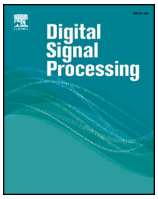
Digital Signal Processing 40 (2015) 53–62

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/© 2015 Elsevier Inc. All rights reserved.

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 ith row of X. The cost function φ 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 Δ 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