Available online at www.sciencedirect.com
ScienceDirect
Progress in Natural Science 18 (2008) 1009–1014
Progress in Natural Science
www.elsevier.com/locate/pnsc

Eliminate indeterminacies of independent component analysis for chemometrics

Zhixiang Yao [a,b], Kai Zhang [c,*], Huanbin Liu [b], Hui Su [a]

[a] Department of Biological and Chemical Engineering, Guangxi University of Technology, Liuzhou 545006, China
[b] State Key Laboratory of Pulp and Paper-making Engineering, South China University of Technology, Guangzhou 510640, China
[c] State Key Laboratory of Heavy Oil Processing, China University of Petroleum, Beijing 102249, China

Received 17 January 2008; accepted 31 January 2008

Abstract

An improved method has been proposed to eliminate the indeterminacies of independent component analysis (ICA) for chemometrics. Following the arrangement of principal components analysis (PCA), the ICA mixing matrix is selected as signal content indexes, and ICA output are sorted and directed. After many times reputations, independent components (ICs) are paired according to the maximum correlation coefficient, and then the mean values of each IC substitutes the original ICs. This indicates that the ICA indeterminacies are eliminated. A simulation example is tested to validate this improvement. Finally, a set of experimental LC–MS data is processed without any prior knowledge or specific limitation and the results show that the improved ICA can directly separate the mixed signals in chemometrics, and it is simpler and more reasonable than the simple to use interactive self-modeling mixture analysis (SIMPLISMA).

Keywords: Independent component analysis; Indeterminacy in ICA; Chemometrics

1. Introduction

Independent component analysis (ICA) has been regarded as one of the most important algorithms to deal with blind source separation (BSS) because it can extract the profile of latent components from the composite signals [1–2]. This method has been used to separate independent signal from mixed multi-component signals in the area of chemometrics, such as simultaneously determination in multi-components system [3], signal purification and overlap spectrogram resolution [4–5].

In ICA calculation, there are the threefold indeterminacy of permutation, sign and shape [6]: (1) permutation indeterminacy: it is random in nature that the order of independent components (ICs) relates to the size of independent criterion; (2) sign indeterminacy: each output component phase is indefinite, i.e., the component phase is completely random in the normal or opposite position; and (3) shape indeterminacy: this kind of indeterminacy includes random error based on seeking optimization, systematic error resulted from the non-orthogonal of original sources, and indefinitely corresponding error between output and original component if the output component number is inconsistent with the original component number. In ICA applications, the troubles caused by these indeterminacies include the rearrangement of all the output components after each calculation, difficulty for dealing with the relationship between output and original components, and occasional accidents from great errors. The aim of this study is to propose an improved algorithm to eliminate the above indeterminacies in ICA for chemometrics.

* Corresponding author. Tel.: +86 10 89733939; fax: +86 10 69724721.
E-mail address: kaizhang@cup.edu.cn (K. Zhang).

1002-0071/$ - see front matter © 2008 National Natural Science Foundation of China and Chinese Academy of Sciences. Published by Elsevier Limited and Science in China Press. All rights reserved.
doi:10.1016/j.pnsc.2008.01.034

2. ICA indeterminacies

In BSS, the simplest and the most frequently considered mixing model is in a linear form [7]:

$$x = \boldsymbol{A}s \tag{1}$$

where $x = [x_1,\ldots,x_m]^{\mathrm{T}}$ and $s = [s_1,\ldots,s_n]^{\mathrm{T}}$ are the vectors of the measured and source signals, respectively. $\boldsymbol{A}$ is an unknown $m \times n$ mixing matrix. As premise ICA definition, there exists only one criterion for estimating components, i.e., each element in s is statistically independent. However, it is difficult to eliminate the indeterminacies without any prior knowledge [8] as s in Eq. (1) has $2n!$ permutation ways.

In fact, except for the ICs arrangement, how to maintain the ICs shape uniformity and mixing matrix determinacy are also two key issues while ICA would be used in more widespread fields, i.e., in the occasions where each source signal is nonorthogonal mutually. Accordingly, this section is focused on ICA indeterminacies and the solution to these key problems. The causes of ICs indefinite errors are first discussed in Section 2.1. Then, the scheme of accurate pairing and predefined arrangement of the ICs containing indefinite errors are shown in Sections 2.2 and 2.3, and the method to eliminate the ICA indeterminacies is finally presented in Sections 2.4 and 2.5.

