Available online at www.sciencedirect.com
SCIENCE DIRECT®
Talanta 66 (2005) 65–73
Talanta
www.elsevier.com/locate/talanta

Overlapping spectra resolution using non-negative matrix factorization

Hong-Tao Gao[a,b], Tong-Hua Li[a,*], Kai Chen[a], Wei-Guang Li[a], Xian Bi[a]

[a] *Department of Chemistry, Tongji University, 1239 Siping Road, Shanghai 200092, China*
[b] *Department of Chemistry, Jining Teachers College, Jining Shandong 272025, China*

Received 7 May 2004; received in revised form 12 September 2004; accepted 23 September 2004

Abstract

Non-negative matrix factorization (NMF), with the constraints of non-negativity, has been recently proposed for multi-variate data analysis. Because it allows only additive, not subtractive, combinations of the original data, NMF is capable of producing region or parts-based representation of objects. It has been used for image analysis and text processing. Unlike PCA, the resolutions of NMF are non-negative and can be easily interpreted and understood directly. Due to multiple solutions, the original algorithm of NMF [D.D. Lee, H.S. Seung, Nature 401 (1999) 788] is not suitable for resolving chemical mixed signals. In reality, NMF has never been applied to resolving chemical mixed signals. It must be modified according to the characteristics of the chemical signals, such as smoothness of spectra, unimodality of chromatograms, sparseness of mass spectra, etc. We have used the modified NMF algorithm to narrow the feasible solution region for resolving chemical signals, and found that it could produce reasonable and acceptable results for certain experimental errors, especially for overlapping chromatograms and sparse mass spectra. Simulated two-dimensional (2-D) data and real GUJINGGONG alcohol liquor GC–MS data have been resolved soundly by NMF technique. Butyl caproate and its isomeric compound (butyric acid, hexyl ester) have been identified from the overlapping spectra. The result of NMF is preferable to that of Heuristic evolving latent projections (HELP). It shows that NMF is a promising chemometric resolution method for complex samples.
© 2004 Elsevier B.V. All rights reserved.

Keywords: Non-negative matrix factorization (NMF); Chemometric method; Resolution; Overlapping

1. Introduction

Principal Component Analysis (PCA) is a basic technique in chemometrics for multi-variate data analysis, which can find several latent variables to extract information from the data matrices. The measured data matrix can be factorized into two factor matrices (score matrix and loading matrix) by PCA. One major problem in applications of PCA is that the data in factor matrices are both positive and negative, and the data are represented as linear combinations of those variables with positive and negative coefficients. In many cases, the negative components contradict physical realities, so the results with negative intensities cannot be reasonably interpreted.

Some multi-variate data resolution techniques, for example, evolving factor analysis (EFA) [2,3], iterative target transformation factor analysis (ITTFA) [2,4–6], generalized rank annihilation factor analysis (GRAFA) [7,8], window factor analysis (WFA) [9,10], Heuristic evolving latent projections (HELP) [2,11], etc., have made great efforts and developed rapidly in recent years. As a result, hyphenated instruments combined with chemometrics algorithms can make it possible to quantify the complicate analytical system clearly. The methods have been successfully applied to many fields [2–11]. However, all the methods mentioned above have a common problem. When they are applied to resolving overlapping spectra, the degree of peak overlapping must be

* Corresponding author. Tel.: +86 21 65983987; fax: +86 21 65983987.
E-mail address: lith@tongji.edu.cn (T.-H. Li).

0039-9140/$ – see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.talanta.2004.09.017

within a certain limit. If the peaks are overlapping strongly or completely, the resolution results will be not acceptable. Some efforts and attempts have been made to improve the resolution. Non-negative factor analysis was purposed by Paatero and Tapper to cure the resulting negative factors when they performed factor analysis on environmental data. They proposed to use alternating least squares (ALS) and positive matrix factorization (PMF) to resolve the problem [12]. Garrido Frenich et al. applied orthogonal projection approach (OPA), PMF and ALS to resolving multi-component peaks [13]. We have made some successful attempts to extract pure components information from the embedded peaks in chromatogram under certain constraints [14]. However, it is still a problem to resolve overlapping (in particular severely or completely overlapping) signals efficiently and accurately.

