Journal of Chromatography A, 1411 (2015) 110–115

Contents lists available at ScienceDirect

Journal of Chromatography A

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

Band target entropy minimization for retrieving the information of individual components from overlapping chromatographic data

Zhenzhen Xia, Yan Liu, 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 27 May 2015
Received in revised form 30 July 2015
Accepted 31 July 2015
Available online 5 August 2015

Keywords:
Gas chromatography–mass spectrometry
Overlapping signal
Curve resolution
Band target entropy minimization
Singular value decomposition

ABSTRACT

Band target entropy minimization (BTEM) is a self-modeling curve resolution (SMCR) approach relying on non-negative criterion and minimization of Shannon entropy. In this study, BTEM algorithm was applied to retrieving the information of individual components from overlapping gas chromatography–mass spectrometry (GC–MS) data. The algorithm starts with dividing the whole data into bands along the retention time. In each band, singular value decomposition (SVD) is used to decompose the data into scores and loadings. Because the pure chromatographic signal possesses the lowest Shannon entropy, the chromatographic signal of each component can be constructed by optimizing the combination of the loadings with minimal Shannon entropy under non-negative criterion. To show the efficiency of the algorithm, a simulated four-component overlapping GC–MS data and an experimental GC–MS data of 18 organophosphorus pesticide mixture are investigated. The results show that both the chromatographic profiles and mass spectra of the components can be successfully extracted from the overlapping signals.

1. Introduction

Chemometrics has provided an alternative way for improving the efficiency of hyphenated chromatographic techniques for complex sample analysis [1–3]. A large number of methods have been developed based on chemical factor analysis (CFA), such as evolving factor analysis (EFA) [4–6], window factor analysis (WFA) [7,8], heuristic evolving latent projections (HELP) [9,10] and target factor analysis (TFA) [11]. These methods are widely applied to resolution of multicomponent system such as traditional Chinese herbal medicine [12]. However, there are still practical difficulties in CFA when the data matrices are irregular compared with theoretical model, for example when high level noise exists in the analyzing signal. Subsequently, multivariate curve resolution-alternating least squares (MCR-ALS) [13–16] was developed for resolution of overlapping signals via an alternating least square algorithm under the constraints of non-negativity and unimodality. With the development of analytical instrumentation, second and even higher dimensional data were generated and analyzed by high order calibration methods. Parallel factor analysis (PARAFAC) [17,18] and alternating trilinear decomposition (ATLD) [19,20] have been proposed as trilinear methods with the second-order advantage. These methods make it possible to quantify the interested components even in the presence of unknown complex interferences. In addition, immune algorithm (IA) [21,22] and independent component analysis (ICA) [23] have successfully applied for resolution of overlapping GC–MS signals in our previous work. All these works have shown that chemometric techniques can enhance the efficiency of hyphenated instruments in analyzing complex samples.

Band target entropy minimization (BTEM) [24,25] is a self-modeling curve resolution (SMCR) method for recovering the spectrum of the component from overlapping spectra. The overlapping data can be decomposed by singular value decomposition (SVD) into scores and loadings. Loadings contain multiple component abstract spectra information. The spectrum of a pure component can be reconstructed by optimizing the combination of loadings. The key step in the algorithm is the optimization of combination. According to the assumption of the algorithm, the spectrum of a pure component owns the lowest value of Shannon entropy [26] compared with the overlapping spectra. Consequently, the optimization of combination is to minimize Shannon entropy of the reconstructed spectrum under non-negative criterion. The efficiency and accuracy of BTEM algorithm have been tested in the analysis of Fourier transform infrared spectroscopy (FTIR) signals [24,25,27]. Even the signal of a component at trace level of concentration can be correctly extracted [24]. Moreover, BTEM algorithm has been used to analyze the signals of Raman [28], mass spectrometry (MS) [29], nuclear magnetic resonance (NMR) [30,31],

* Corresponding author.
E-mail address: xshao@nankai.edu.cn (X. Shao).

http://dx.doi.org/10.1016/j.chroma.2015.07.124
0021-9673/

powder X-ray diffraction (XRD) [32] and ultraviolet–vis (UV–vis) spectroscopy [33,34]. Compared with the resolution approached introduced above, the advantage of BTEM is no need to know the accurate component number, estimations of starting spectra in the optimization, and the information of component contained in the mixture. For example, when IA is used, the information of component contained in the mixture must be provided because IA is a method based on curve fitting to extract the contribution of each component to the total signal by projection and subtraction. On the other hand, when BTEM is used for high-dimensional data, unfolding into two dimensional data is needed. In such cases, methods like PARAFAC and ATLD are more suitable to directly deal with high dimensional data. ICA can be used for direct resolution of overlapping signals, the number of independent components, however, is needed before the calculation.

In this study, BTEM was used to retrieve the information of individual components from highly overlapping gas chromatography–mass spectrometry (GC–MS) data. SVD is used to decompose the data into scores and loadings. By optimizing the combination of loadings with minimal Shannon entropy under non-negative criterion, the chromatographic signal of a pure component can be calculated. Two datasets, including a simulated and an experimental GC–MS data, are used to investigate the performance of the method.

