrm{d}b(x_i) > m + 2\sigma$ or $\mathrm{d}b(x_i) < m - 2\sigma$, the slope at $x_i$ is abnormal and its influence will be eliminated in calculating $m$ and $\sigma$. $m$ and $\sigma$ are mean and standard deviation of slope of all remaining points respectively. Then the process continues until slopes at all remaining wavelength values are in the normal range $[m - 2\sigma, m + 2\sigma]$. The normal global slope range is larger than $S^{Low}_{global} = m - 2\sigma$ and lower than $S^{High}_{global} = m + 2\sigma$. Because $S^{High}_{global}$ and $S^{Low}_{global}$ are obtained based on the information of the whole spectrum, so they are called global slope information.

To further identify whether a point $x_i$ is in a peak range, a straight line ($L^{High}_{m,global}$) is drawn from the nearest point in the most likely baseline part of the left neighbor baseline region $B_m$ with slope $S^{High}_{global}$. Also, a straight line ($L^{Low}_{m+1,global}$) is drawn from the other neighbor region $B_{m+1}$ with slope $S^{Low}_{global}$. The nearest points in region $B_m$ and $B_{m+1}$ are decided by the normal global and local slope information. Then whether $x_i$ belongs to baselines or peaks is decided by:

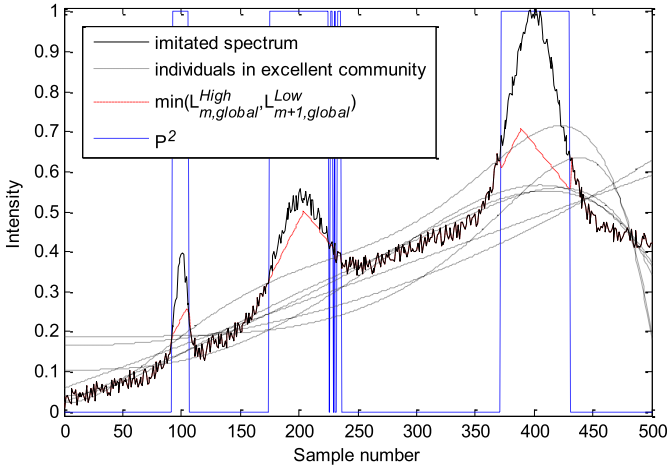


**Fig. 2.** Identify peak areas by global slope information.

$$\begin{cases} P^2(x_i) = 1 & \text{if } y(x_i) > \min(L^{High}_{m,global}(x_i), L^{Low}_{m+1,global}(x_i)) \\ P^2(x_i) = 0 & \text{otherwise} \end{cases} \quad (7)$$

where, $L^{High}_{m,global}(x_i)$ and $L^{Low}_{m+1,global}(x_i)$ are the corresponding value of $L^{High}_{m,global}$ and $L^{Low}_{m+1,global}$ at $x_i$ respectively. $x_i$ is identified again as belonging to a peak region if $P^2(x_i) = 1$. $y(x_i) > L^{High}_{m,global}(x_i)$ means that slopes at some wavelength points in any curve which connects $x_i$ and the baseline region $B_m$ exceeds the normal range $S^{High}_{global}$, so $x_i$ is re-identified in a peak by global slope normal range. Fig. 2 shows the peak distribution ($P^2$) obtained by using global slope information.

### 3.5. Identify peak areas by local slope information

$x_i$ is in $Peak_m$ that is initially identified by $P^1(x_i) = 1$ but not further confirms by global information because $P^2(x_i) = 0$. In this section, local slope information will be used to identify whether it is in a peak range further.

The normal local slope range of $B_m$ is calculated by the same method used in calculating the normal global slope and is larger than $S^{Low}_{m,local} = m - 2\sigma$ and lower than $S^{High}_{m,local} = m + 2\sigma$. $m$ is the mean and $\sigma$ is the standard deviation of slopes at all remaining wavelength points in $B_m$. Because $S^{High}_{m,local}$ and $S^{Low}_{m,local}$ are obtained based on only one preliminary identified baseline band $B_m$ on a spectrum, so they are called local slope information. $S^{High}_{m+1,local}$ and $S^{Low}_{m+1,local}$ are the normal local slope range of $B_{m+1}$.

The effective scopes of local slope information are their neighboring regions and how to decide the exact scope is difficult. So $[C^{low}, C^{high}]$ the normal ranges of $\mathrm{d}^2b(x_i)$ which represents the slope change rate at $x_i$ are used to decide the change rate of $S^{High}_{m,local}$ and $S^{Low}_{m,local}$ when a point is outside the region $B_m$. $[C^{low}, C^{high}]$ are obtained from initially identified baseline areas with the same method used in determining the normal local and global slope range. Here a Savitzky–Gloay filter is used to suppress the fast varying components of the slope $\mathrm{d}b(x_i)$ which are usually caused by remaining noise and then $\mathrm{d}\bar{b}(x_i)$ is got. $\mathrm{d}\bar{b}(x_i) = \text{smoothing}(\mathrm{d}b(x_i), N_2)$, $\mathrm{d}^2b(x_i) = \mathrm{d}\bar{b}(x_{i+1}) - \mathrm{d}\bar{b}(x_i)$, $N_2$ is the number of neighboring data points on the either side of $\mathrm{d}b(x_i)$.