Non-negative matrix factorization (NMF) was introduced by Lee and Seung in a paper on unsupervised learning in 1997. They subsequently developed a simple algorithm for computing the factorization [1,15,16], which was applied in image analysis [17–20]. The algorithmic details will be introduced in the following sections of this paper. Meanwhile, the algorithm and its applications have already been advanced since they were proposed. For instance, Feng et al. proposed a method called local non-negative matrix factorization (LNMF) for learning spatially localized, parts-based subspace representation of visual patterns [21]. And Guillamet et al. introduced a weighted non-negative matrix factorization (WNMF) for image classification to improve the NMF capabilities of representing positive local data [22]. Non-negative matrix factorization is a technique for dimensionality reduction by placing non-negativity constraints on the matrix. Its idea can be interpreted as decomposing a non-negative matrix V into two non-negative factorization matrices W and H. It is assumed that columns in W are far fewer than those in V, and the rows in H are far fewer than those in V, so the approximation can succeed only if it discovers latent structure in the matrix. In other words, NMF is an analytical method for latent variables and can be introduced into multi-variable analysis like PCA. It can possibly overcome the problem that the basis vectors have negative elements.

To our knowledge, NMF has not been applied in chemical signal resolution. In this paper, we applied NMF to resolving simulated two-dimensional (2-D) data (chromatograms were in one dimension, spectra were in the other) and actual experimental data. NMF was modified according to the characteristics of chemical signals, such as smoothness of spectra, unimodality of chromatograms and sparseness of mass spectra. Our results showed that a modified NMF could be introduced to resolve chemical mixed signals, especially overlapping chromatograms and sparse mass spectra. For the resolution of GUJINGGONG alcohol liquor GC–MS data, the result of NMF is preferable to that of Heuristic evolving latent projections. It shows that non-negative matrix factorization will be a promising method for the resolution of chemical mixed signals.

2. Theory and algorithm

2.1. Non-negative matrix factorization

Non-negative matrix factorization is a method to obtain a representation of data using non-negativity restraint. The constraint leads to a parts-based representation because it allows only additive, not subtractive, combinations of the original data [1]. Suppose a bilinear matrix is expressed by $V(n \times m)$, where each column is an n-dimensional non-negative vector of the original matrix (m vectors). It is possible to find two new matrices (W and H) in order to approximate the original analytical matrix $V_{i\mu} \approx (W \times H)_{i\mu} = \sum_{a=1}^{r} W_{ia} H_{a\mu}$. The dimensions of the factorized matrices W and H are $n \times r$ and $r \times m$, respectively. Usually, r is the number of principal components. Each column of matrix W contains a basis vector while each row of H contains the weights needed to approximate the corresponding column in V using the basis from W. It is well known that PCA can decompose the matrix into two factors. However, in the two factorized matrices there are positive and negative entries simultaneously, and these negative components make the result often unacceptable in chemical meanings. In contrast to PCA, NMF does not allow negative entries in the factorized matrices W and H, permitting the combination of multiple basis matrices to represent spectra.

In order to estimate the factorization matrices, an objective function has to be defined. A possible objective function is given by

$$F = \sum_{i=1}^{n} \sum_{\mu=1}^{m} [V_{i\mu} \log (WH)_{i\mu} - (WH)_{i\mu}] \tag{1}$$

This objective function can be related to the likelihood of generating the signals in V from the bases W and encodings H. An iterative approach to reach a local maximum of this objective function is given by the following rules [1]:

(i) Initialize matrices W and H randomly under non-negative condition.

(ii) Calculate the new value of W by H:

$$W_{ia} = W_{ia} \sum_{\mu} \frac{V_{i\mu}}{(WH)_{i\mu}} H_{a\mu} \tag{2}$$

(iii) Normalize W column wisely:

$$W_{ia} = \frac{W_{ia}}{\sum_j W_{ja}} \tag{3}$$

(iv) Calculate the new value of H by W resulted from (ii):

$$H_{a\mu} = H_{a\mu} \sum_{i} W_{ia} \frac{V_{i\mu}}{(WH)_{i\mu}} \tag{4}$$

(v) Repeat from (ii) to (iv) until it converged.

The iterative process stops once the maximum number of iterations reaches or the residual sum of squares between data matrix (V) and reconstituted data matrix ($W \times H$) drops below a certain threshold.