2. Theory and calculations

A GC–MS data **X** of multicomponent mixture can be described as a product of the mass spectra **S** and chromatographic profiles **C** of its components.

$$\mathbf{X}_{n\times m} = \mathbf{SC} + \mathbf{E} \tag{1}$$

where n and m denote the number of m/z channel and retention time sampled in the experiment, respectively. **E** is experimental error.

The data can be decomposed into scores **P** and loadings **V** by SVD,

$$\mathbf{X} = \mathbf{PV}^{\mathrm{T}} \tag{2}$$

where superscript T denotes the transposition. **V** contains multiple component abstract chromatogramphic information. With the combination of the loadings, the chromatogram of a component can be obtained according Eq. (3).

$$\hat{\mathbf{C}}_{1\times m} = \mathbf{R}_{1\times z}\mathbf{V}^{\mathrm{T}}_{z\times m} \tag{3}$$

where $\hat{\mathbf{C}}$ denotes the estimated chromatographic profile of a component and **R** denotes the rotation vector. z is the factor number used for the construction, which was determined according to the cumulative variance proportion. Cumulative variance proportion is a percentage of the variance explained by the first z loadings in the total variance of the data. It reflects the amount of information included in the z loadings.

In general, $\mathbf{R}_{1\times z}$ is initialized randomly, and then optimized by sequential quadratic programming (SQP) algorithm [35] to subject Eq. (4). Global convergence of SQP is established under a reformulation of the complementarity condition combined with a classical penalty function method for solving constrained optimization problems. Therefore, the ambiguities for **R** can be overcome.

$$arg\ \min\left(-\sum_{m=1}^{M} h_m \ln(h_m) + \mathrm{P}\right) \tag{4}$$

The former part of Eq. (4) is the Shannon entropy criterion defined in Eq. (5) and the latter part of Eq. (4), P, is defined as the product of the quadratic sum of the negative elements in $\hat{\mathbf{C}}_{1\times m}$ to ensure non-negativity [24].

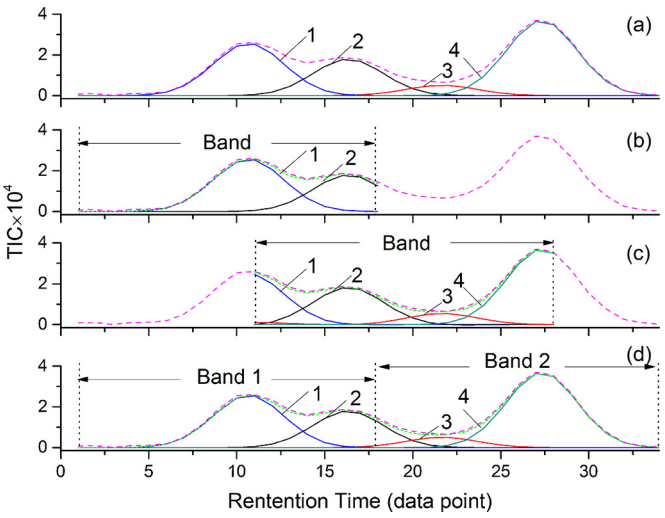

Fig. 1. Simulated (a) and the retrieved chromatographic profiles with different band position (b–d). Retrieved profiles are plotted in different color, the dash line represents the total signal of the four simulated peaks and the short dash line represents a summation of the calculated signals.

$$h_m = \frac{|\hat{c}_m|}{\sum_{m=1}^{M}|\hat{c}_m|} \tag{5}$$

where $\hat{c}_m$ is the element in $\hat{\mathbf{C}}_{1\times m}$, || denotes the absolute value.

When the chromatographic profiles of all the components are constructed, mass spectra can be calculated using Eq. (6).

$$\hat{\mathbf{S}} = \mathbf{X}\hat{\mathbf{C}}^{\mathrm{T}}(\hat{\mathbf{C}}\hat{\mathbf{C}}^{\mathrm{T}})^{-1} \tag{6}$$

In practical application, GC–MS data alone retention time consists of a series of non-negative Gaussian peaks. When the whole data is used in the calculation of BTEM, the presence of multiple local minimum may affect the efficiency of the optimization. Thus, the GC–MS data needs to be divided into several bands along the retention time and the data in each band is used for BTEM analysis.

3. Experimental

3.1. Data simulation

The simulated GC–MS data was used for validation of the method. In the simulation, the GC curves of four components are generated by the Gaussian equation and the mass spectra of the four components are chosen from National Institute of Standards and Technology (NIST) MS database. The simulated total ion chromatograms (TIC) and the mass spectra of the four components are shown in Figs. 1(a) and 2(a), respectively. To show the similarity of the mass spectra, the match ratios are calculated. The match ratios between component 1 and the others are 654‰, 408‰ and 108‰, respectively, between component 2 and components 3 and 4 are 290‰ and 156‰, respectively, and between components 3 and 4 are 161‰.

The GC–MS data were obtained by multiplying the chromatograms and mass spectra. Then, random noise in 5% of maximum signal value was added in the simulated signals.

3.2. GC–MS measurements

The 18 organophosphorus pesticide mixture was obtained by mixing the standards in a solution with a concentration of