Whether $x_i$ belongs to a baseline or peak area is redefined by $P^3(x_i)$. $x_i$ is identified again as belonging to a peak region if $P^3(x_i) = 1$.

$$\begin{cases} P^3(x_i) = 1 & \text{if } y(x_i) > \min(curve_{B_m}(x_i), curve_{B_{m+1}}(x_i)) \\ P^3(x_i) = 0 & \text{otherwise} \end{cases} \quad (8)$$

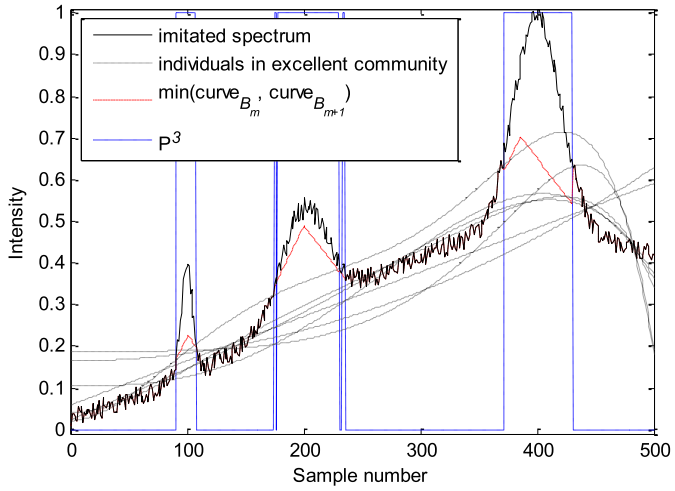


**Fig. 3.** Identify peak areas by the strict boundary definition.

$$s_{B_m}(x_i) = S^{High}_{m,local} + (x_i - x_{B_m})C^{High},$$

$$curve_{B_m}(x_i) = y(x_{B_m}) + \sum_{t=x_{B_m}}^{x_i-1} s_{B_m}(t).$$

$$s_{B_{m+1}}(x_i) = S^{Low}_{m+1,local} - (x_{B_{m+1}} - x_i)C^{Low},$$

$$curve_{B_{m+1}}(x_i) = y(x_{B_{m+1}}) - \sum_{t=x_i+1}^{x_{B_{m+1}}} s_{B_{m+1}}(t).$$

$(x_{B_m}, y(x_{B_m}))$ and $(x_{B_{m+1}}, y(x_{B_{m+1}}))$ are nearest points of the most likely baseline part in $B_m$ and $B_{m+1}$. $s_{B_m}(x_i)$ and $s_{B_{m+1}}(x_i)$ are slopes of $curve_{B_m}$ and $curve_{B_{m+1}}$ at $x_i$. $C^{High}$ and $C^{Low}$ are the upper and lower boundary of $\mathrm{d}^2b(x_i)$.

Because the bounds of the normal range of $\mathrm{d}^2b(x_i)$ and the normal local slope range are used, $curve_{B_m}(x_i)$ and $curve_{B_{m+1}}(x_i)$ give the upper bound of the possible baseline. Some points with a higher value than $\min(curve_{B_m}(x_i), curve_{B_{m+1}}(x_i))$ may also belong to baseline areas, so $P_3(x_i)$ is a strict definition of the peak distribution. Fig. 3 shows the peak distribution ($P^3$) obtained by strict boundary definition.

Looser bounds of the possible baseline are also obtained based on a hypothesis that the effective domains of local information include their adjacent continuous peak range.

A straight line ($L^{High}_{m,local}$) is drawn from the nearest point of the most likely baseline part in $B_m$ with slope $S^{High}_{m,local}$. Also, A straight line ($L^{Low}_{m+1,local}$) is drawn from the other neighbor region $B_{m+1}$ with slope $S^{Low}_{m+1,local}$. Then the looser definition of the possible peak distribution $P^4(x_i)$ are obtained:

$$\begin{cases} P^4(x_i) = 1 & \text{if } y(x_i) > \min(L^{High}_{m,local}(x_i), L^{Low}_{m+1,local}(x_i)) \\ & \text{and } err(x_i) > mean(err) \\ P^4(x_i) = P^3(x_i) & \text{otherwise} \end{cases} \quad (9)$$

$err(x_i) = y(x_i) - Line(Peak_m)$. $mean(err)$ is the mean value of $err(x_i)$. $Line(Peak_m)$ is a straight line which connects both sides of $Peak_m$. $x_i$ is identified belonging to a peak region based on the looser definition $P^4(x_i) = 1$. Fig. 4 shows the looser peak distribution ($P^4$). Looser boundary $\min(L^{High}_{m,local}, L^{Low}_{m+1,local})$ is lower than corresponding values on strict boundary $\min(curve_{B_m}, curve_{B_{m+1}})$.

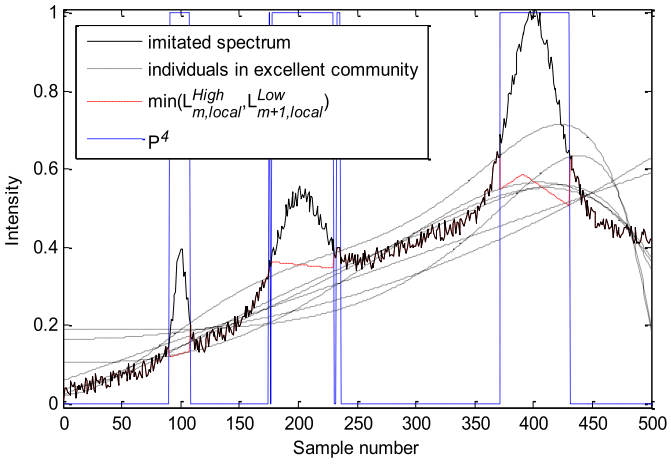


**Fig. 4.** Identify peak areas by the looser peak distribution.

### 3.6. Find the nearest point of the most likely neighbor baseline part

If the result of $P^1(x_i)$ is wrong and $x_i$ is in a baseline area, there must be a curve which connects $(x_i, y(x_i))$ and all points in neighbor baseline area and the slope at each point on the curve should be in the normal global slope range. Looser and strict boundaries also should be defined based on real baseline areas.

However, there must be some peaks which are mistaken as baseline areas because $x_i$ is defined in baseline areas only needs the recognition of one individual curve. So only the most likely baseline parts in a neighbor baseline region will be considered.

