Journal of Chromatography A, 1449 (2016) 89–99

Contents lists available at ScienceDirect

Journal of Chromatography A

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

Simple automatic strategy for background drift correction in chromatographic data analysis

Hai-Yan Fu[a,*], He-Dong Li[a], Yong-Jie Yu[a,b,c,*], Bing Wang[b], Peng Lu[b], Hua-Peng Cui[b], Ping-Ping Liu[b,*], Yuan-Bin She[d]

[a] School of Pharmaceutical Sciences, South-Central University for Nationalities, Wuhan 430074, China
[b] Zhengzhou Tobacco Research Institute of CNTC, Zhengzhou 450001, China
[c] Ningxia Engineering andTechnology Research Center for Modernization of Hui Medicine, Ningxia Medical University, Yinchuan 750004, China
[d] College of Chemical Engineering, Zhejiang University of Technology, Hangzhou 310014, China

ARTICLE INFO

Article history:
Received 13 January 2016
Received in revised form 13 March 2016
Accepted 17 April 2016
Available online 23 April 2016

Keywords:
Background drift correction
Complex sample analysis
Quality control
Metabolic profiling
LC-QTOF

ABSTRACT

Chromatographic background drift correction, which influences peak detection and time shift alignment results, is a critical stage in chromatographic data analysis. In this study, an automatic background drift correction methodology was developed. Local minimum values in a chromatogram were initially detected and organized as a new baseline vector. Iterative optimization was then employed to recognize outliers, which belong to the chromatographic peaks, in this vector, and update the outliers in the baseline until convergence. The optimized baseline vector was finally expanded into the original chromatogram, and linear interpolation was employed to estimate background drift in the chromatogram. The principle underlying the proposed method was confirmed using a complex gas chromatographic dataset. Finally, the proposed approach was applied to eliminate background drift in liquid chromatography quadrupole time-of-flight samples used in the metabolic study of *Escherichia coli* samples. The proposed method was comparable with three classical techniques: morphological weighted penalized least squares, moving window minimum value strategy and background drift correction by orthogonal subspace projection. The proposed method allows almost automatic implementation of background drift correction, which is convenient for practical use.

1. Introduction

Chromatographic instruments are widely employed for complex sample analysis in various scientific fields, including metabolomics, proteomics, and quality control of natural products [1–5]. In particular, the obtained chromatogram in the quality control of natural products and metabolic analysis may involve complexity because of the large amount of compounds eluted in finite time [6]. The traditional chromatographic data analysis strategy that inspects useful chemical information by visualization is a challenging and time-consuming task. Therefore, several data mining technologies that focus on peak picking and time-shift alignment have been developed to completely use the information in the chromatographic data and improve data analysis [7–11]. However, only few studies have been conducted on the influences of chromatographic background drift. The majority of current baseline correction methods have been developed for spectroscopic data analysis, such as NMR spectroscopy [12,13], which has not been demonstrated for chromatographic data analysis.

Chromatographic background drift correction is a critical stage for chromatographic data analysis because the background drift in the chromatogram can seriously affect the peak picking results and lead to invalid time shift alignment [6], which may provide an unacceptable conclusion. Several techniques [6,12–29] have been developed for background drift correction. These methods can be classified into penalized least squares based methods [13,14,18,24,27,28] and moving window based methods [12,21,23]. The widely known penalized least squares based method is represented by asymmetrical least squares (ALS), which estimates background drift with a penalty smooth strategy. Zhang et al. [24] improved the performance of ALS by introducing an adaptive iteratively reweighted strategy, namely, adaptive iteratively reweighted penalized least squares (airPLS).

* Corresponding authors at: School of Pharmaceutical Sciences, South-Central University for Nationalities, Wuhan 430074, China; Zhengzhou Tobacco Research Institute of CNTC, Zhengzhou 450001, China.
E-mail addresses: yongjie.yu@163.com (Y.-J. Yu), Liu_pingping2012@163.com (P.-P. Liu).

http://dx.doi.org/10.1016/j.chroma.2016.04.054
0021-9673/

The performance of ALS and airPLS largely depends on the weights assigned to useful signal parts. These methods can provide reasonable solutions for cases in which useful signals, such as peaks, in the chromatogram are properly weighted. To eliminate the influence of chromatographic peaks, Bao et al. [13] proposed an iterative technique based on peak detection to remove baselines from the NMR data. Li et al. [25] developed the morphological weighted penalized least squares (MPLS) method. Baek et al. [14] employed a logistic function to provide weights for penalized least squares. Among the successful applications of these parametric methods, the penalized parameter must be pre-optimized before use, which may be time consuming and tedious [30].

The moving window based methods [12,26] are mainly used for baseline correction in spectroscopic data analysis, such as infrared spectroscopy and Raman spectroscopy. Friedrichs [26] employed a median value in the window coupled with Gaussian filter to estimate baselines in NMR spectroscopy. Golotivin and Williams [12] subsequently developed baseline correction based on the removal of underlying signal points by comparing the difference of the maximum and minimum points with the pre-estimated instrumental noise. The efficiency of these methods depends on the window size used for data analysis. A large window may not remove the influence of background drift, whereas a small window may eliminate certain chromatographic peaks.

Yaroshchyk and Eberhardt [23] developed the moving window minimum value (MWMV) strategy for baseline correction in NMR data. The principle of MWMV is similar to that of Friedrichs' method. The minimum values in the window were initially obtained, and a moving averaging was employed by smoothing these values to estimate the underlying baseline. The performance of the MWMV still depends on the window size. Theoretically, MWMV is suitable for background drift correction of chromatograms with an acceptable baseline separation of chromatographic peaks. However, in extremely complex sample analysis, chromatographic peaks often co-elute, making MWMV lose certain useful information after baseline correction. Nevertheless, MWMV provides a new strategy for chromatographic background drift correction.

Given the numerous applications of hyphenated chromatographic techniques, such as liquid chromatography–mass spectrometry (LC–MS), in complex sample analysis, background drifts in 2D chromatographic datasets have received considerable attention [5,31–39]. Kuligowski et al. [35] employed orthogonal projection for background drift correction by taking full advantage of the bilinear structure of the hyphenated chromatographic dataset. Orthogonal subspace projection is discussed in more detail in our previous work [39]. Recently, Harrington's research group [37,38] used a similar strategy for baseline correction of hyphenated datasets. The principles of these methods are mathematically equivalent, and the only difference lies in the construction of a projection matrix of the background spectra. In complex sample analysis, the spectral profiles of blank samples or background may vary across samples, thereby presenting a challenge for these orthogonal subspace projection-based methods. Meanwhile, the signal-to-noise ratio (SNR) may significantly decrease when the spectral profiles of the components seriously overlap with those of the blank backgrounds [39]. These methods were not employed to eliminate background drift in liquid chromatography quadrupole time-of-flight (LC-QTOF) samples. Improved methods for background drift correction in high resolution LC–MS should be urgently developed [40].

In this work, a novel automatic background drift correction method was developed using local minimum values coupled with robust statistical analysis (LMV-RSA). The local minimum values in a chromatogram were immediately selected after data collection and then organized as a new baseline vector. A robust statistical strategy was employed to detect the outlier data points that may correspond to the unseparated peaks; these points were subsequently estimated using linear interpolation to obtain a new baseline vector. This procedure was repeated until convergence. The obtained baseline vector was finally expanded into the original chromatogram, and linear interpolation was employed to estimate the background drift in the chromatogram. The principle of the proposed method was completely investigated using a simulated chromatographic dataset. The new method was subsequently demonstrated using a complex gas chromatographic dataset. The performance of the proposed method was compared with two classic baseline correction techniques, MPLS and MWMV. The proposed method was employed for background drift correction in a LC-QTOF metabolic dataset for the first time to eliminate background drift; a comparison with orthogonal subspace projection was also provided.

2. Methodology

The objective of chromatographic background drift correction is to estimate baselines without removing useful information corresponding to chromatographic peaks. The development of an automatic strategy to identify the starting positions of chromatographic peaks and instrumental noise in the chromatogram entails difficulty. However, Yaroshchyk and Eberhardt [23] indicated that the majority of the minimum values in the chromatogram are not affected by chromatographic peaks; thus, the minimum values were used in our proposed method.

Fig. 1 depicts the workflow of LMV-RSA, which consists of three stages, namely, initialization (stage 1), iterative optimization (stage 2), and estimation of background drift (stage 3). The succeeding sections elaborate on each stage.

2.1. Initialization stage

The local minimum values were initially detected and their positions in the original chromatogram were recorded. A data point c_i was recognized as the local minimum for as long as the following equation was satisfied:

$$c_{i-1} > c_i \text{ and} c_i < c_{i+1} \quad (1)$$

where c_i is the ith data point in the chromatogram and c_{i-1} and c_{i+1} are the $i-1$th and the $i+1$th data points in the chromatogram, respectively. Fig. 2A presents the extracted local minimum values, which were organized as a new minimum vector (Fig. 2B). The data points in the new minimum vector can be classified into chromatographic peak points (minimum points in the peak elution range) and instrumental noise points. The chromatographic peak points can be more readily recognized from noise. The following part aims to separate these chromatographic peak points from instrumental noise.

