Journal of Chromatography A, 1359 (2014) 262–270

Contents lists available at ScienceDirect

Journal of Chromatography A

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

Chemometric strategy for automatic chromatographic peak detection and background drift correction in chromatographic data

Yong-Jie Yu[a,*], Qiao-Ling Xia[a], Sheng Wang[a], Bing Wang[a], Fu-Wei Xie[a], Xiao-Bing Zhang[a], Yun-Ming Ma[b], Hai-Long Wu[c,*]

[a] Zhengzhou Tobacco Research Institute of CNTC, Zhengzhou 450001, China
[b] Leiyang Tobacco Branch of Hengyang Tobacco Co., Leiyang 421800, China
[c] State Key Laboratory of Chemo/Biosensing and Chemometrics, College of Chemistry and Chemical Engineering, Hunan University, Changsha 410082, China

ARTICLE INFO

Article history:
Received 20 April 2014
Received in revised form 8 July 2014
Accepted 16 July 2014
Available online 29 July 2014

Keywords:
Automatic chromatographic peak detection
Background drift correction
Chemometrics

ABSTRACT

Peak detection and background drift correction (BDC) are the key stages in using chemometric methods to analyze chromatographic fingerprints of complex samples. This study developed a novel chemometric strategy for simultaneous automatic chromatographic peak detection and BDC. A robust statistical method was used for intelligent estimation of instrumental noise level coupled with first-order derivative of chromatographic signal to automatically extract chromatographic peaks in the data. A local curve-fitting strategy was then employed for BDC. Simulated and real liquid chromatographic data were designed with various kinds of background drift and degree of overlapped chromatographic peaks to verify the performance of the proposed strategy. The underlying chromatographic peaks can be automatically detected and reasonably integrated by this strategy. Meanwhile, chromatograms with BDC can be precisely obtained. The proposed method was used to analyze a complex gas chromatography dataset that monitored quality changes in plant extracts during storage procedure.

1. Introduction

Chromatographic fingerprints have been widely used in a wide range of scientific fields, such as metabonomics studies and traditional Chinese medicine analysis [1–5]. Analyzing complex samples, including botanical extracts and urine samples, is a challenging and time-consuming task because hundreds of peaks may present in a chromatographic profile with insufficient chromatographic resolution [6–8]. Therefore, methods that can automatically process chromatographic signals to extract useful information are urgently required in practical applications. Data analysis procedure mainly consists of four stages, namely, peak detection, background drift correction (BDC), time shift alignment, and data mining. The current study focused on the first two stages of chromatographic data analysis.

Peak detection and BDC are key steps in chromatographic data analysis. Inaccurate peak detection or background correction will certainly influence the final data mining results, including classification, ANOVA, and biomarker identification. Therefore, a number of methods capable of automatically detecting chromatographic peaks and removing background drift have been developed [9–28]. Torres-Lapasió and colleagues [13,14] proposed an automatic chromatographic peak detection (ACPD) and deconvolution method. Brereton's research group [15,16] developed automated peak detection and matching algorithms to analyze gas chromatography (GC)–mass spectrometry datasets. Liang et al. [17–20] proposed a series of algorithms for peak detection and baseline correction in 1D chromatographic data. Du's research group [21,22] developed an automated data analysis pipeline for metabonomics studies.

Many studies have demonstrated the successful applications of the aforementioned methods for chromatographic peak detection and BDC in chromatographic data analysis. However, in practical applications, several trivial parameters should accurately be preestimated for these methods. These methods include the threshold value for the first-order derivative of a chromatogram [13], the proper selection of 'mother' wavelet, and the wavelet transform levels [17,25–28], which may be inconvenient for users in analyzing large-scale samples. In an extremely complex sample analysis, accurate detection of peaks and removal of background drift are still very challenging when multiple components co-elute [15]. The major limitation of the aforementioned methods is that a single method only focuses on a specific purpose in the analysis

* Corresponding authors. Tel.: +86 731 88821818.
E-mail addresses: yongjie.yu@163.com (Y.-J. Yu), hlwu@hnu.edu.cn, hlwu529@gmail.com (H.-L. Wu).

http://dx.doi.org/10.1016/j.chroma.2014.07.053
0021-9673/

procedure. Most of the existing BDC methods do not provide valuable chromatographic peak information and current peak detection methods may not provide useful chromatograms with BDC. Considering peak detection and background correction simultaneously is more useful in practical applications. Therefore, developing new methods that automatically and accurately detect chromatographic peaks and efficiently perform BDC is important.

This study developed a new chemometric strategy for simultaneous automatic chromatographic peak detection and background drift correction (ACPD-BDC) for chromatographic data analysis. Simulation and real liquid chromatographic data were used to verify the performance of this new approach. Results indicated that ACPD-BDC can accurately provide peak information, including the start and end positions, retention time, peak area, and peak height, as well as a reasonable BDC. The results of the background correction from the proposed method were comparable with the adaptive iteratively reweighted penalized least squares (airpLS) [18]. ACPD-BDC was also used to analyze complex GC data to monitor the quality changes of plant extracts during the storage procedure.