Region $B_m$ is on the left of $Peak_m$ and $x_i \in Peak_m$. In the region $B_m$, the most likely baseline parts locate at $x_{m,low}$ and its neighborhood. $y(x_{m,low})$ is the lowest value at all wavelength points in the region $B_m = \{x_{m,1}, x_{m,2}, \cdots, x_{m,low}, \cdots, x_{m,l}, \cdots, x_{m,L}\}$. The neighborhood is decided by normal global and local information:

(1) Make straight lines from $(x_{m,low}, y(x_{m,low}))$ to every point on the right side of it in the region $B_m$ respectively, because $Peak_m$ is on the right side of $B_m$.
(2) Calculate slopes of these straight lines and get $S_{low,low+1}$, $S_{low,low+2}, \cdots, S_{low,l}, \cdots S_{low,L}$. Here, $S_{low,l}$ is the slope of the straight line which connects $(x_{m,low}, y(x_{m,low}))$ and $(x_{m,l}, y(x_{m,l}))$. $x_{m,l}$ is in $B_m$ and on the right side of $x_{m,low}$.
(3) $x_{m,k}$ is the nearest wavelength value in the most likely baseline part of $B_m$ to $Peak_m$, if $S_{low,k} \le \min(S^{High}_{global}, S^{High}_{m,local})$, $S_{low,l} > \min(S^{High}_{global}, S^{High}_{m,local}), l \in [k+1, L]$ and $\mathrm{d}\bar{y}(x_{m,k}) \le \min(S^{High}_{global}, S^{High}_{m,local})$ at $x_{m,k}$.

The nearest wavelength value of the most likely baseline part in $B_{m+1}$ is also decided by the above method. It is needed to replace every point on the right side of $(x_{m,low}, y(x_{m,low}))$ with every point on the left side of $(x_{m+1,low}, y(x_{m+1,low}))$ in $B_{m+1}$ and using $S^{Low}_{global}$ and $S^{Low}_{m+1,local}$.

### 3.7. Identify peak area

Peak regions which are identified by preliminary information $P^1(x_i)$ are verified again by global slope information $P^2(x_i)$ or local slope information $P^3(x_i)$. Then the peak region is decided by a strict definition:

$$\begin{cases} x_i \in \text{baseline regions} & \text{if } P_{Gen}(x_i) = 0 \\ x_i \in \text{peak regions} & \text{else if } P_{Gen}(x_i) = 1 \end{cases} \quad (10)$$

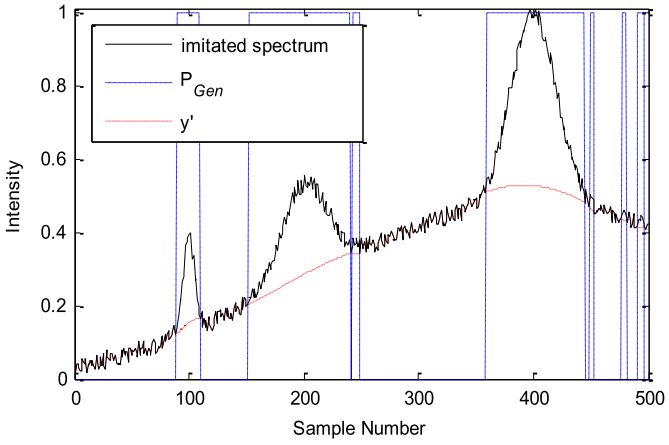


**Fig. 5.** Result of interpolation.

where $P_{Gen}(x_i) = P_{Gen-1}(x_i) \vee (P^1(x_i) \wedge (P^2(x_i) \vee P^3(x_i)))$. Then modified $P_{Gen}$ by this way: $P_{Gen}(x_i) = 1$ requires not only that the point $x_i$ satisfies $P_{Gen}(x_i) = 1$ but also that two of its neighbors do.

The peak region is decided by a loose definition:

$$\begin{cases} x_i \in \text{baseline regions} & \text{if } P'_{Gen}(x_i) = 0 \\ x_i \in \text{peak regions} & \text{else if } P'_{Gen}(x_i) = 1, \end{cases} \tag{11}$$

where $P'_{Gen}(x_i) = P_{Gen-1}(x_i) \vee (P^1(x_i) \wedge (P^2(x_i) \vee P^4(x_i)))$. Modified $P'_{Gen}$ by the same method as used in $P_{Gen}$.

### 3.8. Interpolation

After peak detection, the data in peak regions should be replaced with some interpolated values which are the estimates of baseline. Slope information is used to get interpolated values in peak regions.

Supposing a peak region defined by $P_{Gen}(x_i)$ locates between baseline wavelength values $x_b$ and $x_e$ which locate in $B_m$ and $B_{m+1}$. Slopes of the interpolated curve at $x_b$ and $x_e$ are $\mathrm{dc}(x_b) = \min(\mathrm{d}\bar{y}(x_b), S^{High}_{m,local})$ and $\mathrm{dc}(x_e) = \max(\mathrm{d}\bar{y}(x_e), S^{Low}_{m+1,local})$. On the curve, there must be a point where the slope is the slope ($\mathrm{d}_{line}$) of the straight line ($line_{x_b,x_e}$) which connects $(x_b, y(x_b))$ and $(x_e, y(x_e))$. Then the slope ($\mathrm{dc}(x_i)$) of the interpolated curve uniformly changes from $\mathrm{dc}(x_b)$ to $\mathrm{d}_{line}$ and then to $\mathrm{dc}(x_e)$. By this way, the slope ($\mathrm{dc}(x_i)$) in the interpolated curve in a peak are gotten. Interpolated values are obtained by integrating the estimated slopes. However direct integration of $\mathrm{dc}(x_i)$ makes some mismatched values since the $\mathrm{dc}(x_i)$ is only an estimate of slope information, especially on boundaries of a peak. So linear interpolation also used.

$$s(x_i) = \mathrm{dc}(x_i) - mean\big(\mathrm{dc}(x_i)\big) \tag{12}$$

$$y'(x_i) = line_{x_b,x_e}(x_i) + \sum_{t=x_b}^{x_i-1} s(t) \tag{13}$$

After interpolation, the spectrum is modified as

$$y'(x_i) = \begin{cases} y'(x_i) & x_i \in \text{peak regions} \\ y(x_i) & x_i \in \text{baseline regions} \end{cases} \tag{14}$$

and the result of interpolation is presented in Fig. 5.

