Liu and Lin *BMC Bioinformatics* (2019) 20:611
https://doi.org/10.1186/s12859-019-3188-4

BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

A general-purpose signal processing algorithm for biological profiles using only first-order derivative information

Yuanjie Liu[1,2*] and Jianhan Lin[1,2]

Abstract

Background: Automatic signal-feature extraction algorithms are crucial for profile processing in bioinformatics. Both baseline drift and noise seriously affect the position and peak area of signals. An efficient algorithm named the derivative passing accumulation (DPA) method for simultaneous baseline correction and signal extraction is presented in this article. It is an efficient method using only the first-order derivatives which are obtained through taking the simple differences.

Results: We developed a new signal feature extracting procedure. The vector representing the discrete first-order derivative was divided into negative and positive parts and then accumulated to build a signal descriptor. The signals and background fluctuations are easily separated according to this descriptor via thresholding. In addition, the signal peaks are simultaneously located by checking the corresponding intervals in the descriptor. Therefore, the eternal issues of parsing the 1-dimensional output of detectors in biological instruments are solved together. Thereby, the baseline is corrected, and the signal peaks are extracted.

Conclusions: We have introduced a new method for signal peak picking, where baseline computation and peak identification are performed jointly. The testing results of both authentic and artificially synthesized data illustrate that the new method is powerful, and it could be a better choice for practical processing.

Keywords: Passing accumulation, First derivative, Baseline correction, Signal extraction

Background

In profile-based bioinformatics data analysis, digitized signals have aroused universal interest with a broad range of applications. Extraction of qualitative and quantitative information in a large number of analytical signals from the background noise, however, poses significant challenges. In order to obtain accurate and clear results, some effective methods should be proposed and implemented to perform signal extraction before further data analysis. For instance, mass spectrometry is one of the most used tools to analyze large biological molecules, where the meaningful conclusion of the proteomic studies depends on the extracted signal peaks. In health care, chemical sensing relies on various spectroscopic techniques, which are not meaningful until the signals are extracted [1]. In agricultural applications, such as audio sensing for animal monitoring [2], the situation is exactly the same. Examples in bioelectrical activity measurements, such as electrocardiograms (ECG) and electroencephalograms (EEG) [3], mainly depend on wavelet analysis for signal processing [4].

Two parameters of the signal are often studied: peak position and peak area. Sometimes, the shape of signal is further studied through detailed analyses. In general, we can decompose a signal peak detection procedure into three consequent parts: smoothing, baseline correction and peak finding. Baseline removal and signal extraction are the core problems in signal processing. The baseline drift comes from the background fluctuation that appears as slow but large-scale ups and downs. It is a kind of low-frequency noise. When the signal peaks are selected according to their height, width or shape, the distortion and

* Correspondence: yjliu@cau.edu.cn
[1]College of Information and Electrical Engineering, China Agricultural University, Haidian, Beijing 100083, People's Republic of China
[2]Key Laboratory of Agricultural Information Acquisition Technology, Ministry of Agriculture and Rural Affairs, China Agricultural University, Haidian, Beijing 100083, People's Republic of China

© The Author(s). 2019 **Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

vertical shift caused by the baseline drift result in a significant interference. The high-frequency noises introduce rapid small-scale fluctuations. When peaks are identified as the signals, these noises harm the integrity of signal peaks while generating many interfering peak points. Various methods to smooth the noise and correct the drift with signal reception have been developed. However, due to randomness and complexity, the robust and accurate signal picking remains a challenging task. In this context, we propose a new signal feature extraction algorithm from a raw profile. After comparison with several classical methods, using various kinds of spectra and synthesized profiles, the proposed method was found to be accurate, flexible and easy to use.

Previous works

Baseline drift and random noise are common degradation problems in signal detection [5, 6]. Several methods based on various theories have been developed to solve these two problems. For pure computational methods, the need to extract key signal features that enable advanced processing algorithms to discover useful contextual information has led to the development of a wide range of algorithmic approaches. These include using Fourier analysis, wavelet analysis, the least squares method, computational geometry, neural networks and so on. Generally, after the removal of baseline drift, the signal peak identification could be focused on extraction using its width or area against the noise interference. Baseline correction thus is used as the main component of signal processing. This is especially true when instruments are being used to detect chemical reactions. Actually, an important series of algorithms has been developed in analytical chemistry in which numerous types of spectra are primitively used. There is a long history of developing numeric algorithms for processing the mass, fluorescent, or infrared spectroscopies. Shirley backgrounds [7], airPLS [8], AIMA [9], and Orthogonal Basis [10] are classic techniques playing important roles in different applications and subjects. The unified variation model [11], LMV-RSA [12] and the techniques based on neural networks [13], singular analysis [14], optimization method [15] and computational geometry [16] can also achieve improved effectiveness or efficiency, and some are able to perform joint baseline-correction and denoising.

In electronic signal processing, wavelet methods are widely applied. For example, in audio processing, multilevel 1-D wavelet analysis [17] is typically performed for denoising. In ECG data processing, the EMD method plays a dominant role [18]. It has made significant contributions to the development of wearable health care systems for breathing, cardiology and temperature measurements.

Our work

In this article, we proposed a fully automatic scheme using only the first-order derivative. Others attempted to use the derivative for signal processing [19], where both first and second derivatives were utilized. However, the effect was not good enough and the derivative method was not widely accepted.

In our algorithm, we used only the information of the first derivative and built a straightforward procedure that was able to simultaneously remove the baseline drift and extract the peak signals. It was named derivative passing accumulation (DPA) since it was based on the application of accumulation on the first derivative. The proficiency of this new method was mainly driven by the excellent properties of the derivative. Compared to the previous ones, this new algorithm is cleaner, more vigorous and more efficient.

Results

We selected three representative classical algorithms for comparison with our DPA method, i.e. wavelet [20], EMD [18] and airPLS [8]. These three algorithms were acknowledged as the most commonly used methods in processing electronic and spectroscopic signals.

For the testing data, we chose mass spectroscopy, Raman spectroscopy, audio, ECG and infrared spectroscopies. These were the typical data forms in biological measurements. As had been explained above, in one-dimensional profile processing, the baseline correction is the most important step. For methods based on thresholding separation, the signal peak picking is close to a completion of whole processing after the effective baseline removal. Because we do not know the precise signal information on these authentic data, to better present the comparison, we will only implement the baseline detection procedure in this part of testing. The results intuitively illustrated the performance of the algorithms.

Then, we presented the analysis of the signal location accuracy and peak area loss based on artificially synthesized data to numerically measure the effectiveness.

Testing on authentic data

The results of the four algorithms with respect to the Raman spectroscopy curve were shown in Fig. 1. The testing data-trace was a spectrum from RRUFF database [21]. It could be claimed from the results that the DPA method apparently outperformed the EMD and the Wavelet methods. At the same time, the DPA method generated a very close result to the airPLS method, which was the most widely used method specialized for baseline detection in spectroscopy processing.

The results with respect to the mass spectroscopy data are shown in Fig. 2. The testing data-trace is a MALDI-TOF mass spectra produced in bacteria protein analysis.

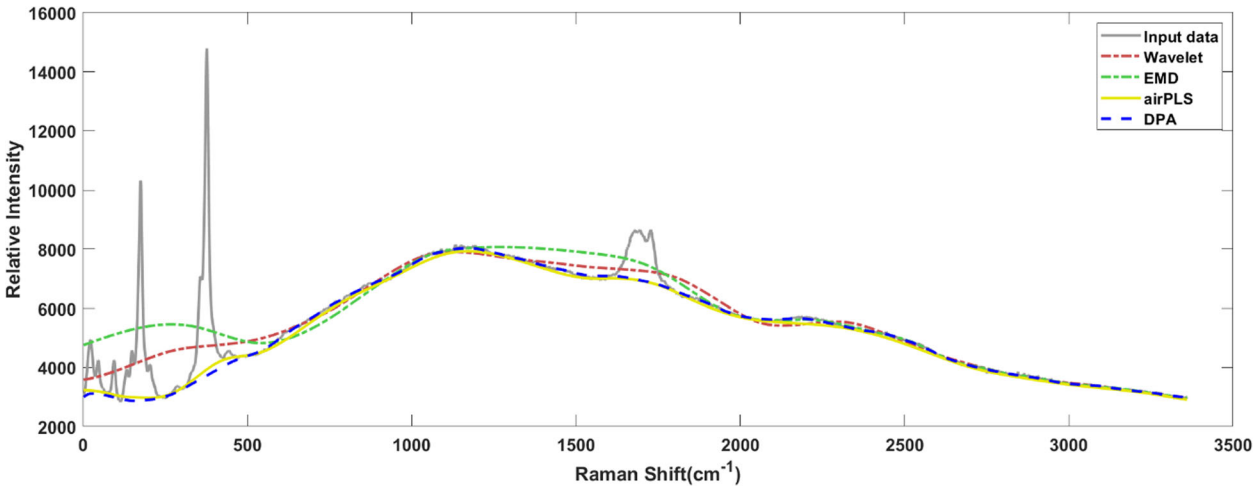

Fig. 1 The methods are implemented on the Raman spectroscopy for baseline detection. The DPA method generates a baseline similar to the airPLS method and much better than the EMD and wavelet method

The EMD and the Wavelet methods generated overestimated baselines. The airPLS method generated a dental baseline which was not preferred. The DPA method captured the basic trend of the drift better.

The results of the four algorithms with respect to the energy curve of a piece of audio are shown in Fig. 3. The audio data were taken from the monitoring of a pig farm.

The results on the ECG data are shown in Fig. 4. The data were downloaded from the MIT-BIH Arrhythmia Database [22, 23]. From Fig. 4, it is apparent that the airPLS performed poorly when processing ECG data. The corresponding baseline corrected results are shown in Fig. 5. It is reasonable to conclude that the waveform was more stable after the baseline removal using the DPA method.

Moreover, on the motor imagery EEG signal, which had been getting more attractive along with the rising of Brain Computer Interface, the DPA method also outperformed its companions. The results tested on part of motor imagery signal of the left-hand movement are presented in Fig. 6. The testing data were taken from the Project BCI - EEG motor activity data set.