2.1. Indefinite errors

The ICs indeterminacies of permutation and sign can be processed artificially according to prior knowledge when data quantity is limited. When ICA is used in the area of chemometrics, however, the ICs indeterminacy of shape must be paid more attention because it directly affects the feasibility and accuracy of the quantitative analysis. As it is known that the mixing matrix $\boldsymbol{A}$ in Eq. (1) represents the contribution of source signals in the overall system [9], this matrix is still indefinite resulting from the shape indeterminacy although the amplitudes of ICs are normalized in ICA methods [1]. It should be noticed that the instability of the matrix $\boldsymbol{A}$ is remarkable. Therefore, this matrix cannot be used directly for quantitative analysis.

Each IC is corresponding to each source signal in the data space. Fig. 1 shows their relations in the two-dimensional space. The components must be orthogonal if they are independent to each other according to the definition of statistical independence. Because of ICs orthogonality, ICs superpose on source signals if the source signals are statistical independence. Otherwise, the vector of the source signal will keep a fixed deviation of least square with that of IC.

For binary-component system, there are four possible combinations:

$$\mathrm{ICa} = \boldsymbol{a}' \pm \boldsymbol{b}'' \text{ and } \mathrm{ICb} = \boldsymbol{b}' \pm \boldsymbol{a}'' \tag{2}$$

The randomness of "+" or "−" in Eq. (2) results in the ICs indeterminacies of shape and range.

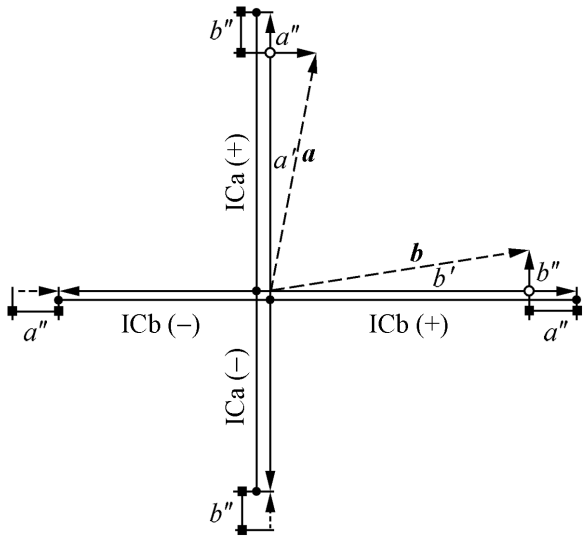

Fig. 1. The relations of ICs and source signals in the two dimensional space. $\boldsymbol{a}$ and $\boldsymbol{b}$ represent for the source vectors, and ICa and ICb for their corresponding ICA vectors. $\boldsymbol{a}'$ (or $\boldsymbol{b}'$) is the projection of $\boldsymbol{a}$ (or $\boldsymbol{b}$) on ICa (or ICb), whilst $\boldsymbol{a}''$ (or $\boldsymbol{b}''$) is projection of $\boldsymbol{a}$ (or $\boldsymbol{b}$) on ICb (or ICa). The direction of IC may be positive or negative in random. ICa is signed as ICa (+), i.e., ICa (+) $= \boldsymbol{a}' + \boldsymbol{b}''$, when ICa synchronizes with $\boldsymbol{a}$. Otherwise, ICa is signed as ICa (−), i.e., ICa (−) $= \boldsymbol{a}' - \boldsymbol{b}''$.

If there are n components in the system, the ICs will have $2(n-1)n$ possible combinations. $\tilde{s}$, the set of ICs can be expressed as

$$\tilde{s} = s' + \varepsilon^{\mathrm{sys}} + \varepsilon^{\mathrm{rand}} \tag{3}$$

where s' is the projection of source signals on ICs corresponding to $\tilde{s}$, $\varepsilon^{\mathrm{sys}}$ and $\varepsilon^{\mathrm{rand}}$ represent systematic and random errors, respectively.