### 3.9. Parameter estimation

The structure of an estimated baseline model is constructed by GP. After that, the weighted least square method is applied to estimate unknown parameters. The structure of an estimated baseline model can be expressed as:

$$\hat{b}(x_i) = c_1 f_1(x_i) + c_2 f_2(x_i) + \cdots + c_k f_k(x_i) \tag{15}$$

Here $\hat{b}(x_i)$ is an estimated baseline value at $x_i$. The structure of $f_k$ and the number of $k$ are defined by genetic programming automatically. Minimize $Q$ and find $C$.

$$\min: Q = \sum_{i=1}^{I} \beta_i \big(y'(x_i) - \hat{b}(x_i)\big)^2. \tag{16}$$

Suppose there are $M$ continuous peak areas ($Peak_m$, $1 \le m \le M$) identified by $P_{Gen}$. $x_{R,Peak_m}$ and $x_{L,Peak_m}$ are baseline wavelength values on both sides of $Peak_m$ based on $P_{Gen}$. Then $\beta_i$ is defined as:

$$\beta_i = \begin{cases} 1 & \text{if } P_{Gen}(x_i) = 0 \text{ and } P'_{Gen}(x_i) = 0 \\ 0.1\frac{L(peak_m)}{L(spectrum)} & \text{if } x_i \in Peak_m \text{ and } P'_{Gen}(x_{R,Peak_m}) = 0 \\ & \text{and } P'_{Gen}(x_{L,Peak_m}) = 0 \\ 0 & \text{if } x_i \in Peak_m \text{ and } (P'_{Gen}(x_{R,Peak_m}) = 1 \text{ or} \\ & P'_{Gen}(x_{L,Peak_m}) = 1) \end{cases}$$

$L(peak_m)$ is the number of points in the peak, and $L(spectrum)$ is the number of all points on the spectrum.

The minimization of $Q$ leads to the following system of equations:

$$\frac{\mathrm{d}Q}{\mathrm{d}C} = 0 \Rightarrow C = \big(F^T \beta F\big)^{-1} F^T \beta Y \tag{17}$$

$$C = \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_K \end{bmatrix}, \qquad F = \begin{bmatrix} f_1(x_1) & f_2(x_1) & \cdots & f_k(x_1) \\ f_1(x_2) & f_2(x_2) & \cdots & f_k(x_2) \\ & & \cdots & \\ f_1(x_I) & f_2(x_I) & \cdots & f_k(x_I) \end{bmatrix},$$

$$\beta = diag(\beta_1, \beta_2, \cdots, \beta_I), \qquad Y = \begin{bmatrix} y(x_1) \\ y(x_2) \\ \vdots \\ y(x_I) \end{bmatrix}.$$

### 3.10. Fitness function

The fitness function provides criterion for each estimated baseline and offers guidance on searching direction. Whether an individual is better or worse is judged by it. So the fitness function has a crucial influence on a GP's performance. Generally, the fitness is designed as:

$$\min: Fit = \sum_{i=1}^{I} \big(e(x_i)\big)^2 \tag{18}$$

where, $e(x_i) = \hat{b}(x_i) - y(x_i)$, $i = 1 \cdots I$ is number of points. $\hat{b}(x_i)$ is the value of the estimated baseline curve at $x_i$. This kind of fitness function is very sensitive to peaks, so it is not suitable for baseline correction.

Generally peak areas provide little useful information for baseline modeling, so in these regions, their effect should be reduced, smoothness is more important than the fitting error. But in baseline regions, the fitting error needs more attention.

Therefore the fitness function of an estimated baseline consists of the following parts:

$$Fit(\hat{b}_j) = \sum_{i=1}^{I} \beta'_i \big(y'(x_i) - \hat{b}_j(x_i)\big)^2 + \sum_{i=1}^{I} F\big(\hat{b}_j(x_i)\big) \tag{19}$$

Here,

$$\beta_i = \begin{cases} 1 & \text{if } P_{Gen}(x_i) = 0 \text{ and } P'_{Gen}(x_i) = 0 \\ 0.5 & \text{if } P_{Gen}(x_i) = 0 \text{ and } P'_{Gen}(x_i) = 1 \\ 0.1\frac{L(peak_m)}{L(spectrum)} & \text{if } x_i \in Peak_m \text{ and } P'_{Gen}(x_{R,Peak_m}) = 0 \\ & \text{and } P'_{Gen}(x_{L,Peak_m}) = 0 \\ 0 & \text{if } x_i \in Peak_m \text{ and } (P'_{Gen}(x_{R,Peak_m}) = 1 \\ & \text{or } P'_{Gen}(x_{L,Peak_m}) = 1) \end{cases}$$

$\sum F$ is a penalty term based on how much the value of $\hat{b}_j$ exceeds the possible upper boundary of baseline scope in peak areas and how much the value of $\hat{b}_j$ is larger than the corresponding value on the spectrum in baseline areas. In a peak area, there is only little information for baseline estimation. But with global slope information, local slope information, and the normal slope change rate, the possible upper boundary of baseline scope in peak regions can be roughly estimated.

$$F\left(\hat{b}_j(x_i)\right) = \begin{cases} (\hat{b}_j(x_i) - Upper(x_i))^2 & \text{if } \hat{b}_j(x_i) > Upper(x_i) \\ & \text{and } P_{Gen}(x_i) = 1 \\ (\hat{b}_j(x_i) - y(x_i))^2 & \text{if } \hat{b}_j(x_i) > y(x_i) + \Delta \\ & \text{and } P_{Gen}(x_i) = 0 \end{cases} \quad (20)$$

Here, $\Delta$ is standard deviation of $\hat{b}_j(x_i) - y(x_i)$ on baseline areas defined by $P_{Gen}$. $Upper(x_i)$ is the upper bound of possible baseline curve fluctuation at peaks.

$$Upper(x_i) = \min\left(L^{High}_{m,global}(x_i), L^{Low}_{m+1,global}(x_i), curve_{B_m}(x_i), curve_{B_{m+1}}(x_i), y(x_i)\right) \quad (21)$$

### 3.11. Steps of the proposed algorithm

(1) A random population of size $N$ is created. Each individual is an estimate of baseline for the spectrum.
(2) Estimate parameters and calculate the fitness value.
(3) Get excellent community, obtain information $P^1(x_i)$.
(4) Reconfirm these peak regions with global slope information and obtain $P^2(x_i)$.
(5) Reconfirm these peak regions with local slope information, obtain $P^3(x_i)$ and $P^4(x_i)$.
(6) Obtain the present strict identified peak regions with $P_{Gen}(x_i)$ and loose peak definition $P'_{Gen}(x_i)$. $P_{Gen-1}(x_i) = P_{Gen}(x_i)$.
(7) Use the selection, recombination, and mutation operator to generate a child population.
(8) Back to step 2 until the final conditions are satisfied. Here final condition is considered to be reached when there are no new peak points founded and the fitness of the best individual doesn't improve for several consecutive generations.
(9) Remove the estimated baseline $\hat{b}(x_i)$ from $y(x_i)$.