2.2. Characteristics of NMF

Non-negative constraints have been widely used in chemometrics for curve resolution. Conventional methods, such as force to zero and non-negativity least squares (NNLS) are introduced to constrain the calculation results in the iterative procedure or optimization algorithm. It can be called "ex-constraint", while it is not the case in non-negative matrix factorization algorithm. NMF can directly obtain a representation of nonnegative data by using multiplies update rules, which can be regarded as "in-constraints". NMF implementation is based on elements, not on vectors. It is different from the conventional factor analysis. Just due to it, NMF can learn the local representations of data, and the factorization results have realistic physical meaning and can be directly understood without additional operations, such as "rotational transformation".

We note that the algorithm contains a "serious" feature: the algorithm does not allow that any element of matrix W or matrix H becomes non-zero if one element of original matrix V becomes zero. Due to the feature, the iteration cannot converge or reach correct results. But it will not be a problem when NMF is applied for chemical curve resolution. If the initial matrices W and H do not contain zero, the feature can be avoided. In addition, the smoothing (see the following section) can be contributed to settling down the shortcoming.

Another feature is that the speed of the algorithm is slow. Typical, some thousands or even tens of thousands of steps will then be needed for "good" convergence to the optimum of objective function F. Practically, it is impossible to expect the residual sum of squares between the original data (V) and the reconstituted data (WH) near zero. But, it is not difficult to reach an acceptable threshold under certain experimental error, for example, 10^{-6}. In that case, hundreds of steps will produce satisfactory results. And we take 400 as the maximum number of iterations in practice.

In summary, there are several issues in NMF discussed above, and the original NMF cannot be introduced into analytical chemistry for curve resolution. Only if nature restraints of chemical curves, such as smoothness of spectra, unimodality of chromatograms, sparseness of mass spectra, etc., are imposed, can NMF be used for curve resolution. This will be discussed in the following sections.

2.3. Algorithm improvement

With the development of hyphenated instruments in the latest decades, such as HPLC-DAD, GC–MS, etc., two-dimensional data become available. The measured two-dimensional data (GC–MS data were taken as an example in this paper) usually can be expressed by a matrix $X(m \times n) = C(m \times r) \times S(r \times n)$, where X is a bilinear matrix, C represents chromatograms measured at m retention time points and S represents spectra measured at n signal channels (wavelength or m/z). The data often contain abundant qualitative and quantitative information and non-negativity is the basic property of the chemical data. There are at least three issues we must address when NMF is introduced to resolve the chemical data. The first is that chromatograms or spectra are often smooth, in other words, the move trend of the curves is gentle, except for mass spectra and IR spectra. The second is the unimodal nature of chromatograms, for there is only a maximum in the chromatographic profiles of each pure component. The last is the sparseness of mass spectra, which is seldom considered. Sparseness means the absorbance is zero at some mass values of m/z (mass/charge), and this is the basic property of mass spectra. So, when NMF is introduced to resolve the chemical data matrix, it must be modified according to the characteristics of signals: smoothness and unimodality of chromatograms and sparseness of mass spectra.

One modification of NMF is to introduce curve smoothing in the algorithm. The calculation of W and H is carried out element by element in the iterative process. Because the values of W and H are initialized randomly, the elements of W and H at the beginning have no relationship each other. So W and H obtained in the iterative process are zigzag but not smooth, which contradict the smooth property of continuous spectral curves. Measures should be taken to ensure that W and H obtained in iterative steps have "spectra" shape. So five-point mean curve smoothing was integrated into the iterative procedure in order to get reasonable and acceptable resolution. For mass spectra, due to the sparse character of mass spectra, the curve smoothing was unnecessary.

Further modification of NMF is to impose unimodality constraints of chromatograms. The resulting chromatographic profiles (C) were inspected to find the global peak maximums, one for each profile. Searching in both the forward and backward directions from the global peak maximums for each constituent, unimodality constraints must be added at point $i+1$, $C_{i+1,\mu} > C_{i,\mu}$, or at point $i-1$, $C_{i-1,\mu} > C_{i,\mu}$, if local maximums were encountered, for example, $C_{i,\mu} < C_{i+1,\mu}$ or $C_{i,\mu} < C_{i-1,\mu}$, then $C_{i+1,\mu} = C_{i,\mu}$, or $C_{i-1,\mu} = C_{i,\mu}$.

