Journal of Chromatography A, 1407 (2015) 203–207

Contents lists available at ScienceDirect

Journal of Chromatography A

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

Generalized window factor analysis for selective analysis of the target component in real samples with complex matrices

Pao Li, Wensheng Cai, Xueguang Shao*

Research Center for Analytical Sciences, College of Chemistry, Nankai University, Tianjin Key Laboratory of Biosensing and Molecular Recognition, State Key Laboratory of Medicinal Chemical Biology, Collaborative Innovation Center of Chemical Science and Engineering (Tianjin), Tianjin 300071, China

ARTICLE INFO

Article history:
Received 16 April 2015
Received in revised form 21 June 2015
Accepted 22 June 2015
Available online 30 June 2015

Keywords:
Gas chromatography–mass spectrometry
Quantitative determination
Generalized window factor analysis
Complex matrices
Peak shift

ABSTRACT

In chromatographic analysis of multicomponent real samples, peak overlapping, high level of noise and background are frequently encountered, making the qualitative and quantitative analysis difficult or even impossible. In this work, an algorithm named as generalized window factor analysis (GWFA) was proposed for quantitative analysis of the target components in the samples with complex matrices by gas chromatography–mass spectrometry (GC–MS). The theory and calculation of GWFA are just similar with the conventional window factor analysis (WFA), but the "window" is defined as the selected channels (mass-to-charge ratios) in the mass spectral dimension of the data matrix, instead of a continuous region in chromatographic dimension along the retention time. Therefore, the generalized window for a target component can be easily determined with the help of the mass spectrum. Then, the calculated mass spectrum can be obtained with the window and quantitative determination can be achieved with the help of the standard. Both simulated and experimental data were investigated with the proposed method. Whether or not a peak shift occurs during the test, accurate results were obtained from the overlapping GC–MS signals with high level of noise and background.

1. Introduction

Analysis of multicomponent real samples has been a great challenge in analytical chemistry, because background and peak overlapping are frequently encountered in analyzing the samples with complex matrices. In the chromatographic analysis, including the hyphenated techniques such as gas chromatography–mass spectrometry (GC–MS) and high-performance liquid chromatography with photodiode-array detector (HPLC-DAD), background and peak overlapping have been serious interferences in the signals [1–3]. Besides, due to the high level of noise, the quantitative determination cannot be easily achieved with the signal of low signal-to-noise ratio. Peak shift due to the matrix effects and variations in chromatographic system is also an impediment to characterization and quantitative analysis [4].

For the analysis of real samples with complex matrices, a variety of chemometric methods base on "window" such as evolving factor analysis (EFA) [5–7], window factor analysis (WFA) [8,9], heuristic evolving latent projections (HELP) [10,11], and subwindow factor analysis (SFA) [12] have been developed. A window means the existing period of a component along the retention time. With the submatrix in the selective window, the concentration profiles and spectra of the components in a complex sample can be obtained from the overlapping signals. However, the determination of the window is difficult for signals with high level of noise and background. Moreover, since the results by these methods are strongly dependent on the window, bias may arise if an incorrect window is provided [13].

Among these methods, WFA is widely used and has been adopted to extract concentration profiles of the chemical species from data matrix of an evolutionary process [8,9,14,15]. The key of WFA method is to specify the window of a target component, i.e., the existing period of the component along the retention time. The best window for a particular component can be determined by the fact that the profile generated by the window (1) fits nicely inside the specified window, neither too large nor too small, (2) exhibits a single maximum and (3) is close to zero at all points outside the window [9]. In our previous works [16–18], efforts have been made to improve the method for the case of noisy signals. However, determination of the window for the calculation is difficult or even impossible when high level of noise and background exist in the signals [19]. Thus, the problem limits the application of WFA in the analysis of real samples [13].

* Corresponding author at: Tel.: +86 22 23503430; fax: +86 22 23502458.
E-mail address: xshao@nankai.edu.cn (X. Shao).

http://dx.doi.org/10.1016/j.chroma.2015.06.059
0021-9673/

In this work, a generalized window was proposed for analysis of GC–MS signals using WFA. For a GC–MS signal, the generalized window is defined as the selected channels (mass-to-charge ratios) in the mass spectral dimension of the data matrix, instead of a continuous region in chromatographic dimension along with the retention time. Therefore, the generalized window for a target component can be easily determined with the help of the mass spectrum, either from the library or from the experimental measurement. Furthermore, the interferences of background, peak overlapping, and even the peak shift will not affect the determination of the window. The method was named as generalized window factor analysis (GWFA) in this study.

2. Algorithm and calculation

WFA is a chemometric method based on chemical factor analysis and used for obtaining the concentration profiles of individual components from evolutionary processes. Generally, **D** is used to represent a measured GC–MS data matrix of multi-component chemical system, where each column is a mass spectrum recorded during an evolutionary process, and each row is a record of evolutionary profile. In WFA, the window of the nth component is defined as the existing period of the component along the retention time, though the concentration profiles of other components may exist inside the window. Let matrix $\mathbf{D}^0$ represent a submatrix of **D** that remains after removing all columns within the window, and perform principal component analysis (PCA) to $\mathbf{D}^0$, then

$$\mathbf{D}^0 = \sum_{j=1}^{n-1} \boldsymbol{s}_j^0 \boldsymbol{c}_j^{0\mathrm{T}} = \mathbf{S}^0\mathbf{C}^{0\mathrm{T}} + \mathbf{E} \quad (1)$$

where superscript "T" denotes the transpose; $\boldsymbol{s}_j^0$ is a vector representing the spectra of component j, and $\boldsymbol{c}_j^0$ is a vector representing the concentration profile of component j. $\mathbf{S}^0$ and $\mathbf{C}^0$ are the matrices of scores and loadings, respectively. $\mathbf{S}^0$ contains $n-1$ orthonormal spectral vectors $\boldsymbol{s}_j^0$ and $\mathbf{C}^0$ contains $n-1$ orthogonal non-normalized concentration profile vectors $\boldsymbol{c}_j^0$. The residual matrix **E** contains the part of variation that is unexplained by the model.

Using the $\mathbf{S}^0$ obtained with the window, the information of the nth component can be obtained by [9,18],

$$\mathbf{X}_n = \beta_{nn}\boldsymbol{s}_n^0\boldsymbol{c}_n^{\mathrm{T}} = (\mathbf{I} - \mathbf{S}^0\mathbf{S}^{0\mathrm{T}})\mathbf{D} \quad (2)$$

where vectors $\boldsymbol{s}_n$ and $\boldsymbol{c}_n$ are the spectrum and concentration profile of the nth component, respectively, and β_{nn} is a linear coefficient. Because rows of $\mathbf{X}_n$ are proportional to each other, the average of the row vectors is the uncalibrated concentration profile of the nth component.

To enhance the ability of WFA for noisy data, an improved version was developed [19]. The information of the nth component can be calculated by

$$\mathbf{Y}_n = k\boldsymbol{c}_n\boldsymbol{c}_n^{\mathrm{T}} = \mathbf{D}^{\mathrm{T}}(\mathbf{I} - \mathbf{S}^0\mathbf{S}^{0\mathrm{T}})\mathbf{D} \quad (3)$$

Because the average of the row or column vectors in $\mathbf{Y}_n$ is identical, the uncalibrated concentration profile of the nth component can be obtained.

The key of WFA method is to specify the existing period of a component along the retention time. However, trials are almost always needed to find the suitable window by examining the calculated concentration profile. It is more difficult when high level of noise and background exist in the experimental data [19]. For GC–MS data, it may be an easier way to define the window along the charge-to-mass dimension, because the mass spectrum is comparatively fixed and it is easy to be obtained from the libraries. Therefore, the concept of "generalized window" is proposed, i.e., the channels in the mass spectrum of the analyzing component are used to define the window, though the spectra of other components may have the same mass channels. It will be very easy to determine the generalized window when the analyzing target is known or there is a pure standard sample. In this method, the transpose of **D** is used. Let $\mathbf{D}^{0\mathrm{T}}$ represent the submatrix of $\mathbf{D}^{\mathrm{T}}$ after removing all columns in the window, and perform PCA to $\mathbf{D}^{0\mathrm{T}}$, then

$$\mathbf{D}^{0\mathrm{T}} = \sum_{j=1}^{n-1} \boldsymbol{c}_j^0 \boldsymbol{s}_j^{0\mathrm{T}} = \mathbf{C}^0\mathbf{S}^{0\mathrm{T}} + \mathbf{E} \quad (4)$$

Using the generalized window, Eq. (3) can be written as

$$\mathbf{Y}_n = k\boldsymbol{s}_n\boldsymbol{s}_n^{\mathrm{T}} = \mathbf{D}(\mathbf{I} - \mathbf{C}^0\mathbf{C}^{0\mathrm{T}})\mathbf{D}^{\mathrm{T}} \quad (5)$$

The average of the row vectors in $\mathbf{Y}_n$ is the calculated mass spectrum of the nth component.

Moreover, with a sample of known concentration, the β in Eq. (2) and the k in Eq. (3) or (4) can be obtained and used for quantitative analysis. In this study, standard addition method was used for quantitative calculation using the square root of the calculated mass spectrum. Let **m** represent the average of the row vectors in $\mathbf{Y}_n$, i.e., the uncalibrated mass spectrum of the nth component, then the square root of the calculated mass spectrum (Q) can be obtained by

$$Q = \sqrt{||\mathbf{m}||} \quad (6)$$

where $||\cdot||$ is the operator of L1 norm. The value of Q can be determined with a sample of known concentration. When standard addition method is used, a linear equation can be obtained between the Q values and the added concentrations.

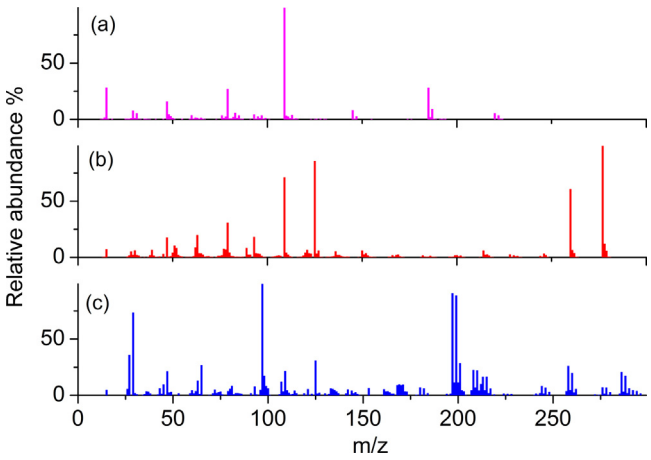

Fig. 1. Mass spectra used in the simulation.

3. Experimental

3.1. Data simulation

A three-component GC–MS data was simulated for validating the method. In the simulation, the GC profiles of three components are generated by commonly used exponentially modified Gaussian (EMG) equation [20]. The mass spectra of three organophosphorus pesticides shown in Fig. 1a–c were used in the simulation, which were selected from the standard mass spectral database (NIST05). By multiplying the two matrices of the chromatogram and mass spectra, the simulated GC–MS data matrix was obtained. In order to make the results more realistic, a constant background and high level of white Gaussian noise were added. The magnitude