## 4. Experimental results

Three simulated spectral datasets and experimental spectra are used to validate the performance of the proposed method.

### 4.1. Simulated data

The simulated data imitate real spectral datasets that contain varying baseline, peaks, and random noise. Broad Gaussian peaks and linear lines are treated as baselines. Narrow Gaussian peaks are treated as peak signals. Spectra with a curved baseline, a double curved baseline and a sloping baseline are used to test the effectiveness of our proposed algorithm. The simulated peaks are three narrow Gaussian peaks, i.e:

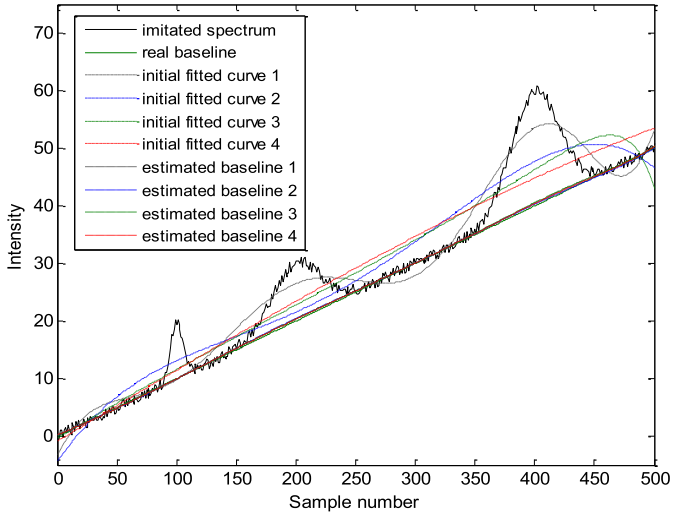


**Fig. 6.** Performance with different initial situations for a sloping baseline.

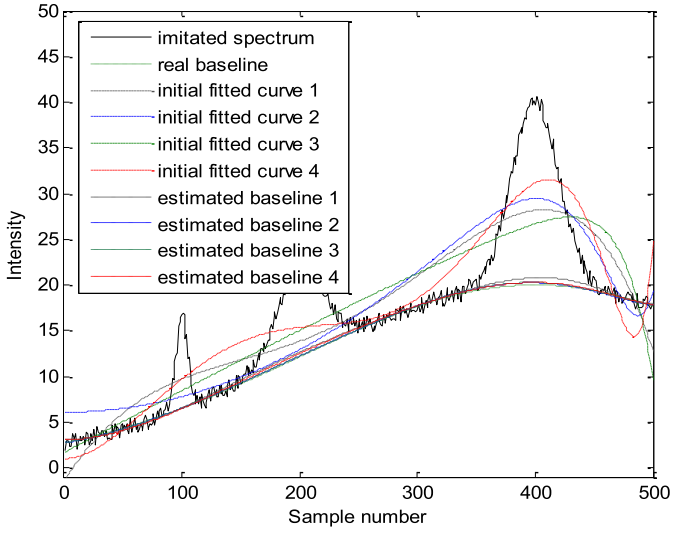


**Fig. 7.** Performance with different initial situations for a curve baseline.

$$s(x_i) = 10\exp\left(-\frac{(x_i - 100)^2}{2 \times 5^2}\right) + 10\exp\left(-\frac{(x_i - 200)^2}{2 \times 20^2}\right) + 20\exp\left(-\frac{(x_i - 400)^2}{2 \times 20^2}\right). \quad (22)$$

Curved baseline is $b_1(x_i) = 100\exp(-\frac{(x_i-400)^2}{2\times 200^2})$, the sloping baseline is represented as $b_2(x_i) = 0.3x_i$, and the doubly curved baseline is represented as $b_3(x_i) = 10\exp(-\frac{(x_i-450)^2}{(200\times 150)}) + 10\exp(-\frac{(x_i-20)^2}{(200\times 150)})$. $x_i = 1, 2, \ldots, 500$. Noise is generated using random numbers between $-1$ and 1.

Fig. 6 shows performances with different initial situations for a sloping baseline. The values of the spectrum around the points 0, 300, and 480 are higher than the corresponding values of the initial estimated baseline 1. The values of the spectrum in the regions around the point 500 are higher than the corresponding values of the initial estimated baseline 2 and 3. But these regions belong to baseline regions in fact. No baseline areas are mistaken as peaks by the initial estimated baseline 4. Fig. 6 shows that Although our method starts with different fitted curves but all estimated baselines are satisfied in the end of the 4 correction processes.

In Figs. 7 and 8, the performance with different initial situations is shown on simulated spectra with a curve baseline or a