$\varepsilon^{\mathrm{sys}}$ is caused by the random possible combinations. ICA is an algorithm based on seeking optimization, so there are random errors ($\varepsilon^{\mathrm{rand}}$). When source signals do not meet the statistical independence, the size and direction of $\varepsilon^{\mathrm{sys}}$ are dependent on the deviation of source signal away from the statistical independence, and the randomness of $\varepsilon^{\mathrm{sys}}$ is limited within $2(n-1)n$ possibilities. Both $\varepsilon^{\mathrm{sys}}$ and $\varepsilon^{\mathrm{rand}}$ errors can be eliminated

$$\frac{1}{p}\lim_{p\to\infty}\sum_{i=1}^{p}\varepsilon^{\mathrm{sys}} \to 0 \text{ and } \frac{1}{p}\lim_{p\to\infty}\sum_{i=1}^{p}\varepsilon^{\mathrm{rand}} \to 0 \tag{4}$$

where p is the repetition times. ICA indeterminacies can be eliminated by many times ICA repetitions, but ICs always deviate from real sources while the sources are not orthogonal.

2.2. Definitive arrangement of independent components

The ICs must be accurately paired in order to obtain automatically a set of reliable mean values of ICs from many times repetitions, and the sign of ICs must be also fixed. In principal component analysis (PCA), there are two natural criteria which keep arranging and pairing the

components without indeterminacies, one of which is that the covariance matrix eigenvalues are not less than zero, and the other is that the arrangement of principal components descends along the covariance matrix eigenvalues sequence. The proportion of principal components in the whole system decreases with decreasing the sequence of the covariance matrix eigenvalues. However, such natural criteria do not exist in ICA, which leads to the indeterminacies of ICs arrangement. Followed by PCA, one stable characteristic can be obtained and used as a criterion for regulating arrangement and direction of components in ICA. Therefore, Eq. (1) may be written as

$$x_i|_{i\in m} = \sum_{j=1}^{n} a_{ij}s_j|_{j\in n} \tag{5}$$

where a_{ij} is an element of $\boldsymbol{A}$, which means "the content" of the jth source in the ith observed value [9]. a_{ij} will remain nonnegative if the signals maintain the behavior of additivity, which is commonly accepted in chemometrics. As shown in Eq. (5), x_i is confirmable and the sign of s_j is fixed once confirming the sign of a_{ij}. In Eq. (1), the jth row in the matrix $\boldsymbol{A}$ corresponds to the jth column in the set of ICs, s. Because a_{ij} is nonnegative, sign of a_{ij} is dictated as

$$\sum_{j=1}^{m}(a_i)_j \geqslant 0 \tag{6}$$

In this situation, the sign of s_j is confirmed.

In PCA, the more fronted component contributes more to the whole system than behind those. By analogy, the mixing matrix $\boldsymbol{A}$ is a content scale of the source within whole system in ICA. The contributions of ICs or source components are also scaled by the distribution of each element in the matrix $\boldsymbol{A}$. The sum of squares of each row in the matrix $\boldsymbol{A}$ is written as

$$SS = \sum_{i=1}^{m}(a_i^2)_j \tag{7}$$

The larger SS value represents the more contribution from its corresponding IC. Therefore, Eqs. (6) and (7) may be selected as predefined criteria for arranging and identifying the ICA output.

2.3. Accurate pairing of independent components

The pairing of ICs based on the mixing matrix $\boldsymbol{A}$ of each repetition is not always trusty, therefore, a constant criterion must be reselected for the pairing of ICs. Sample signal, x, is decomposed into two parts by ICA, one of which is the mixing matrix $\boldsymbol{A}$ and the other is $\tilde{s}$. Generally, $\tilde{s}$ has a greater amount of information than $\boldsymbol{A}$, thus the pairing of ICs based on $\tilde{s}$ is more accurate than that based on A.