2.2. Iterative optimization stage to eliminate outliers

As depicted in Fig. 1, the iterative optimization stage consists of two independent parts. This first part (left part in Stage 2 of Fig. 1) is based on a moving window strategy and is the core principle of our method, whereas the second part (right part in Stage 2 of Fig. 1) is complementary. The estimated local minimum vector is a combination of results from two parts. In this work, the moving window strategy is discussed in more detail.

To detect outliers (chromatographic peak points) in the minimum vector ($\boldsymbol{x}$), the noise level in the minimum vector was first

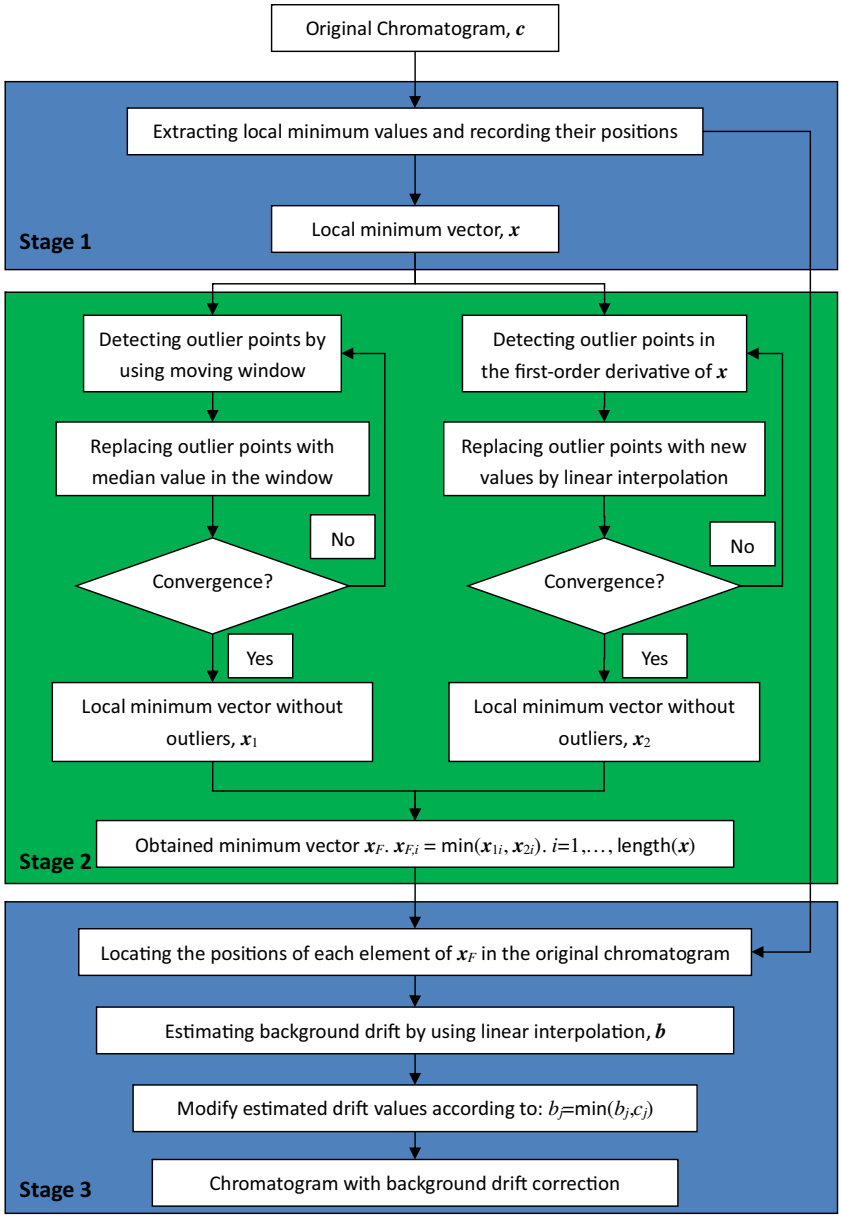

Fig. 1. Workflow of LMV-RSA.

estimated using the following median absolute deviation equation in the first-order derivative of $\boldsymbol{x}$:

$$\sigma = 1.483 * median|dx_j - median\,(\boldsymbol{dx})\,|\,(j = 1, \ldots, N) \quad (2)$$

where σ represents the estimated instrumental noise; the constant value, 1.483, is a correction factor that makes the σ belong to the normal distribution [41]; $\boldsymbol{dx}$ is the first-order derivative of the minimum vector $\boldsymbol{x}$; dx_j is the jth data point in $\boldsymbol{dx}$; and N is the number of data points in $\boldsymbol{dx}$.

Second, a moving window strategy was used to verify whether a data point in the minimum vector, $\boldsymbol{x}$, belongs to instrumental noise or a chromatographic peak. The *SNR* was calculated as follows:

$$SNR_i = max\left(|\,(x_i - median\,(\boldsymbol{x}_w))\,/\sigma|, |\,(x_i - x_{i-1})\,/\sigma|\right) \quad (3)$$

where SNR_i is the ith *SNR* that belongs to the ith data point, x_i, of the minimum vector $\boldsymbol{x}$, and $\boldsymbol{x}_w$ is an extracted vector in the window. The window size is a parameter that should be pre-estimated before use, and a window width of 30 satisfies most situations. Meanwhile, the median value of $\boldsymbol{x}_w$ is more robust compared with the average value. $x_i - x_{i-1}$ is the increment of the ith data point. A large increment usually suggests that x_i is the local minimum belonging to the chromatographic peaks.

The data point x_i is treated as an outlier when SNR_i is larger than 2.5 (this value corresponds to the confidence level of 99% under normal distribution), and replaced with the median value:$\hat{x}_i = median\,(\boldsymbol{x}_w)$. When each data point of $\boldsymbol{x}$ is verified, a new minimum vector can be obtained, which is represented by $\boldsymbol{x}_{new}$.

The above procedure is repeated until convergence. In this work, the convergence criterion is expressed as follows:

$$\|\boldsymbol{x}_{new} - \boldsymbol{x}\|/\|\boldsymbol{x}\| < 10^{-4} \quad (4)$$

where $\|\cdot\|$ represents the Frobenius norm. The iterative optimization stage is depicted in Fig. 1. Fig. 2C shows the updated minimum vector in each iterative step, in which the outliers are properly adjusted when the optimization stage reaches convergence.

The second part adopts a strategy similar to that of the first part, except that only the first-order derivative of $\boldsymbol{x}$, i.e. $\boldsymbol{dx}$, was employed. Elements of $\boldsymbol{dx}$ with absolute value larger than 2.5 are

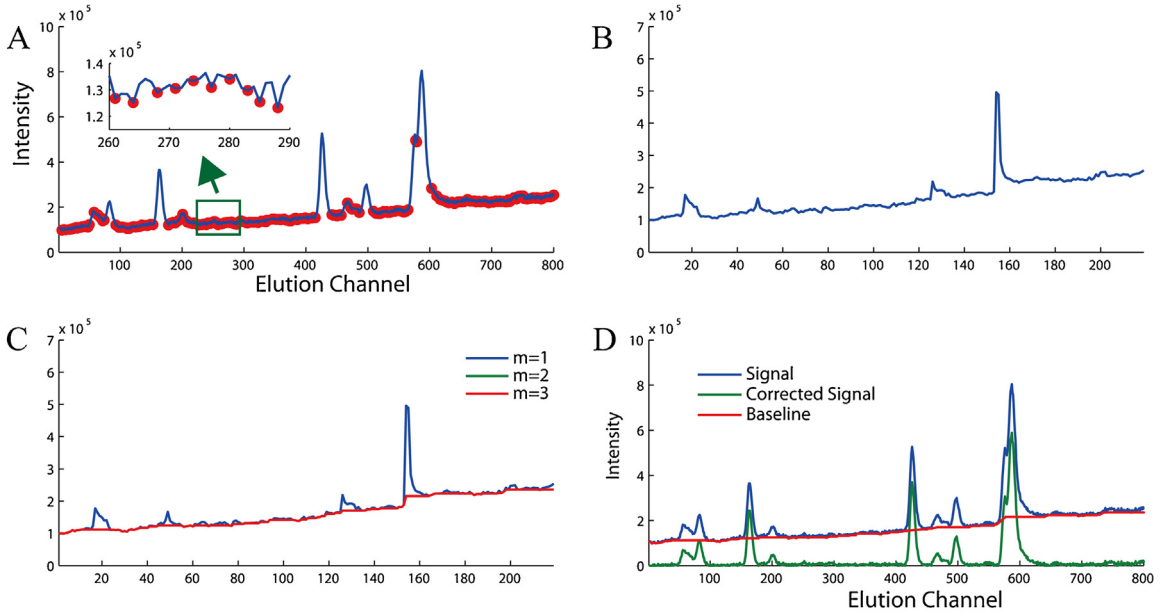

Fig. 2. Moving window strategy in LMV-RSA for background drift correction. (A) Chromatogram with local minimum values detected by LMV-RSA. (B) Extraction of the minimum values into a new local minimum vector. (C) Elimination of the outliers in the local minimum vector by iterative optimization. (D) Chromatogram with background drift corrected by LMV-RSA. The inserted plot of (A) shows the detected local minimum values in greater detail. Profiles with m = 1, m = 2 and m = 3 in (C) represent the updated vectors in each iterative step.

considered as outliers and estimated using a linear interpolation strategy built in MATLAB. The obtained vector is the combination of results derived from the two aforementioned parts, expressed as follows:

$$\boldsymbol{x_F} : \boldsymbol{x}_{F,i} = min\left(\boldsymbol{x}_{1,i}, \boldsymbol{x}_{2,i}\right) i = 1, 2, \ldots, \text{length}(\boldsymbol{x}) \quad (5)$$

where $\boldsymbol{x}_F$ represents the optimized vector. $\boldsymbol{x}_{F,i}$ is the ith element of $\boldsymbol{x}_F$. $\boldsymbol{x}_{1,i}$ and $\boldsymbol{x}_{2,i}$ represent the ith elements in the modified vectors from the first and second part, respectively. length($\boldsymbol{x}$) represents the length of the original local minimum vector $\boldsymbol{x}$.