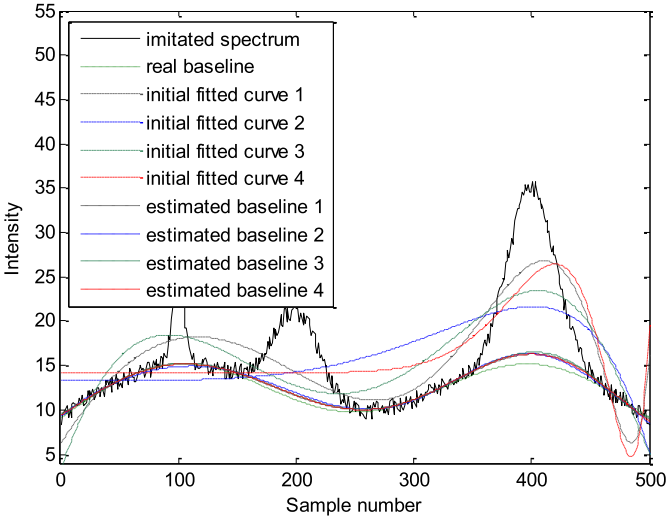


**Fig. 8.** Performance with different initial situations for a double curve baseline.

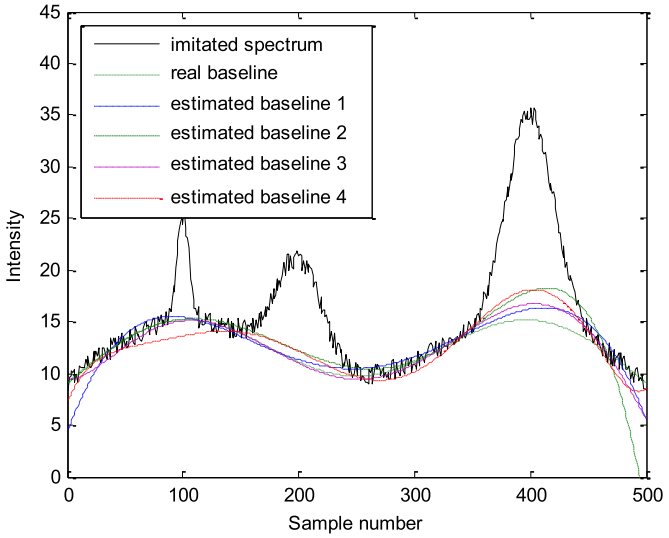


**Fig. 9.** Results when using different polynomial orders.

double curve baseline. Although some different regions are falsely identified by the initial fitted curves, but all results are satisfied. the proposed GPEXI shows its effectiveness for different conditions.

The performances of the proposed GPEXI in the three kinds of baseline demonstrate its effectiveness. Although the proposed GPEXI starts with improper fitted curves, but with the help of the community information, all processes end with good results.

The results of other polynomial curve fitting methods are influenced by the polynomial orders and the threshold which can be chosen according to the noise level and the ratio between the points belonging to the baseline and those belonging to the peaks. Generally optimal values of the two parameters exist but it is difficult to define without priori knowledge of spectra. For some kinds of baseline, slight deviation from the optimum value may cause worse results.

Fig. 9 shows the results of polynomial curve fitting method with an asymmetrical cost function [12] on a simulated spectrum with a double curve baseline. Estimated baseline 1, 2, 3 and 4 are obtained when using 4, 5, 6, 7 polynomial order and their respective corresponding optimal thresholds 0.34, 0.30, 0.084, 0.029. The results are best when using 6 order and 4 order polynomial curve and the results are not guaranteed when using other orders. No-signal areas around 0 and 500 are mistaken as peaks by estimated baseline 1, areas around 500 are also mistaken as peaks by estimated baseline 3, and artificial peaks appear there. These errors around 500 become more severe when estimated baseline 2 is used. When estimated baseline 4 is used, some errors occur in the vicinity of 100 where some no-signal areas near the narrow peak are mistaken as peak areas. There are also some errors around 0 and 500 when using estimated baseline 4. The results grow worse when using other polynomial order. Different thresholds also influence the results, the results are less optimal when using improper thresholds.

Another polynomial curve fitting method with an asymmetrical truncated cost function [12] is also used on this simulated spectrum. The performance of this method is similar to the method with an asymmetrical cost function. Moreover, its performance is more sensitive to the value of thresholds. Other curve fitting methods [11,13,16] have almost the same performances.

From these tests, we can see the proposed baseline correction method is a robust and good choice, especially when there is no information about the spectra and smoothness of their baselines.

### 4.2. Experimental data

Dataset 1 is a FTIR spectrum of the Chinese liquor Guotai sample. The IR spectra in the region 4000–650 $cm^{-1}$ have been recorded with a Perkin-Elmer Spectrum GXFTIR spectrometer, equipped with the Universal ATR Sampling Accessory (ZnSecell). The spectral resolution is set at 4 $cm^{-1}$. Dataset 2 is a NIR spectra of corn measured on spectrometers mp5. The spectrum is collected in 2 nm intervals within the spectral range 1100–2498 nm. The data set is available via Internet http://www.eigenvector.com/data/index.htm.

The performance of the proposed baseline correction algorithm on two real spectra is shown in Figs. 10 and 11. The values of the initial fitted curve 1 around 4000, the values of the initial fitted curve 2 around 3900 and 2500, the values fitted curve 3 around 1800, and the values of fitted curve 4 around 3900 and 2200 are lower than the corresponding values of the spectrum. These four correction processes of the proposed GPEXI start with different fitted curves but all processes end with almost same good results. We can see almost the same performance in Fig. 11. From the result in experimental data, the same conclusion as in the simulated data can be gotten. Without any information of the smoothness of a baseline in a spectrum and without predefining the smoothness of an estimated fitted curve, the proposed GPEXI can get a satisfied result in the end although it starts with improper fitted curves.

The performances of polynomial curve fitting method with an asymmetrical truncated cost function on FTIR spectrum of the Chinese liquor Guotai sample are shown in Figs. 12 and 13.

In Fig. 12, polynomial 3-order is selected because its performance is much better than the other polynomial orders. Estimated baseline 1, 2, 3 and 4 are obtained when using 3 polynomial order with respective thresholds 0.1, 0.05, 0.02, 0.01. Some values of estimated baseline 1 and 2 around 1900, 1100, and 900 are higher than their corresponding value of the spectrum. Compared with estimated baseline 1 and 2, estimated baseline 3 is better because fewer points have a bigger value on it than on the spectrum around areas 1900, 1100, and 900. The results mistake some peak areas as baseline areas when the thresholds become larger. But if the threshold becomes smaller, the performance changes greatly which can be shown by estimated baseline 4 when the threshold is 0.01. So the value of the threshold is closely related to the result, and it is not always easy to set.