The results of the infrared spectroscopy are shown in Fig. 7. Infrared spectroscopy is a standard method for detecting organic matters. The organic compound could be qualitatively analyzed through infrared spectroscopy whether it was a gas, liquid or solid. In biochemical measuring, infrared spectroscopy is a basic and necessary technique. The position, number, absorption intensity and shape of the peaks in the infrared spectrum are related to the structure and state of the compound. In

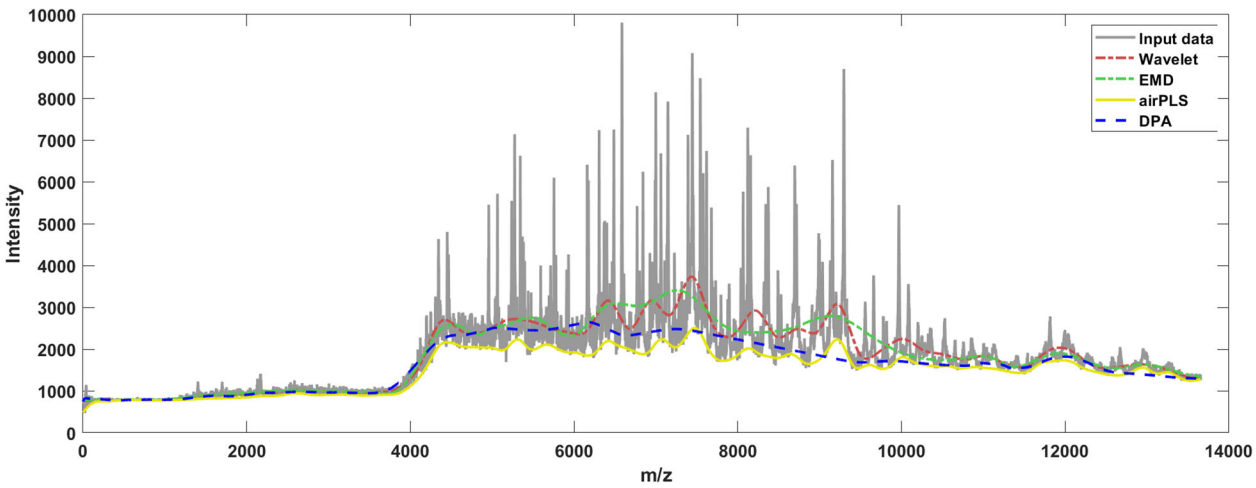

Fig. 2 The methods are implemented on the mass spectrum to detect the baseline. The DPA method captured the basic trend of the drift

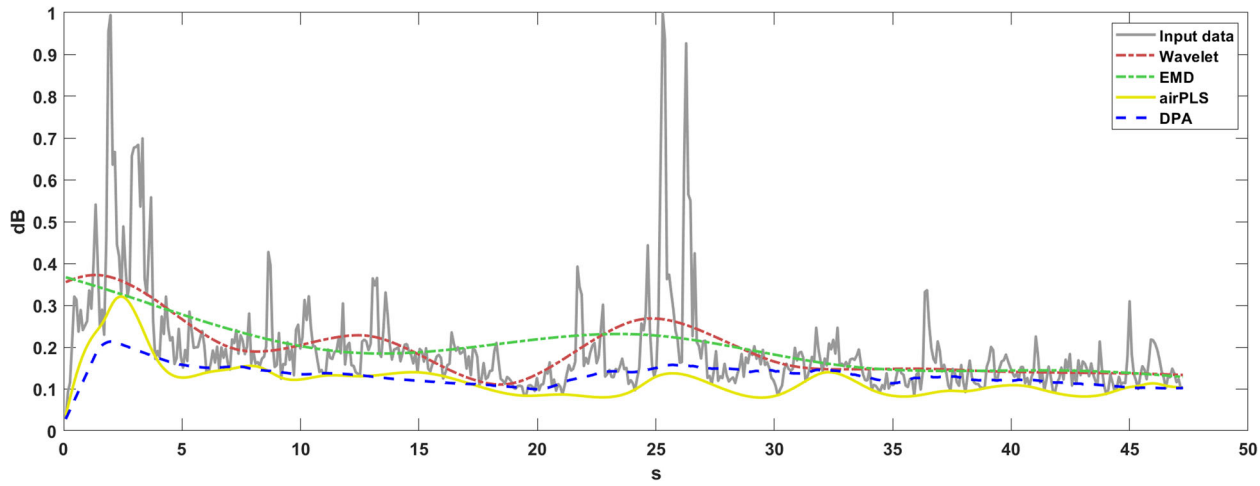

Fig. 3 The baselines detected by the comparison algorithms on the energy curve of pig farm audio data

addition, the baseline drift is ubiquitous with the infrared spectra, which dramatically affects the peak detection. As such, the processing algorithm is important. From the results, we claim that the airPLS and DPA methods performed similarly and were better than two others, while the wavelet method produced undercut baselines and the EMD method produced overcut baselines.

Testing on synthesized data

In addition to the authentic data that can be used to illustrate the practical performance of the new method, we carried out detailed analysis on the peak identification via artificially synthesized data in which the information of the simulated signal peaks was precisely known.

The synthesized data were generated by adding three parts together: the long, softly fluctuating waveform simulating the baseline drift; the sharp spikes with different widths and heights to simulate the signal peaks; and the white noise. The curve simulating the baseline was constructed by using the Fourier series. According to the mathematical theorem, any function $f(x)$ could be represented by the Fourier series expansion.

$$f(x) = a_0 + \sum_{k=1}^{\infty}(a_k \cos k\omega_0 x + b_k sin\, k\omega_0 x)$$

We randomly selected some long periods with random coefficients in the summation. In this way, a slowly undulating waveform was produced. We could theoretically guarantee that the simulating baselines were sampled from a broad scope. The signals were simulated using

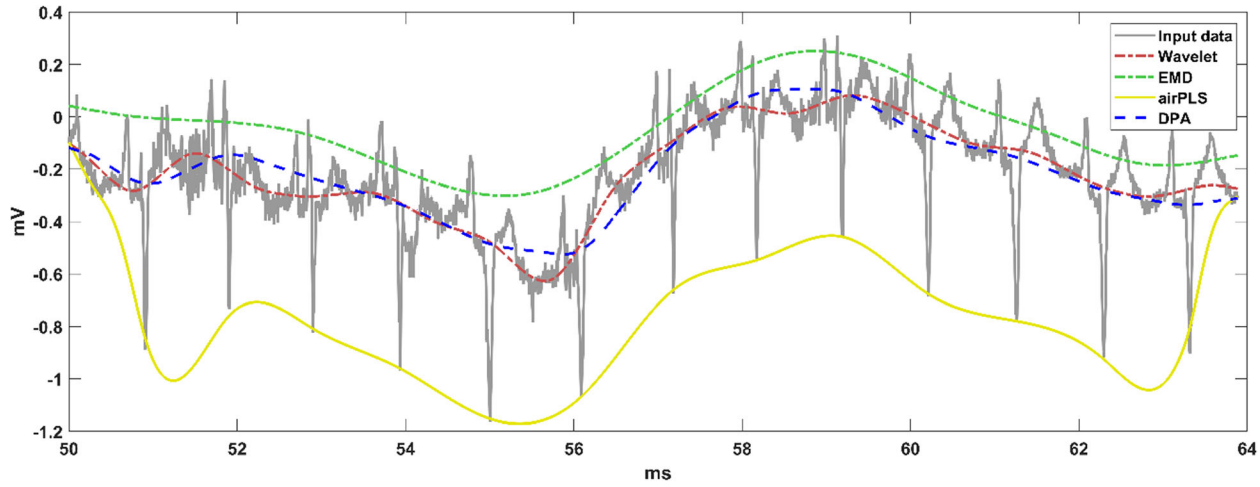

Fig. 4 The baseline detections implemented on the ECG data using the comparison algorithms. The DPA method works well and the result of the airPLS method on this data-trace is invalid

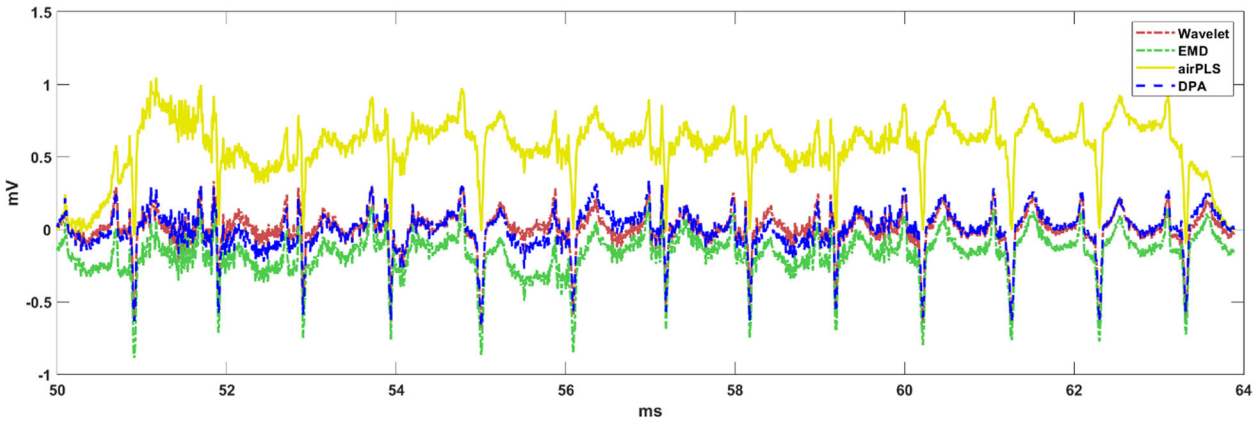

Fig. 5 The baseline corrections implemented on the ECG data using the comparison algorithms. The waveform appears more stable after the baseline removal using the DPA method

the Gaussian peak, which was a widely applied standard model. Universal peaks with different heights and widths could be produced by randomly choosing the mean and variance in the Gaussian model. The noise was generated by sampling in a uniform distribution. The typical synthesized data are shown in Fig. 8.

We implemented the four methods to remove the baseline drift. The performance of the methods was measured by the peak area loss rate. Since the position and area of the signal peak were precisely known, the peak area in the corrected trace could be calculated at the preset peak location. The loss rate of each algorithm is given by comparing the results with the preset peak area. Since the classical wavelet, EMD and DPA method were able to locate the signals directly, we examined the peak identification performance in addition.

The results were graphically presented in Fig. 9. From the results, it could be concluded that the DPA method performed better than the other comparative methods in the baseline detection. The error of the baseline that unfortunately occurred in the signals' interval caused the peak area loss. In this regard, the DPA outperformed the wavelet and EMD methods but not the airPLS. As the famous baseline correcting method, the airPLS method was outstanding in most occasions. Its disadvantages were that it was not able to process the ECG data and the signal peak could only be extracted a step behind rather than simultaneously.

