Chemometrics and Intelligent Laboratory Systems 139 (2014) 97–108

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemolab

Baseline correction of high resolution spectral profile data based on exponential smoothing

Xinbo Liu [a], Zhimin Zhang [a,*], Yizeng Liang [a,**], Pedro F.M. Sousa [b], Yonghuan Yun [a], Ling Yu [c]

[a] Institute of Chemometrics and Intelligent Analytical Instruments, College of Chemistry and Chemical Engineering, Central South University, Changsha 410083, PR China
[b] Faculty of Sciences and Technology, University of Algarve, Portugal
[c] Shanghai Tobacco Group Co., Ltd., Shanghai, PR China

ARTICLE INFO

Article history:
Received 28 April 2014
Received in revised form 9 September 2014
Accepted 30 September 2014
Available online 20 October 2014

Keywords:
Baseline correction
Exponential smoothing
High resolution spectra
Fast preprocessing

ABSTRACT

Extraction of qualitative and quantitative information from large amounts of analytical signals is difficult with drifted baselines, especially in multivariate analysis. Baseline drift obscures, "fuzzy" signals, and even deteriorates analytical results. In order to obtain accurate and clear results, some effective methods should be proposed and implemented to perform baseline correction before further data analysis. However, most of the classic methods require user's intervention or are prone to variability, especially with low signal-to-noise signals in large data. In this study, a novel baseline correction algorithm based on two-side exponential smoothing algorithm and iterative fitting strategy is proposed. In addition, the iteratively smoothing strategies were creatively implemented in progressively smoothing the residuals between fitted baseline and original signals. This method, named Automatic Two-side Exponential Baseline correction algorithm (ATEB), does hardly require user intervention and prior information, such as peak detection. It's worth noting that the innovative ATEB algorithm has some obvious advantages, especially, when it comes to the processing speed and corrected accuracy of high resolution spectral data with large scale dataset. After a series of benchmarks with high resolution spectral datasets and comparisons with several other popular methods, using various kinds of analytical signals (including hepatocellular carcinoma, MALDI-TOF mass spectrometry, coronary heart disease serum, NMR spectrum and GC–TOF-MS data), the proposed method is found to be accurate, fast, flexible and easy to use on real datasets.

1. Introduction

In general, the baseline drift is usually one of the main issues in chromatograms, mass spectra, Nuclear Magnetic Resonance (NMR) spectra and other spectral data analyses, especially for chemometric multivariate analysis, since the signals from these analytical instruments commonly consist of chemical information, baseline and random noises. Moreover, the baseline drift affects significantly some fundamental chemometric algorithms. Therefore it is necessary to fit the baseline and subtract the background from the analytical signal to alleviate its negative influence. It is worth noting that the influence of the background becomes more difficult to fit and subtract from extremely high resolution datasets, such as NMR spectra and Matrix-Assisted Laser Desorption/Ionization Time of Flight Mass Spectra (MALDI-TOF-MS). According to literature, the classic baseline correction method consists of manually selecting the start and end of a signal peak, and using a piecewise linear approximation to fit a curve as the baseline [1]. However, piecewise approximation is obviously time-consuming and requires much work especially for large scale dataset, and the accuracy depends on the users' experience. As a consequence, several flexible algorithms have been proposed for baseline fitting. Thus, literature from many fields has been published, mainly involving chromatography, vibrational spectroscopy, MALDI-TOF MS, NMR, digital signal processing and statistics.

First of all, let's start from some classic corrected measures for common spectrum. It was Pearson and Walter who proposed the first often cited baseline correction estimation method in 1970 [2]. This classic algorithm works iteratively and inspects which points lie in a specific interval related to their standard deviation, distinguishing the peak points from baseline points simultaneously. Although the algorithm is computationally efficient, it relies on the choice of two parameters (denoted by μ and ν), convergence criterion, and finally the use of a type of smooth curve fitted to the estimated baseline points. Slight mistake in the parameters would lead to unacceptable results. Following the research step of Pearson, many excellent researchers focused their views on improving the baseline correction methods. Liang et al. [3] introduced the roughness penalty method to decrease the influence of the measurement noise, and consequently improved the signal detection and resolution of chemical components with very low concentrations. Later, Shao et al. proposed another novel approach, focusing on the determination of the component number of overlapping

* Corresponding author. Tel.: +86 731 88830824, +86 13973110646 (Mobile).
** Corresponding author. Tel.: +86 731 8830824; fax: +86 731 8830831.
E-mail addresses: zhangzhimin.csu@gmail.com (Z. Zhang), yizeng_liang@263.net (Y. Liang).

http://dx.doi.org/10.1016/j.chemolab.2014.09.018
0169-7439/

chromatograms and baseline corrections, relying on wavelet transform for de-noising [4–8]. In order to correct the background of the measured spectra during elution in chromatograms, asymmetric least squares (ALS) was also introduced by Boelens et al. [9]. Subsequently, Cheung et al. advocated a similar method for preprocessing pyrolysis–gas chromatography–differential mobility spectrometry (Py–GC–DMS) data, via asymmetric least squares (ALS) to eliminate any unavoidable baseline drift [10]. A new idea of morphological weighted penalized least squares (MPLS) algorithm was recommended by Li and Zhan [11], which was successfully applied in the baseline correction of GC–TOF-MS datasets.

Pay attention to the area of vibrational spectroscopy, there also exist a great number of researchers who have proposed a series of algorithms for baseline fitting in it. Firstly, Lieber et al. proposed an approach using least-squares polynomial fitting technique to avoid defects of simple curve fitting [12]. Then, Mazet et al. modified Lieber's method, designing it to minimize a non-quadratic cost function, which was proved to be faster and simpler [13]. Regarding near infrared spectroscopy analysis, Schechter introduced a useful method for the fluctuating non-linear background [14]. Morháč developed a non-linear iterative peak clipping algorithm to correct the baseline of various kinds of spectra, such as IR, NIR and Raman [15]. Zhang et al. succeeded in suppressing fluorescent background in Raman spectroscopy using wavelet and penalized least squares algorithm [16,17]. Liland K.H. proposed a customized baseline correction method which successfully applied in Raman spectra on melted fat from pork adipose tissue [18]. Moreover, lifting wavelet has been applied in baseline corrections for Raman and NMR datasets by Liu and Shao [19].

To the best of our knowledge, many methods previously proposed by other analysts could be effectively applied to small datasets, such as low resolution spectra. However, when it comes to the large scale dataset with high resolution spectra, the research progress has kept rather a slow pace. As early as 1990s, Dietrich et al. applied the second derivative to the signal for peak detection and successfully fitted a NMR baseline with a fifth degree polynomial [20]. Soon afterwards, Moore and Jorgenson recommended a method using a median filter with a very broad window [21]. Even though Moore's method was simple and practical, only peaks with wide baseline segments can be successfully fitted in NMR signals. In 2005, Mirre E. et al. [22] innovatively applied modified asymmetric least-squares algorithm to analyze the reliability of human serum protein profiles generated with C8 magnetic beads assisted MALDI-TOF mass spectra. A practical algorithm designated as adaptive iteratively reweighted Penalized Least Squares (airPLS) was also promoted by Zhang et al. [23,24], by iteratively changing the weights of sum-squared errors between fitted baseline and original signals. Recently, Marcelo R. et al. [25] developed a simple orthogonal background correction (OBGC) method to correct the complex diode array detector (DAD) background signals in fast online comprehensive two-dimensional liquid chromatography (LC × LC). Subsequently, Kuoching Wang et al. [26] presented a novel Distribution-Based Classification method, Baseline Corrector, for automatically estimating the baselines of metabolomics 1D proton NMR spectra. Liu et al. innovatively proposed a novel baseline correction method combing statistic quantile regression algorithm with iterative strategy named selective iteratively reweighted quantile regression (SirQR) [27] which was successfully applied to large datasets, such as GC–TOF-MS and NMR signals. In general, these baseline estimators have been proven fast and flexible in some extent, and some methods can be effectively implemented to different kinds of analytical signals as well.

As mentioned above, many different kinds of chemometric algorithms have been proposed and implemented for treating different kinds of analytical signals, including both classic methods and novel algorithms. Thus, it might be a good idea to change our view to the other analytical fields, for instance, learning something from statistics and digital signal processing. It is noteworthy that Roger Koenker proposed a general approach by employing l_1 regularization methods to estimate quantile regression models for longitudinal data [28]. Eilers et al. developed a fast and effective smoothing algorithm based on penalized quantile regression for the Comparative Genomic Hybridization (CGH) signals [29]. Yu et al. suggested a novel quantile-based Bayesian maximum entropy (QBME) method to account for the non-stationary and non-homogeneous characteristics of ambient air pollution dynamics [30]. In addition, Mencía and Sentana et al. promoted a new algorithm using a location-scale mixture of normal representation of the asymmetric Laplace distribution, transferring different flexible modeling concepts from Gaussian mean regression to Bayesian semiparametric quantile regression [31,32]. Simon Luo and Dave Hale proposed a new digital signal analytical method where a vector shift field was used to represent non-vertical deformations in a seismic image flattening [33]. Robert G. Brown proposed an exponential smoothing method for predicting demand inventory control by an electric computing system [34]. In practice, the exponential smoothing algorithm was first suggested by Robert Goodell Brown in 1956 [35], and then expanded by Charles C. Holt in 1957 [36]. Although the estimates of this exponential smoothing method proposed by Robert are not statistically efficient, they are economically efficient considering the cost of computation. Meanwhile, Prajakta S. Kalekar [37] introduced Holt-Winters Exponential Smoothing algorithm that concentrates on the analysis of seasonal time series data to analyze two models including the Multiplicative Seasonal Model and the Additive Seasonal Model. Furthermore, Joseph J. LaViola Jr. successfully presented a novel Filter-Based Predictive tracking algorithm “double exponential smoothing” for predictive tracking of user position and orientation [38]. When compared against Kalman and extended Kalman filter-based predictors with derivative free measurement models, this method runs approximately 135 times faster with equivalent prediction performance and simpler implementations [39].