2.3. Estimating the background drift stage

When the iterative optimization stage converges, a linear interpolation step is immediately performed to estimate the background drift in the chromatogram. The optimized local minimum vectors are first marked in the original chromatogram in accordance with their positions, and a linear interpolation is introduced to the estimated background drift in each elution channel. The interpolated values between two successive local minimum values, such as x_{i-1} and x_i, are calculated as follows:

$$\hat{b}_j = \frac{x_i - x_{i-1}}{p_i - p_{i-1}}(j - p_{i-1}) + x_{i-1} \quad (6)$$

where $\hat{b}_j$ is the estimated background drift of the jth elution channel in the chromatogram. p_{i-1} and p_i represent the positions of the elements x_{i-1} and x_i in the original chromatogram, respectively. $x_i - x_{i-1}/p_i - p_{i-1}$ is the increment per elution channel and $j - p_{i-1}$ is the distance (number of elution channels) between the jth elution channel and x_{i-1}. Finally, chromatogram with baseline correction can be obtained by subtracting the estimated background drift, $\boldsymbol{b}$, from the original chromatogram, $\boldsymbol{c}$, as follows:

$$\boldsymbol{c}_{cor} = \boldsymbol{c} - \boldsymbol{b} \quad (7)$$

where $\boldsymbol{c}_{cor}$ is the corrected chromatogram without baseline drift. Fig. 2D shows the corrected chromatogram.

2.4. LMV-RSA for background drift in the LC-QTOF dataset

The employment of LMV-RSA for background drift in LC-QTOF is based on each chromatogram recorded in mass spectral channels. Fig. 3 provides an illustration. The LC-QTOF data set was transformed into the mzData.xml format and then organized into a matrix in MATLAB, where each row records a TOF spectrum at an elution point and each column contains a chromatogram under a mass-to-charge (m/z) ratio. The total ion chromatogram (TIC) of an LC-QTOF sample is the sum of all chromatograms in the data, Fig. 3A, which can be expressed as follows:

$$TIC = \sum_{i=1}^{I} \boldsymbol{c}_i \quad (8)$$

where $\boldsymbol{c}_i$ is the recorded chromatogram at the ith mass-to-charge ratio. The chromatograms of all mass spectral channels are depicted in Fig. 3B. LMV-RSA is employed to estimate the background drift, $\boldsymbol{c}_{bk,i}$, in each $\boldsymbol{c}_i$. Fig. 3D shows the estimated background drift under each mass spectral channel from LMV-RSA. Fig. 3E provides the corrected chromatogram, $\boldsymbol{c}_{cor,i}$, by subtracting the background drift from the original chromatograms. Finally, a new TIC is generated by summation of the modified chromatograms, $TIC_{cor} = \sum_{i=1}^{I} \boldsymbol{c}_{cor,i}$, and Fig. 3C plots the corrected TIC, where background drift is successfully removed.

2.5. MPLS

MPLS was proposed by Li et al. [25] to eliminate the influence of chromatographic peaks by introducing an additional morphological strategy for penalized least squares. Background drift was calculated according to the following equation:

$$\boldsymbol{z} = \left(\boldsymbol{W} + \lambda \boldsymbol{D}^T \boldsymbol{D}\right)^T \boldsymbol{W}\boldsymbol{c} \quad (9)$$

where $\boldsymbol{z}$ is the estimated background drift for the chromatogram $\boldsymbol{c}$, and $\mathbf{W}$ is a diagonal matrix with estimated weights on the diagonal line. In MPLS, the weights for the data points of the peaks are automatically set to zero by using a morphological strategy. $\mathbf{D}$ is the derivative matrix, and λ is the penalized parameter that

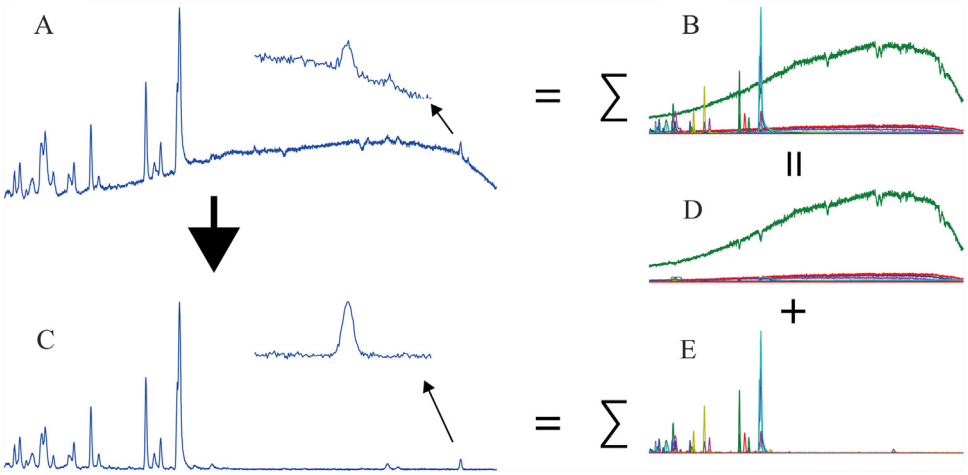

Fig. 3. LMV-RSA for background drift correction in LC-QTOF. The original TIC (A) is the sum of chromatograms recorded under each mass spectral channel (B). (C) TIC without background drift. Chromatograms in (B) can be regarded as the sum of the estimated baseline (D) and signal parts (E).

controls the degree of smoothness of the estimated background drift. In practical applications, the penalized parameter should be estimated before use. In this study, the penalized parameter was optimized as 10^5 for data analysis.

2.6. MWMV strategy

Yaroshchyk and Eberhardt [23] developed the MWMV method to address the baseline in the NMR data. In the present work, this technique was introduced for background drift in the chromatographic data for the first time. MWMV first replaced the central point with the minimum value in the window to obtain an initial baseline estimation. A convolution strategy based on rectangular operation was then employed to smoothen the initial baseline. The performance of MWMV mainly depends on the window size. A small window may eliminate chromatographic peak information, whereas a large window may not completely remove the influence of the background drift. In the current work, the window size was set to 60 elution channels. A detailed discussion of MWMV was provided in the study by Yaroshchyk and Eberhardt [23].

2.7. Background drift correction by orthogonal subspace projection (BD-OSP)

Theoretically, the response of a single sample obtained using LC-QTOF satisfies a bilinear structure under low m/z precision such as 0.1 and 1 Da:

$$\boldsymbol{X} = \boldsymbol{C}\boldsymbol{S}^T + \boldsymbol{E} \tag{10}$$

where **X** is the recorded response surface of a dataset, **C** is the pure chromatographic profile of the components in the sample, **S** is the pure spectral profile in the sample, and **E** is the instrumental noise not accounted by the model. BD-OSP first extracts the spectral profiles of the background ($\mathbf{S}_b$) and eliminates the background drift in the data by projecting the matrix **X** into an orthogonal subspace: $\mathbf{X}(\mathbf{I}–\mathbf{S}_b\mathbf{S}_b{}^+)$.

2.8. Simulation

The performance of our proposed method, LMV-RSA, was evaluated with respect to noise levels and elution conditions by simulation. Fig. 4A presents a typical chromatogram with complex background drift. Ten chromatographic peaks with various baseline separation conditions were provided in the chromatogram. Specifically, the first compound was successfully separated from the others. The second and third compounds co-eluted without baseline separation. The last two chromatographic peak clusters contain three and four peaks, respectively. Fig. 4A indicates that the chromatographic peaks in the last two chromatographic peak clusters seriously overlapped. Various noise levels were designed to investigate the performance of the proposed method.

$$noise = noiselevel \times \max(\boldsymbol{c}) \times randn \tag{9}$$

where the *noise level* employed in this work ranged from 0.5% to 2% of the maximum value of a chromatogram ***c***, and *randn* represents the Gaussian noise.