As mentioned above, except for the airPLS method, the classical wavelet, EMD and our DPA methods were implemented based on transforming. The three methods were able to directly locate the signal peaks by setting

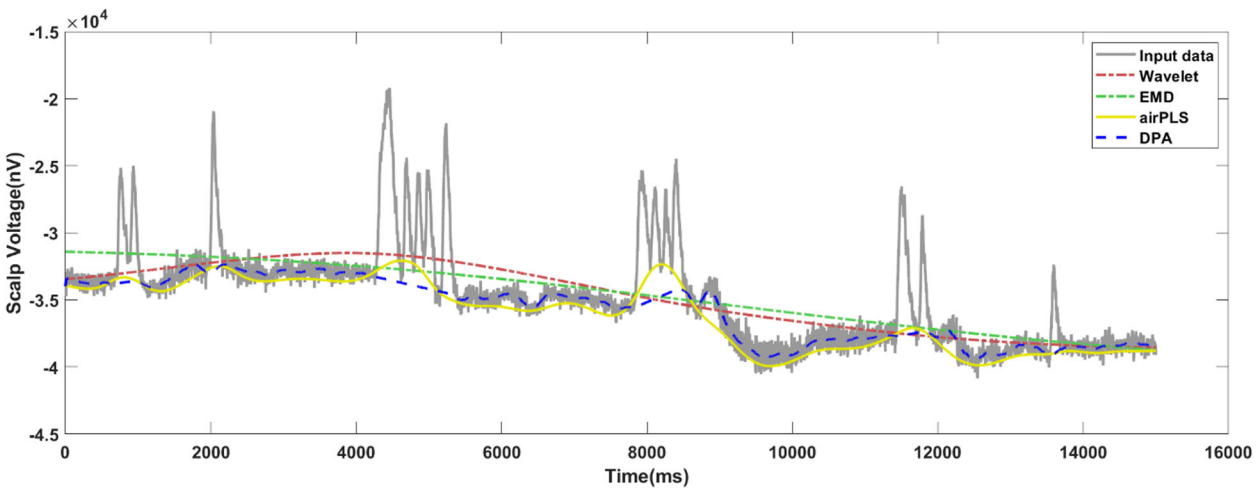

Fig. 6 The baseline detection results for the motor imagery EEG data using the comparison algorithms. The DPA method generates a more accurate result

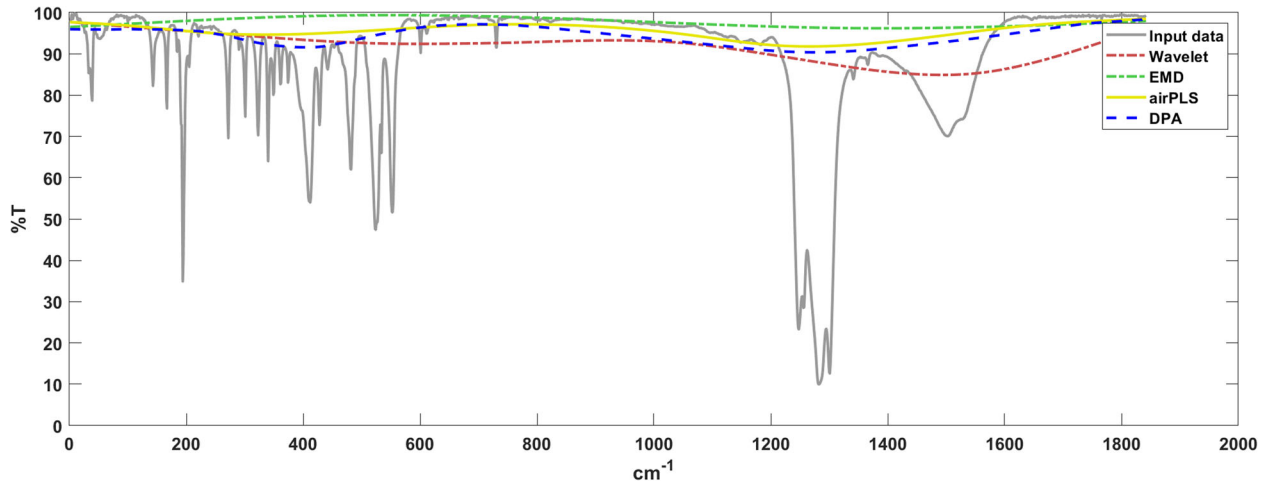

Fig. 7 The baseline detection results for the infrared spectrum using the comparison algorithms. The airPLS and DPA methods performed similarly and are better

the threshold in the transformed intermediate representation. We carried out a set of experiments on these three methods to test the detection accuracy (as measured by the peak missing rate) and area loss rate on the correctly recognized peaks. The numerical results are presented in Table 1. From this set of results, we claim that the DPA method performed better than its two companions with respect to the signal extraction precision.

The limit of SNR for the algorithm

We added noises with different level of signal-to-noise ratio (SNR) to give the limit bound of the algorithm.

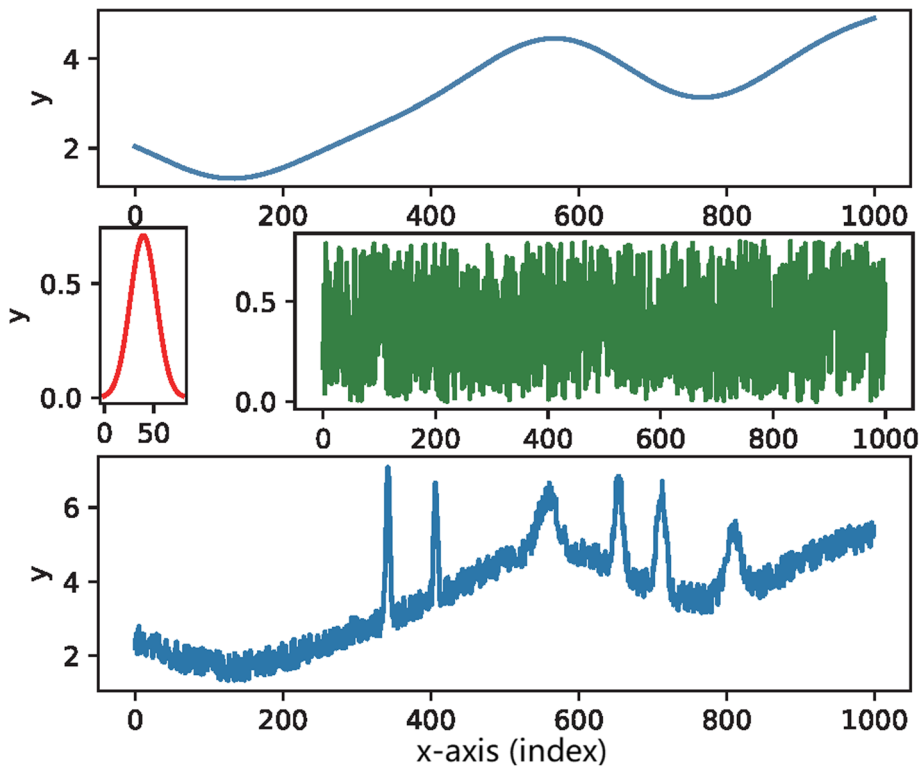

Fig. 8 Upper: the baseline, Middle left: the signal peak, Middle right: the noise, and Bottom: the synthesized data-trace

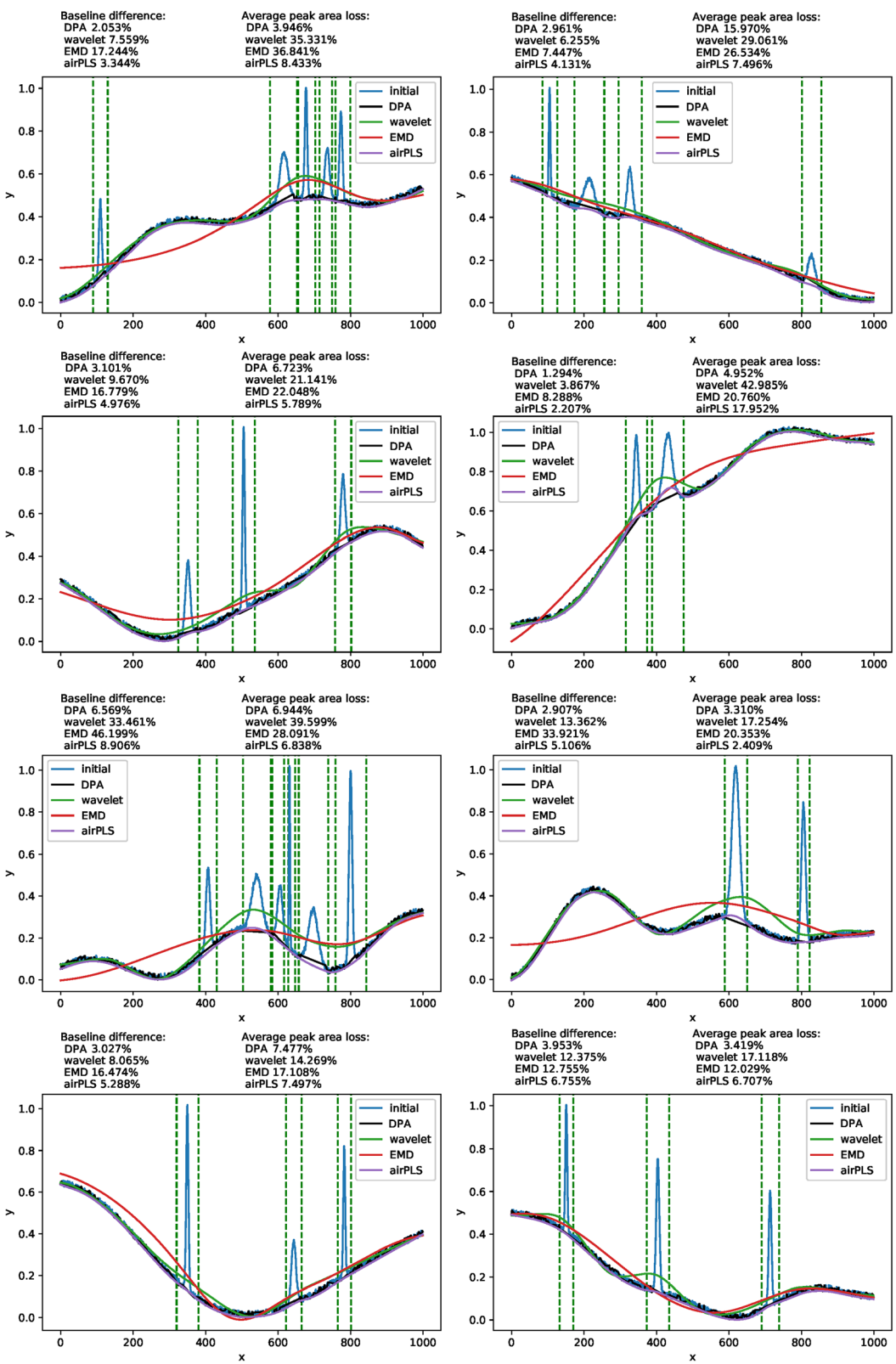

Fig. 9 The testing results of the four comparison methods implemented on randomly synthesized simulated signal traces

Table 1 the peak extraction accuracy. For both peak missing rate and area loss rate, the lower the better

Testing Number	Method					
	DPA		Wavelet		EMD	
	peak missing rate	area loss rate	peak missing rate	area loss rate	peak missing rate	area loss rate
1	0.00%	7.29%	0.00%	15.74%	0.00%	11.82%
2	0.00%	9.86%	0.00%	15.29%	0.00%	12.09%
3	0.00%	13.47%	0.00%	21.04%	0.00%	19.25%
4	0.00%	17.20%	10.71%	13.76%	3.57%	19.18%
5	0.00%	15.68%	0.00%	36.56%	0.00%	9.27%
6	0.00%	19.16%	0.00%	26.74%	0.00%	19.07%
7	3.45%	2.57%	0.00%	19.55%	0.00%	24.07%
8	0.00%	3.13%	0.00%	26.02%	18.18%	17.21%
9	0.00%	8.62%	0.00%	36.43%	0.00%	36.02%
10	0.00%	4.72%	0.00%	35.12%	0.00%	28.44%
11	0.00%	11.26%	0.00%	47.87%	0.00%	16.22%
12	0.00%	3.60%	0.00%	44.72%	0.00%	8.37%
13	0.00%	9.21%	0.00%	39.72%	0.00%	15.70%
14	0.00%	8.78%	3.33%	35.35%	0.00%	10.60%
15	0.00%	6.83%	0.00%	16.24%	0.00%	38.08%

Experiments were also carried out both on authentic and synthesized data.