According to the previous literature, polynomial fitting, penalized or weighted least square, wavelet, derivatives, and robust local regression have been widely adopted in analytic chemistry for baseline corrections. However, none of these algorithms are entirely perfect for all the practical applications. Each of them has some drawbacks in certain aspects. Firstly, simple manual polynomial fitting methods depend on the analysts' experience for accuracy. Although modified polynomial fitting method is suitable for the most cases, it cannot work well in low signal-to-noise and signal-to-background ratio signals. Secondly, the baseline correction algorithms based on wavelet only remove the baseline successfully when the transformed domain of the signal is well-separated. However, most of the real-world signals do not consent this hypothesis. Thirdly, robust local regression not only demands the specification of the bandwidth and tune parameters by the user, but also requires that the baseline should be smooth and vary slowly. Adaptive iteratively reweighted penalized least square (airPLS) seems to be the optimal automatic baseline correction method. However, airPLS depends on the penalized least squares, which is not robustness enough. Last but not least, the most important problem is that when the analysis signals become an extremely large scale with high resolution, many algorithms cannot offer to process them efficiently and effectively. On the other side, in the smoothing strategies, a classic smoothing algorithm designated penalized least squares was proposed by Whittaker in 1923 [40] , without setting zeroes to the weight vectors at positions corresponding to peak segments. A detailed baseline correction treatment with several related applications has been presented by Eilers [41]. However, the error value of minimized Q_2 is not robustness enough, and can be enlarged by square especially for real-world signals. Moreover, asymmetric least squares, which means asymmetric weights of least squares, has been widely applied to different kinds of baseline correction algorithms, such as ALS method by Eilers [42] and EBS(eliminate the background spectrum) method by Boelens et al. [9] Although the ALS algorithm is effective and useful to some extent, it has some drawbacks. On the one hand, two parameters, namely asymmetric and smoothing parameters, need to be optimized to obtain satisfactory results. On the

other hand, asymmetric parameters are all the same for all the baseline region points, however, it is more reasonable that the weights of baseline region should be set with different values according to the differences between previously fitted baseline and original signals [41].

In this study, we propose a fast, flexible, robust and automatic baseline correction algorithm designated Automatic Two-side Exponential Baseline correction algorithm (ATEB). An advanced iteratively smoothing procedure is executed to gradually approximate the complex baseline of spectral datasets. Meanwhile, fast and valid two-side exponential smoothing algorithm can offer a quite useful baseline correction method, which can fit the desired baseline; in addition, it can be smoothed repeatedly to eliminate the influence from peaks. In order to control the smoothness of fitted baseline, a smoothing factor α is introduced to offer a more flexible approach for the user, and then selectively configure the fitting degree according to different datasets with different number of variables or different order of magnitude. Moreover, the fitting termination is adaptively managed through the sum of the second derivative of the whole spectrum. If the smoothness of fitted result reaches a relatively flat trend, the whole fitting procedure will automatically be terminated and output the corrected complete result. It has been successfully implemented in MATLAB® programming language based on sparse matrices and sparse linear algebra with less than 40 lines code, which can fit the baseline of massive signals in acceptable time. It is worth mentioning that the innovative idea originates from Dave Hale's ten-line program code using standard C language [43]. He successfully applied the two-side exponential smoothing algorithm in preprocessing and detection of seismic signals. One can see the core codes modified by MATLAB® programming language in Supplementary materials.

2. Theory

2.1. The basic two-side exponential smoothing algorithm

Exponential smoothing algorithm is a useful technique that can be widely applied to many kinds of analytical signals to produce smoothed data with high signal to noise ratio. Nearly all the real digital signals themselves are a sequence of observations with noise. The observed phenomenon may be an essentially random process, or it may be an orderly, but noisy, process. Whereas in the simple moving average the past observations are weighted equally, exponential smoothing assigns exponentially decreasing weights over increasing variables. As mentioned previously, Dave Hale [33] described that the exponential smoothing filter can normalize the weights so that they sum to one, which means that when this filter is applied to a sequence with constant input value $x_i = constant$ (already as smooth as can be), the output values will be the same $y_i = constant$. Meanwhile, Fig. 1 illustrates the weights for the exponential filter algorithm and two alternative smoothing filters algorithms with comparable widths.

The formulation below, which is the one commonly used, is attributed to Brown and is known as "Brown's simple exponential smoothing". Moreover, this algorithm is not just commonly applied to statistics data and digital signal processing, but can also be used with any discrete set of repeated measurements. The raw data sequence is often represented by $\{x_t\}$ and the output of the exponential smoothing algorithm is commonly written as $\{S_t\}$, which may be regarded as the best estimate of what the next value of x will be. When the sequence of smoothing observations begins, for the first spectral variable (namely $t = 0$), we set the starting value to be: $S_0 = x_0$. Then the simplest form of exponential smoothing is given by the formula:

$$S_t = \alpha x_t + (1-\alpha)S_{t-1} \quad , \quad t>1 \tag{1}$$

where α is the smoothing factor, and $0 < \alpha < 1$. In other words, the smoothed signal value S_t is a simple weighted average of the present observation x_t and the previously smoothed element S_{t-1}. The term smoothing factor applied to α here might be something of a misnomer, as larger values of α actually reduce the level of smoothing. Note that in the limiting case with $\alpha = 1$ the output series is just the same as the original series (with lag of one variable unit). Simple exponential smoothing is easily and broadly applied, simultaneously it produces a smoothed signal result as soon as two observations are available.

When the value of α is close to one it will have a less smoothing effect and give a greater weight to recent changes in the data, while values of α closer to zero have a greater smoothing effect and are less responsive to recent changes. In order to optimize the estimated process with α and observations S_t, the two-side exponential smoothing algorithm has been introduced into our further smoothing procedure.

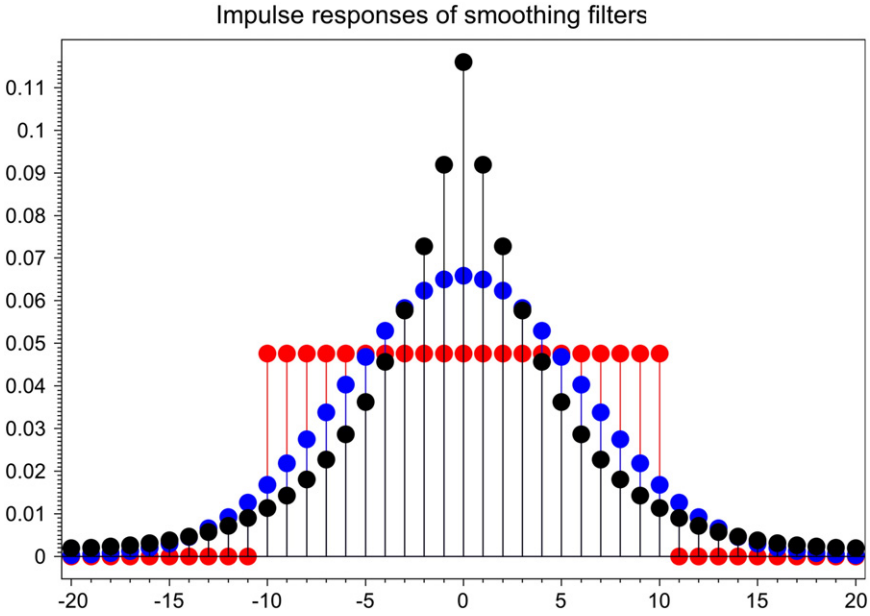

Fig. 1. Impulse responses of exponential (black), Gaussian (blue) and boxcar (red) smoothing filters. For these filters, each output sample is a weighted average of nearby input samples; shown here are the weights.

Similarly as in the exponential smoothing algorithm described above, the raw data sequence of observations is still represented by $\{x_t\}$, beginning at the first spectral variable (namely $t = 0$). Meanwhile, S indicates the exponential smoothing values, therefore $S_t^{(1)}$ and $S_t^{(2)}$ represent the first exponential smoothing values and the second exponential smoothing values with the period of t. In brief, the recursive formula of this two-side exponential smoothing algorithm could be written as follows:

$$\begin{cases} S_t^{(1)} = \alpha x_t + (1-\alpha) S_{t-1}^{(1)} \\ S_t^{(2)} = \alpha S_t^{(1)} + (1-\alpha) S_{t-1}^{(2)} \end{cases}. \tag{2}$$

In which α is also the data smoothing factor, and the value is: $0 < \alpha < 1$. In addition, unlike the traditional double exponential algorithm, we set the initial values of each-side estimation as: $S_1^{(1)} = x_0$, $S_n^{(2)} = S_n^{(1)}$. This means that the algorithm starts with a forward direction in the first smoothing, while it revises the smoothing direction backwards in the second smoothing process. Thus, it starts with the first smoothing result $S_t^{(1)}$ at $t = n$ in the second smoothing.