The last modification of NMF is to impose constraints of sparseness for mass spectra. The data matrix obtained from instrument coupled with mass detection (such as GC–MS) is sparse. This means that there are many zero values in mass spectra direction. In this case, the resolution of NMF will get absurd numerical values because data divided by zero occurs in the iterative process. What we have done to deal with the problem was to set $\frac{V_{i\mu}}{(WH)_{i\mu}} = 0$ when $(WH)_{i\mu} = 0$ appeared.

After these three modifications, NMF could be applied to resolving chemical spectra data matrix. It should be noted that NMF cannot give a unique solution when it is used to decompose a matrix into two factorization matrices, and the

obtained results are in a region. The solutions region varies now and then, depending on the degree of peaks overlapping, so these make the factorization matrix hardly interpretable. The algorithm improvements proposed in this section not only help us to obtain a meaningful solution but also to cut down the feasible region. This will be discussed in detail in the later section.

3. Experimental

3.1. Data

3.1.1. Two-dimensional data simulated by Gaussian profiles

Gaussian profiles were used as models to simulate two-dimensional data for theoretical studies. A 2-D data matrix was reasonably modeled by generating Gaussian peaks on the first dimension (several vectors, for example, two vectors) and then making a cross product multiplication with Gaussian peaks generated on the second dimension (other vectors).

Three cases of HPLC-DAD data matrices of a two-component system were simulated. These cases were different in chromatographic dimension, which was partially, severely and completely overlapping peaks, respectively.

Two-dimensional MS data matrices were modeled by extracting mass spectra of ethylbenzene and *p*-xylene from the NIST MS database and then a cross product multiplication with two-component peaks was made to produce a 2-D data matrix.

In order to compare the results with the original data, the columns of the simulated and real data were both normalized by dividing by the maximum element of each column.

3.1.2. Real experimental data (GC–MS)

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

3.1.2.2. Experimental condition. About 150 ml of GUJINGGONG liquor was treated with CH_2Cl_2 three times in turn by volumes of 50 ml, 30 ml and 20 ml, respectively. The extracted liquids were combined together and concentrated to about 20 ml; 3% Na_2CO_3 (10 ml) was added to 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.

3.1.2.3. Instrumental 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 V; solvent delay, 3 min.

3.2. 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 M 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. Non-uniqueness

NMF implementation is based on elements. Non-uniqueness exists in NMF, which is similar to rotational ambiguity of factor analysis. Before the results of NMF are demonstrated, the issue of multi-solution may be discussed at first. Because the restriction of NMF is only non-negativity, which is not a strong constraint, one will obtain different solutions from the same original matrix in different runs. This issue does not affect the explanation of the results when NMF is used for image analysis, and it has been seldom mentioned in literature. David and Victoris discussed about the assumptions and conditions when NMF could give a correct decomposition [23].

When NMF is used to resolve chemical spectra, the results may be linear combinations of pure components spectra. For example, the obtained solutions of W are linear combinations of the pure components (C), as denoted in the equation: $W = C \times B$, in which B is a transformation matrix ($r \times r$). If B is not a diagonal matrix, W is a linear combination of C, as shown in Fig. 1. There is a feasible region obtained from the resolution of two-component overlapping chromatograms by NMF, in which all the solutions satisfy the constraint of non-negativity.

There is another represent format of multiple solutions in the resolution of the chemical curve. W and H obtained from NMF sometimes may be zigzag and not smooth. The reason is that the estimation of W and H is carried out element by element in an iterative process. Because the values of W and H are usually initialized randomly, the points of W and H at the beginning have no relationship with each other. Zigzag results are also consistent with the constraints; and the residual sum of squares between the original data matrix (V) and

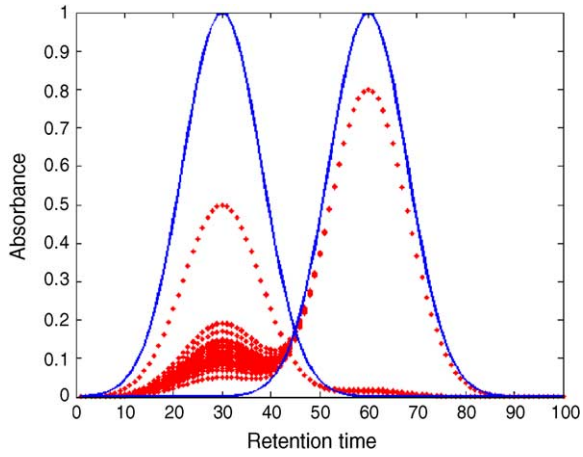

Fig. 1. Feasible solutions obtained from NMF. The solid lines denote the simulated chromatograms; the dot lines denote the resolved chromatograms.