Figures 10, 11 and 12 present the results on motor imagery EEG data corresponding to the raw signal, added noise with SNR of 7 dB and with SNR of 6 dB respectively. From Fig. 11 where SNR is 7 dB for the added noise, the baseline detected by the DPA method was close to the one in the raw data. For the data with the SNR of 6 dB, there was a bump (marked by the red frame in Fig. 12) which implied some overcut. Therefore, the limit of the DPA in the data was set as 7 dB.

Figures 13, 14 and 15 present the results of X-ray diffraction data. The DPA algorithm generated accurate baseline on the raw X-ray diffraction data (shown in Fig. 13) and it was still valid when 6 dB noise was added (shown in Fig. 14). An overcut appeared when the SNR declined to 5 dB (shown in Fig. 15).

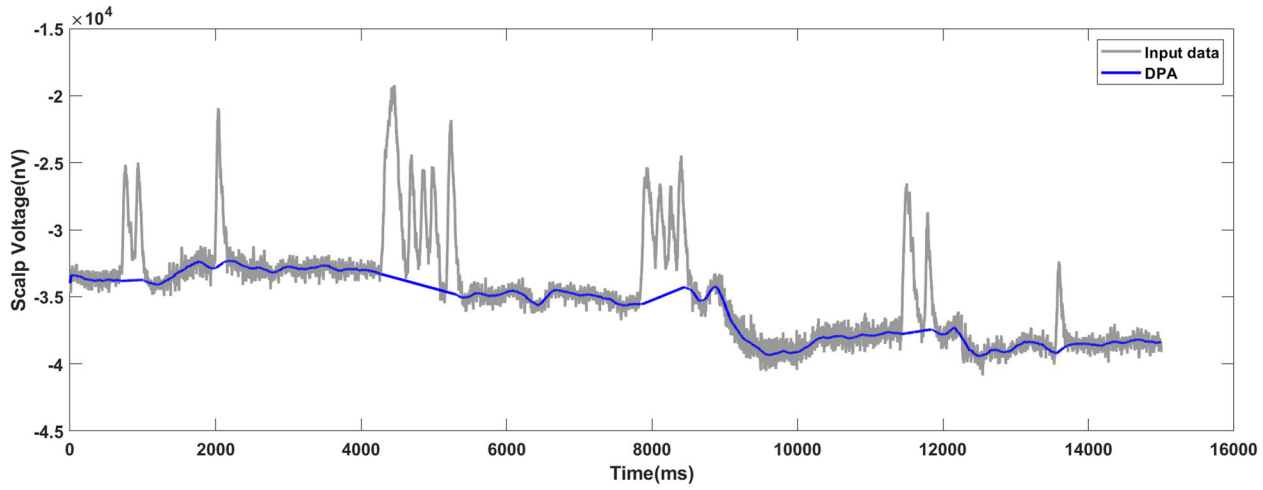

Fig. 10 The DPA method applies on the raw data of motor imagery EEG and generates an accurate baseline

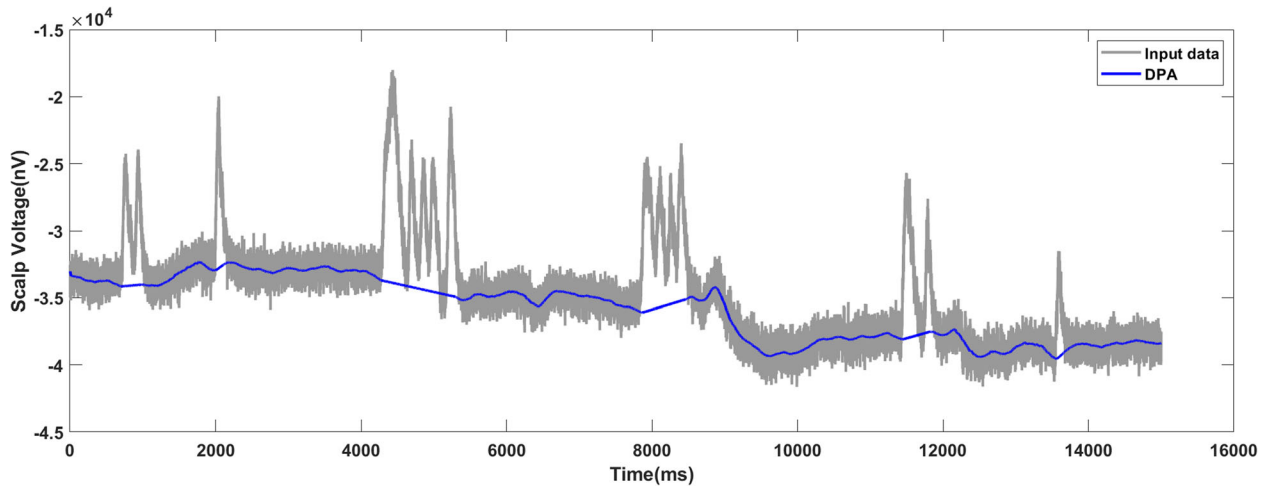

Fig. 11 The DPA method generates a valid baseline on the data-trace with 7 dB SNR

Figure 16 presents the results generated on synthesized data with varying SNR. Subfigures from the first to the third row show the baselines detected on the data with 7 dB, 6 dB, and 5 dB SNR. The results showed that from 7 dB to 6 dB, DPA algorithm output reasonable baselines. When implemented on 5 dB data, DPA failed to produce accurate results due to the overcut marked in the red frame.

From these random testing results, we conclude that the DPA algorithm could perform well if the level of SNR was better than 7 dB. Since this value is an acceptable limit, we claim that the newly developed method is practical.

Discussion

The testing results on both authentic and synthesized data indicated that the newly developed Derivative Passing Accumulation (DPA) method in this article outperformed other classical baseline detection methods. For the signals with random peaks oriented to the same direction, airPLS and DPA methods output similar results which were better than the others. For the data where signal peaks oriented towards both positive and negative directions (for example, the ECG signals), airPLS failed while EMD, wavelet and DPA methods performed well. Generally speaking, the DPA method was wider applicable and more stable. It generated accurate baselines in most cases. We have also tested the limitations of the new method. Noises in different ratios were added to the raw signals and it was found that DPA worked well under at least 7 dB, which was a practical level.

The DPA method was not able to produce valid results when the signal peaks were not fully recovered in the

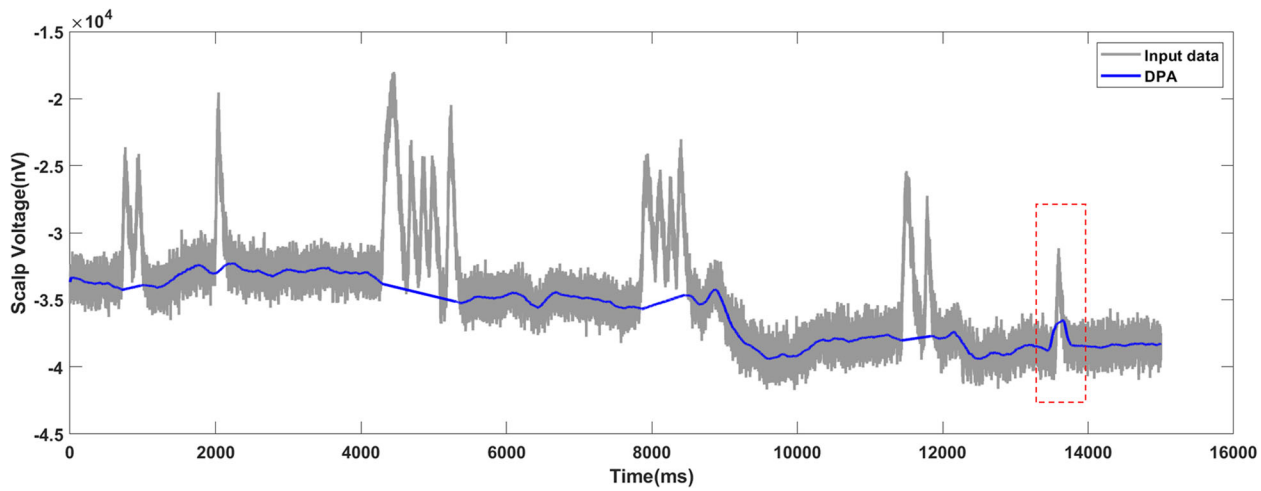

Fig. 12 The DPA method applies on the motor imagery EEG data with 6 dB SNR. The detected baseline is with an inappropriate bump (marked by the red frame)

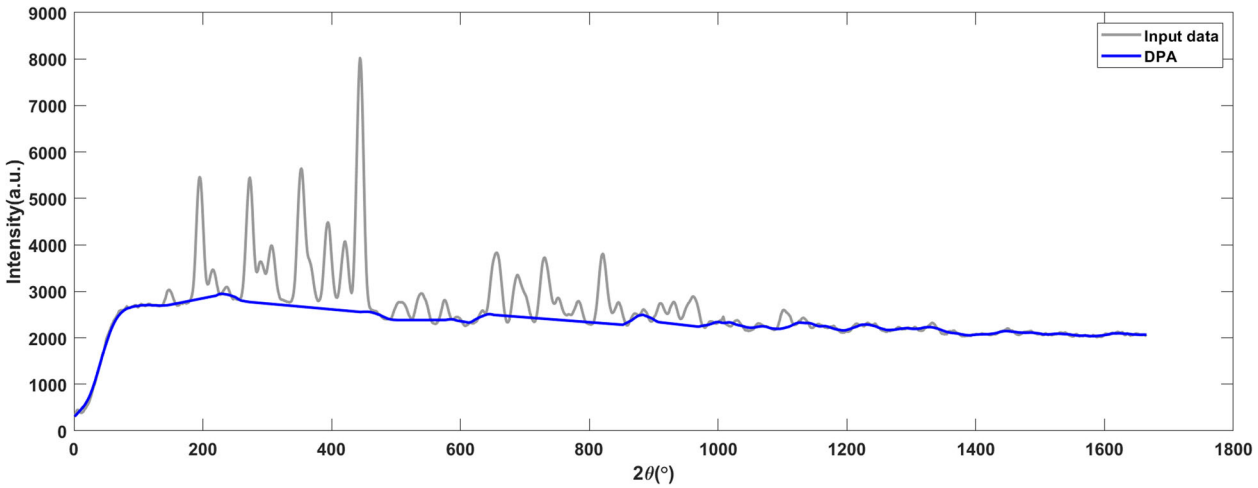

Fig. 13 The DPA algorithm generates an accurate baseline on the raw X-ray diffraction data

accumulated derivative form. In this case, the position belonging to the signal interval might be identified as the background point. That's why, when the DPA algorithm failed, the result was always overcut. Sometimes the accumulated form was sufficient. However, the slight misplacement of the identified signal interval led to an obvious inappropriate bump in the generated baseline. The uncertainty may be a main drawback of the DPA method.