2. Theory

2.1. ACPD-BDC

Fig. 1 shows the ACPD-BDC workflow. This new strategy mainly consisted of two main stages: ACPD and BDC. A data smoothing stage was previously performed to smooth the signals with low signal-to-noise ratio. However, data smoothing was not a circuital step because of the strategy of this new method (see below). Our investigation showed that three-point moving-window averaging was suitable for most cases. The following subsections will discuss automatic peak detection and background correction in detail.

2.1.1. ACPD stage

2.1.1.1. Preliminary chromatographic peak estimation. Preliminary peak estimation was performed. Signals that continuously increased/decreased were considered as evidence of components. Therefore, the peak start and end positions were independently estimated in ACPD according to Eqs. (1) and (2), respectively.

$$x_i < x_{i+1} < x_{i+2} < x_{i+3} \tag{1}$$

$$x_j > x_{j+1} > x_{j+2} > x_{j+3} \tag{2}$$

where x_i and x_j are the recorded signals at the ith and jth elution channels, respectively. Finally, two vectors that indicated the peak start and end positions can be obtained.

$$\boldsymbol{a} = [\, a_1 \quad a_2 \quad \dots \quad a_p \,] \tag{3}$$

$$\boldsymbol{b} = [\, b_1 \quad b_2 \quad \dots \quad b_q \,] \tag{4}$$

where $\boldsymbol{a}$ is a vector whose elements represent the peak start positions and p is the number of estimated start peaks, whereas $\boldsymbol{b}$ is the end vector whose elements denote the peak end positions and q is the number of estimated end peaks. Notably, p and q values need not be equal to each other. The elution ranges of the peaks can be readily obtained when $\boldsymbol{a}$ and $\boldsymbol{b}$ are considered together. For instance, the combination $[a_m\ b_n]$ is regarded as a chromatographic peak elution range as long as the following equation is satisfied:

$$b_{n-1} < a_m < b_n < a_{m+1} \tag{5}$$

2.1.1.2. Automatic instrumental noise level estimation. The instrumental noise level was automatically calculated in a robust statistical manner. A new signal $\boldsymbol{x}_{new}$ was initially obtained by removing previously estimated chromatographic peaks in the original signal ($\boldsymbol{x}_0$), and the first-order derivative $d\boldsymbol{x}_{new}$ was then calculated. Afterward, the outliers of $d\boldsymbol{x}_{new}$ were removed in an iterative manner ($d\boldsymbol{x}_{new}$ was updated in each iteration). These outliers were estimated according to the following formula:

$$\frac{|d_i - \overline{d\boldsymbol{x}_{new}}|}{\sigma} > 3 \tag{6}$$

where d_i is the ith element of $d\boldsymbol{x}_{new}$, $\overline{d\boldsymbol{x}_{new}}$ represents the mean value of $d\boldsymbol{x}_{new}$, σ denotes the corresponding standard derivation, and $|\cdot|$ represents the absolute value. When the iteration converged, the value of 3σ can be used as instrumental noise level and as the threshold value of the first-order derivative in Fig. 1.

2.1.1.3. Pseudo-peak elimination. The number of estimated peaks will be more than the underlying ones. Thus, pseudo-peak elimination was performed to validate whether a peak was true or simply an artifact after the instrumental noise level estimation. The following two criteria should be satisfied to be considered as a valid peak:

1. The number of times in which the absolute value of the first-order derivative is larger than the threshold value is counted. When this number is less than 5, the peak will be removed. (Note that the parameter of 5 is an experience value representing the number of recorded points that can be accurately quantified in a chromatographic peak.)
2. The number of times in which the second-order derivative crossed zero lines is counted. When this number is larger than 8 and the signal-to-noise ratio is less than 10σ, the peak is rejected. (Note that a chromatographic peak with second-order derivative crossed zero lines larger than 8 times and signal-to-noise ratio less than 10σ can usually be treated as a very noisy peak.)

2.1.1.4. Chromatographic peak clustering and peak baseline estimation. Clustering peaks with end positions are close to the start position of the following chromatographic peak possibly provide more accurate results for peak integration (no more than five data points for this application). The baseline under a clustering peak was estimated by linearly interpolating the values between the start and end points of the peak. Finally, the peak area and the corresponding average signal-to-noise ratio (ASNR) were obtained as follows:

$$\text{Area} = \sum_{i=p_s}^{p_s} (x_i - l_i) \tag{7}$$

$$\text{ASNR} = \frac{\text{Area}}{p_e - p_s + 1} \tag{8}$$

where x_i represents the recorded signal, l_i is the corresponding linear baseline, and p_s and p_e denote the start and end positions of the peak, respectively.

2.1.1.5. Small peak elimination. In practical applications such as metabonomics studies, small peaks that cannot be accurately quantified are removed in the succeeding analysis. Therefore, this study adopted a small peak elimination stage. Peaks with ASNR smaller than 10-fold of the instrumental noise level will be eliminated. Peak information, including start and end positions, as well as the areas and the heights of peaks, can be potentially used for quantitative analysis.

A significant advantage of this new chromatographic detection strategy is that only the analyzed signal is required to pick up valuable chromatographic peaks, which will be convenient for the analysis of large-scale samples in practical applications, such as quality control of herbal medicine and metabonomics studies.