2.9. Data analysis

All data analysis was performed in MATLAB environment. A computer with Win7 (64-bit) Intel® Core™ i5CPU (2.8 GHz, 8 G RAM) was used. The LMV-RSA program and some simple examples were provided as a supplementary material and can also be obtained from the authors on request.

3. Experimental

3.1. GC samples

A total of 30 flavor samples were analyzed to monitor the changes in quality during the storage procedure. These samples were divided into three batches and stored in three cities in China, namely, Zhengzhou, Guangzhou, and Changchun. These cities are located in the north, middle, and south zones of China, with slightly different climatic conditions. These samples were analyzed monthly on an Agilent gas chromatography coupled with a flame ionization detector. A DB-5MS (30 m × 0.25 mm, 0.25 μm) chromatographic column was selected with a 1:15 split ratio of the injector at 280 °C. Helium carrier gas with a constant flow rate of 1.5 mL min^{-1} was used. The column temperature was maintained constant at 40 °C for 2 min and then increased to 250 °C at a rate of 6 °C min^{-1}. This temperature was maintained for 10 min. The hydrogen and air flows were 40 and 450 mL min^{-1}, respectively. A chromatogram containing more than 20,000 data points was finally collected for each sample.

3.2. LC-QTOF samples

Fifteen metabolic samples of *Escherichia coli* were analyzed on an Agilent 6540 LC-QTOF. An ultra-high pressure Agilent C_{18} column (200 mm × 4.6 mm, i.d., 1.7 μm particle size) was used

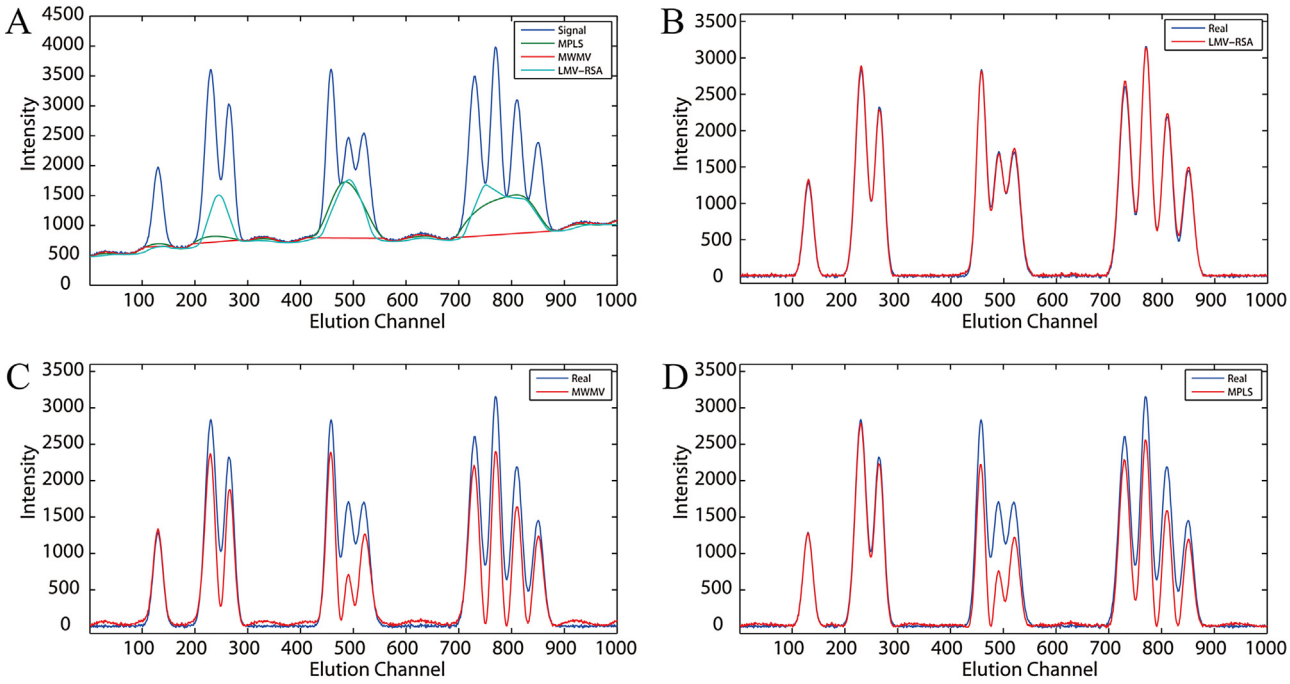

Fig. 4. (A) Simulated chromatogram with complex background drift and various elution conditions. (B), (C), and (D) Estimated baselines obtained using from LMV-RSA, MWMV, and MPLS.

for separation. The mobile phase was composed of (A) 1% aqueous formic acid and (B) methyl alcohol. A gradient elution was employed. The aqueous formic acid was initialized at 95% and then decreased to 5% within 20 min. The instrumental parameters for TOF were set as follows: gas temperature, 350 °C; gas flow, 12 L min^{-1}; nebulizer, 40 psi; sheath gas temperature, 350 °C; sheath gas flow, 10 L min^{-1}; Vcap, 3500 V; and mass range, 50–1000. Finally, the dataset recording more than 3600×1000 (elution channel × mass channel) data points was obtained for each sample. The resolution of each m/z channel is 0.006 min.

4. Results and discussion

The results and discussion are organized as follows. The influence of co-elution condition and noise level on LMV-RSA was investigated by simulation. The performance of LMV-RSA in various types of background drift in complex samples was evaluated using GC samples, together with the influence of the moving window width. By using the LC-QTOF dataset, the influence of m/z precision on background drift correction was discussed in more detail, and the differences between single chromatogram based background drift correction and matrix based background drift correction were identified. Comments and suggestions regarding the proposed method were provided.

4.1. Simulation results

Fig. 4A provides an example of a typical chromatogram with complex background drift and various degrees of overlapped chromatographic peaks. The first peak was successfully separated from the others, whose background drift can be readily estimated. However, with an increment in the complexity of overlapped peaks, such as three or four severely co-eluted peaks, background drift correction becomes a challenging task because local minimum values of chromatographic peaks are difficult to recognize from instrumental noise. Fig. 4B depicts the estimated chromatogram from LMV-RSA without background drift, together with the real one. The estimated linear background drift under overlapped peaks part is a little difference with the real one; nevertheless, LMV-RSA still obtains a high-quality solution. The correlation coefficient of the estimated chromatogram and the real one is larger than 0.9999, which is acceptable for practical applications.

Fig. 4C and D shows the results obtained using MWMV and MPLS. These two methods were influenced by overlapped peaks, and some information was eliminated after background drift correction. As LMV-RSA can be treated as a modification of MWMV, the comparison of the two methods could conclude that the performance of MWMV depends on the window width. By contrast, the influences of the window width can be greatly reduced using local minimum values instead of the original chromatogram. The performance of MPLS is affected by chromatographic peaks if they are not properly weighted. Chromatographic background drift usually changes gradually, which is certainly consistent with the principle of ALS based methods. However, the characteristics of chromatographic peaks that change rapidly within a short elution time may distort the estimated curves. Therefore, the performance of these methods depends on the elimination of chromatographic peaks. Based on our experience, background drift can be estimated using only a few data points across the entire chromatogram and a large percentage of data points can be weighted to zero. Therefore, the performance of ALS based methods can be improved when a heavily-weighted strategy is employed [14].

The influence of various noise levels was investigated. Table 1 summarizes the results obtained using LMV-RSA, MWMV, and MPLS in a statistical manner. Peak areas under various peak clusters were calculated after background drift correction and evaluated using the real ones. The values in Table 1 indicate that results obtained using LMV-RSA are acceptable and could be the best of the three methods overall.