Conclusions

Signal processing plays an important role in biological data analysis. It has a strong impact on the accuracy of downstream operations leading up to the analysis output. The new method developed in this article was able to simultaneously implement baseline removal and peak detection, which constituted the main content of the signal processing stage. Relying on the simple passing accumulation procedure and with the aid of the non-maximum suppression strategy, the proposed DPA method could conduct rapid and automatic calculations.

The results of comparison with the different algorithms that were applied to real-life biological data showed that the new method was more robust in a wide range of applications. We further measured the processing performance by testing the peak area loss rate for the synthesized data. The results also indicated that the DPA method had a superior accuracy. In addition, the operation under the passing accumulation also revealed its potential value for processing higher dimensional scenarios beyond the data stream.

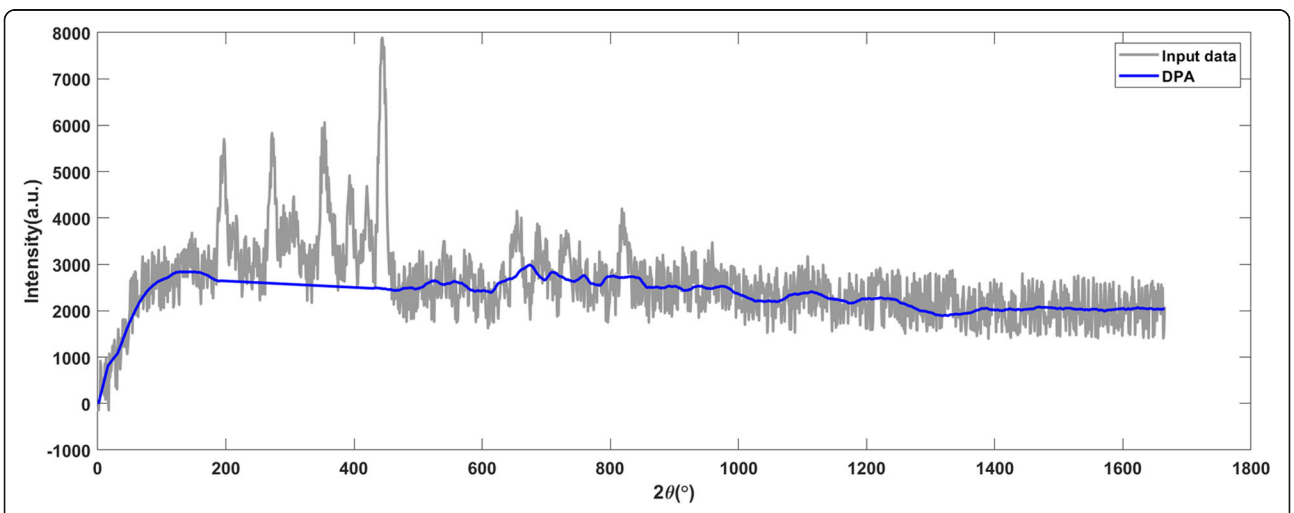

Fig. 14 The DPA algorithm generates an accurate baseline on the raw X-ray diffraction data with added noises. The signal to noise ratio is 6 dB

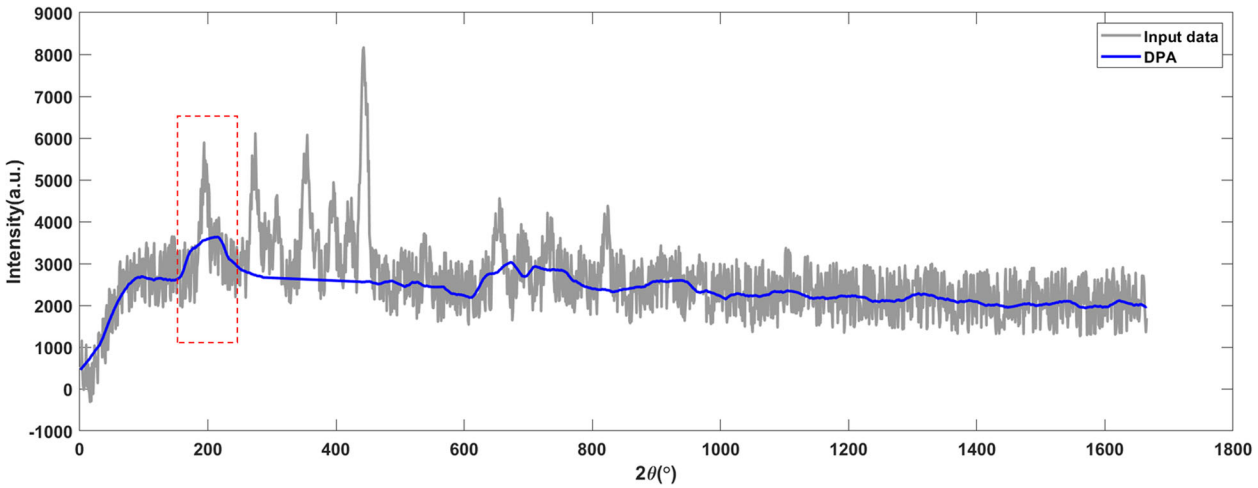

Fig. 15 The DPA method applies on the X-ray diffraction data with 5 dB SNR. The detected baseline is with an inappropriate bump (marked by the red frame)

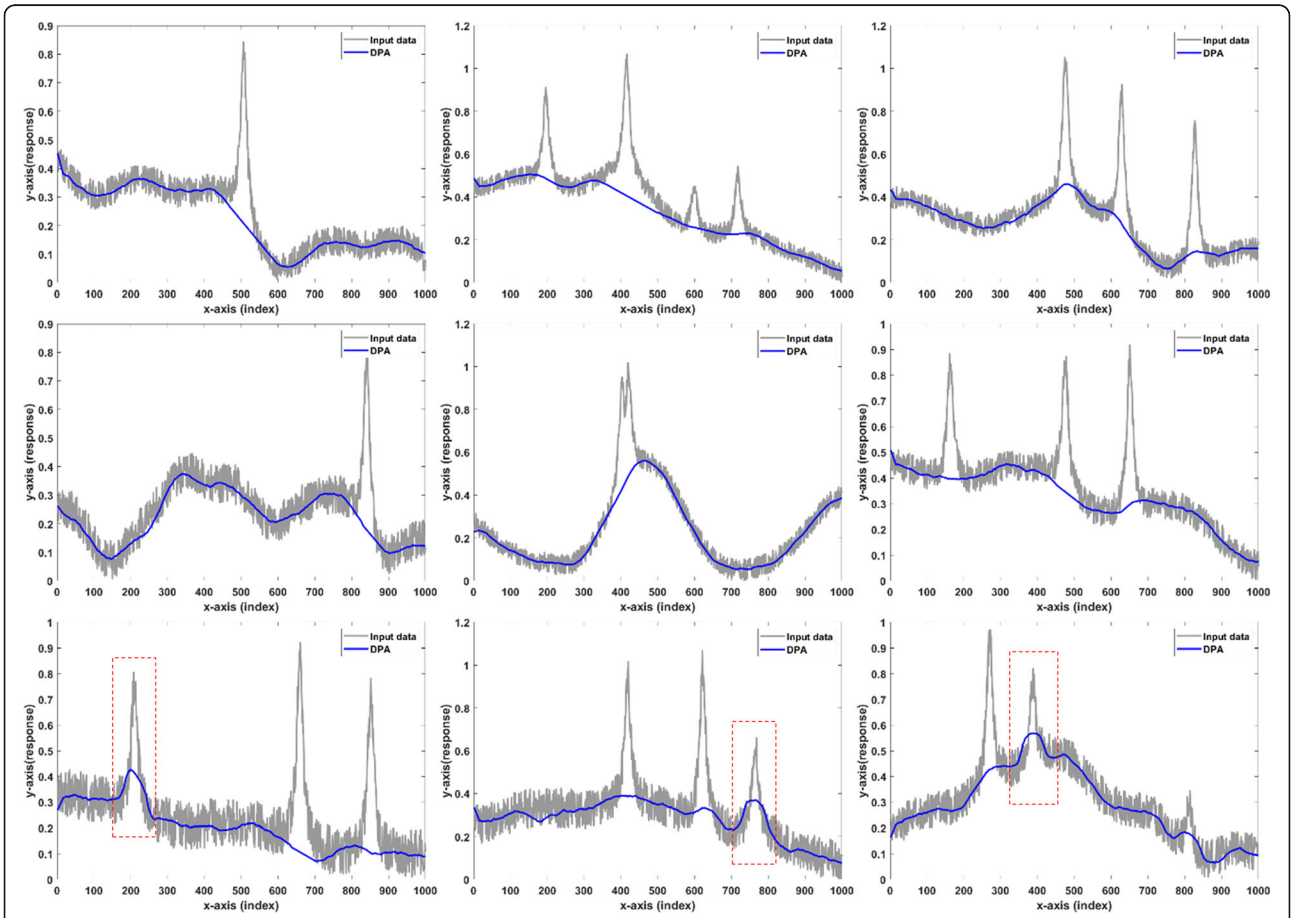

Fig. 16 From 7 dB to 6 dB, DPA algorithm outputs reasonable baselines. When implemented on 5 dB data, DPA fails to produce accurate results due to the overcut marked in red frame. The first row: the baselines detected on the data with 7 dB SNR. The second row: results on the data with 6 dB SNR. The third row: results on the data with 5 dB SNR

Further applications in image processing and the comprehension of higher dimensional mathematical meanings could be studied in future.

Methods

The derivative is a local variable quantity that is not influenced by baseline drift. Explicitly, the background fluctuation is very small in a narrow interval, which could be omitted when considering the derivative. Only the local rise and fall apparently affect the derivative, so it is mainly dominated by the signal peaks. Here, we adopted the easiest operation of simple differences to obtain the discrete derivatives of the waveform. When we applied the differential operator on the data-trace, the rising interval had a positive ascent and the falling interval had a negative descent. The crest was the watershed that laid between the rising and falling parts. A practical data-trace illustrating this basic fact is presented in Fig. 17.

