Available online at www.sciencedirect.com

SCIENCE DIRECT®

Chemometrics and Intelligent Laboratory Systems 82 (2006) 137 – 144

Chemometrics and intelligent laboratory systems

www.elsevier.com/locate/chemolab

A primary study on resolution of overlapping GC-MS signal using mean-field approach independent component analysis

Guoqing Wang [a,b], Wensheng Cai [a], Xueguang Shao [a,*]

[a] *Department of Chemistry, University of Science and Technology of China, Hefei, Anhui, 230026, P.R. China*
[b] *Department of Applied Chemistry, Zhengzhou University of Light Industry, Zhengzhou, Henan, 450002, P.R. China*

Received 10 January 2004; received in revised form 8 April 2005; accepted 26 May 2005
Available online 19 September 2005

Abstract

Independent component analysis (ICA) has been found to be powerful to separate complex signals. However, chemical signals are generally correlated, instead of independent as hypothesized in ICA. In this study, mean-field independent component analysis (MF-ICA) was investigated to resolve the overlapping gas chromatographic-mass spectrometric (GC-MS) signal. In MF-ICA, the sources are estimated from the mean of their posterior distribution. The mixing matrix and noise level are found through the maximum a posterior (MAP) solution. By simulated signals, results show that for cases of the slightly correlated (or overlapped) sources, both the sources (MS) and mixing matrix (chromatogram) can be almost correctly estimated by specification of the nonnegative (positive) priors for the mixing matrix and sources. However, when the sources are highly correlated, no good results can be obtained, although acceptable estimated sources can be obtained somehow for database matching. For experimental overlapping GC-MS data, reasonable results are obtained, because MS spectra of different homologous compounds in GC-MS analysis of a mixture are not generally correlated very much. Therefore, ICA should be an alternate tool for resolution of overlapping chemical signals, although further works are still needed.

Keywords: Independent component analysis (ICA); Mean field; Resolution of overlapping signal; Gas chromatography-mass spectrometry (GC-MS)

1. Introduction

Gas chromatography-mass spectrometry (GC-MS) is being extensively used in scientific research and practical applications. For samples of complex mixtures, incomplete GC separation is a common problem. This will make the quantitative and qualitative analysis of the specific component very difficult. Therefore, the resolution of the overlapping GC-MS signal is a challenging and active research field [1–13].

Independent component analysis (ICA) is a statistical signal/data processing technique that aims at recovering the source signals under the assumption that the source signals are statistically independent [14]. It has the potential applications on blind source separation of multivariate signals and drawn great attention. ICA has been applied to analytical chemical data analysis [1,15,16], medical signals processing [17], speech recognition [18,19], fault detection [20], statistic process monitoring [21], and batch processes monitoring [22], etc.

In our recent work [1], the overlapping GC-MS signal was resolved using an immune algorithm (IA) [12,13] combined with maximum likelihood or Infomax ICA [23,24]. However, in the further studies, we found that the estimated sources may be negatively correlated with the real ones, in some cases without constraint priors for the mixing matrix and sources [16,25]. Furthermore, the intensity values of the different variables in an estimated source may have positive and negative values at the same time. This is obviously not correct for chemical signals because the estimated sources are MS of the pure components and the mixing matrix represents the concentration of the components in the overlapping signal. It is likely that the estimated sources and mixing matrix located outside the practical range, is equivalent to a local optimization. However, such constraints are usually ignored by most of the commonly used modeling methods, such as the principal component analysis (PCA) [25] and conventional ICA [24,25].

[*] Corresponding author. Tel.: +86 551 3606160; fax: +86 551 3601592.
E-mail address: xshao@ustc.edu.cn (X. Shao).

0169-7439/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2005.05.009

There are many different ICA algorithms in the applications. The FastICA algorithm [26] uses a deflation scheme to compute components sequentially. The JADE algorithm [27] is a cumulant-based method that uses joint diagonalization of a set of fourth-order cumulant matrices. The extended Infomax algorithm [24] is a variation on the Infomax algorithm [23] that can deal with either sub-Gaussian or super-Gaussian components by adaptively switching between two nonlinearities. The mean-field approach ICA (MF-ICA) algorithm [25,28] finds the mean of the sources and their covariance matrix, and uses them to describe the sources, mixing matrix and the noise covariance matrix. It can be seen that the principle of ICA is different with factor analysis [2,3] or pure variable approach (PVA) [7]. All these ICA algorithms are reported to be successfully applied in certain application areas.