4.2. Results of GC samples

Fig. 5A presents the chromatograms of 30 samples. The background drifts were quite different across samples. Moreover, the

Table 1
Statistical parameters used to evaluate the performance of LMV-RSA, MWMV and MPLS at varying noise levels.

noise level		LMV-RSA				MWMV				MPLS			
		p1[a]	p2	p3	p4	p1	p2	p3	p4	p1	p2	p3	p4
2.13%	mean[b]	**113%**[d]	**101.2%**	**82.4%**	**91.1%**	121.3%	85.7%	66.9%	77.0%	89.7%	65.5%	48.4%	70.6%
	std[c]	**6.2%**	**5.8%**	10.3%	10.0%	12.1%	6.8%	**5.7%**	**6.1%**	9.4%	15.8%	9.8%	9.2%
1.44%	mean	**110.1%**	**102.3%**	**96.2%**	**101.2%**	110.4%	79.1%	62.8%	72.5%	92.4%	85.2%	52.3%	73.2%
	std	**4.4%**	**2.8%**	**2.7%**	**2.9%**	6.6%	5.3%	5.1%	5.7%	7.3%	10.3%	8.7%	8.3%
1.06%	mean	107.3%	**101.2%**	**98.7%**	**102.4%**	**105.2%**	76.1%	61.1%	71.1%	93.9%	89.7%	52.7%	74.4%
	std	**3.3%**	**1.9%**	**1.5%**	**1.4%**	4.9%	5.0%	4.0%	5.5%	6.0%	7.9%	7.2%	7.2%
0.85%	mean	106.3%	**101.2%**	**99.9%**	**102.3%**	**103.1%**	74.6%	60.4%	71.2%	95.8%	93.5%	52.0%	74.2%
	std	**2.9%**	**1.3%**	**1.2%**	**1.0%**	4.5%	4.5%	4.0%	5.3%	4.1%	4.1%	8.0%	7.9%
0.69%	mean	105.1%	**101.3%**	**100.3%**	**102.2%**	**101.3%**	75.5%	60.7%	70.4%	96.9%	94.9%	53.3%	73.8%
	std	**2.0%**	**1.0%**	**0.9%**	**0.9%**	3.7%	4.4%	3.9%	4.9%	3.2%	4.5%	7.4%	7.8%

[a] p1 is the first peak eluted around the 110th elution channel; p2 is two overlapped peaks; p3 is the overlapped peak of three chromatographic peaks; p4 is the overlapped peak of four peaks.
[b] Represent mean value of recovery under 10,000 repeats.
[c] Represent standard derivation of recovery under 10,000 repeats.
[d] Bold values represent the best results.

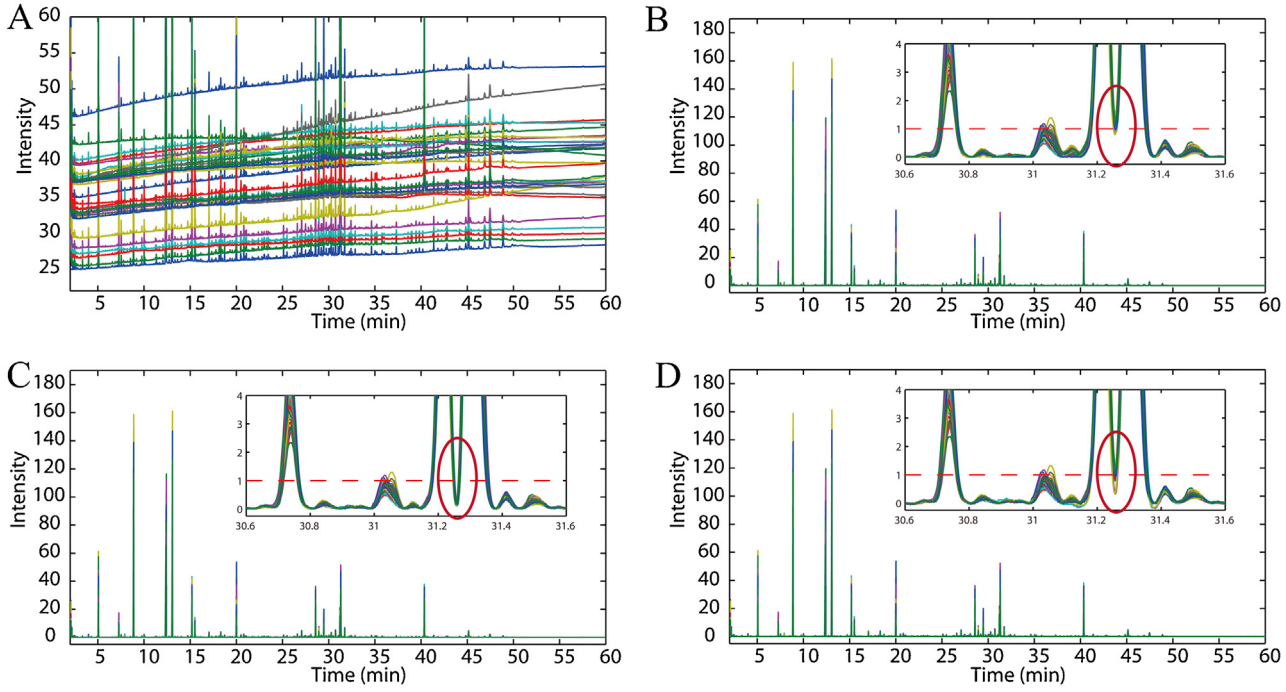

Fig. 5. (A) Original chromatogram of the 30 flavored samples. Results obtained using LMV-RSA (B), MWMV (C), and MPLS (D). Inserted plots show the selected elution range in greater detail. Circles marked with useful information are eliminated using MPLS and MWMV.

background significantly changed within a single sample. The profiles shown in Fig. 5A can be regarded as typical complex examples in quality control and metabolic profiling analysis with hundreds of chromatographic peaks present in a single chromatogram, rendering background drift correction difficult.

The profiles in Fig. 5B indicate that various kinds of background drift are eliminated using LMV-RSA, and chromatographic peaks can be visualized clearly. Corrected chromatograms are more suitable for downstream data analysis, such as peak extraction and time shift alignment, which are frequently used in quality control and metabolic profiling analysis. The results generated using MWMV and MPLS are presented in Fig. 5C and D, respectively, revealing satisfactory solutions. The inserted plots in Fig. 5B–D suggest that background drift correction from LMV-RSA is more acceptable than those of the other two methods; the information provide by overlapped chromatographic peaks can be maintained in LMV-RSA, which is consistent with those of the simulation, Fig. 4.

LMV-RSA is almost an automatic method; regardless, a parameter should be still pre-estimated, i.e. moving window width. The simulation results suggest that the influence of the window width can be greatly reduced by replacing the original signal with local minimum values. Fig. 6A shows estimated chromatograms from LMV-RSA with various window widths, where correlation coefficients are almost equal to 1.000, implying that LMV-RSA is insensitive to window width. If the analyzed chromatogram has a low SNR, the results of LMV-RSA are slightly changed with respect to the window width. Fig. 6B provides the background drift correction results from LMV-RSA when the chromatogram has a lower SNR. The illustration suggests that resolved chromatograms slightly vary under various moving window widths. However, these differences are acceptable in routine data analysis.

The effects of window width on MWMV are summarized in Fig. 6C and D. Comparison of Fig. 6B and D confirms that LMV-RSA is more robust to window width compared with MWMV. The results shown in Fig. 6B also support replacing the original signal with

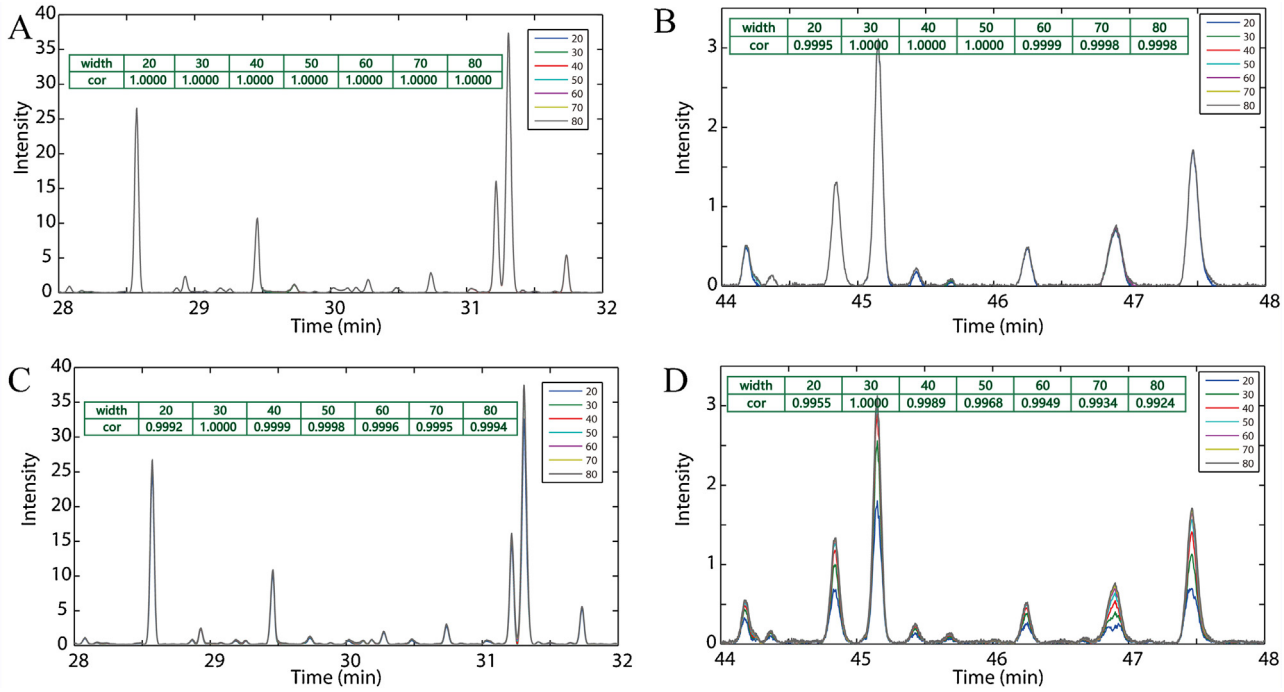

Fig. 6. Influence of window width on LMV-RSA and MWMV. (A) Results obtained using LMV-RSA for chromatogram with high signal-to-noise ratio. (B) Results obtained using LMV-RSA for chromatogram with low signal-to-noise ratio. (C) Results obtained using MWMV for chromatogram with high signal-to-noise ratio. (D) Results obtained using MWMV for chromatogram with low signal-to-noise ratio. The correlation coefficient (*cor*) is calculated as follows: $cor = \left(\boldsymbol{x}-\hat{\boldsymbol{x}}\right)^T\left(\boldsymbol{y}-\hat{\boldsymbol{y}}\right) / \sqrt{\left(\boldsymbol{x}-\hat{\boldsymbol{x}}\right)^T\left(\boldsymbol{x}-\hat{\boldsymbol{x}}\right)\left(\boldsymbol{y}-\hat{\boldsymbol{y}}\right)^T\left(\boldsymbol{y}-\hat{\boldsymbol{y}}\right)}$, where $\boldsymbol{x}$ and $\boldsymbol{y}$ are two chromatograms, and $\hat{\boldsymbol{x}}$ and $\hat{\boldsymbol{y}}$ are the averages of $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively. Legends in figures represent the moving window width.