Let $\tilde{s}^{(1)}$ and $\tilde{s}^{(2)}$ be, respectively, two ICA repetitions outputs. After standardization processing, their correlation coefficient is

$$\begin{aligned} r_{kl} &= \frac{1}{n-1}\mathrm{cov}(\tilde{s}_k^{(1)}\tilde{s}_l^{(2)}) \\ &= \frac{1}{n-1}\mathrm{cov}[(s_k^{(1)}+\varepsilon_k^{(1)})(s_l^{(2)}+\varepsilon_l^{(2)})] \\ &= \frac{1}{n-1}[\mathrm{cov}(s_k^{(1)}s_l^{(2)}) + \mathrm{cov}(s_k^{(1)}\varepsilon_l^{(2)}) + \mathrm{cov}(s_l^{(2)}\varepsilon_k^{(1)}) \\ &\quad + \mathrm{cov}(\varepsilon_k^{(1)}\varepsilon_l^{(2)})] \end{aligned} \tag{8}$$

If $\|s\| > \|\varepsilon\|$, only when $s_k^{(1)} = \pm s_l^{(2)}$, $|r_{kl}|$ is maximum.

Therefore, the maximum correlation is a criterion for pairing the components between each ICA repetitions. For some special occasions of stable and random ICs, Eq. (8) approaches to zero, which suggests that components pairing is invalid based on the correlation coefficient. So these stable and random variables need to be tagged before ICs pairing.

2.4. Tagging stable and random components

In PCA, principal components in the front of arrangement have described the most characteristics of the system, whilst the rest is regarded as noise. In ICA, stable and random components are possibly arranged at the front position if the system has more noise and components are arranged in the matrix $\boldsymbol{A}$, which leads to the pairing ICs invalid based on the criterion of the maximum correlation. This implies that these components need to be tagged and then put in the end of arrangement.

Can stable component be judged by the autocorrelation of time series? In entirely random time series, the autocorrelation value, R, equals to 0 when the number of sample points is close to infinite, whilst the number of sample points, p, is limited, $R \sim N\ (0, p^{-1})$, $|R| < 4p^{-1/2}$ may be selected as a criterion for assessing the stable component.

2.5. The method for arranging and pairing independent components

Following the aforementioned discussion, we propose a method for arranging and pairing independent components. The details of the method can be summarized as the following three steps:

First step (preliminary arrangement):

1. conducting ICA for the sample data to obtain the independent components and the mixing matrix $\boldsymbol{A}$;
2. arranging and directing the independent components in the first time calculation on the basis of the mixing matrix $\boldsymbol{A}$;
3. computing autocorrelation values of ICs;
4. estimating whether there are stable components in ICs or not; and
5. putting the stable components into the end positions of ICs arrangement.

Second step (repeating computation):

1. repeating ICA and obtaining ICs;
2. tagging the stable components, and then putting them into the end of the arrangement or abandoning them from the ICs;
3. pairing these effective components with the preliminary output; and
4. repeating for adequate times, and attaining the mean value of each IC as a verified IC.

Third step (final arrangement):

1. regressing a new mixing matrix from the verified ICs, this new matrix is determinate, i.e., its indeterminacies have been eliminated;
2. rearranging the ICs based on this new matrix; and
3. outputting verified matrix $\boldsymbol{A}$ and a set of ICs, s.

3. Case study

3.1. Simulation data set

A set of simulation data for binary-component overlapped chromatograph signals (see Fig. 2) was used to confirm the determinacy of the algorithm in this study. As shown in Fig. 3, it is obviously that single FastICA [2] calculation deviates from the real source signals (true value), which is indeterminate in each calculation. In the meantime, each deviation of single calculations is also obvious and indeterminate (see Fig. 4), which indicates the columns of the mixing matrix $\boldsymbol{A}$ and the real content of source signals.

In order to check the reliability of the improved method, 10,000 times repetitions have been processed and each of repetitions has not any mistake in pairing and identifying

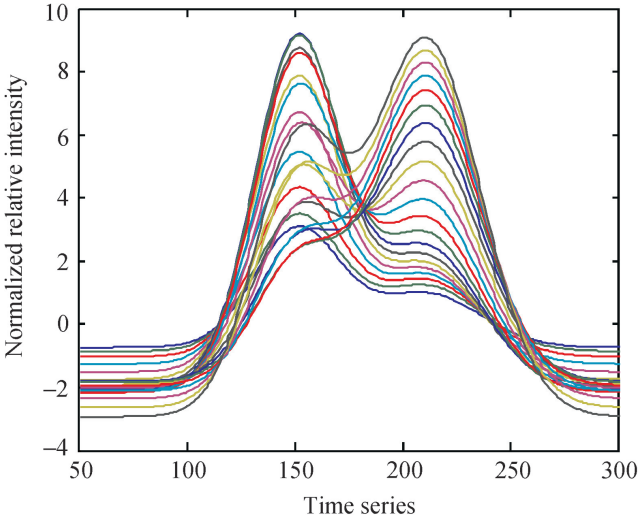

Fig. 2. Composite signals mixed with different contents of binary-source signals.

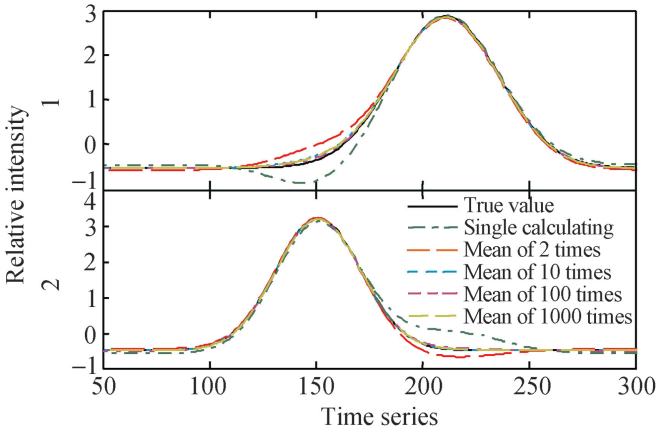

Fig. 3. Binary-source signals and ICs via FastICA.

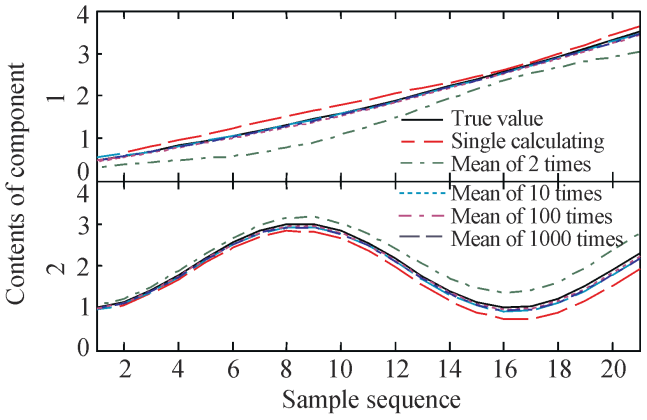

Fig. 4. Contents of binary-source signals and the rows of the mixing matrix of ICA.

ICs without manual intervention. Algorithm determinacy can be represented by the interval of correlation coefficients of ICA output. For this set of simulated data, the interval is (0.94, 0.99) by normal ICA method, whilst the interval, respectively, becomes (0.997, 1) or (0.999, 1) after 10 or 100 times repetitions by this improved ICA method. These results indicate that the improved ICA can eliminate the threefold indeterminacy of permutation, sign and shape.

This improved method not only eliminates ICA indeterminacies but also reduces the errors caused by the mutual influence of components because the maximum correlation coefficient between source and normal ICA output is 0.9978, while the correlation coefficient between source and improved output is up to 0.9997; and because by the normal ICA the minimum and maximum relative errors of "content" are 1.8% and 15%, respectively, but the relative error decreases to 1.6% by the improved ICA. These results indicate that accuracy of the improved ICA method also meets the requirement of quantitative analysis.

3.2. Experimental data set

Phalp et al. [10] used simple to use interactive self-modeling mixture analysis (SIMPLISMA) to purify mass

spectra from LC–MS spectrograms of mixture, and to separate the overlapped chromatograph peaks. Without selecting the purity curve, the improved ICA method can also be used to process the data set used by Phalp et al. [10]. The purified mass spectra and separated chromatograph peaks are directly worked out by the improved ICA, and its output is determinate. The above results for three components are in fair agreement with those of the reference spectra as shown in Fig. 5, which shows that the improved ICA neither needs prior knowledge nor has any limitation to sample data comparing with the SIMPLISMA.

It is disappointing that the mass spectrum of the fourth component could not be found out in the paper by Phalp et al. [10] and the component D could not be directly determined by the SIMPLISMA method. The argument was

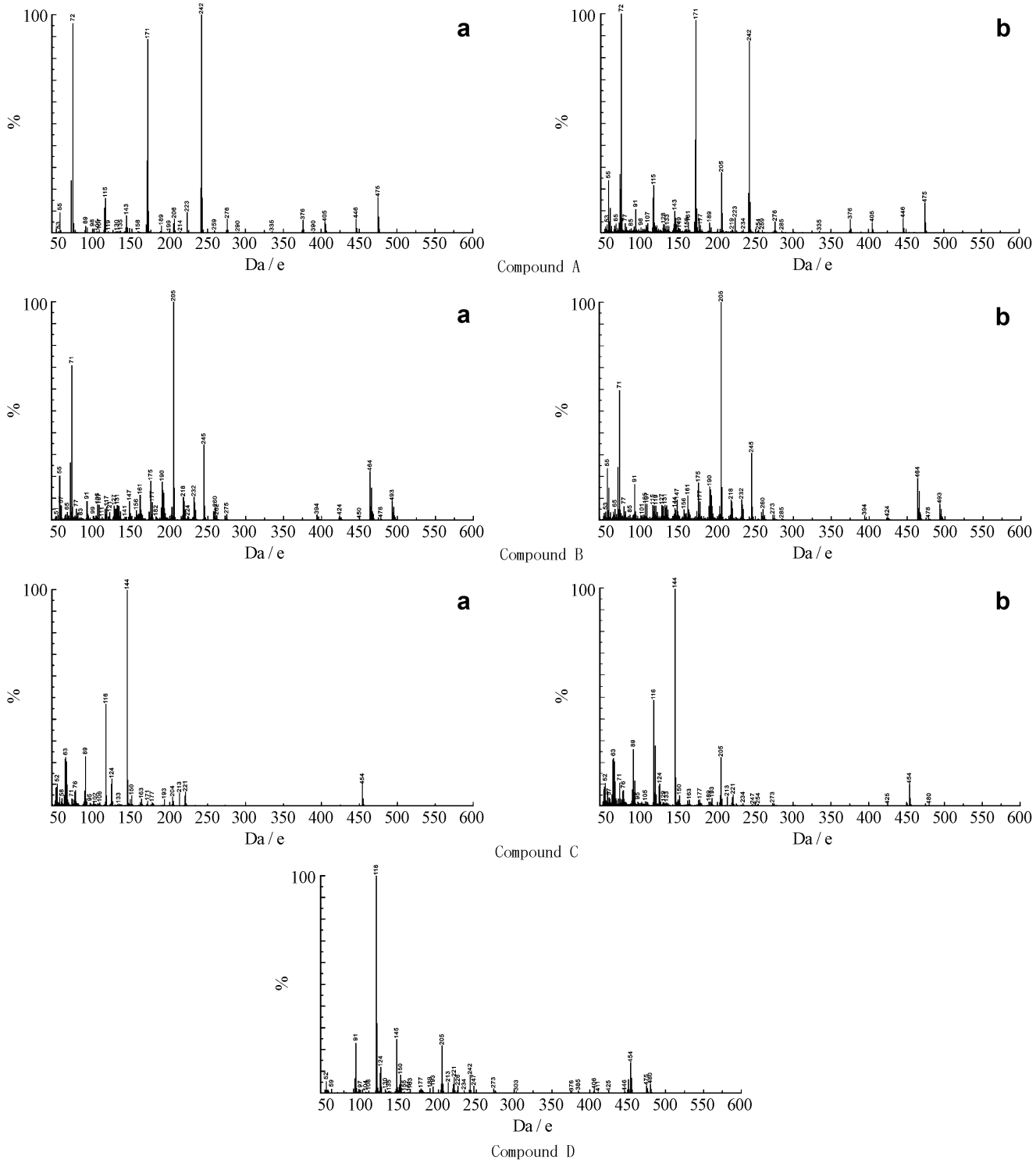

Fig. 5. Spectra for the components detected by improved ICA in sample data set [10]. (a) For resolved spectra, and (b) for reference spectra.

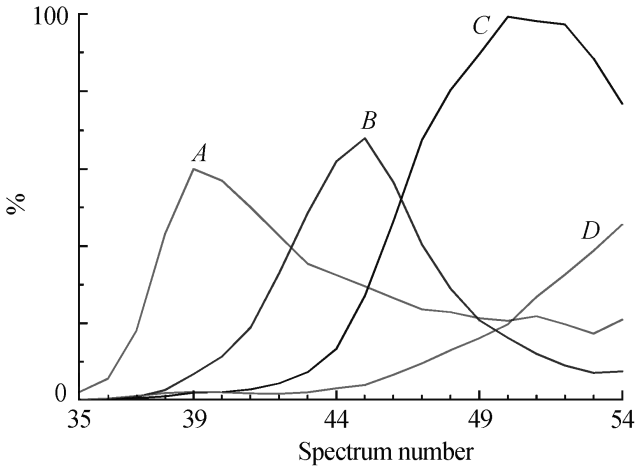

Fig. 6. Resolved purity curves for four components (curves represent their purity peaks within overlapped chromatograms).

that the component D was most intense for the component A. But the component D was the closest to the component C in the chromatogram [10], but furthest away from component A. Their contradict conclusion was a guess from the remaining information after all other components were drawn out.

In ICA, four spectra have been worked out simultaneously, and the component D no longer is influenced by the error cumulated in the remaining information as shown in Figs. 5 and 6. It is obvious that the component D is one independent matter. Known from 454, 144/145, and 116/118 (Da/e), the component D is likely to be an isomeride of the component C, so they are difficult to be separated by chromatogram. Furthermore, it can be found that the reference spectrum in Fig. 53-b has included the impressions of the components C and D, which may not be separated away (or the reference spectrum is a mixture of the components C and D), by comparing the reference spectrum for the component C in Fig. 53-b with the spectra of the component C in Fig. 53-a and the component D in Fig. 54-a. The above facts indicate that the results from ICA are reasonable and the improved ICA method has higher capability for separating mixed signals than the SIMPLISMA method.

4. Conclusion

After the indeterminacies are eliminated, ICA is able to separate and recognize the composite signals in the chemometrics. An improved ICA has been presented for eliminating the threefold indeterminacy of ICA, and the accuracy of ICA also is better than before. The improved ICA can be directly used to separate the mixing signals in chemometrics. As a typical example, a set of experimental LC–MS data is processed without any prior knowledge or specific limitation. The results show that the improved ICA proposed in this paper can directly separate the mixed signals in chemometrics, and it is simpler and more effective than SIMPLISMA.

Acknowledgement

This work was supported by Major State Basic Research Development Program of China (Grant No. 2005CB221205).

References

[1] Cichocki A, Amar S. Adaptive blind signal and image processing. West Sussex: John Wiley and Sons Press; 2002, 231–271.
[2] Hyvarinen A, Karhunen J, Oja E. Independent component analysis. New York: John Wiley and Sons Press; 2001, 3–5.
[3] Gustafsson MG. Independent component analysis yields chemically interpretable latent variables in multivariate regression. J Chem Inf Model 2005;45(5):1244–55.
[4] Yao ZX, Huang H, Liu HB. Processing GC–FTIR by the blind source separation. Spectrosc Spect Anal 2006;26(8):1432–6, [in Chinese].
[5] Visser E, Lee TW. An information–theoretic methodology for the resolution of pure component spectra without prior information using spectroscopic measurements. Chemometr Intell Lab Syst 2004;70(2):147–55.
[6] Ciaramella A, Tagliaferri R. In amplitude and permutation indeterminacies in frequency domain convolved ICA, Portland, OR, United States, 2003; Institute of Electrical and Electronics Engineers Inc.: Portland, OR, United States, 2003, 708–13.
[7] Amari SI. In: Blind signal separation: mathematical foundations of ICA, sparse component analysis and other techniques, Orlando, FL, United States, 2005. International Society for Optical Engineering, Bellingham, WA 98227-0010, United States: Orlando, FL, United States, 2005, 1–10.
[8] Tichavsky P, Koldovsky Z. Optimal pairing of signal components separated by blind techniques. IEEE Signal Proc Lett 2004;11(2):119–22.
[9] Yao ZX, Liu HB. Estimations to state variable in multivariable system by independent component analysis and wavelet. Control Decis 2006;21(1):88–92, [in Chinese].
[10] Phalp JM, Payne AW, Windig W. The resolution of mixtures using data from automated probe mass spectrometry. Anal Chim Acta 1995;318:43–53.