Note this point, formula (2) can be simplified to:

$$\begin{cases} S_t^{(1)} = \alpha \sum_{i=0}^{t-1} (1-\alpha)^i x_{t-i} + \alpha(1-\alpha)^t x_0 \\ S_t^{(2)} = \alpha \sum_{i=0}^{t-1} (1-\alpha)^i S_{t-i}^{(1)} + \alpha(1-\alpha)^t S_0^{(2)} \end{cases}. \tag{3}$$

Since the initial value of the second smoothing is:

$$S_n^{(2)} = S_{t=n}^{(1)} = \alpha \sum_{i=0}^{n-1} (1-\alpha)^i x_{n-i} + \alpha(1-\alpha)^n x_0. \tag{4}$$

Combing with formula (4), the two-side exponential smoothing result can be summarized as:

$$\begin{aligned} S_t^{(2)} = {} & \alpha \sum_{i=0}^{t-1} (1-\alpha)^i \left[\alpha \sum_{j=0}^{t-i-1} (1-\alpha)^j x_{t-j} + \alpha(1-\alpha)^{t-i} x_0 \right] \\ & + \alpha(1-\alpha)^t \left[\alpha \sum_{i=0}^{n-1} (1-\alpha)^i x_{n-i} + \alpha(1-\alpha)^n x_0 \right]. \end{aligned} \tag{5}$$

An important fact described by Eq. (5) is that all of the raw data points and the corresponding weighted values are considered in estimating each smoothing point during the whole process.

2.2. Baseline correction based on exponential smoothing

Automatic Two-side Exponential Baseline correction method includes a two-step process. First, the original dataset can be smoothed and used to fit the baseline with the two-side exponential smoothing algorithm in an iterative fitting process. Subsequently, the corrected final baseline can be automatically determined, when the approximate fitting result reaches the corrected termination in the iterative procedure.

2.2.1. Iterative smoothing process

In this study, Automatic Two-side Exponential Baseline correction algorithm turns the problems of chemical baseline recognition into problems of digital signal processing. Before fitting the baseline, we first assume that there are two types of points in an original dataset: "noise points" and "signal points". A "noise point" is defined as an unprocessed data point when the signal intensity is in the range of $x \le (\mu - 3\sigma)$; and a "signal point" is as an unprocessed data point when the value of signal intensity is in the range of $x > (\mu - 3\sigma)$. Therefore, after subtracting those "noise points", the processed signals will be smoothed via Two-side Exponential algorithm. During this process, the smoothing factor α can effectively make this smoothed result close to background signals. The iterative smoothing process is clearly illustrated in Fig. 2. The blue line means original signals. The green lines represent the successive approximation process, and the red line is the final corrected result. Moreover, the value of α (smoothing factor) can be flexibly adjusted by user according to different kinds of datasets.

2.2.2. Fitting termination determination

After classifying and smoothing the signal points, the next step is to fit an approximate baseline. Because signal intensities will obscure the position of baseline, it is difficult to estimate the real baseline drift that occurs in correspondence to the signal segments, especially for automatically determining the fitting termination. However, if the peak points (high intensity) and noise points (low and random intensity) can be identified accurately and robustly, a suitable repeatedly iterative process through the two-side exponential smoothing algorithm can well estimate the baseline and preserve the signals. Therefore, immediately following the smoothing process, the second derivative of the smoothed result is taken to confirm the smoothness of the background. Then, the sum of absolute derivative value will be taken into fitting termination determination, as follows:

$$\begin{aligned} d_s &= \sum \left| \Delta^2 x_i \right| = \sum |\Delta(\Delta x_i)| = \sum |(x_i - x_{i-1}) - (x_{i-1} - x_{i-2})| \\ &= \sum |x_i - 2x_{i-1} + x_{i-2}|. \end{aligned}$$

With the fitted curve tending to be a smooth background, the smoothness of the fitting baseline will be close to a minimum and stabilized value in a decreasing function. Once the value of the decreasing function tends to be a stabilized value (the difference between d_{s-1} and d_s was below 5.0×10^{-5}), it can be determined that the whole correcting process is reaching the final termination. As is shown in Fig. 3, the blue descending curve represents the variation tendency of the smoothness of the fitting baseline with iterative times (namely the value of d_s versus the number of iterations). And the red triangle indicates the fitting termination. On the contrary, if the value of the decreasing function is not close to a stabilized value, it means that the fitted baseline is not smooth enough to fit a better background in the smoothing process. So the smooth process is going to be repeated with the last iterative result until the termination. It means that the red triangle in Fig. 3 will shift rightwards.

For a better overview of the framework of the proposed baseline correction algorithm, the flow structure chart of the ATEB algorithm is illustrated in Fig. 4.

3. Materials and applications

In order to test the performance of the ATEB algorithm in practical application, datasets of several broadly used analytical instruments were selected to reveal its performance, especially large spectral datasets, such as NMR spectra and MALDI-TOF mass spectra. In most cases, baseline drift and random noise influenced badly the analytical results. In the following section, artificially designed simulated data was taken as an example, and then extended to actual spectra.

3.1. Simulated data

To construct the desirable dataset, three parts of the simulated data were combined, including linear and curved baselines, standard Gaussian peak signals and random noise, as shown in the following equation:

$$M(x) = p(x) + l(x) + n(x)$$

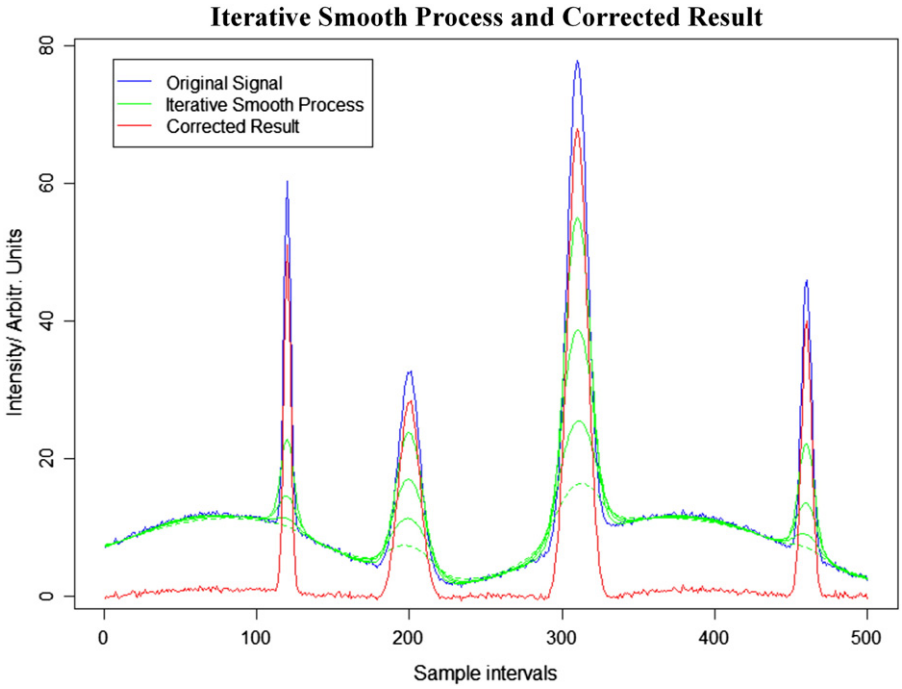

Fig. 2. Iterative smooth process and corrected results of the ATEB algorithm. Blue line means the uncorrected original signal, red line means the corrected result and the green lines represent the successive approximation process with several iterative times.

where $M(x)$ represents the simulated dataset, $p(x)$ stands for the pure standard Gaussian peak, $l(x)$ refers to the standard simulated baseline in curve mode and $n(x)$ represents the random noise.

In order to simulate a good curved test baseline, we adopted a sinus curve, and linear baselines were also introduced. Then, four standard Gaussian peaks for linear baseline (and six standard Gaussian peaks for curved baseline including three overlapping peaks) were composed as the purely standard signals, whose variances and averages were distinct in their intensities. As listed in Table 1, the constructed dataset is also illustrated in Fig. 5. The $n(x)$ is generated by a random function via MATLAB® with the data fluctuating between 0 and 1% of the synthetic signals. One can get the generated dataset from the Supplementary material.

Meanwhile, the factorial design method was also used to generate the benchmark dataset with different peak heights, baseline types and noise levels for testing the stability of corrected result by the ATEB algorithm. As listed in Table 2, the factors of the factorial design were $p(x)$, $l(x)$ and $n(x)$ which were mentioned in the previous paragraph. $p(x)$ represents the peak height of the first three pure Gaussian peaks with different values, as mentioned in Fig. 5. $l(x)$ means the standard simulated background including two different styles (line and curve). $n(x)$ represents the two different random noise levels of the simulated background, including low and high levels. In addition, the stability of the proposed algorithm was represented by *similarity* between the standard pure simulated analytical signals and corrected result using different analytical factors. One can see the more detailed information in Table 2.

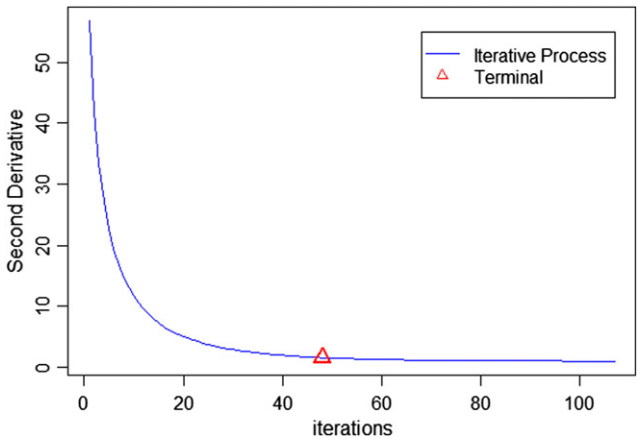

Fig. 3. Fitting termination determination of the ATEB algorithm. The blue decline curve indicates the variation tendency of the smoothness of the fitting baseline (namely d_s) with iterative times and the red triangle represents the fitting termination.

3.2. Experimental data

3.2.1. MALDI-TOF mass spectrometry dataset

Hepatocellular carcinoma (HCC) data. In 2006, Ressom et al. set out a study to help discover early markers for hepatocellular carcinomas triggered by viral infections. The samples were obtained from the Kasr El-Aini Hospital (Cairo, Egypt), where this carcinoma is a primary health problem. After removing proteins greater than 50 kDa (including albumin), the spectra are generated by a MALDI-TOF instrument. The dataset includes 36,802 (m/z) final readings for 46 samples, 23 affected and 23 non-affected controls. The original data of MALDI-TOF mass spectra is represented in Fig. 6(a), where one affected hepatocyte sample signal (the lower red line) and one non-affected hepatocyte sample signal (the upper blue line) are represented. This dataset is an open access set, one can obtain it from the address in Ref. [44].

3.2.2. Nuclear Magnetic Resonance dataset

Coronary heart disease serum (CHS) data. In order to analyze the metabolites in patients with coronary heart disease, Yanpeng An et al. collected ten patients' serum samples with blood stasis symptoms from Shandong Analysis and Test Center (Jinan, China). These spectrum datasets were acquired on a Varian INOVA AS600 (Varian, Inc. US) 600 MHz NMR. Proton chemical shifts at 298 K were obtained from depurated human serum. Every serum sample was tested under the

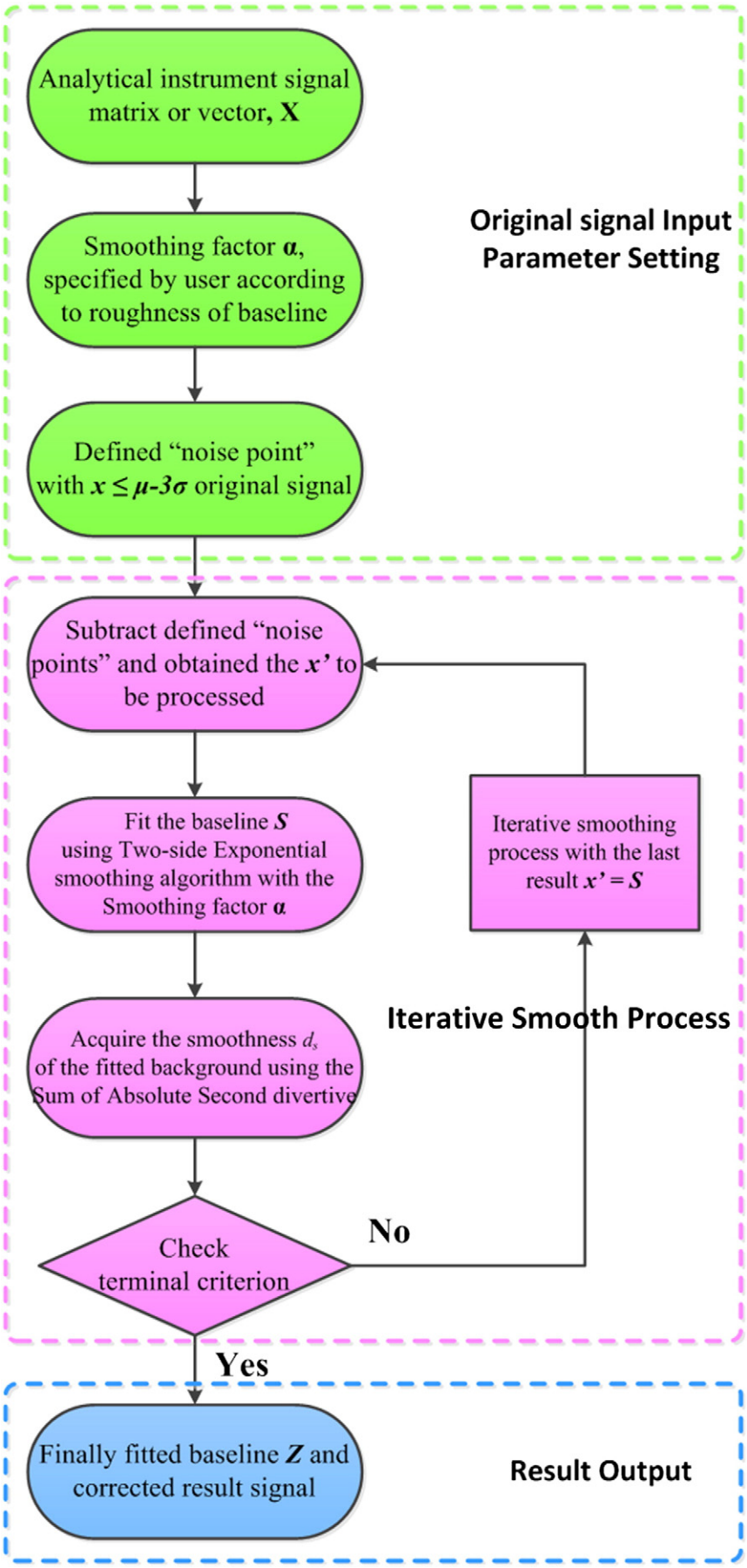

Fig. 4. Flow chart of the Automatic Two-side Exponential Baseline correction (ATEB) algorithm's framework.

Table 1
The analysis and comparison result of baseline correction for the simulated dataset with the expected heights.

Baseline type	Peak no.	Peak height						
		Uncorrected	Expected	ALS[a]	FABC[b]	airPLS[c]	Wavelet [d]	ATEB[d,e]
Linear	Peak 1	57.01	50.00	50.55	50.44	50.36	50.45	50.34
	Peak 2	40.64	30.00	29.06	29.64	29.34	30.21	29.79
	Peak 3	86.29	70.00	68.42	69.86	68.17	69.01	69.31
	Peak 4	63.85	40.00	40.21	40.06	40.09	40.11	39.95
Curved	Peak 1	49.95	40.00	41.06	40.83	40.68	40.70	40.73
	Peak 2	124.50	120.00	119.23	119.08	119.68	119.75	119.70
	Peak 3	29.19	30.00	27.92	27.89	28.89	27.80	29.23
	Peak 4	77.89	80.00	77.21	76.23	78.89	77.85	79.13
	Peak 5	23.95	30.00	27.43	27.56	28.55	28.12	28.35
	Peak 6	80.83	90.00	90.08	88.88	90.15	89.02	90.16

[a] For the ALS method, the parameters are as follows: $\lambda = 10, p = 10^{-5}$ and $d = 2$.
[b] For the FABC method, the parameters are as follows: $\lambda = 10$ and $a = 10$.
[c] For the airPLS method, the parameter is as follows: $\lambda = 8$.
[d] For the Wavelet method, the parameters are as follows: $h = 10$ and $l = 8$.
[e] For the ATEB method, the parameter is as follows: $\alpha = 0.923$.

same conditions, and each spectrum data included 65,483 chemical shifts for these 10 samples. The original dataset of ^{1}H NMR signals is illustrated in Fig. 7(a), where ten different patients' serum signals are represented in ten different colors.

3.2.3. Chromatographic data

Raw tobacco leaves data. Chromatograms of the analyses of tobacco smoke using GC–TOF-MS, whose raw tobacco leaves were collected from Yunnan province, were selected to test the proposed sirQR method. The weight of each cigarette was 0.700 ± 0.015 g, which was filled by CMB-120 cigarette tube filling machine (Burghart, Germany). The plant perfumes from herb extractions were injected into the cigarette via CIJECTOR cigarette injection machine (Burghart, Germany). Twenty cigarettes are smoked simultaneously by the smoking machine (Borgwaldt, Germany), and the cigarette smoke was collected by a Cambridge filter. An extraction solvent (80 mL, dichloromethane: methanol = 2:1 (v/v)) was used to elute the compounds enriched in the Cambridge filter. After extraction, evaporation and concentration to 1 mL, the sample was injected into GCT Premier™ GC–TOF-MS. A DB-35MS (30 m × 0.25 mm, 0.25 μm) chromatographic column was used, with a split ratio of the injector of 1:30 at 250 °C. Helium was used as carrier gas at a constant flow rate of 1.5 mL/min. The column temperature was programmed from 50 °C to 280 °C. Mass spectra from 40 to 400 m/z were collected. The ionization voltage was 70 eV and the ion source temperature was 220 °C. One can get this dataset from the Supplementary material.