the local minimum vector. The results presented in Fig. 6 suggest that the influence of the window width is not a critical parameter to LMV-RSA. Our investigation of LMV-RSA with various complex chromatographic datasets suggested that a window width of 30 would be suitable for most situations.

Simulation and real GC samples were conducted to verify the performance of our proposed method. The results indicated that LMV-RSA can provide satisfactory background drift correction. In the subsequent LC-QTOF data analysis, only LMV-RSA was used for background drift correction, and the retrieved solutions were compared with those obtained using BD-OSP.

4.3. Results of LC-QTOF samples

The TIC obtained from LC-QTOF is the sum of the chromatograms recorded under each mass-to-charge ratio. The background drift may be caused by two factors: the underlying baseline and the overlapped chromatographic peaks. If background drift correction is forcedly implemented on the TIC mode, some useful information from overlapped components may also be eliminated. Therefore, performing the background drift correction on each mass spectral channel is necessary.

Data analysis using chromatography coupled with high-resolution mass spectrometry is presented with the challenge of data analysis efficiency. The *m*/*z* precision of chromatographic peaks from TOF is known to be as high as 0.0001 Da, which indicates that, theoretically, more than one million chromatograms generated for a single sample if the mass spectra varied from 1 to 1000. However, most of these chromatograms contain useless information. If a chromatogram with *m*/*z* precision of 0.0001 Da is strictly extracted, the chromatographic peak shape in TIC may not be maintained.

Fig. 7A shows a typical chromatogram of *m*/*z* 279, which is the sum of chromatograms recorded between *m*/*z* of 278.5001 and 279.4999. Background drift can be evidently observed in Fig. 7A, together with two significant peaks. If the chromatogram is split into 0.1 Da *m*/*z* precision, the chromatogram in Fig. 7A can be divided into two separate chromatograms with meaningful information i.e. *m*/*z* 279.1 and 279.2 whereas the others contain useless information. However, background drift is still present in two recorded chromatograms. Fig. 7C and D reveal chromatograms with *m*/*z* precision of 0.01 Da and 0.001 Da, respectively, in which background drift cannot be eliminated with high *m*/*z* precision. By contrast, part of the useful information is lost under. Fig. 7D suggests that the shape of the chromatographic peak eluted about 6 min becomes a number of dispersed vertical lines, and the traditional peak shape of the continuous curve is no longer maintained. The profiles in Fig. 7 indicate that peak detection is not necessarily performed based on high *m*/*z* precision. In fact, chromatographic peaks are detected in chromatograms with low *m*/*z* precision, such as, 1 and 0.1 Da, and then precisely marked with high-precision *m*/*z*, such as 0.0001 Da, according to their retention time. In conclusion, a data compression strategy is necessary to reduce the computational task by rounding off the mass-to-charge ratio.

The background drift correction results with an *m*/*z* precision of 1 Da are plotted in Fig. 8A, whereas those with *m*/*z* precision of 0.1, 0.01, and 0.001 Da are plotted in Figs. 8B, 8C, and 8D, respectively. The results clearly indicate that background drift has been satisfactorily corrected. Moreover, the profiles in Fig. 8A–C are almost identical, but not that in Fig. 8D. Although background drifts with 0.001 Da *m*/*z* precision are reasonably removed, the profiles in Fig. 8D differ from the others because some useful data are lost. Fig. 8E shows the original chromatograms with various *m*/*z* precision. Notably, chromatograms with *m*/*z* precision of 0.1, 0.01, and 0.001 Da in Fig. 8E are the sum of those in Fig. 7B–D, respectively. Fig. 8F shows the corresponding chromatograms without background drift in which only chromatographic peaks can be observed obviously. Profiles in Fig. 8E and F indicate that the results from data

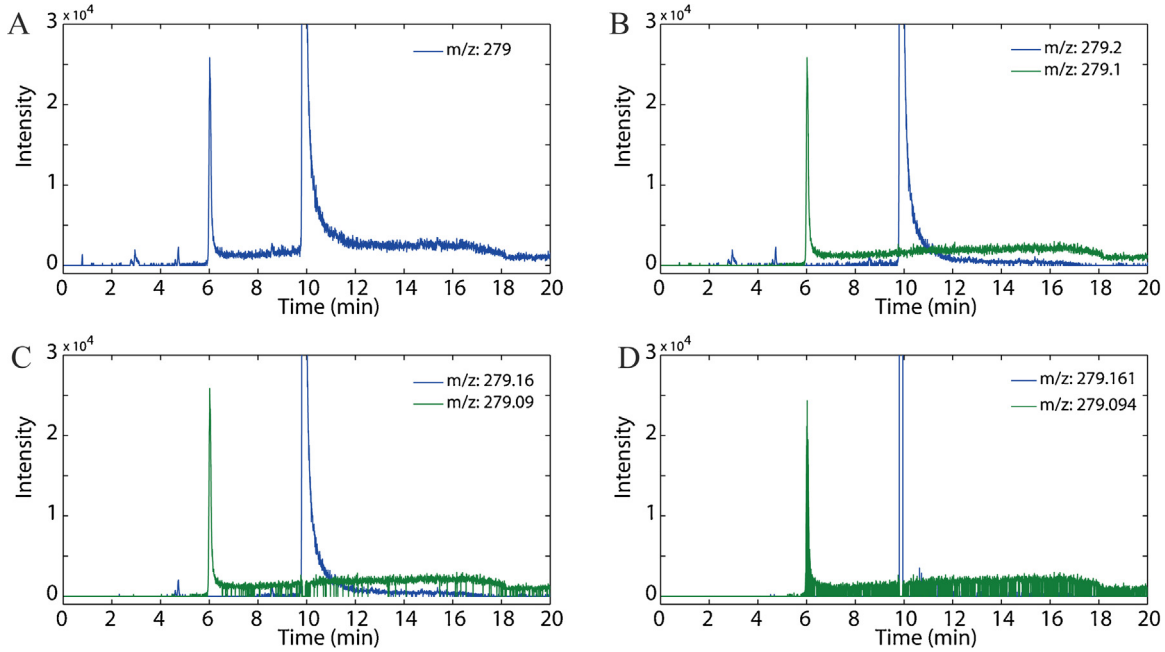

Fig. 7. (A) Original chromatogram with 1 Da *m*/*z* precision. (B) Chromatograms with 0.1 *m*/*z* Da precision. (C) Original chromatograms with 0.01 *m*/*z* Da precision. (D) Original chromatograms with 0.001 Da *m*/*z* precision. *m*/*z* 279 means MS range from 278.5 to 279.5; *m*/*z* 279.2 means MS range from 279.15 to 279.25; *m*/*z* 279.1 means MS range from 279.05 to 279.15; *m*/*z* 279.16 means MS range from 279.155 to 279.165; *m*/*z* 279.09 means MS range from 279.085 to 279.095; *m*/*z* 279.161 means MS range from 279.1605 to 279.1615; *m*/*z* 279.094 means MS range from 279.0935 to 279.0945.