Due to the fact that in the GC-MS analysis, the content of the overlapping multicomponents and the abundance at every mass-to-charge ratio position (variable) is impossible to be a negative value, the nonnegative (positive) constraints should be specified to the sources and the mixing matrix. The estimation of the sources and mixing matrix based on the constraint priors should be more reasonable. Therefore, in this study, the MF-ICA with nonnegative constraints for the sources and mixing matrix was investigated to resolve the overlapping GC-MS signal. Results show that acceptable results can be obtained when the sources are slightly correlated (or overlapped). However, when the sources are highly correlated, only acceptable estimation of the sources (MS) can be obtained, whereas the correct chromatogram cannot be obtained. Even though reasonable results for experimental GC-MS data can be obtained, MS spectra of different homologous compounds are not generally highly correlated in practical analysis.

2. Theory and algorithm

2.1. Mean-field approach independent component analysis (MF-ICA)

Generally, there are three kinds of mean-field methods [25]: the variational approach, linear response correction, and adaptive Thouless, Anderson and Palmer (TAP) approach [28].

The general model for ICA is that the mixed signals are generated through a linear combination of pure signals, where additive noise can be present,

$$\mathbf{X} = \mathbf{AS} + \mathbf{\Gamma} \tag{1}$$

or,

$$\mathbf{X}_{m,n} = \sum_{k=l}^{d} \mathbf{A}_{m,k}\mathbf{S}_{k,n} + \mathbf{\Gamma}_{m,n} \tag{2}$$

where $\mathbf{X}=[\mathbf{x}_1, \mathbf{x}_2,\ldots, \mathbf{x}_m]^{\mathrm{T}}$ represents the matrix holding the m mixed or observed signals (GC-MS data matrix) in each row with n variables, $\mathbf{A}=[\mathbf{a}_1, \mathbf{a}_2,\ldots, \mathbf{a}_m]^{\mathrm{T}}$ represents the $m \times n$ combination coefficients or mixing matrix, and $\mathbf{S}=[\mathbf{s}_1, \mathbf{s}_2,\ldots, \mathbf{s}_m]^{\mathrm{T}}$ represents the matrix holding d independent source signals (mass spectra) in rows of n variables. The noise is added by the $m \times n$ matrix $\mathbf{\Gamma}$ that is generally defined to be Gaussian or fully neglected. In the overlapping GC-MS data, $\mathbf{X}$, $\mathbf{A}$, $\mathbf{S}$ and $\mathbf{\Gamma}$ correspond to the overlapping MS signals, the relative concentration information of the components, the MS information of the pure components and the experimental (instrumental measuring) noise, respectively.

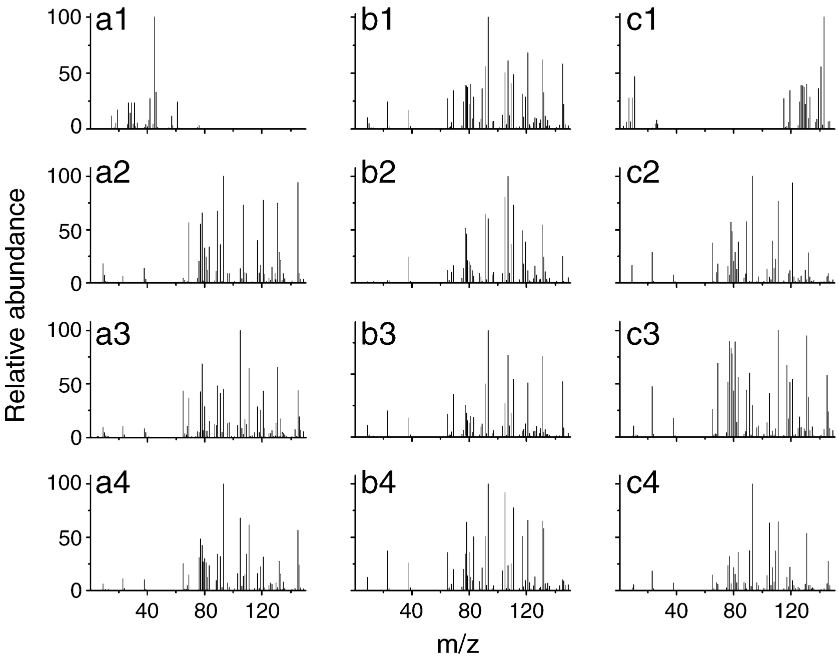

Fig. 1. Four groups simulated mass spectra (relative abundance) a_i, b_i, and c_i (i=1, 2, 3, 4) denote three components in each group.