

Selective iteratively reweighted quantile regression for baseline correction

Xinbo Liu · Zhimin Zhang · Pedro F. M. Sousa ·
Chen Chen · Meilan Ouyang · Yangchao Wei ·
Yizeng Liang · Yong Chen · Chaoping Zhang

Received: 13 September 2013 / Revised: 11 December 2013 / Accepted: 30 December 2013
© Springer-Verlag Berlin Heidelberg 2014

Abstract

Extraction of qualitative and quantitative information from large numbers of analytical signals is difficult with drifted baselines, particularly in multivariate analysis. Baseline drift obscures and “fuzzies” 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 conducting further data analysis. However, most of the classic methods require user intervention or are prone to variability, especially with low signal-to-noise signals. In this study, a novel baseline correction algorithm based on quantile regression and iteratively reweighting strategy is proposed. This does not require user intervention and prior information, such as peak detection. The iteratively reweighting strategy iteratively changes weights of residuals between fitted baseline and original signals. After a series of tests and comparisons with several other popular methods, using various kinds of analytical signals, the proposed method is found to be fast, flexible, robust, and easy to use both in simulated and real datasets.

Electronic supplementary material The online version of this article (doi:10.1007/s00216-013-7610-x) contains supplementary material, which is available to authorized users.

X. Liu · Z. Zhang (✉) · C. Chen · M. Ouyang · Y. Wei ·
Y. Liang (✉)

Institute of Chemometrics and Intelligent Instruments, College of
Chemistry and Chemical Engineering, Central South University,
Changsha 410083, China
e-mail: zhangzhimin.csu@gmail.com
e-mail: yizeng_liang@263.net

P. F. M. Sousa
Faculty of Science and Technology, University of Algarve,
Faro, Portugal

Y. Chen · C. Zhang
Shanghai Tobacco Group Co., Ltd, Shanghai, China

Keywords Baseline correction · Quantile regression ·
Iteratively reweighted · Robustness

Introduction

The baseline drift is one of the primary issues in chromatography, nuclear magnetic resonance (NMR) spectra and Raman spectra data analysis, especially in chemometric multivariate analysis. The signals from these analytical instruments commonly consist of chemical information, baseline, and random noises. However, the baseline drift will significantly affect some fundamental chemometric algorithms. Therefore, it is necessary to fit the baseline and subtract it from the analytical signal to alleviate its negative influence. According to some 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, and the accuracy depends on the user's experience. As a consequence, several algorithms have been proposed to fit the baseline. Meanwhile, literature on this issue was spread among many fields, mainly involving chromatography, NMR, vibrational spectroscopy, and statistics.

Pearson 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 requires the choice of two parameters (denoted μ and ν), convergence criterion, and finally the use of a type of smooth curve fitted to the estimated baseline points. If the selection of these parameters has any slight mistake, the data could render unacceptable results.

After Pearson's research, many 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. Following these steps, another novel approach was recommended by Shao et al. [4], which focused on the determination of the component number of overlapping chromatograms and baseline corrections, relying on wavelet transform for de-noising. In order to correct the baseline of the measured spectra during elution in chromatograms, asymmetric least squares (ALS) was also introduced by Boelens et al. [5]. Cheung et al. advocated a similar method for preprocessing pyrolysis-gas chromatography-differential mobility spectrometry data, via ALS to eliminate any unavoidable baseline drift [6]. Li and Zhan [7] proposed the morphological weighted penalized least squares (MPLS) and applied it in the baseline correction of gas chromatography–time-of-flight mass spectrometry (GC-ToF-MS) datasets.

For the baseline issues in vibrational spectroscopy, a great number of researchers have proposed a series of algorithms to fit the baseline. Ruckstuhl et al. proposed a novel robust baseline estimation to correct the baseline of original spectra using pulse laser-induced fluorescence detection of gas-phase hydroxyl (OH) vibration-rotation populations, created by the reaction of hydrogen atoms with ozone ($\text{H} + \text{O}_3 \rightarrow \text{OH} + \text{O}_2$), with robust local regression technique [8]. Regarding the analysis of near-infrared spectroscopy, Schechter introduced a useful method for the fluctuating nonlinear background [9]. To avoid defects of simple curve fitting, Lieber et al. suggested an approach using least squares polynomial fitting technique [10]. Mazet et al. modified Lieber's method, designing it to minimize a nonquadratic cost function, which was proved to be faster and simpler [11]. Morháč developed a nonlinear iterative peak clipping algorithm to correct the baseline of various kinds of spectra, such as IR, NIR, and Raman [12]. By using wavelet and penalized least squares algorithm Zhang et al. succeeded in suppressing fluorescent background in Raman spectroscopy [13, 14]. Kristian proposed a customized baseline correction method successfully used in Raman spectra on melted fat from pork adipose tissue [15]. Lifting wavelet has been applied in baseline corrections for Raman and NMR dataset by Liu and Shao [16].

For the NMR instruments, background drift is also a serious issue in metabolomics with massive NMR spectra. 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 [17]. Soon afterwards, Moore and Jorgenson recommended a method using median filter with a very broad window [18]. Even though Moore's method is simple and practical, it can only be successfully applied to NMR signals with peaks with wide baseline segments. Golotvin advocated a novel approach for baseline

correction using a smoothed NMR spectrum for both baseline area recognition and modeling [19]. A continuous wavelet transform derivative in signal-free regions, combined with penalized least squares has been presented by Carlos Cobas et al. [20]. Recently, a robust baseline correction method for dense signal NMR spectra has been proposed by Chang et al. [21]. A practical algorithm designated as adaptive iteratively reweighted penalized least squares (airPLS) has also been promoted by Zhang et al., by iteratively changing weights of sum squares errors between fitted baseline and original signals [22, 23]. This baseline estimator has been proved fast and flexible, being successfully implemented to various analytical signals.

As mentioned earlier, 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 field, for instance, learn something from statistics. It is worth noting that Koenker proposed a general approach by employing l_1 regularization methods to estimate quantile regression models for longitudinal data [24]. Eilers et al. developed a fast and effective smoothing algorithm based on penalized quantile regression for the comparative genomic hybridization signals [25]. Hong et al. presented a new technique using these quantiles as an indicator of the X-ray color hardness of the source to classify spectral properties of X-ray sources with limited statistics [26]. Callister et al. proposed a normalization approach, investigating locally weighted regression and quantile techniques, for removing systematic biases associated with mass spectrometry and label-free proteomics [27]. In addition, Chernozhukov and Hansen developed robust inference products for an instrumental variable model defined as $Y = D'\alpha(U)$, which is computationally convenient in typical applications and can be carried out using available software for ordinary quantile regression [28]. In the same year, Jun advocated a weakly robust identification test method to assess the quality of an instrumental quantile model, which consists of a two-step Anderson–Rubin (AR) statistic and an orthogonal decomposition of the AR statistic, and was applied to reduce the computational burden [29]. It was also Jun who proposed a simple two-step estimator for the coefficients evaluated at particular values of the latent variables, using a control variable idea and quantile regression methods in which the instruments were locally relevant to the established the consistency and asymptotic normality [30]. Wunderli introduced a method to penalty examine the nondifferentiable quantile regression and defined an appropriate weak solution of the time flow [31]. Yu suggested a novel quantile-based Bayesian maximum entropy method to account for the nonstationary and nonhomogeneous characteristics of ambient air pollution dynamics [32]. At the same time, Waldmann et al. promoted a new algorithm using a location-scale mixture of normals representation of the asymmetric Laplace distribution, transferring different flexible

modeling concepts from Gaussian mean regression to Bayesian semiparametric quantile regression [33, 34].

According to the previous literature, polynomial fitting [10], penalized or weighted least square [3, 5, 6, 10, 11, 20], wavelet [4, 13, 16], derivatives [13, 22], and robust local regression [28, 29] have been widely adopted in analytic chemistry for baseline corrections. However, none of these algorithms are entirely perfect for practical applications. Each of them has drawbacks in certain aspects. First, 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. Second, the baseline correction algorithms based on wavelet will remove the baseline successfully, if 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. airPLS seem to be the optimal automatic baseline correction method. However, airPLS depends on the penalized least squares, which is not robust enough. Last but not the least, quantile regression method could fit a very smooth and similar baseline, but for the practical signals with noise and narrow peak signals, desirable results are hard to achieve.