The first-order derivative trace was divided into positive and negative parts, which are denoted as P and N, respectively. Both parts carried the information of the initial data-trace reflecting the signal shape and location, but the baseline drift was eliminated. The basic idea of the new algorithm was utilizing this derivative to reconstruct the signal peaks and estimate their locations to discriminate the signal and background intervals. The procedure was explained as follows and formally described in Algorithm-1.

For the negative part N, its absolute value was used to flip the trace to be positive, as shown in Fig. 18 (the green line in the upper left of Fig. 18 is the negative part that is denoted as N, and the cyan line in the upper right of Fig. 18 is the flipped absolute curve). The flipped part is denoted as N^+.

While overlaying the two vectors P and N^+, it is easy to see that the derivative traces look like steady peaks that are split from the initial trace at the crests (as shown in Fig. 19). Moreover, the interval of these peaks falls in the range of the corresponding initial peaks.

Derivative Passing Accumulation

The key operation for rebuilding the signals is to shift and accumulate the trace of P and N^+.

The procedure is named the passing accumulation operation and the schematic diagram is presented in Fig. 20. The detailed calculation is given as follows. A shift width k is designated. Set $\alpha = \{a_0, a_1, \cdots, a_m\}$, and initialize a_i, where $i = 0, \cdots, m$, to 0. Move $P = \{p_0, p_1, \cdots, p_m\}$ and $N^+ = \{n_0, n_1, \cdots, n_m\}$ to each other $w + 1$ times and accumulate the result as α. In each step j, $a_i = a_i + p_{i-j} + n_{i+j}$. Thus, the result of our passing accumulation is presented by Eq. (1)

$$a_i = \sum_{j=0}^{w} p_{i-j} + \sum_{j=0}^{w} n_{i+j} \tag{1}$$

where w is the maximum shifting width and is manually set up. The newly defined computation is named as the derivative passing accumulation and denoted as DPA for short. The effectiveness of this computation is presented in Fig. 21, in which it could be intuitively seen that the data-trace of a Raman spectroscopy with serious baseline drift is straightened and the signal peaks are kept and augmented. The accumulating procedure is summarized in Algorithm-1.

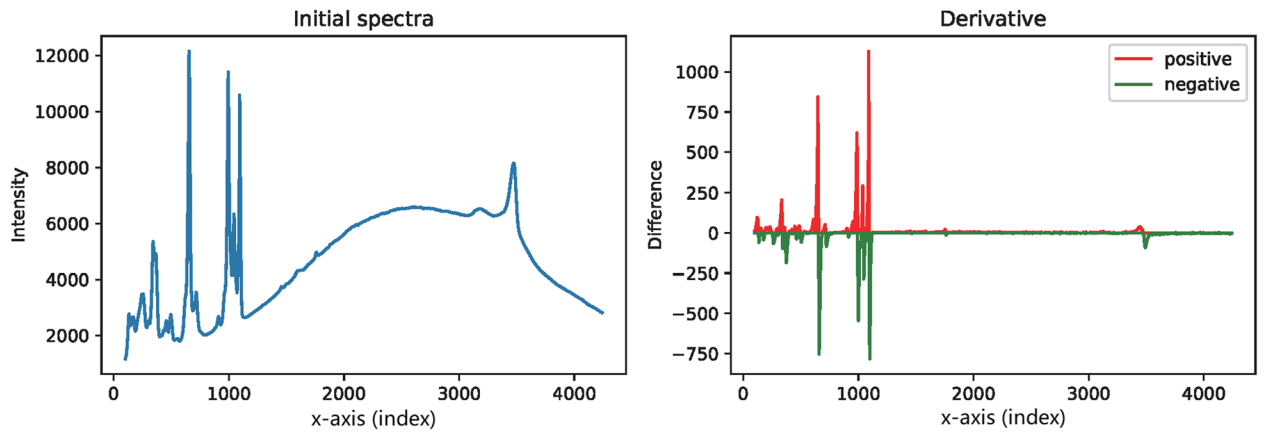

Fig. 17 The plot of the derivative. Left: the initial raw spectra data (The x-axis can be assigned with any independent variable, or simply use the index of the point in the data sequence). Right: the plot of the derivative on each data point. The derivatives are categorized into positive part and negative part

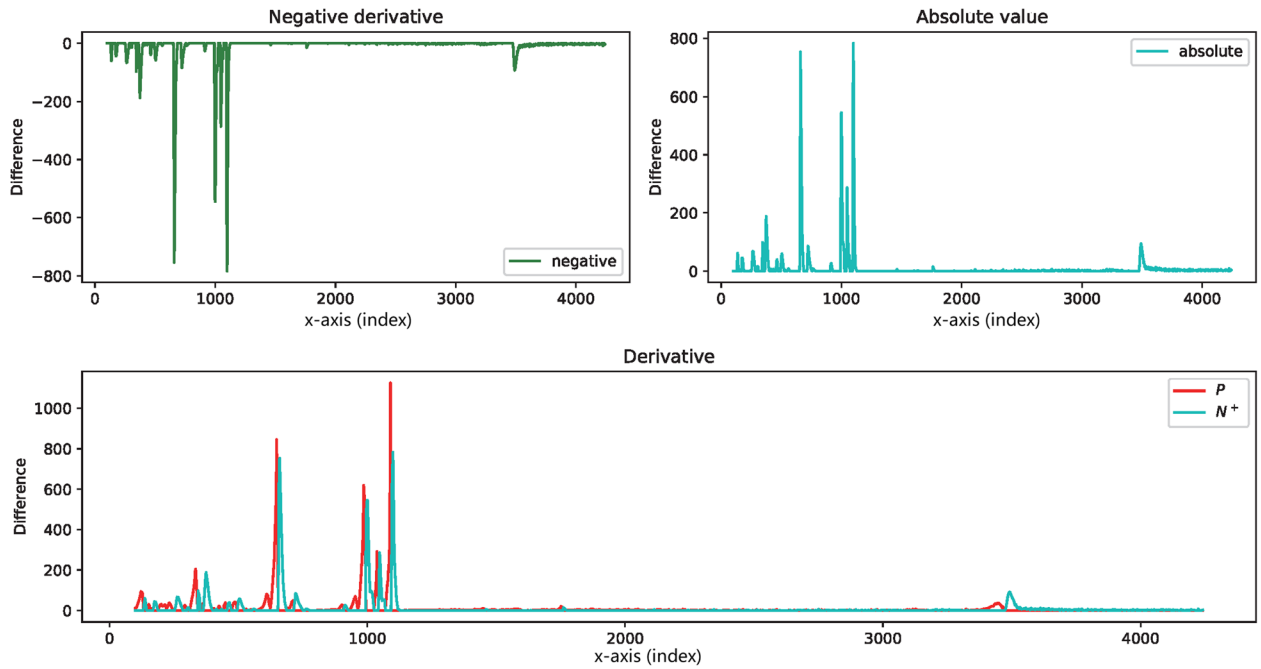

Fig. 18 Upper left: the negative part N, upper right: flip N to get its positive version of N^+, and bottom: overlay the trace of P and N^+ at the same position

Algorithm 1 Derivative Passing Accumulation Operation

Input: $I = \{(x_0, y_0), (x_1, y_1), \cdots, (x_m, y_m)\}$, the data points of the data-trace with length $m+1$; an integer w, indicating the maximum shift width.

Output: A vector representing the result of the data trace under the transformation.

1: Calculate the derivative trace using the first-order difference quotient to obtain an array of $d = [d_0, d_1, \cdots, d_{m-1}]$, where $d_k = \frac{y_{k+1} - y_k}{x_{k+1} - x_k}$.

2: Split d into the positive part and negative part, which are denoted as P and N, respectively:

$$P = [p_0, p_1, \cdots, p_{m-1}]$$

$$N = [n_0, n_1, \cdots, n_{m-1}]$$

where $p_k = \begin{cases} d_k, & d_k \geq 0 \\ 0, & d_k < 0 \end{cases}$ and $n_k = \begin{cases} d_k, & d_k \leq 0 \\ 0, & d_k > 0 \end{cases}$.

3: Set at a zero vector $\boldsymbol{\alpha}$ with length $m+1$ to record the accumulated quantity $\boldsymbol{\alpha} = [a_0, a_1, \cdots, a_m]$, where $a_i = 0, k = 0, \cdots, m$.

4: **for** i from 0 to m **do**

5: **for** j from 0 to w **do**

6: $a_i = a_i + p_{j-i} - n_{j+i}$

7: **return** $\boldsymbol{\alpha}$ as the result of the transformation.

Automating

In the preliminary version of the algorithm, the maximum shift width w was set manually and uniformly. In the following upgraded version, the non-maximum suppression was applied to automatically determine w for each subscript in the derivative array, which thereby improved the whole algorithm to become fully self-driven. Non-maximum suppression (NMS) is a widely used technique in computer vision tasks, such as the Faster R-CNN [24]. The basic idea is to select the object according to its descriptive value, i.e. to examine if its value was the maximum among all those intersecting with it. In our case, this principle could be explained as follows.

On a fixed point (x_i, y_i) in the data-trace, for each width w, we could compute the accumulation according to Eq. (1) to get a function with respect to w,

$$A(w) = \sum_{j=0}^{w} p_{i-j} + \sum_{j=0}^{w} n_{i+j} \tag{2}$$

We would like to determine the width w_i automatically for position i (corresponding to the data point (x_i, y_i)). According to the principle of non-maximum suppression, we select w_i which meets the requirement of Eq. (3),

$$A(w_i) = \max_{|t-i| \leq w_i} A(w_t) \tag{3}$$

Substituting (2) into (3), we get that the width w_i must satisfy the condition in (4),