In Fig. 13, the influences of different polynomial order with corresponding optimal thresholds are tested. Estimated baseline 1, 2,

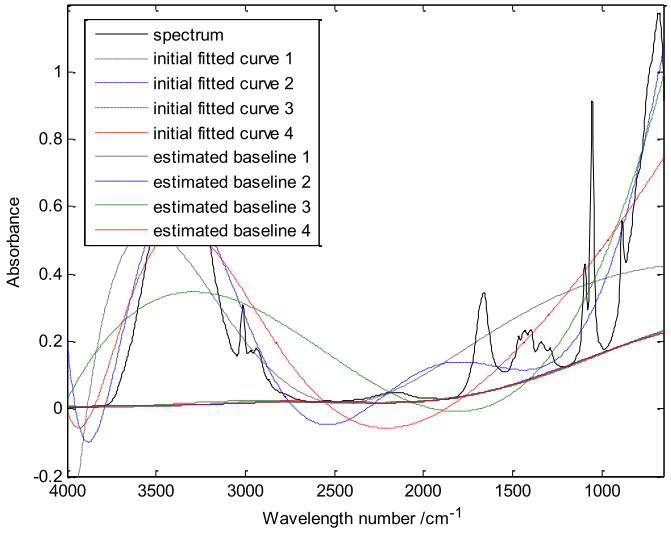


**Fig. 10.** Performance with different initial situations for dataset 1.

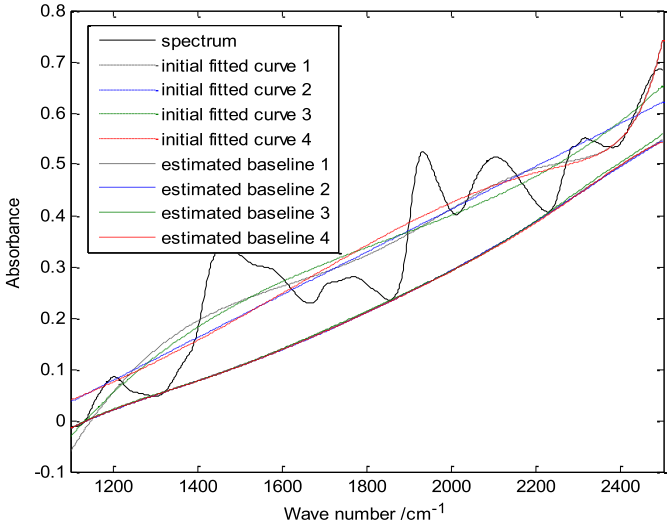


**Fig. 11.** Performance with different initial situations for dataset 2.

3 and 4 are obtained when using 2, 3, 4 and 5 polynomial order with respective thresholds 0.03, 0.02, 0.03, 0.05. Some values of estimated baseline 2 and 3 around 1900, 1100 are higher than their corresponding value of the spectrum. Regions around 2700 and 1800 are identified as peaks by estimated baseline 1, but these regions 2700 are identified as no signal areas by estimated baseline 2, 3, and 4. Some areas around 4000 are identified as significant peaks when using 5-order. When selecting other polynomial orders, the performances change significantly and the results are not guaranteed.

Other polynomial curve fitting methods [11,13,16] generally have the same problem as polynomial curve fitting method with an asymmetrical truncated cost function and an asymmetrical cost function.

In the proposed method the parameter $N_1$ and $N_2$ in noise remove are robust for the results. These tests list above have different noise level, different kinds of baseline, and different ratio between the points belonging to peaks and all points on the spectrum, but the same values of $N_1$ and $N_2$ are used and all processes end with satisfied results. The processes with he value of $N_1$ from 30 to 60 and the value of $N_2$ from 30 to 60 can get good results. Furthermore, with spikes removing, the parameter which is

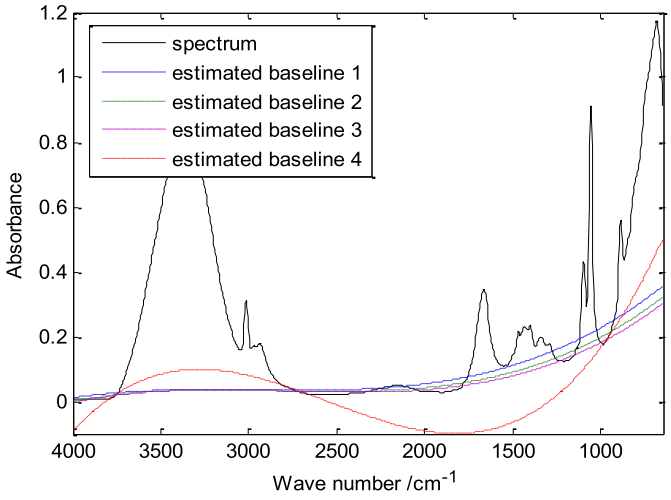


**Fig. 12.** Polynomial curve fitting method using an asymmetrical truncated cost function with different thresholds.

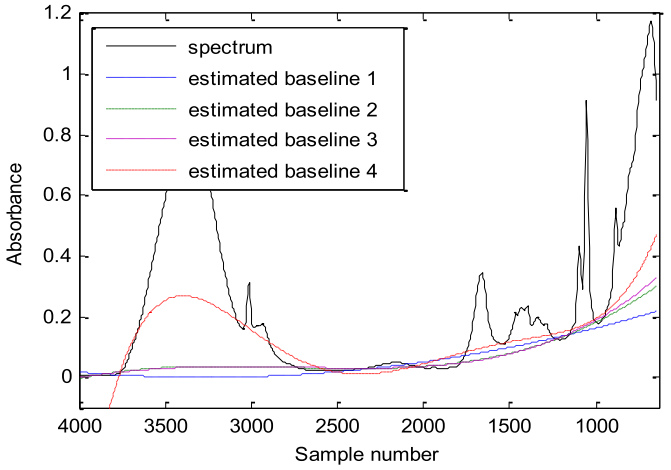


**Fig. 13.** Results of polynomial curve fitting method using an asymmetrical truncated cost function with different polynomial orders.

related to the noise level needn't to be considered in our proposed method.

## 5. Conclusion

The interpretation of spectroscopic data is largely hampered by the baseline or trend problem. Many proposed automatic curve fitting methods need to predefine the curve order and thresholds. But how to set these parameters is based on the user's experience. Generally the performance of these automatic correction algorithms needs to be tested more times and then the final result is chosen from these tests based on the user's experience.

In this paper, a novel spectra baseline correction algorithm based on community information is presented. By exploiting the common characteristic of excellent community, peaks are distinguished from baseline areas. By relying on the excellent community, the global, local slope information is used to further identify the location of peaks. In addition, a weighted least squares algorithm is proposed to estimate the parameters of the fitting curve, which can help the algorithm to distinguish the different influence of the peak and the baseline region on the learning process.