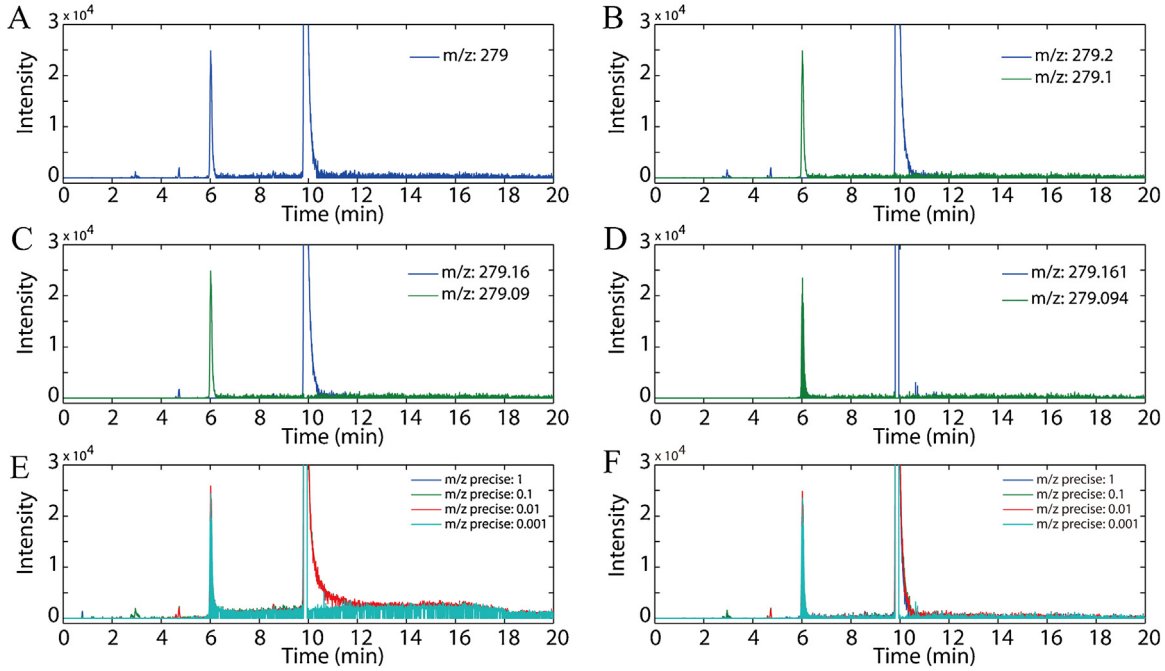

Fig. 8. (A) Chromatograms with a 1 Da *m*/*z* precision after background drift correction by LMV-RSA. (B) Chromatograms with 0.1 Da *m*/*z* precision after background drift correction by LMV-RSA. (C) Chromatograms with 0.01 Da *m*/*z* precision after background drift corrected by LMV-RSA. (D) Chromatograms with 0.001 Da *m*/*z* precision after background drift correction by LMV-RSA. (E) Total ion current chromatogram with various *m*/*z* precision and (D) results obtained using LMV-RSA. *m*/*z* 279 means MS range from 278.5 to 279.5; *m*/*z* 279.2 means MS range from 279.15 to 279.25; *m*/*z* 279.1 means MS range from 279.05 to 279.15; *m*/*z* 279.16 means MS range from 279.155 to 279.165; *m*/*z* 279.09 means MS range from 279.085 to 279.095; *m*/*z* 279.161 means MS range from 279.1605 to 279.1615; *m*/*z* 279.094 means MS range from 279.0935 to 279.0945.

compression with *m*/*z* precision of 1 and 0.1 Da are almost identical. On the basis of the results shown in Fig. 7 and 8, we employed the *m*/*z* precision of 1 Da, in the present study, to maximally improve the data analysis efficiency. One should keep awareness that LMC-RSA can be also used for data with *m*/*z* precision of 0.1 Da, if necessary.

The 15 metabolic chromatograms of the *E. coli* samples are presented in Fig. 9A. The background drifts in the LC-QTOF samples slightly differ from those in the GC samples in Fig. 5A.

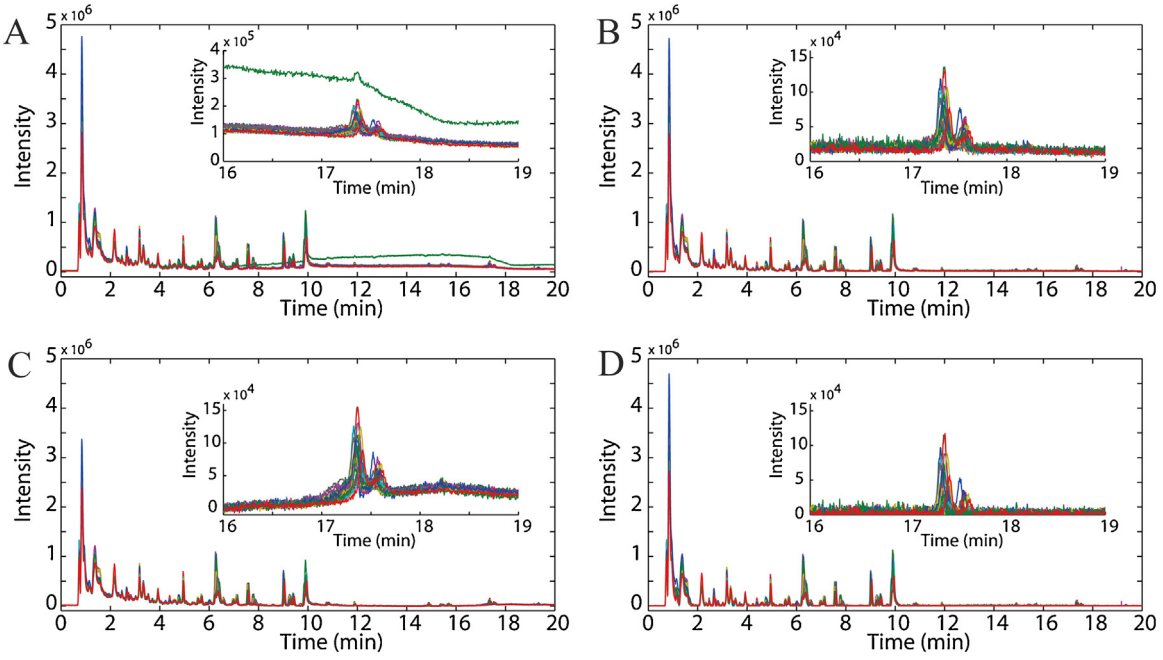

Fig. 9. Comparison of LMV-RSA with BD-OSP. (A) Original TICs of 15 samples. (B) Results obtained using LMV-RSA. (C) Results obtained using BD-OSP. (D) Results obtained using LMV-RSA on TICs directly. Inserted plots present the results of background drift correction results in greater detail.

The background drifts obtained from LC-QTOF are almost identical across samples except for one simple. Fig. 9B presents the chromatograms with background drifts satisfactorily corrected by LMV-RSA. Based on the bilinear structure of each sample, the background drifts corrected using BD-OSP were implemented in this work for comparison. The background spectra were initially obtained by performing singular value decomposition on elution ranges without components i.e., between 13 and 14 min. The chromatograms with background drift correction can be obtained by projecting the response matrix of each sample into the orthogonal subspace spanned by the background spectra. Fig. 9C depicts the chromatograms with background drift correction using BD-OSP. However, the baselines are still present.

The results of background drift correction based on the implementation of LMV-RSA on the original TICs of 15 samples are plotted in Fig. 9D. Although the background drifts across the samples are removed, a part of the useful information eluted around 2 min is eliminated. Comparison of Fig. 9B–D indicates that among the three methods, LMV-RSA obtained the optimal results.

The fundamental principal of LMV-RSA is based on local minimum values in the recorded chromatogram. In certain extreme cases with no local minima, the use of LMV-RSA is not recommended. By contrast, MPLS and MWMV remain applicable. Fortunately, the requirement of LMV-RSA that local minimum values caused by instrumental noise be scattered across the entire chromatogram can be satisfied for most applications. Efficiency of data analysis is another important parameter when large datasets are analyzed. Our study indicated that LMV-RSA requires no more than 15 s (12.13 s for MWMV and 14.92 s for MPLS) to analyze the GC chromatogram dataset containing 23,500 × 30 data points(elution channel × samples), which is acceptable for practical application.

Complex sample analysis involves a large amount of overlapped chromatographic peaks. If a large percentage of seriously overlapped chromatographic peaks are preset, background drift correction should be performed carefully. Additionally, although background drift correction is indeed useful for data analysis, several pseudo peaks can also emerge [6], which should be considered in practical applications.

The results obtained using simulated and real GC chromatographic datasets indicated that our proposed method, LMV-RSA, can provide a satisfactory solution for background drift. The performance of LMV-RSA is comparable with those of the two classical methods MWMV and MPLS. In addition, the results obtained using the LC-QTOF datasets indicated that employing LMV-RSA to eliminate background drift is applicable for complex sample analysis based on hyphenated chromatographic instruments. A significant advantage of LMV-RSA is that background drift correction can be almost automatically implemented, which is convenient for downstream data analysis. Based on background drift correction results, chromatographic peak extraction can be performed readily, which is under way in our laboratory.

5. Conclusion

An automatic method for background drift correction in chromatographic data analysis was developed in this work, which is a modification of classical MWMV. The performance of the proposed method was confirmed using two extremely complex chromatographic datasets that involve a GC dataset that monitors quality changes in the flavored sample and a LC-QTOF dataset used for metabolic profiling analysis. The results indicate that the proposed method is comparable with the three classical methods employed for chromatographic data analysis.