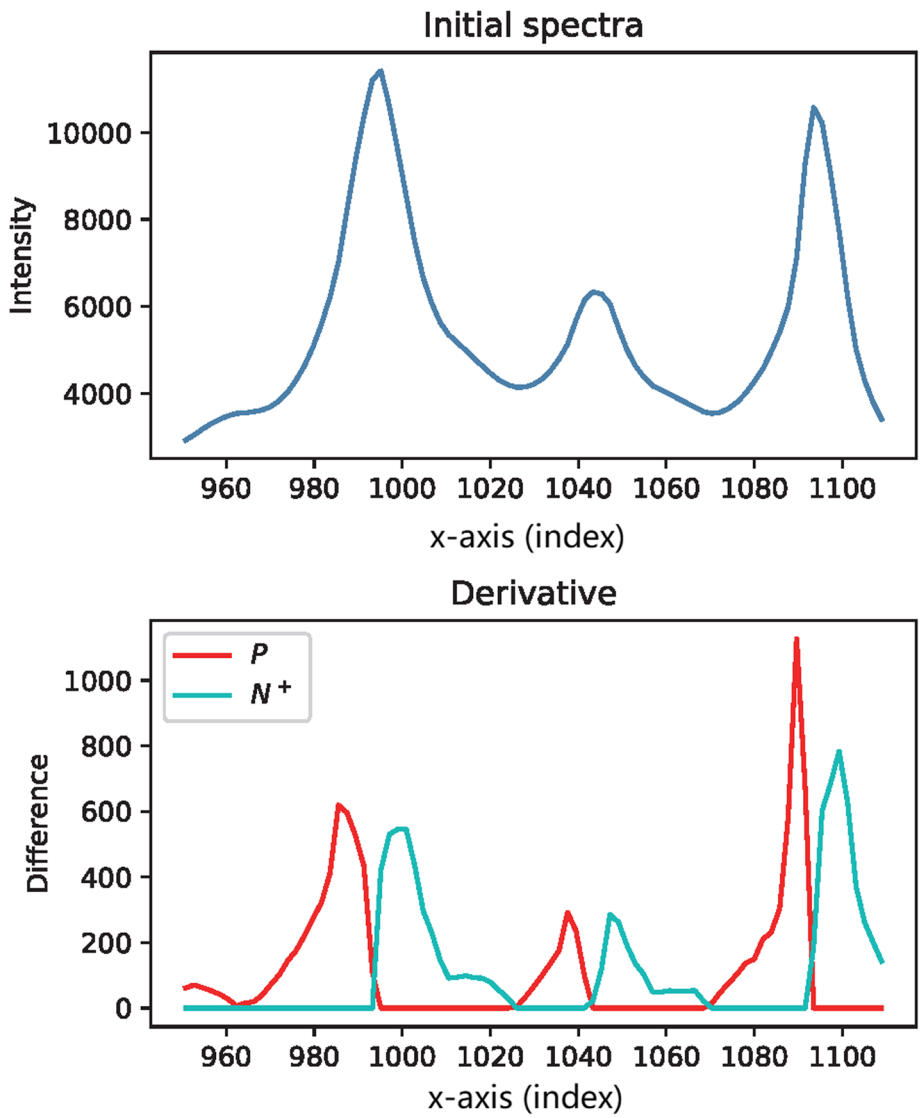

Fig. 19 Compare the initial trace with the overlaid derivatives. Upper: an arbitrary trace representing the general raw signal data. The x-axis could be assigned with any independent variable, or simply use the index of the point in the data sequence. Bottom: the overlay of the positive part P and the negative part N^+ of the derivative. Remark: the overlaid diagram forms a profile similar to the initial trace. The fact inspired the idea of using the two parts of the derivative to reconstruct the signals

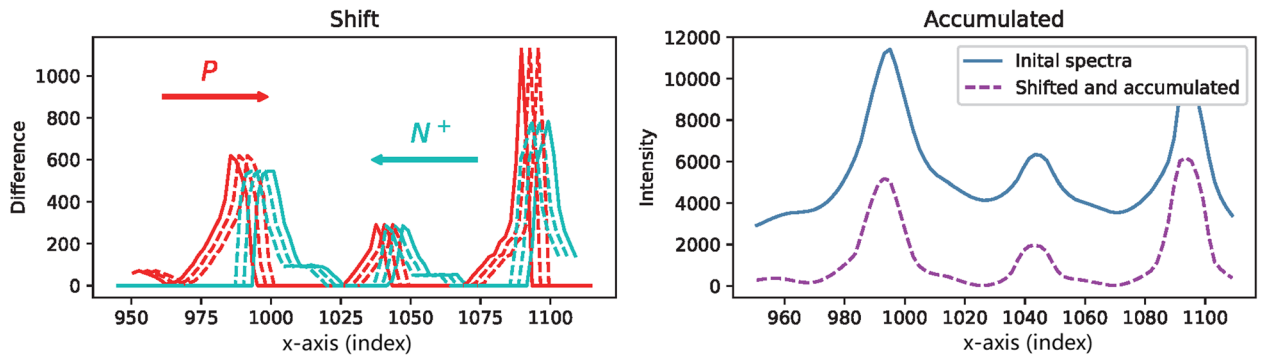

Fig. 20 The passing accumulation operation. Left: slide positive part and negative parts toward each other to pass through. The value is accumulated on each shift. Right: the accumulated result compared with the initial raw spectra

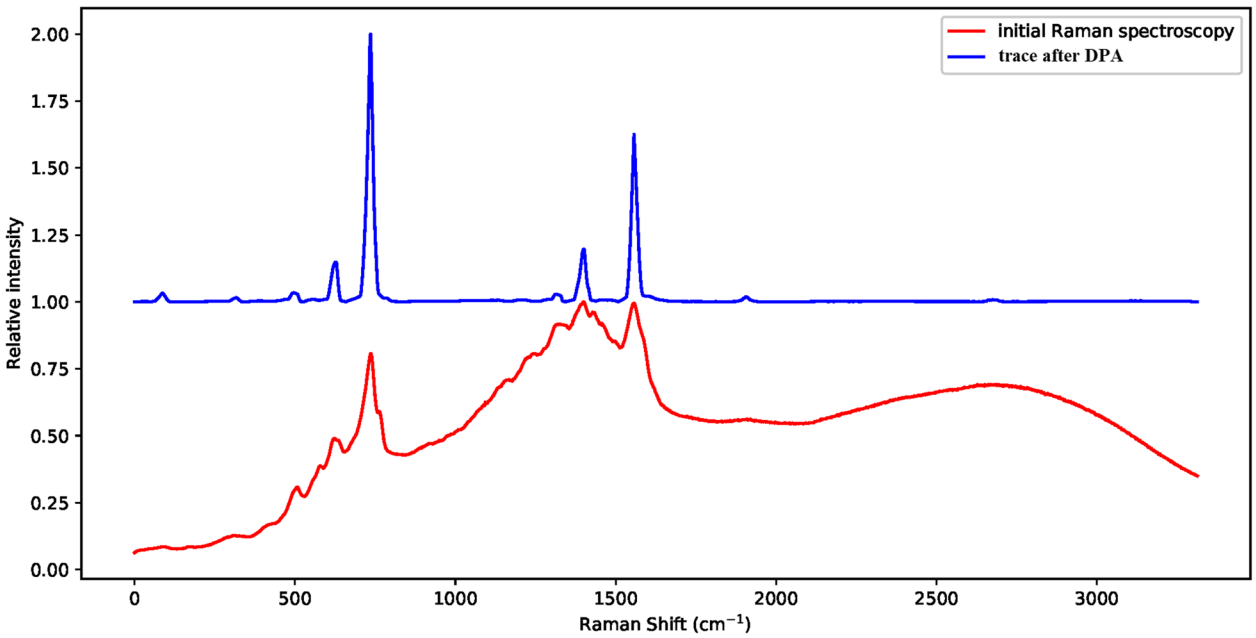

Fig. 21 The effect of DPA computation applied on an authentic Raman spectroscopy. Compared with the initial trace, the peaks are kept and augmented at the same position and the baseline is straightened

$$w_i = \left\{ k | k \in Z^+, \sum_{j=0}^{k} p_{i-j} + \sum_{j=0}^{k} n_{i+j} = \max_{|t-i| \le k} \left(\sum_{j=0}^{w_t} p_{t-j} + \sum_{j=0}^{w_t} n_{t+j} \right) \right\} \tag{4}$$

We just need to determine an appropriate w_i according to Eq. (4) instead of finding out all the possible solutions. To implement the procedure in the program, we can execute the accumulation adaptively by checking if the requirement is met during the passing. In the computation according to Eq. (1), the summation for each index i could be implemented as follows. Set $a_i = 0$ and start a loop to grow j from 0 to $\frac{L}{2}$ (L represents the length of the derivative array). Accumulate the p_{i-j} and

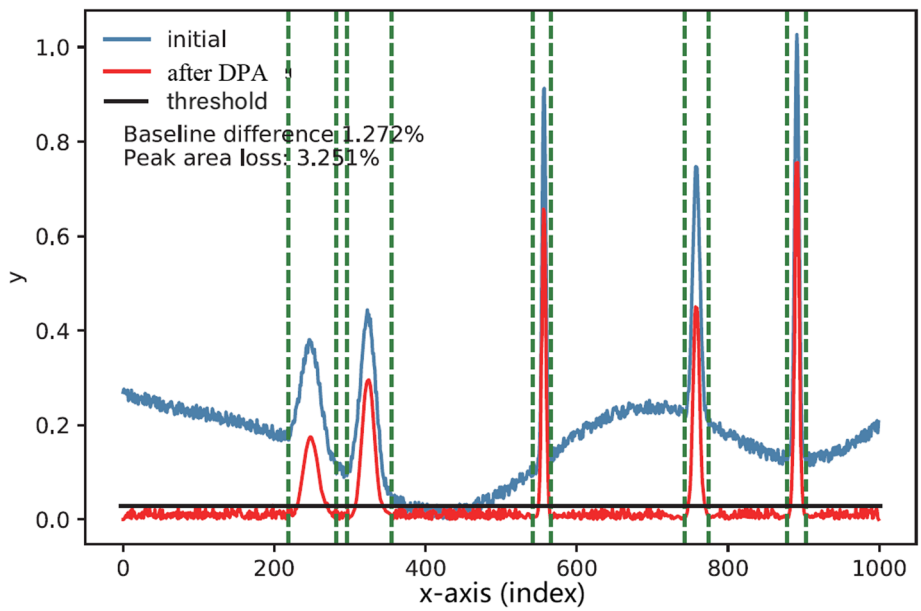

Fig. 22 The peak identification results. Set a threshold on the result after DPA operation, in which the baseline is removed and signals are kept. Then, the positions of peaks are identified and the points belonging to the baseline in the initial trace are extracted

n_{i+j} to a_i. For each step in the loop, examine if the accumulated value was maximum in its j-nearest neighborhood. When the maximum requirement is not satisfied, the accumulation at this index stops and the accumulated value is stored. It also stops if all the j nearest neighbors stop growing. The corresponding width w is set equal to j consequently. The adaptive passing accumulation is accomplished as the termination of the loop or every width w is determined.

We summarized the scheme in Algorithm-2.

Algorithm 2 Auto-DPA with adaptive accumulation width

Input: $I = \{(x_0, y_0), (x_1, y_1), \cdots, (x_m, y_m)\}$, the data points of the data-trace with length $m + 1$.

Output: A vector representing the result of the data-trace under the transformation.

1: Calculate the derivative trace using the first-order difference quotient to obtain an array of $d = [d_0, d_1, \cdots, d_{m-1}]$, where $d_k = \frac{y_{k+1} - y_k}{x_{k+1} - x_k}$.

2: Split d into the positive part and negative part, which are denoted as P and N, respectively:

$$P = [p_0, p_1, \cdots, p_{m-1}]$$

$$N = [n_0, n_1, \cdots, n_{m-1}]$$

where $p_k = \begin{cases} d_k, & d_k \geq 0 \\ 0, & d_k < 0 \end{cases}$ and $n_k = \begin{cases} d_k, & d_k \leq 0 \\ 0, & d_k > 0 \end{cases}$.