In this paper, we propose a robust, fast, flexible and automatic baseline correction algorithm designated selective iteratively reweighted quantile regression (SirQR). An advanced iteratively reweighted procedure is implemented to gradually approximate the complex baseline. quantile regression algorithm can offer a quite useful baseline correction method which can fit the desired baseline, and it can be reweighted to eliminate the influence from peaks. The weights are adaptively and iteratively obtained via the difference value between previously fitted baseline and original signals. Moreover, for datasets with large number of variables and large order of magnitude, the weight condition can be selectively changed to fit the original data instead of the default value. The proposed algorithm is proved to be intuitional and effective with several kinds of analytical signals. It has been implemented in MATLAB[®] programming language based on sparse matrices and sparse linear algebra, which can fit the baseline of massive signals in acceptable time.

Theory

Quantile regression algorithm

The robust quantile regression method was first proposed by Koenker and Bassett in 1978 [35]. This kind of regression analysis was primarily used in statistics and econometrics. As the regression result is robust, it is not affected by outliers in signal. One advantage of quantile regression, relative to the

ordinary least squares regression, is that the quantile regression estimates are more robust against outliers in the response measurements. However, the main attraction of quantile regression goes beyond that. In practice, we often prefer using different measures of central tendency and statistical dispersion to obtain a more comprehensive analysis of the relationship between variables [36]. This algorithm has been widely used in many kinds of fields. Particularly, quantile regression method is suitable for the situation of dependent variables with heteroscedasticity, tail and spike distributions. Essentially, quantile regression algorithm was an expansion of general least squares regression method, while the robust median regression is a special case of quantile regression.

As the least squares algorithm could minimize the value of error sum of square, the least squares regression is also known as sample mean regression. However, the fundamental difference between quantile regression and least squares regression can be assumed that when \mathbf{x} is the value of the independent variable \mathbf{X} , the arbitrary quantile $Q(y; \tau)$ of the corresponding values of \mathbf{Y} should follow a distribution function $\mathbf{X} \sim \mathbf{F}(\mathbf{x})$ as an approximation. At the same time, the principle of the corresponding estimation method is looking for a ζ , making the sum of absolute value of the asymmetry weighted error minimum. It could be summarized as the following equation:

$$\min_{\xi \in R} \sum_{i=1}^n \rho_{\tau}(y_i - \xi) \quad (1)$$

Where $\rho_{\tau}(\mu) = \mu(\tau - I(\mu < 0))$, $I(\cdot)$ is the a simple indicator function;

Assuming that:

$$\tau \sum_{i: y_i \geq \xi} |y_i - \xi| + (1 - \tau) \sum_{i: y_i < \xi} |y_i - \xi| = l(Y, \tau, \xi) \quad (2)$$

Where the y_i were the samples of $y_{(1)}, y_{(2)} \dots y_{(n)}$. Therefore $\sum_{i=1}^n |y_i - \xi|$ is a strictly decreasing function when $n\tau > m$; and instead $\sum_{i=1}^n |y_i - \xi|$ is a strictly increasing function (1) when $n\tau < m$. Consequently, the τ quantile $Q(y; \tau) = y_{([n\tau])}$ of the sample $y_1, y_2, \dots y_n$ is the desired ζ .

As a special case, when τ is equal to $\frac{1}{2}$, the formula can be simplified into: $\min_{\xi \in R} \sum_{i=1}^n |y_i - \xi|$, and the $median(y)$ of the sample $y_1, y_2, \dots y_n$ is the desired ξ . During regression analysis, the distribution function $\mathbf{X} \sim median(x)$ is taken as the approximation $\mathbf{X} \sim \mathbf{F}(\mathbf{x})$ instead of the expected value $\mathbf{X} \sim \mathbf{E}(\mathbf{x})$, and the regression coefficient estimation method has also been named least absolute deviation regression.

In summary, the function of least squares regression algorithm determines the mean value $\mathbf{X} \sim \mathbf{F}(\mathbf{x})$ of the corresponding \mathbf{Y} when the independent variable \mathbf{X} takes the value \mathbf{x} . However, in quantile regression, the main function is to identify the various τ quantile $Q(y; \tau)$ of the value with corresponding \mathbf{Y} . In other words, quantile regression

algorithm can estimate various fractile points of the corresponding \mathbf{Y} when the independent variable \mathbf{X} takes the value \mathbf{x} . Therefore more information can be obtained from this process, for example, determining the correlation between the independent variable \mathbf{X} with the value of \mathbf{x} and the greater (high-fractile point) or lesser (low-fractile point) value for the corresponding \mathbf{Y} . Consequently, it conveys that the result of quantile regression is provided with the features of robustness and richness of information.

Quantile regression was introduced by Koenker and Bassett in 1978, but it was only used for baseline fitting until 2011 by Komsta [34]. Baseline estimation with quantile regression works similarly as polynomial fitting, which fits the baseline with a small quantile at the peak region of the signal (0.01 is proposed). In this way, both the quantile regression and the polynomial are fitted to the lowest values. Consequently, the peaks have little or no effect on the baseline. A more detailed discussion to quantile regression was given by Koenker and Hallock [37] and a comprehensive description was also made by Koenker [36].

Selective iteratively reweighted quantile regression

A classic smoothing algorithm designated penalized least square was proposed by Whittaker in 1923 [38]. If \mathbf{y} is a series of m data points and \mathbf{z} is the smooth series that should approximate \mathbf{y} , we can minimize:

$$Q_2 = \sum_{i=1}^m (y_i - z_i)^2 + \lambda \sum_{i=2}^m (z_i - z_{i-1})^2 \quad (3)$$

The first term measures the fitness of \mathbf{z} to \mathbf{y} , and the second term is used as the penalty term, discouraging changes in \mathbf{z} . Thus, the influence of the penalty is tuned by the parameter λ . The larger λ is chosen, the smoother \mathbf{z} will be, at the cost of a worse fit to the data. A detailed treatment with several related applications has been presented by Eilers [39]. However, the minimized Q_2 is not robust, which may not be ideal for real-world signals. Therefore, the error value of this equation can be absolutely enlarged by square.

The solution for this issue is to change the objective function Q_2 into Q_1 :

$$Q_1 = \sum_{i=1}^m |y_i - z_i| + \lambda \sum_{i=2}^m |z_i - z_{i-1}| \quad (4)$$

The sums of squares of residues have been replaced by the sums of absolute values, which means that the L_2 norm has been modified into L_1 norm. Focusing in the right side of (4), it can be deemed that the functions of \mathbf{y} have a single summation of $2m-1$ absolute values of terms. Thus, a novel method could be proposed by combining the ideas from quantile regression with linear programming to fit the dataset.

Koenker and Bassett proposed in 1984 [40] the following problem: a vector \mathbf{y} , a regression basis \mathbf{B} and n regression coefficients α . With τ a parameter between 0 and 1, we can minimize the following equation:

$$S(t) = \sum_i^m \rho_\tau \left(y_i - \sum_j^n b_{ij} \alpha_j \right) \quad (5)$$

Here, $\rho_\tau(\mu)$ is the check function; it is $\tau\mu$ when $\mu > 0$ and $(\tau-1)\mu$ when $\mu \leq 0$. It will return the weighted absolute values of residuals in (5), τ for the positive one and $1-\tau$ for the negative one. The weights are independent of the sign with $\tau=0.5$. Therefore, solving Eq. (5) is equivalent as solving (4). This idea is the so-called median regression.

It was Portnoy and Koenker again who presented a detailed account of quantile regression and also an efficient algorithm based on the interior point method for linear programming [41]. These methods have been implemented in both R and MATLAB programming languages. However, some modifications should be made for signal smoothing and baseline correction:

$$\mathbf{y}^* = \begin{pmatrix} \mathbf{y} \\ \mathbf{0} \end{pmatrix} \quad \text{and} \quad \mathbf{B} = \begin{pmatrix} \mathbf{I} \\ \lambda \mathbf{D} \end{pmatrix} \quad (6)$$

where \mathbf{y} is original dataset, $\mathbf{0}$ is a $m-1$ zeros vector; \mathbf{I} is the $m \times m$ identity matrix and \mathbf{D} is a matrix so that $\mathbf{D}\mathbf{z} = \Delta\mathbf{z}$. Thus, \mathbf{D} is an adjusted matrix $(m-1) \times m$ whose purpose is to transform \mathbf{z} into differences of neighboring elements. For a desired result, \mathbf{y}^* and \mathbf{B} are the best choices for next steps.

The key product of selectively iteratively reweighted quantile regression is similar to the weighted least squares [42], the iteratively reweighted least squares [43, 44], and the adaptively iteratively reweighted penalized least squares [22]. However, it uses different ways to calculate the weights and utilizes item adjustment to control smoothness of the fitted baseline. As it is shown:

$$S'(t) = \sum_i^m \omega_i \rho_\tau \left(y_i - \sum_j^n b_{ij} \alpha_j \right) \quad (7)$$

where ω_i is the weight vector that selectively acquires the changed values by using the iterative method. The initial value of ω_{i0} should be assigned by 10^{-4} at the starting step, which has been determined by several tests and calculations from 0 to 10^{-10} . After initialization, the ω of each iterative step could be acquired using the following expression:

$$w_i^t = \begin{cases} 10^{-10} & d_i \geq d_m \\ \frac{t(x_i - z_i^{t-1})}{e^{|d^t|}} & d_i < d_m \end{cases} \quad (8)$$

where $d_i = x_i - z_i^{t-1}$. The vector d^t consists of the elements of the differences between x and z^{t-1} , whose value is below $d_m = 5.0 \times 10^{-5}$ for a better fitting in the iteration step of t .

Meanwhile, the value of d_m can be selectively designated by the users for a better fitting to the original signal dataset, not just for the default 5.0×10^{-5} , such as the dataset with many numbers of variables and large orders of magnitude. The corresponding value should be designated bigger than the default one for a better approximation. The fitted value z^{t-1} in the previous ($t-1$) iteration is a candidate for the baseline. If the value of the i th point is greater than the candidate for the baseline, it can be regarded as part of the peak. Thus, the weight of the corresponding ω will be assigned to a tiny value 10^{-10} (cannot be zero after considering the influence of quantile regression method in whole dataset) to nearly neglect it at the next iteration of fitting. In order to obtain the points of the peaks and eliminate them gradually preserving the baseline points in the weight vector ω , the iterative and reweight methods were adopted in this SirQR algorithm.

For the iteration procedure, it reaches the goal either in the maximal iteration times or when the termination criterion is arrived in. The termination criterion is defined by:

$$|d^t| < 10^{-5} \times |x| \quad (9)$$

Here, vector d^t is the same as in Eq. (9), consisting of the negative elements of differences between x and z^{t-1} .

For a better overview of the framework of the proposed baseline correction algorithm, the flow structure chart of the SirQR algorithm is illustrated in Fig. 1.

Experimental and applications

In order to test the performance of the SirQR algorithm in practical application, datasets of several broadly used analytical instruments were selected to reveal its performance, such as chromatography, NIR, Raman, and NMR spectra. In most cases, baseline drift and random noise influenced badly the analytical result. In the following section, artificially designed simulated data was taken as an example at first, and then extended to actual spectra.

Simulated data

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

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

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

In order to simulate the curved baseline, we adopted a sinus curve, and absolutely linear baselines were also introduced. Then, four standard Gaussian 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. S1 (see Electronic supplementary material Fig. S1). The $n(x)$ noise is generated by a random function via MATLAB® with the data is fluctuating between 0 and 1 % of the synthetic signals.

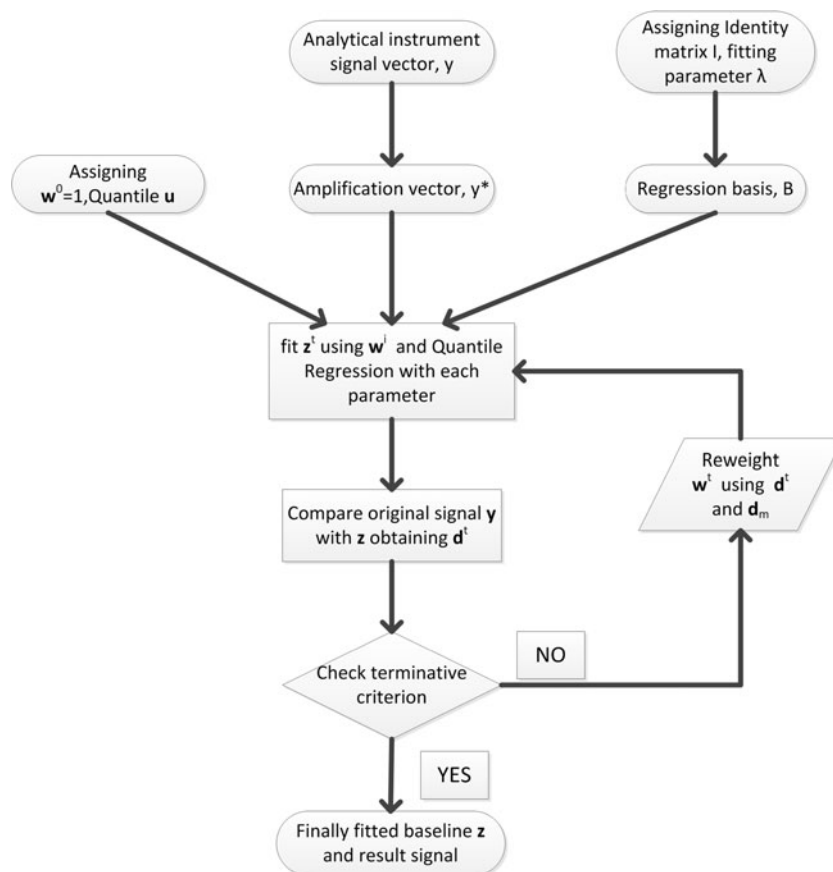
Real chromatograms data

HPLC-DAD dataset

Dendrobium is a quite famous Traditional Chinese Medicinal herb. The HPLC-DAD dataset of *Dendrobium* by a Dionex U3000 HPLC apparatus (Dionex, Sunnyvale, CA, USA) were chosen to test the proposed method. These *Dendrobium* samples were collected from ten different production regions in China, such as Hunan, Hainan, Jilin, Xinjiang, etc. The experiments were performed in a UV spectrometer with a MWD-3000RS multiple wavelength detector. At each step the spectra ranges from 190 to 400 nm with bandwidth intervals of one nm and resulting 210 data points for each UV spectrum. Then, the “most-rich peak” of wavelength 254 nm was selected. This was carried out at the Research Center of Modernization of Chinese Medicines, Central South University. All these chromatograms are represented in Fig. 2a. They include the ten different samples, and the baseline drifts are clearly shown.

GC-ToF-MS data

Chromatograms of the analyses of tobacco smoke using CG-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 to 280 °C. Mass spectra from 40 to 400 m/z were collected. The ionization voltage was 70 eV and ion source temperature was 220 °C.

Fig. 1 Flow chart of the SirQR algorithm's the framework

The GC-ToF-MS chromatograms of the analyses of tobacco smoke with various baselines are illustrated in Fig. 3a.

Raman spectra

Several prednisone acetate tablets (PATs) were analyzed by a BWTEK i-Raman-785 spectrometer using a laser

of 785 nm wavelength for excitation with a 2048 elements thermoelectric cooled linear charge-coupled device arrays. PATs' Raman spectra from ten different pharmaceutical factories were recorded using 5,000 ms integration times. As three tablets for each pharmaceutical factory were measured, a total of 30 Raman spectra were obtained. These Raman spectra are plotted in Fig. 4a,

Table 1 The analysis of the results 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	SirQR ^d
Linear	Peak 1	57.01	50.00	50.55	50.44	50.36	50.31
	Peak 2	40.64	30.00	29.06	29.64	29.34	30.14
	Peak 3	86.29	70.00	68.42	69.86	68.17	70.17
	Peak 4	63.85	40.00	40.21	40.01	40.03	39.53
Curved	Peak 1	59.71	50.00	50.56	50.53	52.11	49.71
	Peak 2	32.45	30.00	29.03	29.08	28.78	29.18
	Peak 3	78.20	70.00	68.39	68.99	68.89	69.92
	Peak 4	45.71	40.00	40.04	38.88	40.05	40.20

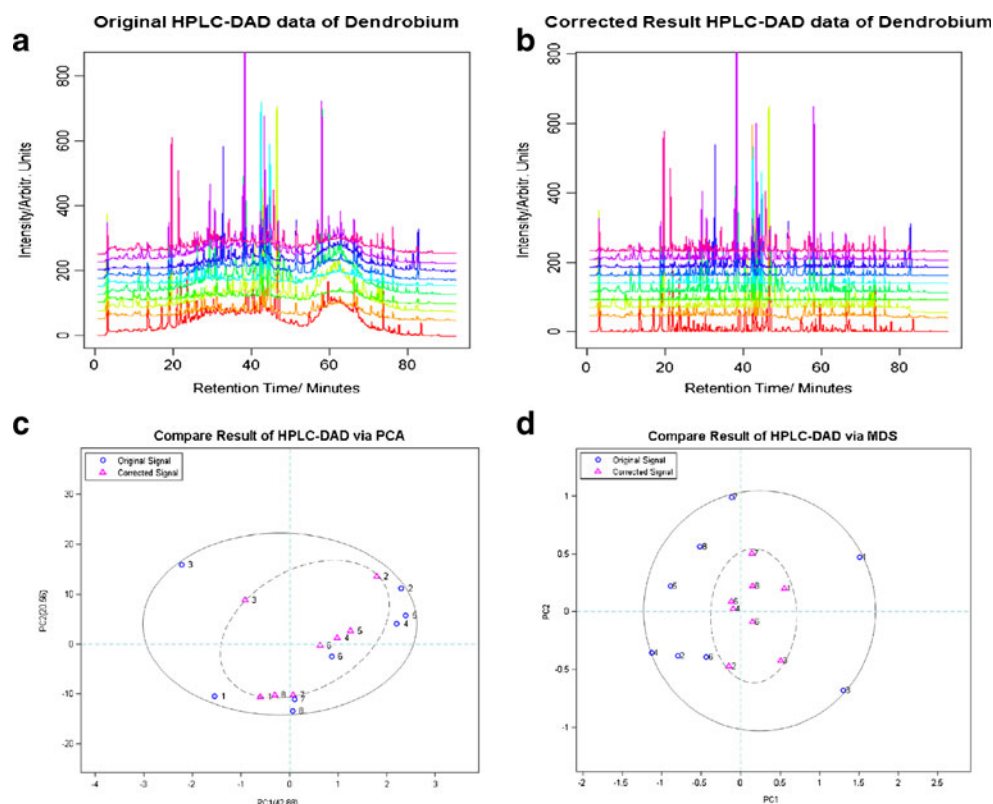
^a For the ALS method, the parameters are as follows: $\lambda=10, p=10^{-5}, d=2$

^b For the FABC method, the parameters are as follows: $\lambda=10, a=10$

^c For the airPLS method, the parameters are as follows: $\lambda=8$

^d For the SirQR method, the parameters are as follows: $\lambda=1.25, \mu=0.03, d_m=4 \times 10^{-4}$

Fig. 2 Chromatographic data of *Dendrobium*: **a** original chromatograms from ten samples and **b** corrected chromatograms. **c** Analyses results of the corrected chromatograms with PCA scores and variance after centralization and normalization using PCA method, and **d** analyses results also after centralization and normalization using multidimensional scaling analysis method. In **(c)** and **(d)**, these blue circles indicate the original chromatograms and the red triangles represent the corrected chromatograms. The compactness of these eight sample points becomes closeness from the solid circle to dashed circle



and one can obviously notice that the baselines vary from sample to sample.

Nuclear magnetic resonance dataset

Nuclear magnetic resonance signals are also influenced by baseline. This paper demonstrates the feasibility of the proposed algorithms in the correction of the baseline with larger numbers of variables.

These spectra were acquired on a Varian INOVA AS600 (Varian, Inc. US) 600 MHz NMR. Proton chemical shifts at 298 K were obtained from depurated plasma and human urine. Five urine samples of the same people collected in different times and one plasma sample were included in this dataset. The original data of ^1H NMR signals is illustrated in Fig. 5a, where one urine signal (the red line and best out of five) and the plasma signal (the cyan line) are represented.

Result and discussion

HPLC-DAD result

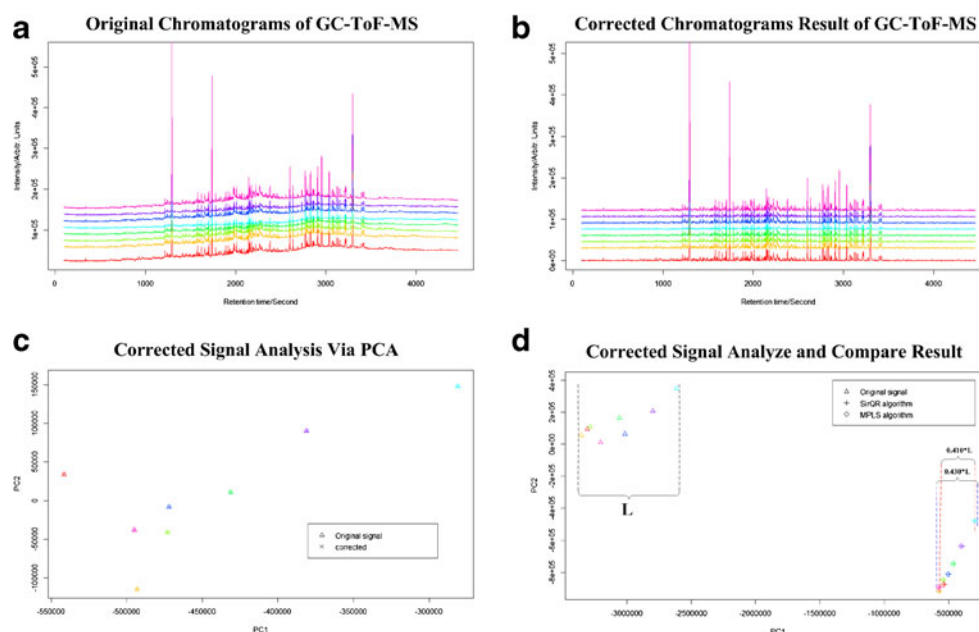
The baselines of HPLC-DAD chromatograms of ten *Dendrobium* samples from different regions have been corrected by SirQR algorithm with $\lambda=1.25, \mu=0.03$. Both original and corrected chromatograms are illustrated in Fig. 2a, b.

Although the influence of the baseline decreased significantly after correction, in order to express its effectiveness, two different dimension reduction methods have been applied to the mean-centered datasets for further research. The first one was a principal component analysis (PCA), applied to investigate the influences of clustering analysis on the proposed SirQR algorithm, comparing the first two principle components between original signals and corrected ones with centralized and normalized dataset. The second one was a multidimensional scaling (MDS) [45–47], which also applies to the matrix, consisting of the original and corrected mean-centered normalization chromatogram signals.

This resulted in a significant improvement in the scores and the variance of the first and second principal components and making the interpretation more clearly and be easy to distinguish the wanted difference and the center of gravity of the model, which are plotted in Fig. 2c, d. In Fig. 2c, d, the blue circles represent the original chromatograms and the red triangles represent the corrected chromatograms. Obviously, the movement and aggregation of red triangles are more compact, which can be easily seen from the plots (Fig. 2c, d) that the size of the sample point circle after correction is much smaller than that of the sample point circle before background correction. Notice that the first two components can explain more than 60 % of total information. In Fig. 2d, especially, the MDS [48] method is also used to indicate the aggregation trend after the correction in a clearer manner. From Fig. 2d, one could

Fig. 3 Background correction results and analysis for the GC-TOF-MS dataset of tobacco smoke by SirQR algorithm.

a Original chromatogram dataset of eight samples with various backgrounds; **b** corrected chromatogram results through the SirQR method; **c** comparison of the distribution of samples before and after background correction in principal component spaces; and **d** first two principal components of the original and corrected chromatograms with first-order numerical differentiation preprocessing and comparison with MPLS method



clearly see that the results tend obviously to the center of gravity of the data after corrected process via SirQR algorithm. The combination of the PCA and MDS plots demonstrates the validity of the SirQR algorithm. Via SirQR algorithm, the corrected chromatograms have a more compact pattern and are closer to the desired chromatograms. The clustering and classification results increased because of the compactness and closeness in principal components pattern space to a certain degree.

GC-ToF-MS discussion

For the GC-ToF-MS chromatogram datasets of tobacco samples, the background was corrected by the proposed SirQR algorithm. The corrected results were further analyzed by principal component analysis (PCA), also comparing these results with other novel algorithm named MPLS [7], as shown in Fig. 3. In Fig. 3a, b one can clearly see that the original and

corrected chromatograms demonstrate that the SirQR algorithm is flexible enough to remove the background drifts even with an excessive number of variables. Moreover, PCA was also implemented to assess the validity of the proposed SirQR algorithm. As known to all, numerical differentiation can eliminate the tardily shifting background [49, 50]. Thus, PCA method has been executed in the original and corrected chromatogram signals with first-order numerical differentiation preprocessing. In Fig. 3c, the triangles represent the original chromatogram signals, and the crosses represent the corrected ones. As illustrated, the good matching in principal component spaces, suggests that the SirQR algorithm does not eliminate the important information from the original chromatogram signals. Moreover, as all eight samples were parallel samples, if the effect of background can be ignored, they should be located closer to each other in principal component spaces. Focusing on Fig. 3d, the triangles represent chromatograms without any background correction, neither by SirQR

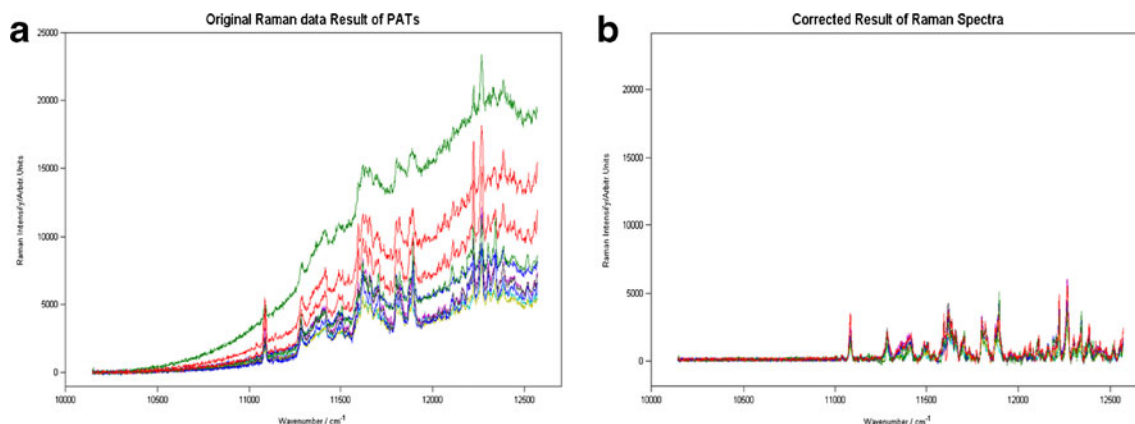
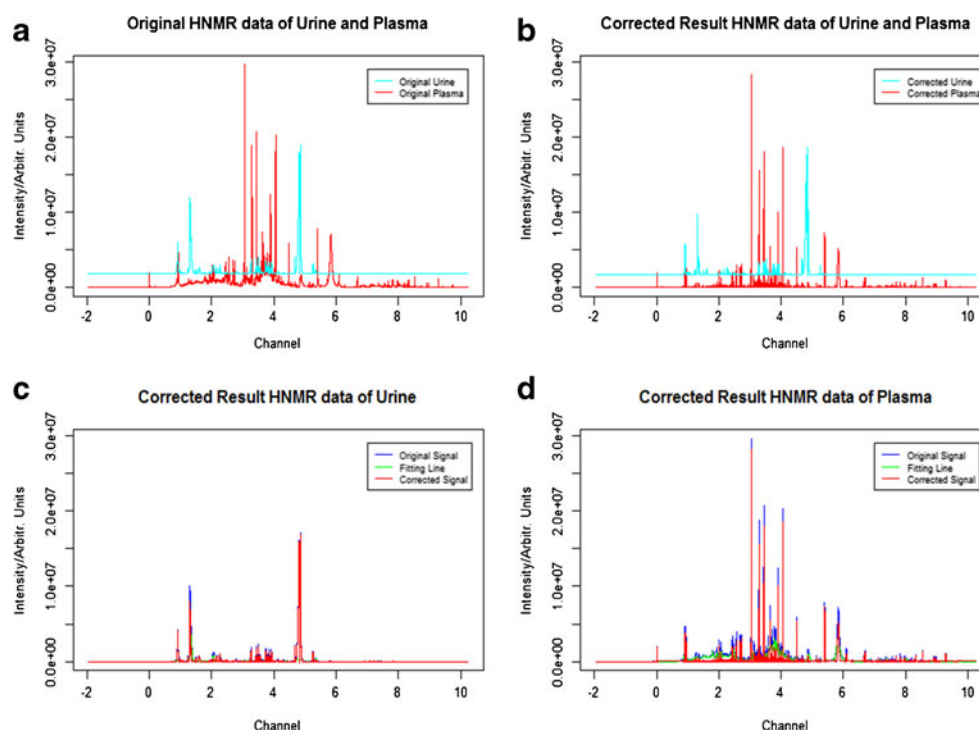


Fig. 4 Baseline correction results of the Raman spectra of PATs. **a** Original spectra dataset. **b** Corrected spectra dataset

Fig. 5 Baseline correction results of the HNMR data of plasma and urine samples. **a** Original HNMR data of a plasma sample (the below red line) and a urine sample (the above cyan line). **b** Corrected HNMR data of a plasma sample (the below red line) and a urine sample (the above cyan line). **c** Original data (blue line) with the fitting baselines (green line) and corrected result (red line) of one urine sample of these ones. **d** Original data (blue line) with the fitting baselines (green line) and corrected results (red line) of one plasma sample



nor by MPLS in principal component spaces; the plus symbols represent the corrected chromatogram signals by the proposed SirQR algorithm; and the diamonds represent the corrected chromatogram signals by the MPLS algorithm. The direction in the first principal component mainly conveyed the sample difference, as the value reaching 92 % of the total variance in first principal component. If we let the scope of these original samples' chromatogram signals be indicated as L in the first principal component direction, it is not difficult to obtain the scope of the corrected samples' chromatogram signals in both SirQR method and MPLS method. Comparing with the two distances, the SirQR method, with $0.410 \times L$, is slightly better than the MPLS method, with $0.430 \times L$. Thus, these values can point out that SirQR algorithm can decrease the variance mostly originated from the background, the same way as the MPLS algorithm does or even better to some extent. According to the analysis above, one can demonstrate that the larger variation in the first principal component direction of the original chromatograms could be due to the variation of background from chromatogram to chromatogram. This proposed SirQR method can remove this background variation among a series of chromatograms without missing useful and important information.

Correction result of Raman spectra

The proposed sirQR algorithm has also been applied to Raman spectra of PATs with highly fluorescence background. All baseline influences on the ten different PATs tablets spectra were favorably removed. These results are represented in

Fig. 4b. Similarity analyses (correlation coefficient) have been applied to the corrected analysis results with original Raman spectra, via comparison with the corresponding mean spectra, which is listed in Table 2.

Before correction, the similarity values of the original Raman signal and the original mean spectra seem unsatisfactory and unacceptable. Even some original signals went far beyond our expectation. However, an obvious improvement was achieved, as shown in the corrected row of Table 2, via the SirQR algorithm proposed correction method. One can see that the improvement of the corrected result is not just a few percent digits but even one or two orders of magnitude. Some corrected results can reach an extremely high similarity, which suggests that the SirQR method has the ability to remove baselines and conserve the useful signals. In order to further confirm this, we took a t test method to test the significance of the improvement of correlation data between with and without baseline correction. A paired t test has been taken, in which the hypothesis testing, say H_0 , is that $\mu_1 = \mu_2$. Here, μ_1 means the mean value of the correlation values listed in Table 2 without baseline correction, while μ_2 indicates the corresponding one with baseline correction. The testing result is also shown in Table 2, that is, they are significantly different at $\alpha = 0.05$ level. That is to say, according to the statistical analysis, the similarity after baseline correction using SirQR algorithm is clearly better than that of the original dataset without baseline correction.

In short, the SirQR algorithm could amend the baseline validly and accurately while retaining and optimizing primary useful information.

Result information of NMR signals

The NMR of the depurated plasma and human urine samples for proton signals, as described in the “[Experimental and applications](#),” has been also corrected by the proposed SirQR algorithm. The main purpose of this correction is to test the performance of the SirQR approach on high-throughput dataset, which exceeded four thousand variables in H NMR signal, and make a preliminary classification between these two samples.

As the number of variables and the value of signals were extremely large, the difference value (d_m) between original signals and fitting values was selectively changed by the user. We designed $d_m = 2.5 \times 10^4$ and other parameters were set by these default ones ($\lambda = 1.25$, $\mu = 0.03$). One can observe the satisfactory corrected result in Fig. 5b, where only three iterations fit the prospective baseline. The cyan line means the corrected plasma signal and the red line indicated the urine signal. To clearly see each sample, the separated figures of these two samples are illustrated in Fig. 5c, d, including the original and corrected signals. In addition, these NMR signals were tested with PCA and MDS methods, as shown in Fig. 6. In Fig. 6a, b, the samples nos. 1 to 5 correspond to the mean-centered and normalized dataset of different five urine samples where the blue circles indicate these original signal and the red triangles correspond to the corrected ones via SirQR algorithm. The two big solid circles present the compactness of the original sample points, while the dashed circles indicate the compactness of the corrected sample points. From both Fig. 6a, b, one can see the clear variation trend and aggregation extent in both PCA and MDS spaces between solid circle and dashed circle, which demonstrates that the urine samples obtained a better aggregation after correction. The PCA variance of the first two components are more than 80 %. From Fig. 6b, one can easily see even clearer compact trend for the red triangles (corrected sample points) in MDS space. In summary, the SirQR algorithm can also be successfully and flexibly implemented in high-throughput real experimental datasets.

Comparison with other algorithms using simulated dataset and real dataset

The corrected baselines of linear and curved synthetic datasets have been implemented by the proposed SirQR algorithm. The corrected results can be seen in Fig. 7a, b. Both the linear and the curved baselines were subtracted successfully. Additionally, the SirQR algorithm could converge rapidly. As the simulated datasets were obtained from compounds with four known standard Gaussian peaks, the expected heights of four peaks are well known. Therefore, one can compare the heights before and after correction with the expected ones. The comparison result is listed in Table 1.

In the light of the expected heights, which are known, four different methods have been taken to make comparisons using datasets with the linear and the curved baseline. The methods used for comparison were the fully automatic baseline correction procedure of Carlos Cobas et al. [20] (FABC algorithm), asymmetric least squares baseline correction of Eilers [51, 52] (ALS algorithm) and adaptive iteratively reweighted penalized least squares of Zhang [22] (airPLS algorithm). The results of the ALS algorithm, the FABC algorithm, the airPLS algorithm, and the SirQR algorithm are listed in Table 1. Combining these results with Fig. 7a, b, one can observe that the SirQR algorithm is accurate and can fit the most reasonable baseline in both linear and curved situations. In the linear baseline, peak heights after baseline correction of the ALS algorithm and the airPLS algorithm were significantly closer to the expected height than the FABC algorithm and the ALS algorithm, especially in some wide and large peak. In addition, the SirQR algorithm is also successful in fitting the curved baseline and much better than the other three algorithms, including some small peak sections. Deducing from Table 1, one can conclude that the SirQR algorithm corrected the baseline as well as the other algorithms and even better them in some extent.

It is worth noting that using the quantile regression method is a clear advantage of our correction algorithm, compared with the ALS correction method, the FABC approach and airPLS algorithm. As it was shown in the first four pictures

Table 2 The similarity of original or corrected Raman spectra with mean spectrum

Sample	Sample1	Sample2	Sample3	Sample4	Sample5	Sample6	Sample7	Sample8	Sample9	Sample10
Similarity										
Original	0.9524	0.8949	0.9449	0.8850	0.9723	0.9588	0.9767	0.9556	0.9266	0.9601
Corrected	0.9824	0.8942	0.9960	0.9975	0.9881	0.9901	0.9910	0.9512	0.9977	0.9991
<i>t</i> test	$\mu_1 = 0.9527$ $\mu_2 = 0.9789$ $n = 10$ $ t = 2.2935 > t_{\alpha/2}(9) = t_{0.25}(9) = 2.2622$, reject H_0									

Ten samples were used to compare the similarity between corrected and before with the corresponding mean signal

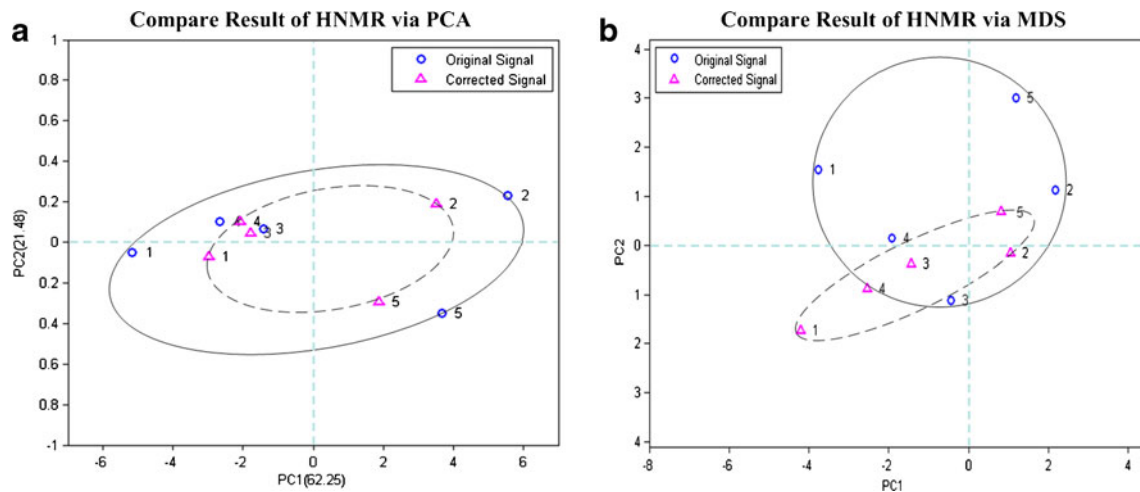


Fig. 6 First two principle component plots using two different methods with the scores of the original and corrected results of the five urine samples after centralization and normalization: **a** using classic PCA method with the variance and **b** using MDS method. In both pictures, these blue circles indicate the original urine HNMR data; the red triangles

represent the corrected urine HNMR data; the solid circle shows the compactness of the original sample points; and the dashed circle represents the compactness of the samples after corrected. The variation trend between two circles also illustrates the improvement of sample compactness after correction

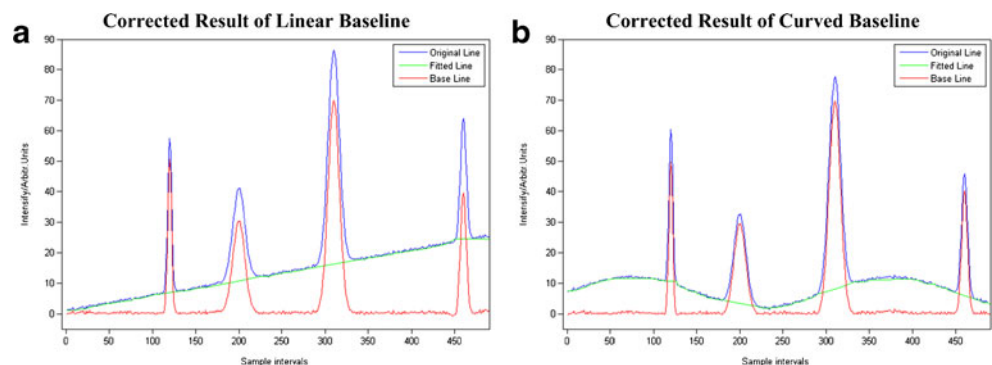
of Fig. 8, these three algorithms have been applied in correcting the noisy data in comparison with the proposed SirQR algorithm. One can clearly see that, although the FABC method could be better to fit the noise data, it could not correct the wider peaks (see Fig. 8a). For the results from ALS (Fig. 8b) and airPLS (Fig. 8c) algorithms, the effect of noise processing is not satisfactory, as the fitting lines are obviously under the original data. On the other hand, the SirQR algorithm could do a better fitting as shown in Fig. 8d. Although the corrected result in high noise is not better than with the FABC method, the fitting line kept quite close to the original data, and also is obviously better than the other two methods (ALS and airPLS). In particular, the SirQR algorithm can be flexible when dealing with these wide peaks notably better than FABC method.

In addition, three other different simpler methods have been taken to make a comparison with the real spectra dataset of milk in Fig. 8 as well. The methods used for comparison were iterative restricted least squares of Coombes et al. [53] (IRLS algorithm), simultaneous peak detection and baseline

correction of Coombes et al. [54] (PeakDetection algorithm), Rolling Ball algorithm for X-ray spectra by Kneen and Annegarn [55] (RollingBall algorithm). The results of the IRLS algorithm, the PeakDetection algorithm, the RollingBall algorithm, and the SirQR algorithm are illustrated in the rest four pictures in Fig. 8, respectively.

It is worth noting that using the quantile regression method is a clear advantage of our correction algorithm, compared with the IRLS correction method, the PeakDetection approach and RollingBall algorithm. According to Fig. 8, one can clearly see that, although the IRLS correction method (Fig. 8e) could be well in fitting the overall baseline trend, it could not do well in high noise part and some parts of the corrected spectra were still drift, which were clearly below the x -axis. For the result from RollingBall algorithm (Fig. 8f) and PeakDetection algorithm (Fig. 8g), the effect of noise processing made a huge influence that the fitted baseline in the range of 0 to 1,000 were significantly distort and unreasonable. Although the corrected result using SirQR algorithm (Fig. 8h) could not obtain a perfect result in every sections,

Fig. 7 Corrected results of simulated data with different baselines. **a** Linear baseline and the corresponding corrected results. **b** Curved baseline and the corresponding corrected results



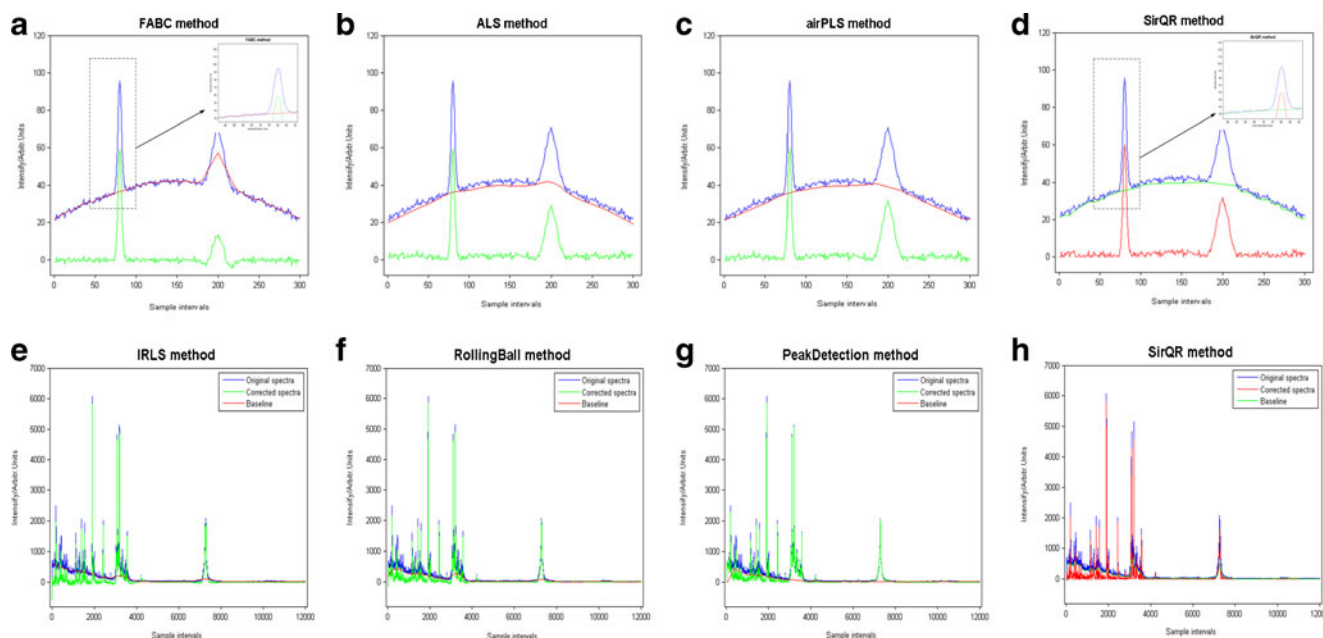


Fig. 8 Test and comparison with other different algorithms using simulated dataset and real dataset. The first four pictures describe the robustness testing result for four algorithms on high noisy simulated data. **a** FABC method, corrected result and the focusing partial figure. **b** ALS method and corrected result. **c** airPLS method and corrected result. **d** SirQR method, corrected result and the

focusing partial figure. The rest four demonstrate the comparison of the corrected results with other four algorithms on real dataset. **e** Using IRLS method and the corrected result. **f** Using RollingBall method and the corrected result. **g** Using PeakDetection method and the corrected result. **h** Using SirQR method and the corrected result

it could do well in the range of high noise section, kept the fitting line close the original spectra and kept the stability of the corrected result. For the analysis above, the SirQR algorithm is obviously better than the other three methods (IRLS, RollingBall, and PeakDetection).

In summary, combining with the two aspects we can say that the SirQR algorithm seems to be a robust correction method. The main reason may be due to L_1 norm, which is the sum of absolute values during the correcting process. These absolute values not only reach the purpose of fitting the aim line, but also, unlike the sums of squares, cannot enlarge the error value. Thus, the robustness of this algorithm can be a significant advantage, obviously better than others in general.

Processing speed and expansibility

As described in the section above, Table 3 describes the number of variables, total calculation times and calculation time per iteration of simulate data, Raman spectra, GC-ToF-MS, HPLC-DAD and NMR signals. From this table one can deduce that the speed of the proposed SirQR algorithm is swift enough even for the large datasets with more than 20,000 variables, such as the NMR signal, that was only 14.611 s for 24,000 variables with two iterations. This achievement can be owed to the sparse matrix [56]. Moreover, in order to control the effectiveness of the fitting baseline, the maximum iteration times and the value of d_m can be manually established by the users. By means of this manual test, a desirable and

reliable fitting baseline can be more effectively amended for further analysis, avoiding for the overfitting phenomenon. Detailed figure of the overfitting phenomenon is given as Fig. S2 in the Electronic supplementary material.

Continuing to investigate Table 3, it is not difficult for one to find the relationship between the number of variables and calculation time per iteration. Obviously, with the increasing of the number of variables, the corresponding calculation time will also increase. Although the time of calculation is higher for larger number of variables; it is fast enough for most general data, especially with less than 10,000 variables. It can be summarized that the usage of sparse matrices and the exponential reweight strategy enables the application of the SirQR algorithm in more high-throughput domains and meet the needs of data analysis.

Table 3 The calculation time of simulated dataset, HPLC-DAD, GC-ToF-MS, Raman spectra, and NMR signal

Resource signal	Number of variables	Total calculation time(s)	Iteration times	Calculation time per iteration(s)
Simulated dataset	500	0.111	1	0.111
Raman spectra	1,715	0.349	2	0.175
GC-ToF-MS	4,464	2.095	2	1.048
HPLC-DAD	11,064	4.974	2	2.487
NMR signal	24,000	14.611	2	7.305

Different datasets were applied to infer the relationship between the number of variables and the calculation time per iteration

Conclusions

In this research, the proposed SirQR algorithm provides a robust, valid, and fast baseline correction method for processing different analytical signals. The proposed algorithm combines quantile regression and reweighted iterative strategy to fit the background as desired. After comparing with several popular baseline correction methods, such as ALS, FABC airPLS and MPLS, the results demonstrate that the proposed algorithm can offer a robust and accurate baseline corrected signals for both simulated data and real analytical signals. Moreover, the successful results of these datasets have proved that this approach can be used as a preprocessing method for many analytical instruments (chromatograms, Raman spectra and NMR signals, even MALDI-TOF).

Acknowledgments This work is financially supported by the National Nature Foundation Committee of P.R. China (grants no. 21075138, 21105129, 21175157, 21275164, and 21305163), China Hunan Provincial science and technology department (grants no. 2012FJ4139), Hunan Provincial Natural Science Foundation of China (Grants No. 14JJ3031). The studies meet with the approval of the university's review board. The authors are grateful to all employees of this institute for their encouragement and support of this research. Simultaneously, the authors want to thank Zhong Li of Yunnan Academy of Tobacco Science, China tobacco Yunnan industrial Co., Ltd for providing the GC-ToF-MS dataset. Yanpeng An of the Shandong Analysis and Test Center, Jinan, China is acknowledged for providing the NMR dataset of the plasma and urine for the regression and classify analysis.

References

- Jirasek A, Schulze G, Yu MML, Blades MW, Turner RFB (2004) Accuracy and precision of manual baseline determination. *Appl Spectrosc* 58(12):1488–1499. doi:10.1366/0003702042641236
- Pearson GA (1977) *J Magn Reson* 27:256–272
- Liang YZ, Leung AKM, Chau FT (1999) A roughness penalty approach and its application to noisy hyphenated chromatographic two-way data. *J Chemom* 13(5):511–524
- Shao XG, Cai WS, Pan ZX (1999) Wavelet transform and its applications in high performance liquid chromatography (HPLC) analysis. *Chemom Intell Lab Syst* 45(1–2):249–256
- Boelens HFM, Dijkstra RJ, Eilers PHC, Fitzpatrick F, Westerhuis JA (2004) New background correction method for liquid chromatography with diode array detection, infrared spectroscopic detection and Raman spectroscopic detection. *J Chromatogr A* 1057(1–2):21–30. doi:10.1016/j.chroma.2004.09.035
- Cheung W, Xu Y, Thomas CLP, Goodacre R (2009) Discrimination of bacteria using pyrolysis-gas chromatography-differential mobility spectrometry (Py-GC-DMS) and chemometrics. *Analyst* 134(3):557–563. doi:10.1039/b812666f
- Li Z, Zhan DJ, Wang JJ, Huang J, Xu QS, Zhang ZM, Zheng YB, Liang YZ, Wang H (2013) Morphological weighted penalized least squares for background correction. *Analyst* 138(16):4483–4492. doi:10.1039/c3an00743j
- Ruckstuhl AF, Jacobson MP, Field RW, Dodd JA (2001) Baseline subtraction using robust local regression estimation. *J Quant Spectrosc Radiat Transf* 68(2):179–193
- Schechter I (2002) Correction for nonlinear fluctuating background in monovariate analytical systems. *Anal Chem* 67(15):2580–2585. doi:10.1021/ac00111a014
- Lieber CA, Mahadevan-Jansen A (2003) Automated method for subtraction of fluorescence from biological Raman spectra. *Appl Spectrosc* 57(11):1363–1367
- Mazet V, Carteret C, Brie D, Idier J, Humbert B (2005) Background removal from spectra by designing and minimising a non-quadratic cost function. *Chemom Intell Lab Syst* 76(2):121–133
- Morhác M, Matoušek V (2008) Peak clipping algorithms for background estimation in spectroscopic data. *Appl Spectrosc* 62(1):91–106
- Zhang ZM, Chen S, Liang YZ, Liu ZX, Zhang QM, Ding LX, Ye F, Zhou H (2010) An intelligent background-correction algorithm for highly fluorescent samples in Raman spectroscopy. *J Raman Spectrosc* 41(6):659–669. doi:10.1002/jrs.2500
- Chen S, Li XN, Liang YZ, Zhang ZM, Liu ZX, Zhang QM, Ding LX, Ye P (2010) Raman spectroscopy fluorescence background correction and its application in clustering analysis of medicines. *Spectrosc Spectr Anal* 30(8):2157–2160. doi:10.3964/j.issn.1000-0593(2010)08-2157-04
- Liland KH, Rukke E-O, Olsen EF, Isaksson T (2011) Customized baseline correction. *Chemom Intell Lab Syst* 109(1):51–56. doi:10.1016/j.chemolab.2011.07.005
- Liu Y, Cai W, Shao X (2013) Intelligent background correction using an adaptive lifting wavelet. *Chemom Intell Lab Syst* 125(0):11–17. doi:10.1016/j.chemolab.2013.03.010
- Dietrich W, Rüdel CH, Neumann M (1991) Fast and precise automatic baseline correction of one- and two-dimensional NMR spectra. *J Magn Reson* (1969) 91(1):1–11
- Moore AW Jr, Jorgenson JW (1993) Median filtering for removal of low-frequency background drift. *Anal Chem* 65(2):188–191
- Golotvin S, Williams A (2000) Improved baseline recognition and modeling of FT NMR spectra. *J Magn Reson* 146(1):122–125
- Carlos Cobas J, Bernstein MA, Mart-Pastor M, Tahoces PG (2006) A new general-purpose fully automatic baseline-correction procedure for 1D and 2D NMR data. *J Magn Reson* 183(1):145–151
- Chang D, Banack CD, Shah SL (2007) Robust baseline correction algorithm for signal dense NMR spectra. *J Magn Reson* 187(2):288–292
- Zhang ZM, Chen S, Liang YZ (2010) Baseline correction using adaptive iteratively reweighted penalized least squares. *Analyst* 135(5):1138–1146. doi:10.1039/b922045c
- Zhang ZM, Liang YZ (2012) Comments on the baseline removal method based on quantile regression and comparison of several methods. *Chromatographia* 75(5–6):313–314. doi:10.1007/s10337-012-2192-x
- Koenker R (2004) Quantile regression for longitudinal data. *J Multivar Anal* 91(1):74–89
- Eilers PH, De Menezes RX (2005) Quantile smoothing of array CGH data. *Bioinformatics* 21(7):1146–1153
- Hong J, Schlegel EM, Grindlay JE (2004) New spectral classification technique for X-ray sources: quantile analysis. *Astrophys J* 614(1):508
- Callister SJ, Barry RC, Adkins JN, Johnson ET, W-j Q, Webb-Robertson B-JM, Smith RD, Lipton MS (2006) Normalization approaches for removing systematic biases associated with mass spectrometry and label-free proteomics. *J Proteome Res* 5(2):277–286
- Chernozhukov V, Hansen C (2008) Instrumental variable quantile regression: a robust inference approach. *J Econ* 142(1):379–398. doi:10.1016/j.jeconom.2007.06.005
- Jun SJ (2008) Weak identification robust tests in an instrumental quantile model. *J Econ* 144(1):118–138. doi:10.1016/j.jeconom.2007.12.006
- Jun SJ (2009) Local structural quantile effects in a model with a nonseparable control variable. *J Econ* 151(1):82–97. doi:10.1016/j.jeconom.2009.02.011

31. Wunderli T (2013) Total variation time flow with quantile regression for image restoration. *Journal of Mathematical Analysis and Applications*.
32. Yu H-L, Wang C-H (2013) Quantile-based Bayesian maximum entropy approach for spatiotemporal modeling of ambient air quality levels. *Environ Sci Technol* 47(3):1416–1424
33. Waldmann E, Kneib T, Yue YR, Lang S, Flexeder C (2013) Bayesian semiparametric additive quantile regression. *Stat Model* 13(3): 223–252
34. Komsta L (2011) Comparison of several methods of chromatographic baseline removal with a new approach based on quantile regression. *Chromatographia* 73(7–8):721–731. doi:10.1007/s10337-011-1962-1
35. Koenker R, Bassett G Jr (1978) Regression quantiles. *Econometrica: journal of the Econometric Society* 46:33–50
36. Koenker R (2005) Quantile regression, vol 38. Cambridge University Press, Cambridge
37. Hallock RKaKF (2001). *J Econ Perspect* 15:143–156
38. Whittaker E (1923) On a new method of graduation. *Proc Edinburgh Math Soc* 41:63–75
39. Eilers PH (2003) A perfect smoother. *Anal Chem* 75(14):3631–3636
40. Koenker RW, Bassett GW (1984) Four (pathological) examples in asymptotic statistics. *Am Stat* 38(3):209–212
41. Portnoy S, Koenker R (1997) The Gaussian hare and the Laplacian tortoise: computability of squared-error versus absolute-error estimators. *Stat Sci* 12(4):279–300
42. Holland PW, Welsch RE (1977) Robust regression using iteratively reweighted least-squares. *Communications in Statistics-Theory and Methods* 6(9):813–827
43. Rubin DB (1983) Iteratively reweighted least squares. *Encyclopedia of Statistical Sciences* 4:272–275
44. Green PJ (1984) Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives. *Journal of the Royal Statistical Society Series B (Methodological)*: 149–192.
45. Torgerson WS (1958) Theory and methods of scaling. Wiley, New York
46. McCune BaG JB (2002) Oregon. *MjM Software Design. Analysis of Ecological Communities*, Gleneden Beach
47. Green PJ (1975) Marketing applications of MDS: assessment and outlook. *J Mark* 39(1):24–31. doi:10.2307/1250799
48. Kruskal JB (1964) Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika* 29(1):1–27. doi:10.1007/bf02289565
49. Leger MN, Ryder AG (2006) Comparison of derivative preprocessing and automated polynomial baseline correction method for classification and quantification of narcotics in solid mixtures. *Applied Spectroscopy* 60(2):182–193
50. Ben-Amotz DMZD (2000) *Appl Spectrosc* 54:1379–1383
51. Eilers PHC (2004) Parametric time warping. *Anal Chem* 76(2):404–411. doi:10.1021/ac034800e
52. Paul HC, Eilers H, MacFie JH (2006) Baseline correction with asymmetric least squares smoothing. *J Magn Reson* 183(1):145–151
53. Goldstein H (1989) Restricted unbiased iterative generalized least-squares estimation. *Biometrika* 76(3):622–623
54. Coombes KR, Fritsche HA, Clarke C, Chen J-N, Baggerly KA, Morris JS, Xiao L-C, Hung M-C, Kuerer HM (2003) Quality control and peak finding for proteomics data collected from nipple aspirate fluid by surface-enhanced laser desorption and ionization. *Clin Chem* 49(10):1615–1623
55. Kneen M, Annegarn H (1996) Algorithm for fitting XRF, SEM and PIXE X-ray spectra backgrounds. *Nuclear Instruments and Methods in Physics Research Section B: Beam Interactions with Materials and Atoms* 109:209–213
56. Gilbert JR, Moler C, Schreiber R (1992) Sparse matrices in MATLAB: design and implementation. *SIAM Journal on Matrix Analysis and Applications* 13(1):333–356