Acknowledgments

The authors would like to acknowledge the financial support of the Foundation of ZTRI (Grant No. 322013CA0290) and the National Natural Science Foundation of China (No. 21205145, 21576297, 21305160).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.chroma.2016.04.054.

References

[1] J.M. Amigo, T. Skov, R. Bro, ChroMATHography: solving chromatographic issues with mathematical models and intuitive graphics, Chem. Rev. 110 (2010) 4582–4605.
[2] M. Goodarzi, P.J. Russell, Y. Vander Heyden, Similarity analyses of chromatographic herbal fingerprints: a review, Anal. Chim. Acta 804 (2013) 16–28.
[3] N.L. Kuehnbaum, P. Britz-McKibbin, New advances in separation science for metabolomics: resolving chemical diversity in a post-genomic era, Chem. Rev. 113 (2013) 2437–2468.
[4] X.-G. Liu, S.-Q. Wu, P. Li, H. Yang, Advancement in the chemical analysis and quality control of flavonoid in Ginkgo biloba, J. Pharm. Biomed. Anal. 113 (2015) 212–225.
[5] H. Parastar, R. Tauler, Multivariate curve resolution of hyphenated and multidimensional chromatographic measurements: a new insight to address current chromatographic challenges, Anal. Chem. 86 (2014) 286–297.
[6] Y.-J. Yu, Q.-L. Xia, S. Wang, B. Wang, F.-W. Xie, X.-B. Zhang, Y.-M. Ma, H.-L. Wu, Chemometric strategy for automatic chromatographic peak detection and background drift correction in chromatographic data, J. Chromatogr. A 1359 (2014) 262–270.
[7] H.P. Benton, J. Ivanisevic, N.G. Mahieu, M.E. Kurczy, C.H. Johnson, L. Franco, D. Rinehart, E. Valentine, H. Gowda, B.K. Ubhi, R. Tautenhahn, A. Gieschen, M.W. Fields, G.J. Patti, G. Siuzdak, Autonomous metabolomics for rapid metabolite identification in global profiling, Anal. Chem. 87 (2015) 884–891.
[8] W.R. French, L.J. Zimmerman, B. Schilling, B.W. Gibson, C.A. Miller, R.R. Townsend, S.D. Sherrod, C.R. Goodwin, J.A. McLean, D.L. Tabb, Wavelet-Based peak detection and a new charge inference procedure for MS/MS implemented in ProteoWizard's msConvert, J.Proteome Res. 14 (2015) 1299–1307.
[9] A. Nordström, G. O'Maille, C. Qin, G. Siuzdak, Nonlinear data alignment for UPLC-MS and HPLC-MS based metabolomics: quantitative analysis of endogenous and exogenous metabolites in human serum, Anal. Chem. 78 (2006) 3289–3295.
[10] R. Tautenhahn, G.J. Patti, D. Rinehart, G. Siuzdak, XCMS online: a web-Based platform to process untargeted metabolomic data, Anal. Chem. 84 (2012) 5035–5039.
[11] Y.-J. Yu, H.-Y. Fu, L. Zhang, X.-Y. Wang, P.-J. Sun, X.-B. Zhang, F.-W. Xie, A chemometric-assisted method based on gas chromatography–mass spectrometry for metabolic profiling analysis, J. Chromatogr. A 1399 (2015) 65–73.
[12] S. Golotvin, A. Williams, Improved baseline recognition and modeling of FT NMR spectra, J. Magn. Reson. 146 (2000) 122–125.
[13] Q. Bao, J. Feng, F. Chen, W. Mao, Z. Liu, K. Liu, C. Liu, A new automatic baseline correction method based on iterative method, J. Magn. Reson. 218 (2012) 35–43.
[14] S.-J. Baek, A. Park, Y.-J. Ahn, J. Choo, Baseline correction using asymmetrically reweighted penalized least squares smoothing, Analyst 140 (2015) 250–257.
[15] A.F.P. Ferreira, H.F.M. Boelens, J.A. Westerhuis, Inline monitoring of butane isomers adsorption on MFI using near-Infrared spectroscopy: drift correction in time based experiments, Langmuir 21 (2005) 6830–6835.
[16] J. Kuligowski, D. Carrión, G. Quintás, S. Garrigues, M. de la Guardia, Cubic smoothing splines background correction in on-line liquid chromatography–Fourier transform infrared spectrometry, J. Chromatogr. A 1217 (2010) 6733–6741.
[17] J. Kuligowski, G. Quintás, S. Garrigues, M. de la Guardia, New background correction approach based on polynomial regressions for on-line liquid chromatography–Fourier transform infrared spectrometry, J. Chromatogr. A 1216 (2009) 3122–3130.
[18] X. Liu, Z. Zhang, Y. Liang, P.F.M. Sousa, Y. Yun, L. Yu, Baseline correction of high resolution spectral profile data based on exponential smoothing, Chemom. Intell. Lab. Syst. 139 (2014) 97–108.
[19] Y. Liu, W. Cai, X. Shao, Intelligent background correction using an adaptive lifting wavelet, Chemom. Intell. Lab. Syst. 125 (2013) 11–17.
[20] M. Mecozzi, A polynomial curve fitting method for baseline drift correction in the chromatographic analysis of hydrocarbons in environmental samples, APCBEE Proc. 10 (2014) 2–6.
[21] B.D. Prakash, Y.C. Wei, A fully automated iterative moving averaging (AIMA) technique for baseline correction, Analyst 136 (2011) 3130–3135.
[22] L. Shao, P.R. Griffiths, Automatic baseline correction by wavelet transform for quantitative open-path fourier transform infrared spectroscopy, Environ. Sci. Technol. 41 (2007) 7054–7059.
[23] P. Yaroshchyk, J.E. Eberhardt, Automatic correction of continuum background in laser-induced Breakdown Spectroscopy using a model-free algorithm, Spectrochim. Acta Part B 99 (2014) 138–149.
[24] Z.M. Zhang, S. Chen, Y.Z. Liang, Baseline correction using adaptive iteratively reweighted penalized least squares, Analyst 135 (2010) 1138–1146.
[25] 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.
[26] M. Friedrichs, A model-free algorithm for the removal of baseline artifacts, J. Biomol. NMR 5 (1995) 147–153.
[27] P.H.C. Eilers, A perfect smoother, Anal. Chem. 75 (2003) 3631–3636.
[28] J.J. de Rooi, O. Devos, M. Sliwa, C. Ruckebusch, P.H.C. Eilers, Mixture models for two-dimensional baseline correction, applied to artifact elimination in time-resolved spectroscopy, Anal. Chinm. Acta 771 (2013) 7–13.
[29] Y. Liu, X. Zhou, Y. Yu, A concise iterative method using the Bezier technique for baseline construction, Analyst 140 (2015) 7984–7996.
[30] L.G. Johnsen, T. Skov, U. Houlberg, R. Bro, An automated method for baseline correction, peak finding and peak grouping in chromatographic data, Analyst 138 (2013) 3502–3511.
[31] R.C. Allen, M.G. John, S.C. Rutan, M.R. Filgueira, P.W. Carr, Effect of background correction on peak detection and quantification in online comprehensive two-dimensional liquid chromatography using diode array detection, J. Chromatogr. A 1254 (2012) 51–61.
[32] 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.
[33] P. de, B. Harrington, J. Kister, J. Artaud, N. Dupuy, Automated principal component-Based orthogonal signal correction applied to fused near infrared-Mid-Infrared spectra of french olive oils, Anal. Chem. 81 (2009) 7160–7169.
[34] 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.
[35] J. Kuligowski, G. Quintas, R. Tauler, B. Lendl, M. de la Guardia, Background correction and multivariate curve resolution of online liquid chromatography with infrared spectrometric detection, Anal. Chem. 83 (2011) 4855–4862.
[36] 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.
[37] Z. Wang, M. Zhang, P.D.B. Harrington, Comparison of three algorithms for the baseline correction of hyphenated data objects, Anal. Chem. 86 (2014) 9050–9057.
[38] Z. Xu, X. Sun, Pde B. Harrington, Baseline correction method using an orthogonal basis for gas chromatography/mass spectrometry data, Anal. Chem. 83 (2011) 7464–7471.
[39] Y.-J. Yu, H.-L. Wu, H.-Y. Fu, J. Zhao, Y.-N. Li, S.-F. Li, C. Kang, R.-Q. Yu, Chromatographic background drift correction coupled with parallel factor analysis to resolve coelution problems in three-dimensional chromatographic data: quantification of eleven antibiotics in tap water samples by high-performance liquid chromatography coupled with a diode array detector, J. Chromatogr. A 1302 (2013) 72–80.
[40] H. Zhang, M. Grubb, W. Wu, J. Josephs, W.G. Humphreys, Algorithm for thorough background subtraction of high-Resolution LC/MS data: application to obtain clean product ion spectra from nonselective collision-Induced dissociation experiments, Anal. Chem. 81 (2009) 2695–2700.
[41] P. Gemperline, Practical Guide to Chemometrics, second ed., Taylor & Francis Group, New York, 2006.