3: Set a zero vector α with length $m + 1$ to record the accumulated quantity: $\alpha = [a_0, a_1, \cdots, a_m]$, where $a_i = 0, k = 0, \cdots, m$. Set $\beta = [b_0, b_1, \cdots, b_m]$ to buffer the intermediate α for the non-maximum suppression operation and set $\gamma = [c_0, c_1, \cdots, c_m]$ as the flag bit for stop criteria, where $b_i = 0$, $c_i = 0$, for $k = 0, \cdots, m$.

4: **for** w from 1 to $\frac{m}{2}$ **do**

5: **for** i from w to $m - w$ **do**

6: **if** c_i is the only 0 in $c_{i-w}, \cdots, c_{i+w}$ **then**

7: $c_i = 1$

8: **if** $c_i == 1$ **then**

9: **continue**

10: **if** a_i is the maximum in $a_{i-w}, \cdots, a_{i+w}$ **then**

11: **for** j from 1 to w **do**

12: $b_i = b_i + p_{j-i} - n_{j+i}$

13: **else** $c_i = 1$

14: **for** i from w to $m - w$ **do**

15: $a_i = b_i$

16: **return** α as the result of the transformation.

The complete algorithm

With the help of non-maximum suppression strategy, the DPA was upgraded into an automatic pipeline. With the converted waveform T, it was straightforward to get the final results. In T, since the baseline drift was removed and the peaks were kept in the corresponding interval, we just needed to select a threshold [25] to divide the array into peak points and baseline points. In this way, the signal peaks were extracted. In addition, the baseline could be constructed by linearly connecting the key points that were selected according to the identified baseline points. Figure 22 illustrated the schematic diagram.

The complete procedure was summarized in Algorithm-3 and named as the derivative passing accumulation method (DPA).

Algorithm 3 Derivative Passing Accumulation Method

Input: $I = \{(x_0, y_0), (x_1, y_1), \cdots, (x_m, y_m)\}$, the data points of the data trace with length m+1.

Output: Recognized peak list and baseline.

1: Convert I to α by applying Algorithm-2.

2: Apply the bi-trapezoid thresholding criteria on α to set a threshold denoted as t. The point with the abscissa of x_k is identified as the background point if $y_k < t$; otherwise, it is identified as the signal point.

3: All the background points are labeled and formed into a sequence $b = \{(x_{u_0}, y_{u_0}), (x_{u_1}, y_{u_1}), \cdots, (x_{u_L}, y_{u_L})\}$, where $y_{u_j} < t$, $j = 0, 1, \cdots, L$.

4: On the x-axis, examine the sequence of b, and all of the absent intervals are the candidate signal peaks' locations. Check these intervals and eliminate the apparently narrow ones and merge the too close ones. The finally recognized peak locations are denoted as $p = \{[x_{c_0 0}, \cdots, x_{c_0 d_0}]_0, \cdots, [x_{c_r 0}, \cdots, x_{c_r d_r}]_r\}$. Update b by removing p from the corresponding abscissa in I.

5: Build a list of key points K as the anchors of the baseline: $K = \{(x_{v_0}, \bar{y}_{v_0}), \cdots, (x_{v_r}, \bar{y}_{v_r})\}$, where x_{vi} is the position averaged from b, and $\bar{y}_{vi}$ is the corresponding averaged y value in I under a density for stability.

6: Use a polyline to sequentially connect the key points in K to get the detected background B.

7: Subtract B from I and select the corresponding signal arrays of p to get the peak list P.

8: **return** the peak list P and baseline B.

Perspectives

In future, studies will be carried out in two directions.

First, we plan to improve the performance based on fractional derivative techniques. Because, currently there is a trend of exploiting fractional derivatives for solving different identification problems. These show better results than standard first-order derivative-based algorithms [26–28]. If we find a way to utilize the fractional derivative for recovering the signals, a more accurate separation of the baseline intervals will be achieved, leading to better results.

Second, we will work on extending the application of the DPA method to a 2-dimensional case. We will study to discover the way of appropriately defining the accumulation of derivatives for 2-D function. Thus, develop the scheme for extracting the background, based on the accumulated results. Consequently, the upgrade to a higher dimension will enable the DPA strategy to process image data and expand the application range.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12859-019-3188-4.

Additional file 1. Testing data of the mass spectroscopy, infrared spectroscopy and energy curve of audio.

Abbreviations
DPA: Derivative Passing Accumulation method; ECG: Electrocardiograms; EEG: Electroencephalograms

Acknowledgements
We thank Dr. Xiaoguang Zhou and Dr. Yuntao Li, Intelligene Biosystems Qingdao, Shandong, China for the data support.
We thank Professor Mateen Khattak, Professor Desheng Jiang, and Ms. Chong Wang for their help with the language editing.

Authors' contributions
YL led the design and the implementation of the overall procedure, its evaluation and drafted the manuscript. JL supervised the design of the procedure, its evaluation and finalized the manuscript. All authors read and approved the final version of the manuscript.

Funding
This work was supported by the National Natural Science Foundation of China (Grant no. 61807032) and the Fundamental Research Funds for the Central Universities of China (Grant no. 2019TC045).
These funding bodies did not play any role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Availability of data and materials
The datasets used to generate the figures presenting results on Raman spectra, ECG profiles were downloaded from public databases which were cited within the manuscript. The data used to generate the figures presenting results on the energy curve of the audio, mass spectroscopy and infrared spectroscopy are available in the Additional file 1. The synthesized data for random testing were generated following the procedure explained within the manuscript.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 14 June 2019 Accepted: 4 November 2019
Published online: 27 November 2019

References
1. Assis MWD, De Fusco DO, Costa RC, Lima KMG, Cunha LC, Teixeira GHD. "PLS, iPLS, GA-PLS models for soluble solids content, pH and acidity determination in intact dovyalis fruit using near-infrared spectroscopy," (in English). J Sci Food Agric. 2018;98(15):5750–5.
2. Kim JT, Jung SH, Cho K-H. Efficient harmonic peak detection of vowel sounds for enhanced voice activity detection. IET Signal Process. 2018;12(8): 975–82.
3. C. Dora and P. K. Biswal, "Robust ECG artifact removal from EEG using continuous wavelet transformation and linear regression," (in English), 2016 International Conference on Signal Processing and Communications (SPCOM), Conference Paper pp. 5 pp.-5 pp., 2016 2016.
4. Kumar A, Komaragiri R, Kumar M. "Design of wavelet transform based electrocardiogram monitoring system," (in English). Isa Transac. 2018;80:381–98.
5. Patel R, Gireesan K, Sengottuvel S, Janawadkar MP, Radhakrishnan TS. Suppression of baseline wander artifact in Magnetocardiogram using breathing sensor. J Med Biol Eng. 2017;37(4):554–60.
6. Su M, Zheng J, Yang Y, Wu Q. A new multipath mitigation method based on adaptive thresholding wavelet denoising and double reference shift strategy. GPS Solutions. 2018;22:2.
7. Liu Y, Yu Y. A survey of the baseline correction algorithms for real-time spectroscopy processing. In: Photonics Asia, vol. 10026. Beijing: SPIE; 2016. p. 100260Q.
8. Kanginejad A, Mani-Varnosfaderani A. Chemometrics advances on the challenges of the gas chromatography–mass spectrometry metabolomics data: a review. J Iran Chem Soc. 2018;15(12):2733–45.
9. Shen X, et al. Study on baseline correction methods for the Fourier transform infrared spectra with different signal-to-noise ratios. Appl Opt. 2018;57(20):5794–9.
10. Wang Z, Zhang M, Harrington Pde B. Comparison of Three Algorithms for the Baseline Correction of Hyphenated Data Objects. Anal Chem. 2014; 86(18):9050–7.
11. Liu H, Zhang Z, Liu S, Yan L, Liu T, Zhang T. Joint baseline-correction and denoising for Raman spectra. Appl Spectrosc. 2015;69(9):1013–22.
12. Fu HY, et al. Simple automatic strategy for background drift correction in chromatographic data analysis. J Chromatogr A. 2016;1449:89–99.
13. Mani-Varnosfaderani A, Kanginejad A, Gilany K, Valadkhani A. Estimating complicated baselines in analytical signals using the iterative training of Bayesian regularized artificial neural networks. Anal Chim Acta. 2016;940:56–64.
14. Sun Z, Wang X, Wang X, Sun K, Tan Q. Removal of Baseline Wander in ECG Signals Using Singular Spectrum Analysis. In: 2019 IEEE 9th International Conference on Electronics Information and Emergency Communication (ICEIEC); 2019. p. 391–4.
15. Picaud V, et al. Linear MALDI-ToF simultaneous spectrum deconvolution and baseline removal. BMC Bioinformatics. 2018;19(1):123.
16. Liu Y, Zhou X, Yu Y. A concise iterative method using the Bezier technique for baseline construction. Analyst. 2015;140(23):7984–96.
17. Davoudabadi MJ, Aminghafari M. "A fuzzy-wavelet denoising technique with applications to noise reduction in audio signals," (in English). J Intell Fuzzy Syst. 2017;33(4):2159–69.
18. Mariyappa N, et al. Baseline drift removal and denoising of MCG data using EEMD: role of noise amplitude and the thresholding effect. Med Eng Phys. 2014;36(10):1266–76.
19. Santos MCD, Morais CLM, Nascimento YM, Araujo JMG, Lima KMG. Spectroscopy with computational analysis in virological studies: a decade (2006–2016). TrAC Trends Anal Chem. 2017;97:244–56.
20. Cetin AE, Tofighi M. Projection-based wavelet Denoising [lecture notes]. IEEE Signal Process Mag. 2015;32(5):120–4.
21. Lafuente B, Downs RT, Yang H, Stone N. The power of databases: the RRUFF project. In: Armbruster T, Danisi RM, editors. Eds Highlights in mineralogical crystallography. Berlin: W. De Gruyter; 2015. p. 1–30.
22. Moody GB, Mark RG. The impact of the MIT-BIH arrhythmia database. IEEE Eng Med Biol Mag. 2001;20(3):45–50.
23. Goldberger AL, et al. PhysioBank, PhysioToolkit, and PhysioNet. Circulation. 2000;101(23):e215–20.
24. Ren SQ, He KM, Girshick R, Sun J. Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell. 2017;39(6):1137–49.
25. Liu Y, Yu Y, Zhou X, Wang C. A new automatic threshold selecting criteria for spectroscopy data processing. Chemometrics Intell Lab Syst. 2017;161:8–14.
26. Zubair S, Chaudhary NI, Khan ZA, Wang W. Momentum fractional LMS for power signal parameter estimation. Signal Process. 2018;142:441–9.
27. Chaudhary NI, Aslam khan Z, Zubair S, Raja MAZ, Dedovic N. Normalized fractional adaptive methods for nonlinear control autoregressive systems. Appl Math Model. 2019;66:457–71.
28. Chaudhary NI, Aslam MS, Baleanu D, Raja MAZ. Design of sign fractional optimization paradigms for parameter estimation of nonlinear Hammerstein systems. Neural Comput & Applic. 2019. https://doi.org/10.1007/s00521-019-04328-0.

Publisher's Note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions