Available online at www.sciencedirect.com
SCIENCE DIRECT®

Talanta 68 (2006) 542–548

Talanta

www.elsevier.com/locate/talanta

A novel approach of initial vectors construction in iterative target transformation factor analysis

Hongtao Gao [a,b], Tonghua Li [a,*], Kai Chen [a], Shufang Lin [a]

[a] *Department of Chemistry, Tongji University, 1239 Siping Road, Shanghai 200092, China*
[b] *College of Chemistry and Molecular Engineering, Qingdao University of Science & Technology, Qingdao 266042, China*

Received 25 November 2004; received in revised form 22 April 2005; accepted 22 April 2005

Abstract

The construction of the favorable initial iterative vectors is the key to iterative target transformation factor analysis (ITTFA). A tentative approach to construct the better initial vectors, which is based on the chromatographic information provided by evolving factor analysis (EFA), is proposed. A region, which contains the peak position at maximum height, is determined. Several elements in the region of each initial vector, instead of one element, are initialized as 1. The elements out of the region are initialized as 0. So it is not necessary to determine the exact peak position at maximum height for the resolution of partly overlapping chromatographic profiles, which may avoid the divergence brought by determination of the peak position at maximum height. In addition, it may give acceptable resolution for the embedded peaks. It is applied to resolve 2D-simulated data and experimental liquor GC/MS data, the resolutions are reasonable and improved.

Keywords: Iterative target transformation factor analysis (ITTFA); Initial vector construction; Initialization

1. Introduction

Iterative target transformation factor analysis (ITTFA), based upon self-modeling curve resolution (SMCR), is an iterative method approximating the results from initial vectors [1,2]. At every iterative step, constraints like non-negativity and unimodality are applied [3].

It has been commonly discussed and applied to resolve 2D hyphenated chromatographic data. Lilley and Wheat [4] used ITTFA to inspect the chromatograms from a capillary electrophoresis/diode array instrument for peak purity. Enric Comas et al. [5] applied ITTFA in second-order liquid chromatographic data for time shift correction. Rodrĭguez-Cuesta et al. [6] have checked the influence of selectivity and sensitivity parameters on detection limits in the resolution of chromatographic second-order data by ITTFA. van Zomeren et al. [7] compared ITTFA with some curve resolution, such as alternating least squares (ALS), evolving factor analysis (EFA), and heuristic evolving latent projection (HELP), for drug impurity profiling using high performance liquid chromatography with diode array detection. It has been proved that ITTFA can attain great accuracy associated with the systems where the concentration distributions of each absorbing species do not overlap seriously [8]. In contrast, for extensive overlapped systems, a proper initial iterative vector, which can lead us to the correct result, is the key to ITTFA [9]. Zhu et al. [10] proposed to construct the initial vectors of ITTFA based on spectral feature of isoabsorptive point between components in the system for the resolution of kinetic-spectral data with an unknown kinetic model.

Since analytical objects, such as environmental, biological and proteomic samples, are becoming more and more complex, there always exist overlapping spectra in experimental spectra, no matter how advanced the analytical instrument are [11,12]. If they are similar in characteristic and behavior of chromatographic profiles, co-eluting will appear in the separation of two or more compounds. The overlapping (even embedded) chromatographic peaks will appear. It is necessary to resolve them for qualitative and quantitative analysis.

[*] Corresponding author. Tel.: +86 2165983987; fax: +86 2165983987.
E-mail address: lith@tongji.edu.cn (T. Li).

0039-9140/$ – see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.talanta.2005.04.033

People have made attempts in modifying and combining existent methods or finding new resolution method to settle the problem. For example, Mason et al. [13] proposed resolving factor analysis (RFA) and applied it for resolution of simulated kinetic and chromatographic data. Gemperline and Cash [14] proposed alternating least squares with penalty functions, which yielded improved results by incorporating soft constraints, for self-modeling curve resolution. Bi et al. [15] used independent component analysis (ICA) to resolve mixed IR spectra. Gao et al. [16] used component analysis based on kurtosis for resolution of overlapped peaks.

Constructing favorable initial vectors is the key to ITTFA, which can lead us to the correct result. A novel strategy of initial vectors construction to ITTFA was proposed in this paper.

EFA can provide us the chromatographic information of “eluting in and eluting out”. It can be used to determine the regions in which the peak maximum occurs. And, that in the initial vectors all values in these regions are set to 1, while remaining values are set to 0. The approach extends conventional initialization of setting one element in each vector. It is successfully applied to the simulated 2D data and the analytical GC/MS data.

2. Theory and algorithm

Assume a bilinear data matrix (**Y**) has been gained from hyphenated instrument (such as HPLC/DAD or GC/MS). The size of the matrix is $n \times m$, where n denotes the number of measurement time and m denotes the number of the corresponding wavelength (or m/z). The data matrix **Y** is allowed to be decomposed to the pure chromatography **C** and the pure absorption spectral matrix $\mathbf{S}^{\mathrm{t}}$. The relationship can be described by:

$$\mathbf{Y} = \sum_{i=1}^{\mathrm{Numc}} c_i s_i^{\mathrm{t}} + \mathbf{E} = \mathbf{C}\mathbf{S}^{\mathrm{t}} + \mathbf{E} \qquad (1)$$

where Numc represents the number of components, c_i (column vector) and s_i^{t} (row vector) represents the chromatography and the spectrum of the ith component, respectively. Superscript t denotes the transpose of a matrix (or a vector), **E** is the errors-related matrix.

2.1. Evolving factor analysis (EFA)

The idea, on which EFA is based, is to follow the change or the evolution of the rank of **Y** with progressing elution by rank analysis of the submatrices $\mathbf{Y}_i$ [2,8]. The following algorithm is based on a paper by Maeder [17]. First of all, submatrices $\mathbf{Y}_i$ are formed, which contain an ever-increasing number of spectra, starting with the first two spectra only and ending with the complete data matrix. Singular value decomposition (SVD) is applied to the submatrices and the logarithm of the eigenvalues (log(EV)) is plotted as a function of the number of spectra. This forward EFA-plot gives information about the appearance of components. Subsequently, submatrices are formed in a similar way, starting with the last two spectra. This results in a backward EFA-plot which gives information about the disappearance of components. Matching the appearance and disappearance of components from the EFA-plots, provides a rank map of the data matrix. This rank map is used to detect selective component windows. A selective component window is a part of the data matrix where one component of interest occurs.

2.2. Iterative target transformation factor analysis (ITTFA)

ITTFA [2] is an iterative method approximating the results from an initial vector. By using principal component analysis (PCA), the data matrix **Y** can be decomposed to two orthogonal matrices: $\mathbf{Y} = \mathbf{T}\mathbf{P}^{\mathrm{t}} + \mathbf{E}$. And **T** can be combined with **C** through a target transformation matrix **R** ($n \times m$), according to:

$$\mathbf{C} = \mathbf{T}\mathbf{R} \qquad (2)$$

R can be computed from the present target (the chromatography or the absorption spectrum of individual component). Taking chromatography for an example, provided only the chromatogram c_i of the ith component is considered, gives:

$$c_i = \mathbf{T} \times r_i \qquad (3)$$

Consequently, if by certain means the chromatogram c_i of the ith component is obtained, the ith column of the transformation matrix can then be calculated by:

$$r_i = (\mathbf{T}^{\mathrm{t}}\mathbf{T})\mathbf{T}^{-1} c_i \qquad (4)$$

However, c_i is unknown and is just the main target of this work. Here is the skeletal procedure of ITTFA:

i. Construct an initial iterative vector c_i^0.
ii. Following Eq. (4), r_i^0 is obtained; then substituting into Eq. (3) yields a new vector, c_i^1.
iii. Repeat step ii until $\left\| c_i^k - c_i^{k+1} \right\|$ is less than a given constant, and the c_i^k is the result.

Obvious, a favorable initial vector is the key of the ITTFA, which can lead us to the correct result. In our work, initial vectors are chosen according to the selective information of chromatography.

2.3. The construction approach of the initial vectors

EFA provides the chromatographic information of eluting in and eluting out. It can be used to determine the region in which the peaks maximum occurs. The initial vectors are constructed according to the information. The values in the determined region are set to 1, while remaining values are set to 0.

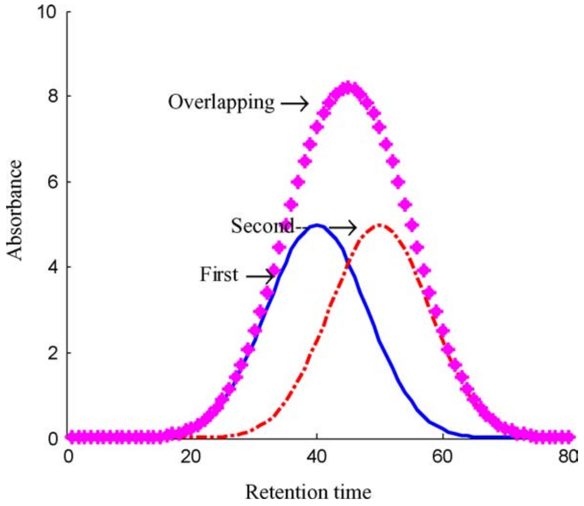

Fig. 1. Two-component simulated overlapping chromatograms. Solid line: the first component chromatogram; dashed line: the other chromatogram; symbol line: the overlapping chromatogram.

2.3.1. Partly overlapping peak

One case of 2D data is taken to illuminate how to construct initial vectors based on the chromatographic information by EFA. There is a two-component partly overlapping peak on the chromatographic dimension of the simulated data. The chromatographic profiles and the EFA-plot of the logarithm of the eigenvalues of the data are shown in Figs. 1 and 2, respectively.

It can be seen from Fig. 1 that the components are agreeable to the chromatographic property of ‘first in and first out’. In Fig. 2, the region between points C_1 and C_3 denotes a region in which the first component exists (C_1 is the point in which it starts to fill in and C_3 is the point in which it elutes out). The region between A_1 and A_4 denotes a region in which the summit span of one-component peak exists. It is the same case that the region between points C_2 and C_4 denotes a region in which the other component exists (C_2 is the point in which it starts to fill in and C_4 is the point in which it elutes out). And the region between A_3 and A_2 denotes a region in which the summit span of one-component peak exists.

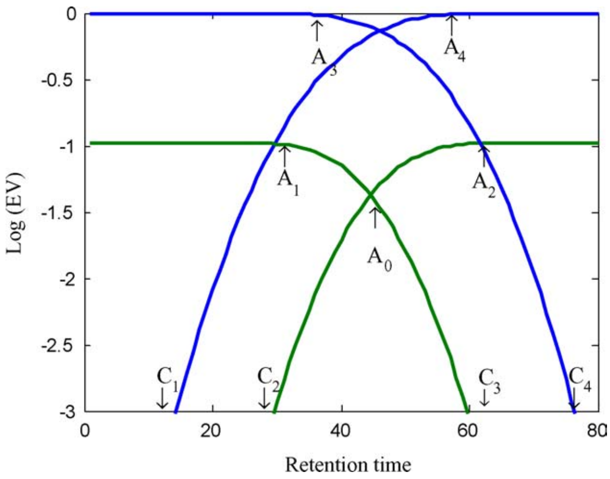

Fig. 2. The change of log(EVs) from EFA.

The construction of initial vectors to conventional ITTFA is to initialize the maximal element as 1, which is often determined by the method of needle search [18], the other elements of each vector are initialized 0. In practice, whether the result is correct or not depends on whether the position, where the peak maximum occurs, can be determined correctly. If there is divergence between the determined position and actual position, the resolution of ITTFA is not acceptable. We attempt to set more than one element, instead of only one element in each vector, as 1. For example, the elements in the region between A_1 and A_4 are initialized as 1, otherwise, those elements out of the region are initialized as 0; similarly, the elements in the region between A_3 and A_2 are initialized as 1, and the elements out of the region are initialized as 0. We realized that the size of the region had little impact on the resolved result. We can also choose the region between A_1 and A_0 as the selective region to construct the initial vectors, and the resolution is also reasonable. Constructing the initial vectors like that, the resolution can be improved. The results are reasonable even when the chromatographic resolution is 0.275, which can be seen from Table 1. It can overcome the divergence brought by determination of the peak position at maximum height in the resolution of overlapping chromatographic profiles, because it does not need to determine exactly where the peak position at maximum height is.

2.3.2. Embedded chromatography

To illustrate how to construct initial vectors based on the chromatographic information, two cases of 2D data which contain completely overlapping chromatographic profiles are simulated. The simulated overlapping chromatographic profiles and the EFA-plot of the logarithm of the eigenvalues of the simulated data are shown in Figs. 3 and 4, respectively.

There is a peak embedded in the other peak in Fig. 3, which does not agree to the chromatographic property of ‘first in and first out’. And the explanation of EFA-plot is different from that of partly overlapping chromatographic profiles. In Fig. 4, the region between points A_1 and A_2 denotes a region in which considerable capacity of the first component exists. It is the same case that the region between points A_1 and A_2 denotes a region in which the summit span of one-component peak exists. There is a flat region between B_1 and B_2 in which the summit span of narrow peak exists. It is specially noted that the region of B_1 and B_2 is always contained in the region between A_1 and A_2. We are able to construct the initial vectors of ITTFA according to this information to resolve the embedded chromatographic peak. The regions, such as the region between A_1 and A_2, the flat between B_1 and B_2, are what we want to determine. The elements in the determined region of initial vectors are set as 1, and the elements out of the region are set as 0.

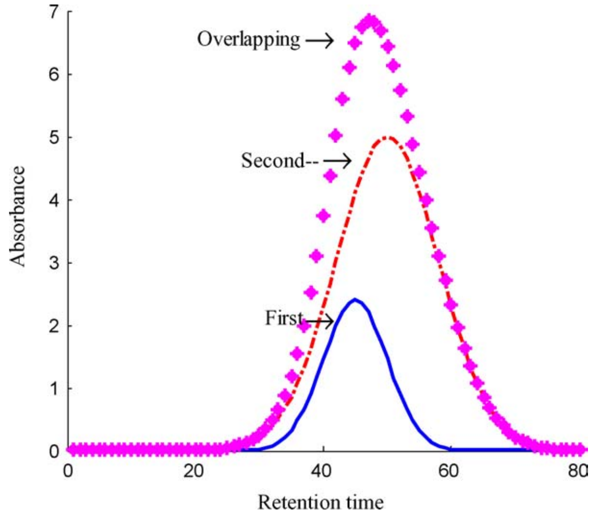

Fig. 3. Two-component simulated overlapping chromatograms. Solid line: the first component chromatogram; dashed line: the other chromatogram; symbol line: the overlapping chromatogram.

The construction of initial vectors is the key to ITTFA, Gemperline constructed iterative vectors by initializing as one maximal element as 1 and the other elements as 0 in each vector. Whereas, the position in which the maximal element exists is difficult to determine when the chromatographic profiles are overlapped severely and completely. Whether the determination is correct or not has a great impact on the resolution of ITTFA. A replaceable approach is proposed to construct the initial vectors based on chromatographic information provided by EFA. Several elements of each initial vector, instead of one element, are initialized as 1. A region, in which the summit span of one-component peak exists, is determined. All values in the region are set to 1, while the remaining values are set to 0.

Theoretically, if the elements in the region between the two points of half-height of chromatographic peak are initialized as 1, the computation converges fast and the resolution is reasonable. But the region is difficult to determine. So a region, which contains the summit span of one-component peak, is taken to replace it to construct initial vectors. This approach of constructing initial vectors to ITTFA has been applied to resolve the overlapping peaks in this paper. The results were reasonable.

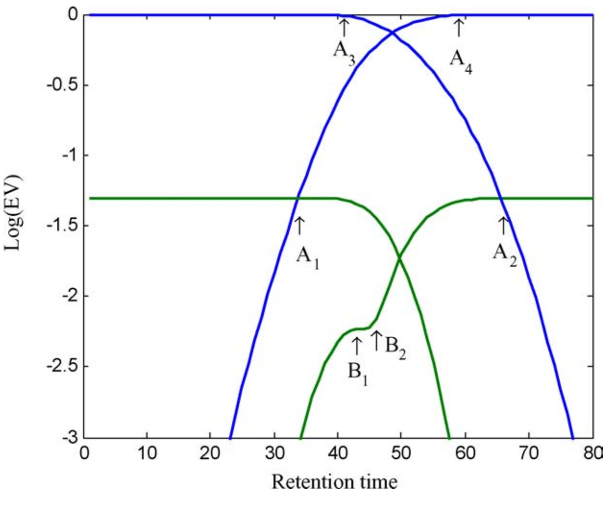

Fig. 4. The change of log(EVs) from EFA.

3. Experimental

3.1. Simulated 2D data

Gaussian profiles are used as models to simulate 2D data for theoretical studies. A 2D data matrix is reasonably modeled by generating Gaussian peaks on the first dimension (chromatographic profiles) and then made a cross product multiplication with Gaussian peaks generated on the second dimension (spectra). The 2D-simulated data of two components with partly overlapping and completely overlapping peaks in the chromatographic dimension are shown in Figs. 1 and 3, respectively.

3.2. Analytical data (GC/MS)

3.2.1. Materials

The sample of GUJINGGONG alcohol liquor (made in Bozhou, Anhwei province, China) was purchased from a supermarket.

3.2.2. Experimental condition

About 150 ml of GUJINGGONG liquor was treated with CH_2Cl_2 three times in turn by volume of 50, 30 and 20 ml, respectively, the extracted liquids were combined together and concentrated into about 20 ml; 3% Na_2CO_3 (10 ml) was added into the residual to back-extract, and then a small quantity of CH_2Cl_2 was added for laving, which was combined into the extracted liquid obtained in the former step. It was dried by using anhydrous Na_2SO_4, and concentrated for examination.

3.2.3. Instruments condition

A Hewlett-Packard 6890 gas chromatograph equipped with a HP5973 mass-selective detection system and a split–splitless injector was used for the analysis of the studied liquor. A non-polar fused-silica capillary column, HP-5 (30 m × 25 mm i.d.) and 0.25 μm film thickness supplied by Agilent Co., was employed, with Helium as carrier gas at 1 ml/min. The column temperature was maintained at 50 °C for 5 min, then programmed at 5 °C/min to 180 °C, held 10 min and programmed at 10 °C/min to 220 °C, held 10 min. The injector port was maintained at 250 °C and a 2 μl volume was injected in the split (1/40) mode. Mass spectrometer parameters: electron impact ionization mode with 70 eV energy, *m*/*z* 50–550; ion source temperature, 250 °C; MS Quad temperature, 150 °C; scan rate, 0.1 s per scan; electron multiplier voltage, 1000; solvent delay, 3 min.

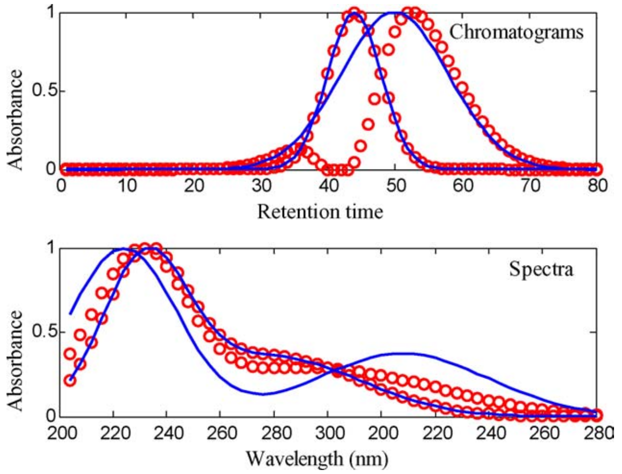

Fig. 5. Resolution of the simulated 2D data by traditional ITTFA. Solid line: the original curves; dashed line: the resolved curves.

3.3. Software

All the calculations were performed by using programs written by the authors in the Matlab environment (The Mathworks, Natick, USA), running on PC with Intel (R) Pentium4 CPU 2.00 GHz and 256 MB RAM. The library searches and spectral matching of the resolved pure components were conducted on the National Institute of Standards and Technology MS database (NIST 98).

4. Results and discussion

4.1. Simulated 2D data

The simulated 2D data was resolved by conventional ITTFA and modified ITTFA, respectively. The resolutions which contain both chromatographic profiles and spectra are shown in Figs. 5 and 6, respectively. In order to estimate the resolution, both the resolved and original curves are normalized by the maximum.

There is little discrepancy, both in peak position and in peak shape, between the original curves and the resolved curves resolved by modified ITTFA. This could be seen from Fig. 6. While the resolution by conventional ITTFA is not acceptable, and there is considerable discrepancy between the resolved and original curves, which could be seen from Fig. 5.

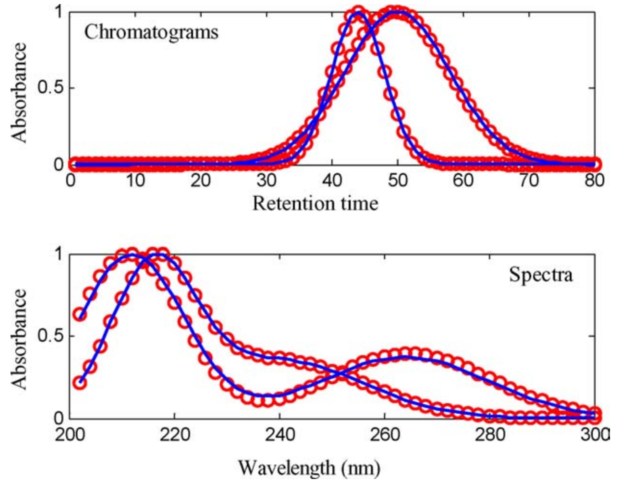

Fig. 6. Resolution of the simulated 2D data by modified ITTFA. Solid line: the original curves; dashed line: the resolved curves.

Table 1
The influence of chromatographic resolution (R) on modified ITTFA

R_chrom	Corr(C1)	Corr(C2)	Corr(S1)	Corr(S2)
0.50	1	1	1	1
0.45	1	1	0.9999	1
0.40	1	1	0.9996	1
0.35	0.9999	0.9999	0.9985	1
0.30	0.9996	0.9996	0.995	1
0.275	0.9992	0.9993	0.9915	1
0.25	0.9987	0.9987	0.9859	1

4.2. The influence of chromatographic resolution and heteroscedastic noise on the resolution

In order to estimate the resolution, we choose the correlation coefficient (correlations) to compare the resolved curves and the original curves. The influences of chromatographic resolution (R_chrom) and heteroscedastic noise on the resolved results were shown in Tables 1 and 2, respectively.

It could be seen from Tables 1 and 2 that reasonable and acceptable results [0] could be obtained with our approach when the R_chrom > 0.25 or noise level is no more than 0.06, and most correlations of the resolved spectra and the original spectra are not less than 0.99.

4.3. Analytical data

Under the experimental condition mentioned in Section 3.2, the total ion chromatograms (TIC) of GUJINGGONG liquor were obtained. Part data ranging from 6.38 to 17.80 min were shown in Fig. 7.

It can be seen from the TIC that most of the chromatograms separated ideally and could be analyzed (qualitative analysis) directly by MSD data analysis software and NIST98 MS database. There were about 11 alcohols, 3 aldehydes and 22 esters that could be identified. However, there were

Table 2
The influence of heteroscedastic noise on modified ITTFA

Noise level	Corr(C1)	Corr(C2)	Corr(S1)	Corr(S2)
0.01	1	1	1	1
0.02	0.9999	1	1	1
0.03	0.9998	1	1	1
0.04	0.9998	1	0.9999	1
0.05	0.9992	0.9984	0.9667	1
0.06	0.9985	0.9999	0.9999	1

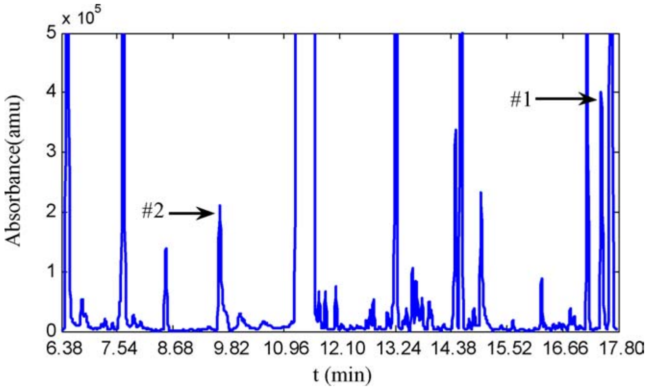

Fig. 7. Total ionic chromatograms of liquor from 6.38 to 17.80 min.

also some overlapping peaks in the TIC, the matches from direct searching with the NIST MS database were quite low for these chromatographic peaks. If these overlapping peaks were not resolved, the simple search with the database would fail, since the mass spectra of mixtures measured could not get a good match with that of a pure component in the NIST MS database. Furthermore, since a 2D data obtained by mass spectral measurement unavoidably contained peaks associated with base line and noise, it was difficult to estimate low content components correctly with the database.

We took the GC/MS data from 8.46 to 8.60 min and the data from 17.38 to 17.52 min as examples to validate the construction initial vectors to ITTFA.

4.3.1. Partly overlapping chromatographic profiles

The #1 peak, which is from 17.38 to 17.52 min in Fig. 7, looks like a one-component peak. But it is actually a two-component peak which can be tested by cross-validation. The resolved chromatograms and mass spectra are shown in Figs. 8 and 9, respectively.

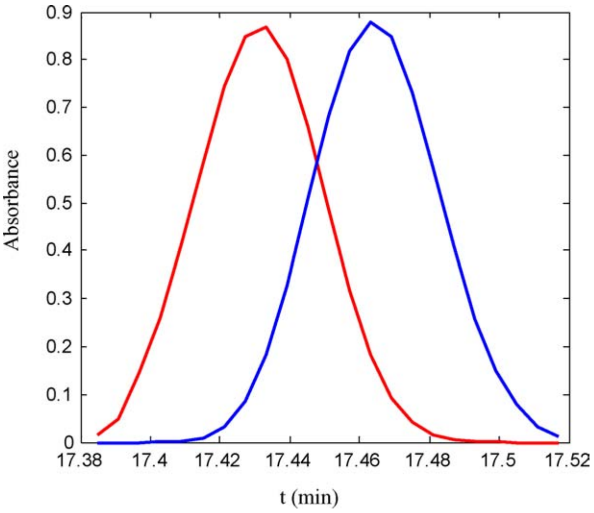

Fig. 8. The resolved chromatographic profiles of the #1 peak.

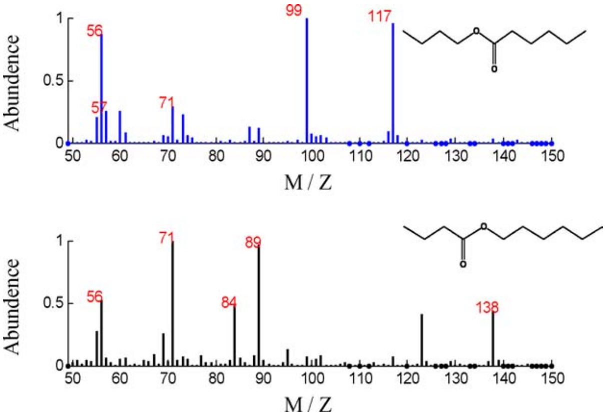

Fig. 9. The resolved mass spectra of the #1 peak.

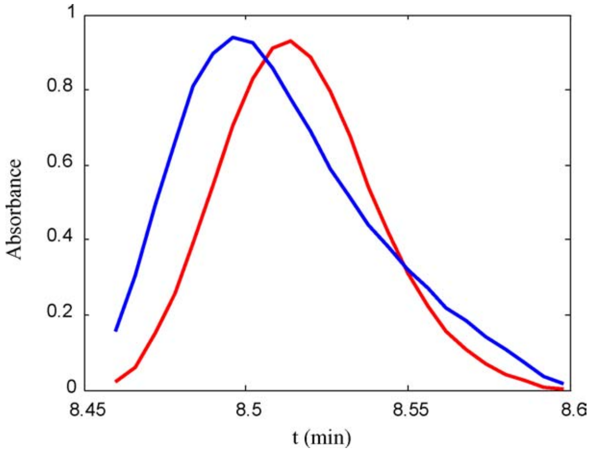

Fig. 10. The resolved chromatographic profiles of the #2 peak.

With modified ITTFA [0], the match quality of $C_{10}H_{20}O_2$ (butyl caproate) increased from 90.9 to 94.5% in the NIST MS database, and that of the isomeric compound (butyric acid, hexyl ester) increased from 86.0 to 91.6%.

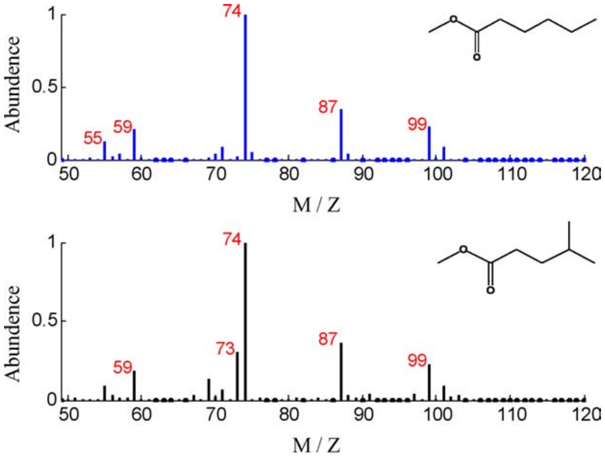

Fig. 11. The resolved mass spectra of the #2 peak.

4.3.2. Completely overlapping chromatographic profiles

The #2 peak, which is from 8.46 to 8.60 min in Fig. 7, looks like a one-component peak, but the result of cross-validation showed that it is actually a two-component chromatography. The resolved chromatograms and mass spectra were shown in Figs. 10 and 11, respectively.

After resolved using modified ITTFA, the match of $C_7H_{14}O_2$ (hexanoic acid, methyl ester) enhanced from 91.3 to 92.0% in the NIST MS database, and that of the isomeric compound (pentanoic acid, 4-methyl-, methyl ester) enhanced from 80.9 to 81.6%.

5. Conclusions

When one recognizes that the construction of the initial iterative vectors is the key to ITTFA, a natural consequence is to improve the initial vectors just as what were proposed in this paper. The merits of our tentative approach are: the region in which the summit span of one component exists is easier to determine than the peak position at maximum height. The resolution is correct; moreover, it can be applied in the resolution for embedded chromatography.

Acknowledgement

The authors thank the National Nature Science Foundation of China (Grant No. 20275026) and the Outstanding Adult-Young Scientific Research Encouraging Foundation of Shandong Province (Grant No. 2005BS1004) for financial support.

References

[1] P.J. Gemperline, J. Chem. Inf. Comput. Sci. 24 (1984) 206.
[2] Y.Z. Liang, R.Q. Yu, Handbook of Analytical Chemistry, Chemometrics, vol. 10, Chemical Industry Press, Beijing, 2000, p. 260.
[3] P.J. Gemperline, Anal. Chem. 58 (1986) 2656.
[4] K.A. Lilley, T.E. Wheat, J. Chromatogr. B: Biomed. Appl. 683 (1996) 67.
[5] E. Comas, R.A. Gimeno, J. Ferré, R.M. Marcé, F. Borrull, F. Xavier Rius, Anal. Chim. Acta 470 (2002) 163.
[6] M.J. Rodrı̆guez-Cuesta, R. Boqué, F. Xavier Rius, Anal. Chim. Acta 476 (2003) 111.
[7] P.V. van Zomeren, H. Darwinkel, P.M.J. Coenegracht, G.J. de Jong, Anal. Chim. Acta 487 (2003) 155.
[8] J. Xu, Z. Guo, Y.Z. Liang, R.Q. Yu, J. Chemomet. 10 (1995) 63.
[9] Y.Z. Liang, Y.L. Xie, R.Q. Yu, Acta Chim. Sin. 49 (1991) 394.
[10] Z.L. Zhu, W.Z. Cheng, Y. Zhao, Chemomet. Intell. Lab. Syst. 64 (2002) 157.
[11] F. Gong, Y.Z. Liang, Q.S. Xu, F.T. Chau, J. Chromatogr. A 905 (2001) 193.
[12] X.G. Shao, M.Q. Li, Chin. J. Chromatogr. 19 (2001) 184.
[13] C. Mason, M. Maeder, A. Whitson, Anal. Chem. 73 (2001) 1587.
[14] P.J. Gemperline, E. Cash, Anal. Chem. 75 (2003) 4236.
[15] X. Bi, T.H. Li, L. Wu, Chem. J. Chin. Univ. 25 (2004) 1023.
[16] H.T. Gao, T.H. Li, K. Chen, X. Bi, S.F. Lin, Chin. J. Anal. Chem. 32 (2004) 993.
[17] M. Maeder, Anal. Chem. 59 (1987) 527.
[18] A. de Juan, B. van den Bogaert, F. Cuesta Sanchez, D.L. Massart, Chemomet. Intell. Lab. Syst. 33 (1996) 133.