The proposed method doesn't require prior knowledge about the sample composition and selection of suitable baseline correction points. Experimental results on experimental spectra and

synthetic spectra show that the proposed method could handle various kinds of baseline completely automatically.

## Acknowledgments

This work is supported by National Natural Science Foundation of China (No. 51177002, 61032007), Doctoral Fund of Ministry of Education of China (No. 20113401120007), and Anhui Provincial Natural Science Project (KJ2012A012).

## References

[1] T. Vickers, R. Wambles, C. Mann, Curve fitting and linearity: data processing in Raman spectroscopy, Appl. Spectrosc. 55 (4) (2001) 389–393.
[2] A. Jirasek, G. Schulze, M.M.L. Yu, M.W. Blades, R.F.B. Turner, Accuracy and precision of manual baseline determination, Appl. Spectrosc. 58 (12) (2004) 1488–1499.
[3] B.F. Liu, Y. Sera, N. Matsubara, K. Otsuka, S. Terabe, Signal denoising and baseline correction by discrete wavelet transform for microchip capillary electrophoresis, Electrophoresis 24 (18) (2003) 3260–3265.
[4] Ł. Górski, F. Ciepiela, M. Jakubowska, W.W. Kubiak, Baseline correction in standard addition voltammetry by discrete wavelet transform and splines, Electroanalysis 23 (11) (2011) 2658–2667.
[5] P.A. Mosier-Boss, S.H. Lieberman, R. Newbery, Fluorescence rejection in Raman spectroscopy by shifted-spectra, edge detection, and FFT filtering techniques, Appl. Spectrosc. 49 (5) (1995) 630–638.
[6] 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. 60 (2) (2006) 182–193.
[7] K.H. Liland, E.O. Rukke, E.F. Olsen, T. Isaksson, Customized baseline correction, Chemom. Intell. Lab. Syst. 109 (1) (2011) 51–56.
[8] P.H.C. Eilers, Parametric time warping, Anal. Chem. 76 (2) (2004) 404–411.
[9] Z.M. Zhang, S. Chen, Y.Z. Liang, Baseline correction using adaptive iteratively reweighted penalized least squares, Analyst 5 (2010) 1138–1146.
[10] J. Peng, S. Peng, A. Jiang, J. Wei, C. Li, J. Tan, Asymmetric least squares for multiple spectra baseline correction, Anal. Chim. Acta 683 (1) (2010) 63–68.
[11] C.A. Lieber, A.M. Jansen, Automated method for subtraction of fluorescence from biological Raman spectra, Appl. Spectrosc. 57 (11) (2003) 1363–1367.
[12] V. Mazet, C. Carteret, D. Brie, J. Idier, B. Humbert, Background removal from spectra by designing and minimising a non-quadratic cost function, Chemom. Intell. Lab. Syst. 76 (2) (2005) 121–133.
[13] Y.H. Fang, W. Xiong, C. Kong, Automatic baseline correction of infrared spectra, Chin. Opt. Lett. 5 (10) (2007) 613–616.
[14] J. Zhao, H. Lui, D.I. McLean, H. Zeng, Automated autofluorescence background subtraction algorithm for biomedical Raman spectroscopy, Appl. Spectrosc. 61 (11) (2007) 1225–1232.
[15] F. Gan, G. Ruan, J. Mo, Baseline correction by improved iterative polynomial fitting with automatic threshold, Chemom. Intell. Lab. Syst. 82 (1–2) (2006) 59–65.
[16] P.J. Rousseeuw, K.V. Driessen, Computing LTS regression for large data sets, Data Min. Knowl. Discov. 12 (1) (2006) 29–45.
[17] S.J. Baek, A. Park, A. Shen, J. Hu, A background elimination method based on linear programming for Raman spectra, J. Raman Spectrosc. 42 (11) (2011) 1987–1993.
[18] P.H.C. Eilers, H.F.M. Boelens, Baseline correction with asymmetric least squares smoothing, Report, Leiden University Medical Centre, 2005.
[19] J. Idier, Convex half-quadratic criteria and interacting auxiliary variables for image restoration, IEEE Trans. Image Process. 10 (2001) 1001–1009.
[20] B. Wolfgang, F.D. Francone, E.K. Robert, N. Peter, Genetic Programming: An Introduction, Morgan Kaufmann, San Francisco, CA, 1998.
[21] A.H. Gandomia, A.H. Alavib, Multi-stage genetic programming: a new strategy to nonlinear system modeling, Inf. Sci. 181 (23) (2011) 5227–5239.
[22] R.A. Davis, A.J. Charlton, S. Oehlschlager, J.C. Wilson, Novel feature selection method for genetic programming using metabolomic 1H NMR data, Chemom. Intell. Lab. Syst. 81 (1) (2006) 50–59.
[23] S.J. Baek, A. Park, J. Kim, A. Shen, J. Hu, A simple background elimination method for Raman spectra, Chemom. Intell. Lab. Syst. 98 (1) (2009) 24–30.
[24] W.H. Press, S.A. Teukolsky, W.T. Vetterling, B.P. Flannery, Numerical Recipes in C, Cambridge University Press, USA, 1992.

**Yanling Wu** was born in Anhui, China, in 1977. She received her Ph.D. degree in control science and engineering from the Zhejiang University in 2009. She is currently an associate professor with the School of Electrical Engineering and Automation, Anhui University. Her research interests are evolutionary computation, robust estimation, and spectral analysis.

**Qingwei Gao** was born in Anhui, China, in 1965. He received his Ph.D. degree in information and communication engineering from the University of science and technology of China, in 2002. He is currently a professor with the School of Electrical Engineering and Automation, Anhui University. His research interests include wavelet analysis, image processing and fractal signal processing.

**Yuanyuan Zhang** was born in Anhui, China, in 1977. She received her Ph.D. degree in test and measurement technology and instrument from HeFei University of Technology, in 2010. She is currently an associate professor with the School of Electrical Engineering and Automation, Anhui University. Her research interests are nonlinear modeling and control, optimization algorithm.