4. Results and discussions

4.1. Simulated data result and comparison with other algorithms

The baseline correction of linear (300 variables) and curved (600 variables) synthetic datasets was implemented by the proposed ATEB algorithm. The graph of corrected results is represented in Fig. 5(a) and (b). Both the linear and the curved baselines were subtracted successfully, which has proven the widely applicability of the ATEB algorithm. Combining with the successive approximation progress in Fig. 2, one can see that the iterative smoothing process, either in linear baseline or in curved baseline, takes rarely few times to fit the prospected baseline (more details were listed in Table 3, Section 4.5). In other words, the ATEB algorithm could automatically converge in an extremely high speed. Since the simulated datasets were composed with four known standard Gaussian peaks for linear data and six standard Gaussian peaks for curved data, these expected

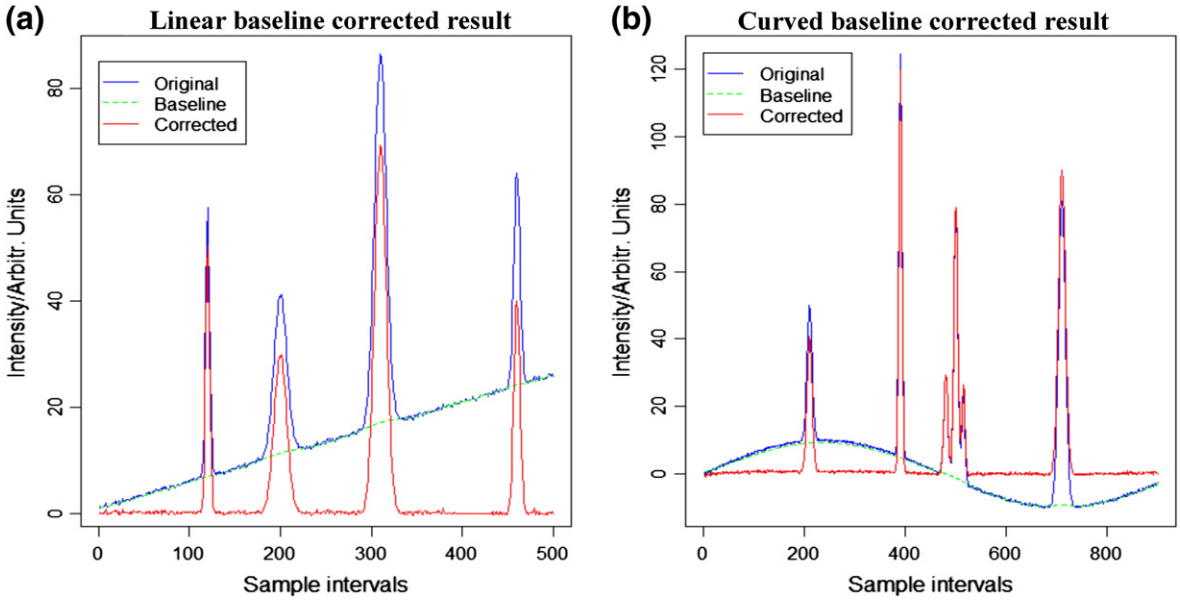

Fig. 5. Corrected results of simulated data with different baselines. (a) Linear baseline (blue line) and the corresponding corrected results. (b) Curved baseline (blue line) and the corresponding corrected results.

Table 2
The similarity value between targeted signals with corrected result using different peak heights, baseline types and noise levels generated by factorial design method.

Group	$p(x)$[a]			$l(x)$[b]	$n(x)$[c]	Similarity[d]
	Peak 1	Peak 2	Peak 3			
1	50	30	70	Line	Low	0.9989
2	80	50	120	Line	Low	0.9991
3	50	30	70	Line	High	0.9968
4	50	30	70	Curve	High	0.9975
5	80	50	120	Curve	High	0.9965

[a] $p(x)$ represents the peak height of the first three pure Gaussian peaks.
[b] $l(x)$ represents the standard simulated baseline that includes line and curve backgrounds.
[c] $n(x)$ represents the random noise intensity level of the simulated background that includes low intensity level and high intensity level.
[d] *Similarity* represents the value of similarity between the standard pure simulated analytical signals and corrected result using different analytical factors.

heights of the peaks were known to us, as well. Therefore one can take the heights before and after corrected to compare the expected heights. The comparison result is listed in Table 1.

In light of the expected height, which was known by us, five different methods were applied for comparison, using the linear and the curved baselines. These were the fully automatic baseline correction procedure of Carlos Cobas [45] (FABC algorithm), asymmetric least squares baseline correction of P. H. C Eilers [22,41] (ALS algorithm), adaptive iteratively reweighted penalized least squares of Zhi-Min Zhang [24] (airPLS algorithm) and baseline drift cancelation using wavelet transform of Tinati [46] (Wavelet algorithm). The consequent calculation of the ALS algorithm, the FABC algorithm, the airPLS algorithm, the Wavelet algorithm and the ATEB algorithm is listed in Table 1. Combining with the corrected result in Fig. 5, one can clearly see that both the linear and curved baselines could be removed successfully and accurately, which proves the flexibility of the ATEB algorithm. In the linear baseline, the fitting result of the ALS algorithm and the airPLS algorithm was significantly worse than the FABC algorithm and the ATEB algorithm, especially in some wide and large peaks. Comparing with the FABC algorithm and the Wavelet algorithm, the ATEB algorithm is slightly better in narrow tip peaks. The ATEB algorithm is also successfully implemented in the curved baseline, and much better than ALS and airPLS algorithms, especially on overlapping peaks. Deducing from Table 1, one can generalize that the ATEB algorithm corrected the baseline for small peaks as well as the other four algorithms, but clearly much better than others for large and overlapping peaks, which is the main influence in both the linear and curved baselines, especially in practical application, such as multivariate analysis.

Furthermore, in order to test the correction stability of the proposed algorithm, a factorial method was applied in our research. As it listed in Table 2, one can see clearly the similarity variation with the different factors. According to group 1 and group 2, one can know that the corrected effect will obtain a better result with the growing of peaks' height and intensity. Comparing group 1 & group 3, group 3 & group 4 and group 4 & group 5, the noise level, simulate baseline type, and peak height will influence the corrected effect more or less. However, since the slight deviation of these three comparing groups (1 vs 3 & 3 vs 4 & 4 vs 5) is considered acceptable. Therefore, we can make a conclusion that the influence from the noise level, simulate baseline type and peak height on the corrected result and accuracy can be neglected if we fit and remove it by the ATEB method. Especially, when comparing with all different factor groups (group 1 and group 5), the deviation value of 0.0024 is also acceptable with the high intensity noise, curve background and over 1.5 times' peak height. This stability test using unique factorial analysis method was proven that the ATEB method is very effective and practical tool even for signals with different peak heights, baseline types and noise levels.

4.2. Corrected spectra and classification result of MALDI-TOF mass spectra

In our study, the proposed ATEB algorithm was effectively applied to MALDI-TOF spectra of affected and non-affected hepatocytes with highly fluorescent baseline. All backgrounds of 46 spectra of hepatocyte samples from different people were removed successfully, designing $\alpha = 0.930$ and testing all spectra with the same parameter. In Fig. 6, two spectra from the two types of 46 experimental samples are represented. To clearly observe each sample, separate figures of these two samples are illustrated in Fig. 6(a) and (b), including the original and corrected signals, where the red lower line represents the affected hepatocyte sample signal and the blue upper line represents the non-affected hepatocyte sample signal. In addition, one can observe the corresponding corrected result of these two picked-out samples in the right figure (Fig. 6(b)) with the same color and the fitted prospective baseline (green dashed line) with original signal in Fig. 6(a). In order to investigate the classification result of the proposed ATEB algorithm, Principal Component Analysis (PCA) and Random Forest (RF) algorithms are adopted for further clustering analysis.

PCA. In the first case, the original data and corrected data were mean-centered by the mean spectrum and normalized by corresponding maximum value. Then, PCA was respectively performed on the matrix consisting of the original signals with mean-centered and normalized results. The first two principal components were taken out and plotted in Fig. 8(a), where the red circles represent the affected

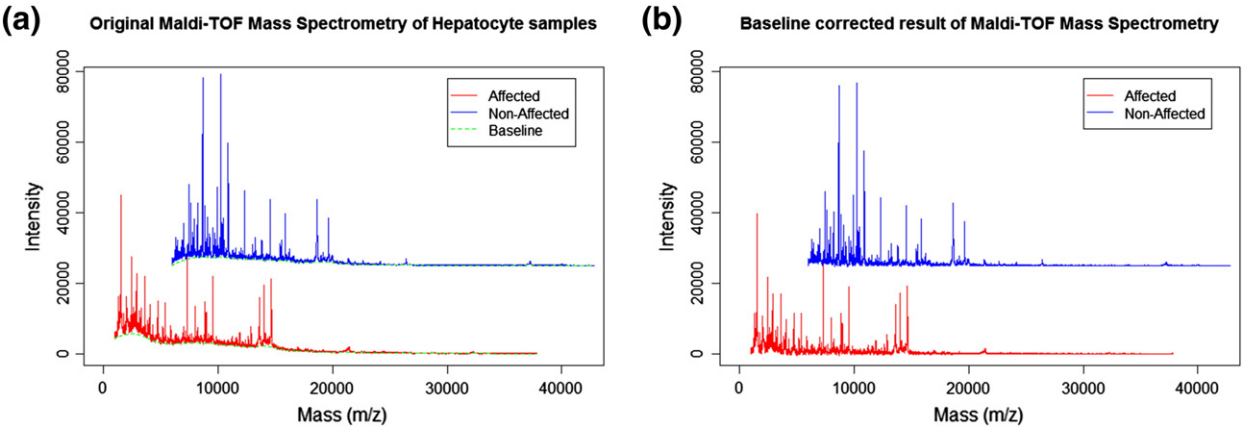

Fig. 6. Baseline correction results of the Matrix-Assisted Laser Desorption/Ionization Time of Flight Mass Spectrometry (MALDI-TOF-MS) of hepatocellular carcinoma. (a) Original uncorrected spectra dataset (upper red line represents the affected samples and lower blue line represents the non-affected samples) with the fitting baselines (green dashed line). (b) Corrected spectra result (upper red line represents the affected result and lower blue line represents the non-affected result).

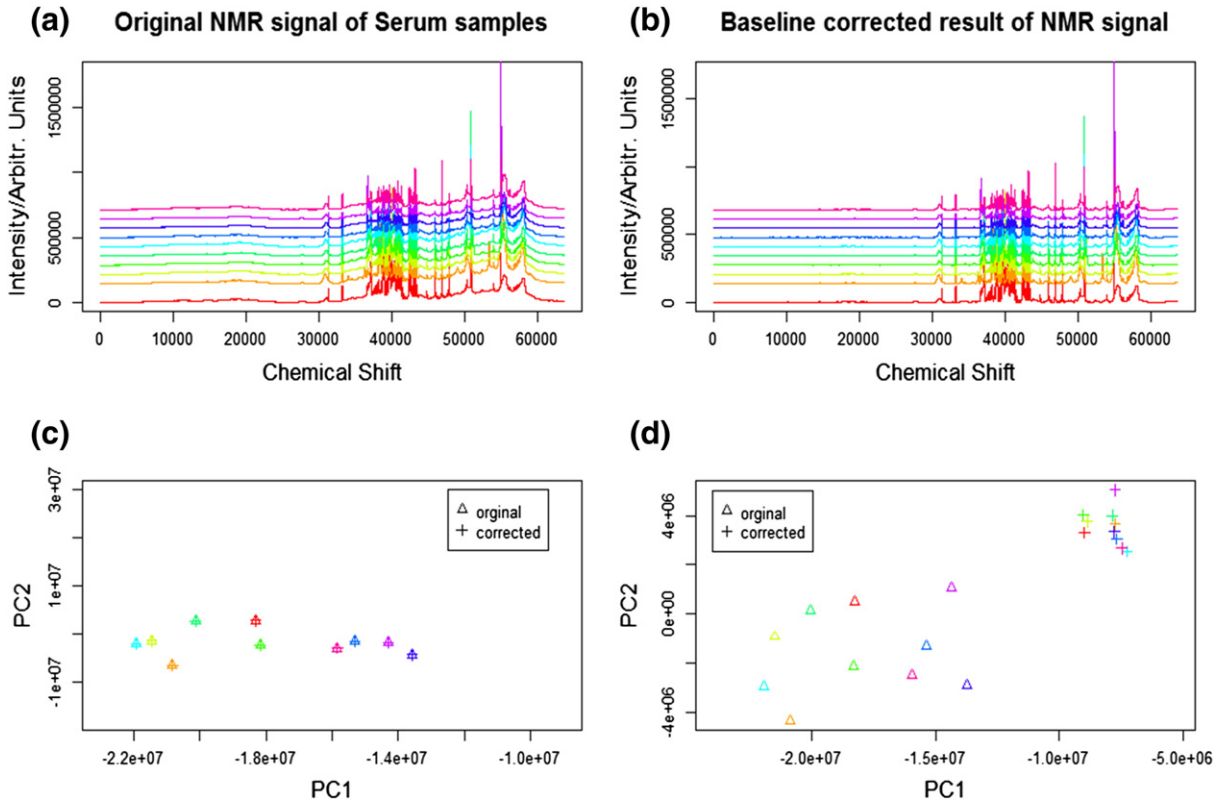

Fig. 7. Background correction results and analysis for the Nuclear Magnetic Resonance (NMR) dataset of coronary heart disease serum by the ATEB algorithm. (a) Original uncorrected NMR dataset of 10 samples with various backgrounds; (b) corrected NMR results through the ATEB method; (c) first two principal components of the original and corrected chromatograms with first-order numerical differentiation preprocessing. (d) Comparison of the distribution of samples before (triangles) and after (plus) background correction in the principal component spaces.

samples' signals, and the blue triangles represent the non-affected ones. One can observe that the affected samples and non-affected samples are mixed in the principal component spaces, which means that the classification result is not satisfactory. Then PCA was also performed with the same spectra with mean-centered and normalized result, but preprocessed by the ATEB algorithm to subtract background. Fig. 8(b) represents the scatter-plots of the two principal components. One can clearly see that the classification result is obviously improved in principal component pattern space, and the PCA variance of the first two components is reaching more than 90%. As a consequence, this correction result with a significant improvement in the separation of the scores and the variance can significantly make the interpretation and classification clearer. In addition, it may be easier to distinguish the wanted difference and the center of gravity of the model, which should be attributed to the ATEB algorithm.

RF. Considering the problem of clustering high resolution large scale dataset in a high-dimensional space and in order to more directly and conveniently observe the patterns in the proximity matrix, supervised RF algorithm [47,48] was employed in the classification of 46 MALDI-TOF mass spectra as well. Firstly, the two parameters for generating a prediction model for the RF classifier were optimized, including the number of classification trees desired (*ntree*) and the number of variables (*mtry*) which are used in each tree to make the tree grow. Secondly, after determining these two parameters with *ntree* = 10,000 and *mtry* = 190, the RF classificatory algorithm was respectively performed on the matrix consisting of the original signals with two types of samples. As illustrated in Fig. 8(c), where the red no. 1 markers represent the affected sample signals and the cyan no. 2 markers represent the non-affected ones, one can clearly observe that the affected samples and non-affected samples were obviously mixed in the multi-dimensional spaces, which means that the classification result is not satisfactory. Following the previous step, the same supervised RF classificatory algorithm was also implemented with the corrected spectra via the ATEB algorithm to remove baseline. The clustering result using the same two parameters is shown in Fig. 8(d), where we can observe a better separation among these two groups, including the affected (the red no. 1 markers) and non-affected (the cyan no. 2 markers) samples which could not be separated before in Fig. 8(c). It also demonstrates that the figures are the visual exploration of the proximity matrix using the multi-dimension space technique, and the distance between these two types can measure their similarity. So it can be inferred that, although there was already a close relationship between the affected samples and the non-affected samples, the small difference could also be discovered and amplified by the ATEB algorithm in large dataset.

The validity of the ATEB algorithm was demonstrated by combining the classification results of the two approaches' plots (PCA and RF). Via the ATEB algorithm, the corrected spectrum of the same type has a more compact pattern and is closer to a standard spectrum. In addition, the clustering and classification results increased because of the

Table 3
The execution time of the simulate data, GC–TOF-MS, MALDI-TOF-MS and HNMR signals.[a]

Algorithms (unit second)	Simulated dataset (900 variables)	Chromatograms (4000 variables)	MALDI-TOF-MS (36,802 variables)	NMR signal (63,483 variables)
ALS	0.0608	0.0885	0.5367	1.4529
FABC	0.0284	0.0669	0.3643	1.5160
airPLS	0.0213	0.0460	0.2413	0.5185
ATEB	0.0172	0.0296	0.2283	0.3662

[a] The dataset with different numbers of variables was applied to test the proposed algorithm and compare with other three different methods. It could also infer the relationship between the number of variables and the execution time (*s*) when using the ATEB algorithm.

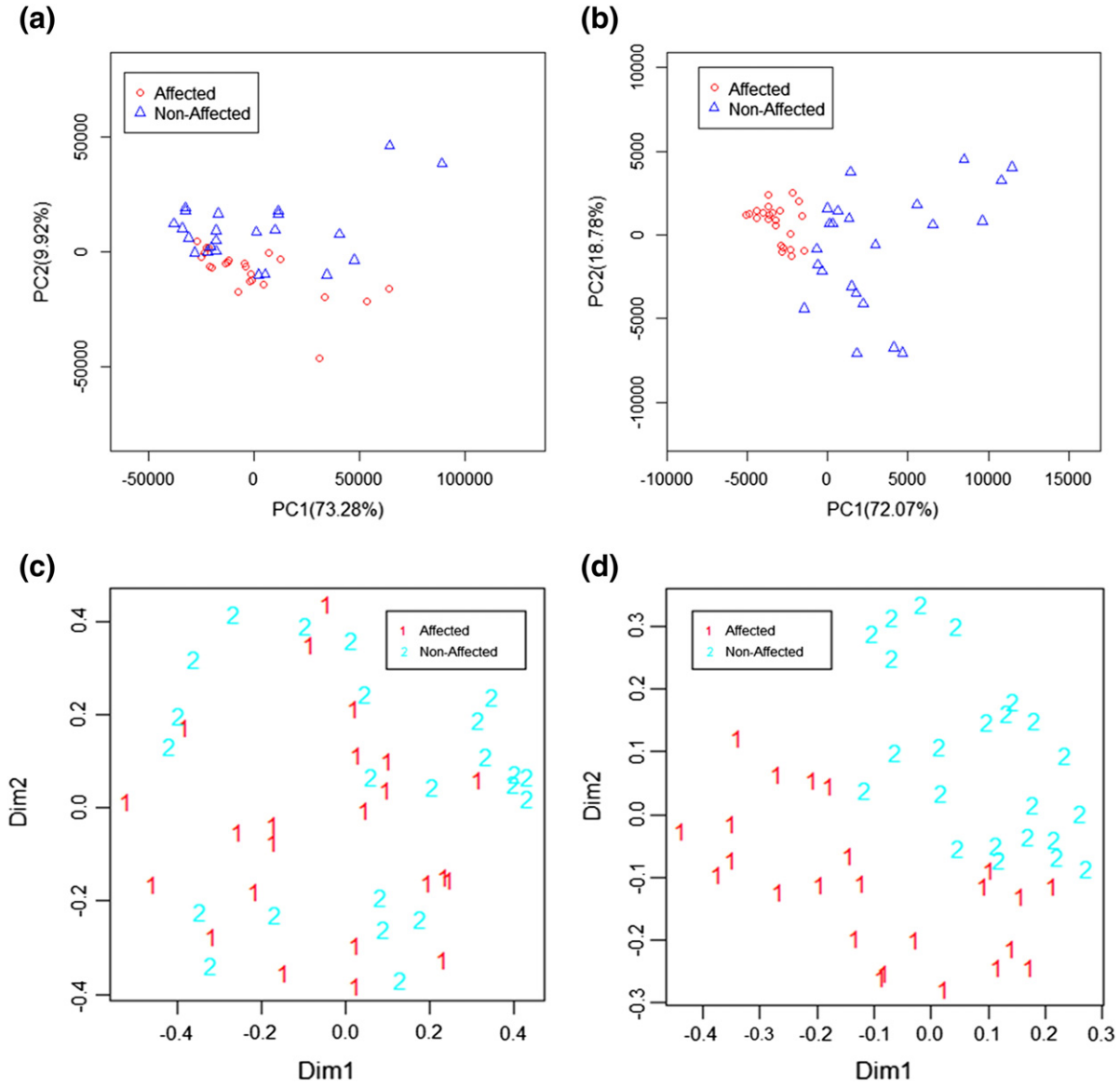

Fig. 8. The classification result of hepatocellular carcinoma samples' MALDI-TOF-MS including the PCA scores and Random Forest (RF) results. (a) First two principal components of the affected (red circles) and non-affected (blue triangles) samples' spectra without any preprocessing. (b) Also first two principal component analysis result of the affected and non-affected spectra via the ATEB correction algorithm. (c) and (d), Classification analysis of Affected (the red no. 1 markers) and Non-affected samples (the cyan no. 2 markers) with supervised learning RF algorithm (*ntree* = 10,000 and *mtry* = 190). (c) Without any preprocessing. (d) Corrected by the ATEB algorithm.

compactness and closeness in principal components pattern space to a certain degree. In summary, the ATEB algorithm could effectively correct the baseline with reserving primary useful information, which has important implications for classification analysis.

4.3. Corrected spectra and analytical result of NMR dataset

For the higher resolution HNMR spectra dataset of the coronary heart disease patients' serum, the background was also corrected by the proposed ATEB algorithm. Meanwhile, these corrected results were further analyzed by the principle component analysis (PCA), as shown in Fig. 7. Comparing Fig. 7(a) with Fig. 7(b), one can clearly observe that the original and corrected HNMR spectra demonstrate that the ATEB algorithm is flexible and available enough to remove the background drifts despite the large number of variables. As shown in the upper right Fig. 7(b), these ten spectra could significantly illustrate the similarity characteristic absorption peaks of the same type. Moreover, PCA was also implemented to assess the validity of the proposed ATEB algorithm as numerical differentiation can eliminate the slowly shifting background. Since those samples are parallel experiment samples, they will be located much closer to each other in the principal component spaces if the influence of the background is ignored. Thus, PCA method has been executed in the original and corrected HNMR signals with first-order numerical differentiation preprocessing. In Fig. 7(c), the triangles represent the original HNMR signals, and the plus symbols represent the corrected ones. As illustrated, the good matching via numerical differentiation preprocessing method in the principal component spaces, suggests that the ATEB algorithm has not eliminated the important information from the original HNMR signals. Moreover, since all these ten samples were similar parallel experiment samples, the process was further carried on the compactness and closeness analysis between the original samples and corrected samples. Focusing on Fig. 7(d), the triangles also represent NMR spectra without any background correction via the ATEB algorithm in the principal component spaces; the plus symbols represent the corrected NMR signals by the proposed ATEB algorithm. The direction in the

first principal component mainly conveyed the sample difference, as the value reached 94% of the total variance in first principal component. In addition, the variation and aggregation extent can be clearly observed in the principle component spaces between the triangles and the plus symbols, which indicates that the serum sample points obtained a better aggregation after correction.

According to the analysis above, it can be demonstrated that the large variation in the first principal component direction of the original HNMR signals could be due to the variation of background from spectrum to spectrum. This proposed ATEB method can remove this variation among a series of NMR spectra background without missing useful and important information.

4.4. Tuning α to obtain better estimation of baseline

In order to obtain a better estimation baseline of different real data, the parameter of α should be flexibly tuned by the users according to their needs for different types of spectra instead of the default initial value in computer program ($\alpha = 0.95$). Since α can be varied from 0 to 1, the common exponential smoothing algorithm proposed before will not work well in the preprocessing situation during fitting process. Both Robert Goodell Brown [34] and Charles C. Holt [36] recommend that searching the optimal smoothing factor α between 0 and 1 on this procedure would influence the whole smoothing effect on a mass scale as previously mentioned. If α is too close to 1, the fitted baseline will be too flat out of our expectation. Otherwise, if α is too close to 0, the fitted baseline will be too flexible to include almost all the peak parts. It was the reason why there are significant differences when smoothing factor α is too large or too small, one can manually optimize the parameter using a method like binary search algorithm. Start with $\alpha = 0.5$, and add α with 0.25 or more when the fitted baseline is too flexible and includes some parts of the peaks. On the contrary, if α is large enough (close to 1) and the fitted baseline is flatter than the real baseline, stop adding α, and fine tune to decrease the value of α searching the optimal α in the region using the step-size of 0.01 until satisfactory. In summary, the parameter α is intuitive to choose, which doesn't need some global optimization method to optimize it.

4.5. Processing speed and expansibility

As described in the section above, the simulated data, MALDI-TOF-MS, HNMR signals and Chromatograms (GC–TOF-MS) with different numbers of variables were used to test the quickness of the proposed algorithm. In addition, this was compared with three different algorithms at the same time, with the same dataset. The execution times of each algorithm (including ALS, FABC, airPLS and ATEB), implementing in different numbers of variables is detailed in Table 3. One can notice that the result of ATEB algorithm's execution time is astonishingly fast.

As represented in Table 3, it is evident that the proposed ATEB algorithm is obviously faster than the other, especially in large datasets like NMR signals. Although the subtle advantage of processing speed is difficult to be observed with small data in application, when it comes to process the large dataset with high resolution, the superiority can be well-reflected immediately. For instance, a fast preprocessing of classification analysis with thousands of different samples including tens of thousand variables. Meanwhile, the relationship between numbers of variables and execution time was also investigated in detail. It was verified that the relationship between the execution time and the number of variables is exactly linear, which can be observed in Fig. 9. Obviously, with the increasing of the number of variables, the corresponding execution time will also increase in a linear trend. The exact linear relationship between number of variables and the whole execution time guarantees the good performance of the ATEB algorithm in data even with a larger number of variables. This is mainly attributed to the usage of the combination of two-side exponential smoothing algorithm. It can be summarized that the usage of the exponential smoothing algorithm enables the application of the ATEB algorithm in more high-throughput domains and effectively meet the needs of data analysis with great speed.

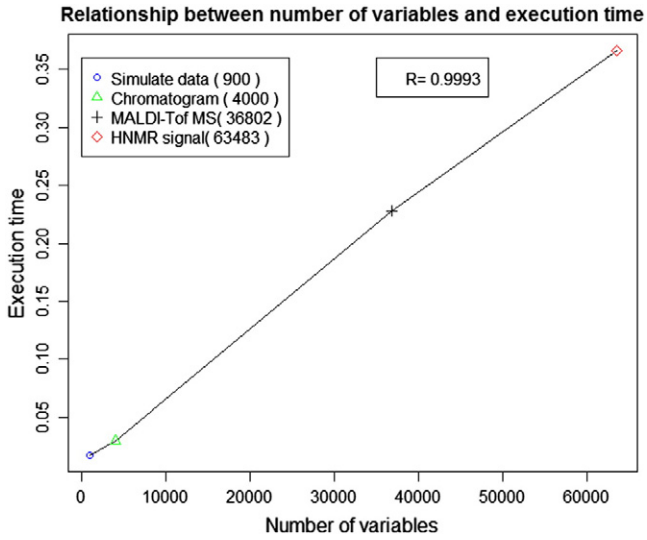

Fig. 9. The relationship between numbers of variables and the total execution time with corresponding fitness.

5. Conclusion

In this research, the proposed ATEB algorithm provides a fast, flexible and valid baseline correction method for processing different analytical signals, especially for the high resolution large scale dataset. The proposed algorithm combines two-side exponential smoothing strategy and iterative smoothing process to fit the background as desired. After comparing with several popular baseline correction methods, such as ALS, FABC and airPLS, the results demonstrate that the proposed algorithm can offer a fast and accurate baseline correction of signals for both simulated data and real high resolution analytical signals. Moreover, the successful results on the tested datasets have proven that this approach can be used as an effective and efficient pre-processing method for many high resolution analytical instruments (such as MALDI-TOF, NMR signals, GC–TOF-MS and HPLC-DAD, even LC–TOF-MS).

Conflict of interest

There is no conflict of interest

Acknowledgment

This work is financially supported by the National Nature Foundation Committee of PR China (Grant No. 21075138, Grant No. 21105129, Grant No. 21175157, Grant No. 21275164 and Grant No. 21305163), National Instrumentation Program of China (Grant No. 2011YQ03012407), China Hunan Provincial Science and Technology Department (Grant No. 2012FJ4139), Hunan Provincial Natural Science Foundation of China (Grant No. 14JJ3031), China Postdoctoral Science Foundation (Grant No. 2014M552146) and the Fundamental Research Funds for the Central Universities of Central South University (Grant No. 2014zzts153 and Grant No. 2014zzts014). The studies meet with the approval of the university's review board. We are grateful to all employees of this institute for their encouragement and support of this research. Simultaneously, the authors want to thank Ressom of the Kasr El-Aini Hospital, Egypt for providing the MALDI-TOF-MS dataset. Yanpeng An of the Shandong Analysis and Test Center, Jinan, China for providing the NMR dataset of the patients' serum samples for the regression and classification analysis.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.chemolab.2014.09.018.

References

[1] A. Jirasek, G. Schulze, M. Yu, M. Blades, R. Turner, Accuracy and precision of manual baseline determination, Appl. Spectrosc. 58 (2004) 1488–1499.
[2] G.A. Pearson, R.I. Walter, Deconvolution of broad-line NMR spectra containing overlapping modulation sidebands, J. Magn. Reson. 16 (1974) (1969) 348–353.
[3] Y.Z. Liang, A.K.M. Leung, F.T. Chau, A roughness penalty approach and its application to noisy hyphenated chromatographic two-way data, J. Chemom. 13 (1999) 511–524.
[4] X. Shao, W. Cai, Z. Pan, Wavelet transform and its applications in high performance liquid chromatography (HPLC) analysis, Chemom. Intell. Lab. Syst. 45 (1999) 249–256.
[5] X. Shao, C. Ma, A general approach to derivative calculation using wavelet transform, Chemom. Intell. Lab. Syst. 69 (2003) 157–165.
[6] X.-G. Shao, A.K.-M. Leung, F.-T. Chau, Wavelet: a new trend in chemistry, Acc. Chem. Res. 36 (2003) 276–283.
[7] H.W. Tan, S.D. Brown, Wavelet analysis applied to removing non-constant, varying spectroscopic background in multivariate calibration, J. Chemom. 16 (2002) 228–240.
[8] O. Sayadi, M.B. Shamsollahi, Multiadaptive bionic wavelet transform: application to ECG denoising and baseline wandering reduction, EURASIP J. Adv. Signal Proc. 2007 (2007) 1–11.
[9] H.F.M. Boelens, R.J. Dijkstra, P.H.C. Eilers, F. Fitzpatrick, J.A. Westerhuis, New background correction method for liquid chromatography with diode array detection, infrared spectroscopic detection and Raman spectroscopic detection, J. Chromatogr. A 1057 (2004) 21–30.
[10] W. Cheung, Y. Xu, C.L.P. Thomas, R. Goodacre, Discrimination of bacteria using pyrolysis–gas chromatography–differential mobility spectrometry (Py–GC–DMS) and chemometrics, Analyst 134 (2009) 557–563.
[11] Z. Li, D.J. Zhan, J.J. Wang, J. Huang, Q.S. Xu, Z.M. Zhang, Y.B. Zheng, Y.Z. Liang, H. Wang, Morphological weighted penalized least squares for background correction, Analyst 138 (2013) 4483–4492.
[12] C.A. Lieber, A. Mahadevan-Jansen, Automated method for subtraction of fluorescence from biological Raman spectra, Appl. Spectrosc. 57 (2003) 1363–1367.
[13] 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 (2005) 121–133.
[14] I. Schechter, Correction for nonlinear fluctuating background in monovariable analytical systems, Anal. Chem. 67 (2002) 2580–2585.
[15] M. Morháč, V. Matoušek, Peak clipping algorithms for background estimation in spectroscopic data, Appl. Spectrosc. 62 (2008) 91–106.
[16] Z.M. Zhang, S. Chen, Y.Z. Liang, Z.X. Liu, Q.M. Zhang, L.X. Ding, F. Ye, H. Zhou, An intelligent background-correction algorithm for highly fluorescent samples in Raman spectroscopy, J. Raman Spectrosc. 41 (2010) 659–669.
[17] S. Chen, X.N. Li, Y.Z. Liang, Z.M. Zhang, Z.X. Liu, Q.M. Zhang, L.X. Ding, P. Ye, Raman spectroscopy fluorescence background correction and its application in clustering analysis of medicines, Spectrosc. Spectr. Anal. 30 (2010) 2157–2160.
[18] K.H. Liland, Multivariate methods in metabolomics — from pre-processing to dimension reduction and statistical analysis, TrAC Trends Anal. Chem. 30 (2011) 827–841.
[19] Y. Liu, W. Cai, X. Shao, Intelligent background correction using an adaptive lifting wavelet, Chemom. Intell. Lab. Syst. 125 (2013) 11–17.
[20] W. Dietrich, C.H. Rüdel, M. Neumann, Fast and precise automatic baseline correction of one-and two-dimensional NMR spectra, J. Magn. Reson. 91 (1991) (1969) 1–11.
[21] A.W. Moore Jr., J.W. Jorgenson, Median filtering for removal of low-frequency background drift, Anal. Chem. 65 (1993) 188–191.
[22] M.E. de Noo, R.A. Tollenaar, A. Özalp, P.J. Kuppen, M.R. Bladergroen, P.H. Eilers, A.M. Deelder, Reliability of human serum protein profiles generated with C8 magnetic beads assisted MALDI-TOF mass spectrometry, Anal. Chem. 77 (2005) 7232–7241.
[23] Z. Zhang, Y. Liang, P. Xie, F. Chau, K. Chan, Chromatographic Fingerprinting and Chemometric Techniques for Quality Control of Herb Medicines, Data Analytics for Traditional Chinese Medicine Research, Springer, 2014, pp. 133–153.
[24] Z.-M. Zhang, S. Chen, Y.-Z. Liang, Baseline correction using adaptive iteratively reweighted penalized least squares, Analyst 135 (2010) 1138–1146.
[25] M.R. Filgueira, C.B. Castells, P.W. Carr, A simple, robust orthogonal background correction method for two-dimensional liquid chromatography, Anal. Chem. 84 (2012) 6747–6752.
[26] K.C. Wang, S.Y. Wang, C.H. Kuo, Y.J. Tseng, Distribution-based classification method for baseline correction of metabolomic 1D proton nuclear magnetic resonance spectra, Anal. Chem. 85 (2013) 1231–1239.
[27] X. Liu, Z. Zhang, P.F. Sousa, C. Chen, M. Ouyang, Y. Wei, Y. Liang, Y. Chen, C. Zhang, Selective iteratively reweighted quantile regression for baseline correction, Anal. Bioanal. Chem. 406 (2014) 1985–1998.
[28] R. Koenker, Quantile regression for longitudinal data, J. Multivar. Anal. 91 (2004) 74–89.
[29] P.H.C. Eilers, H.F.M. Boelens, Baseline correction with asymmetric least squares smoothing, http://www.science.uva.nl/~hboelens/publications/draftpub/Eilers_2005.pdf 2005.
[30] H.-L. Yu, C.-H. Wang, Quantile-based Bayesian maximum entropy approach for spatiotemporal modeling of ambient air quality levels, Environ. Sci. Technol. 47 (2013) 1416–1424.
[31] J. Mencía, E. Sentana, Multivariate location–scale mixtures of normals and mean–variance–skewness portfolio allocation, J. Econ. 153 (2009) 105–121.
[32] T.J. Kozubowski, K. Podgorski, Asymmetric Laplace distributions, Math. Sci. 25 (2000) 37–46.
[33] S.H. Luo, Dave, Non-vertical Deformations for Seismic Image Flattening, 2011 SEG Annual Meeting, 2011.
[34] R.G. Brown, Smoothing, Forecasting and Prediction of Discrete Time Series, Courier Dover Publications, 2004.
[35] R.G. Brown, Exponential Smoothing for Predicting Demand, Arthur D. Little Inc., Cambridge, Massachusetts, 1956. 15.
[36] C.C. Holt, Forecasting Trends and Seasonal by Exponentially Weighted Averages, Office of Naval Research Memorandum, 1957. 52.
[37] P.S. Kalekar, Time Series Forecasting Using Holt-Winters Exponential Smoothing, 4329008Kanwal Rekhi School of Information Technology, 2004. 1–13.
[38] J.J. LaViola, Double Exponential Smoothing: An Alternative to Kalman Filter-based Predictive Tracking, Proceedings of the workshop on Virtual environments 2003, ACM, 2003, pp. 199–206.
[39] J.H. Lee, N.L. Ricker, Extended Kalman filter-based nonlinear model predictive control, Ind. Eng. Chem. Res. 33 (1994) 1530–1541.
[40] E.T. Whittaker, On a new method of graduation, Proc. Edinb. Math. Soc. 41 (1922) 63–75.
[41] P.H. Eilers, A perfect smoother, Anal. Chem. 75 (2003) 3631–3636.
[42] P.H. Eilers, H.F. Boelens, Baseline Correction with Asymmetric Least Squares Smoothing, Leiden University Medical Centre Report, 2005.
[43] D. Hale, My favorite ten-line computer, program, http://inside.mines.edu/~dhale/ 2012 (August 13, 2012).
[44] Ressom, et al., http://www.sc.ehu.es/ccwbayes/members/ruben/ms/node14.html 2006.
[45] J. Carlos Cobas, M.A. Bernstein, M. Mart-Pastor, P.G. Tahoces, A new general-purpose fully automatic baseline-correction procedure for 1D and 2D NMR data, J. Magn. Reson. 183 (2006) 145–151.
[46] M.A. Tinati, B. Mozaffary, A wavelet packets approach to electrocardiograph baseline drift cancellation, Int. J. Biomed. Imaging (2006) 1–9, http://dx.doi.org/10.1155/IJBI/2006/97157 (Article ID 97157).
[47] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32.
[48] F.F. Ai, J. Bin, Z.M. Zhang, J.H. Huang, J.B. Wang, Y.Z. Liang, L. Yu, Z.Y. Yang, Application of random forests to select premium quality vegetable oils by their fatty acid composition, Food Chem. 143 (2014) 472–478.