the reconstructed data matrix ($W \times H$) also may drop below a certain threshold. But it obviously contradicts the smooth property of continuous spectra curve, except for mass spectra. Zigzag results are illustrated in Fig. 2.

As mentioned in Section 2, in order to improve the acceptability and reliability of solutions, we propose to improve the NMF algorithm by imposing the character of chemical spectra, such as smoothness of spectra, unimodality of chromatograms and sparseness of mass spectra. After the three modifications are taken, NMF can be applied to resolving chemical spectra data matrices. It should be noted that the modified NMF generally cannot give a unique solution when it is used to resolve chemical spectra. However, in most cases the obtained results are in a limited region, which is acceptable under certain experimental error (see the next section).

We also find that there are two factors, which affect the size of the feasible region in mixture components analysis.

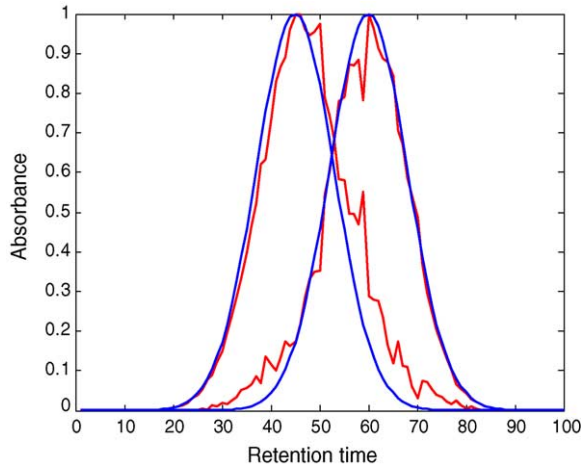

Fig. 2. Zigzag curves obtained from NMF. The zigzag curves denote the resolved results and the smooth curves denote the original data.

One factor is the correlation between the chemical spectra. If the spectra are uncorrelated or mutually independent, the multiple solutions obtained from NMF will be in a very narrow region, which is acceptable under certain experimental error. Most mass spectra satisfy the constraint of independence as exactly as possible. If the data matrix contains one dimension of mass spectra, the resolution will be nearly unique. The other factor is the degree of chromatograms overlapping. Practically, the more the degree of overlapping is, the narrower the feasible region will be gained when NMF is used to decompose a mixed spectral matrix. It makes NMF a powerful tool to resolve severely overlapping peaks, and we will illustrate it in detail in the next sections.

From the above discussions, we can conclude that if the constraints of chemical characteristics are imposed, the problems of multi-solution may be partially solved. Moreover, if one-dimensional data of the matrix are uncorrelated or mutually independent, the multiple solutions will be in a very narrow region, which is acceptable under certain experimental error. And for chromatograms, if two peaks overlap strongly, the results are also still acceptable. So we limit the use of NMF to LC, GC, MS, IR and a part of UV spectra. Moreover, we can introduce further constraints, such as the eluting sequence of chromatograms of different components, which will be discussed in subsequent work.

4.2. Simulated HPLC-DAD data matrix

The advent of hyphenated chromatography-detection systems, such as HPLC-DAD, makes more powerful analyzing approaches in chromatography possible. These hyphenated techniques, being capable of generating huge amounts of 2-D data, allow qualitative and quantitative identification from spectral properties in addition to retention time. Thus, they are extensively applied to many fields [24–29]. However, incomplete separation or overlapping of chromatographic peaks is still a likely occurrence for a complex sample. Consequently, it is an eye-catching task to develop a good resolution in hyphenated chromatography-detection systems in practice.

Three cases of simulated HPLC-DAD data matrices of two-component systems (each chromatogram contains one Gauss peak and each spectrum consists of two Gauss peaks) are discussed, with partially, severely and completely overlapping peaks in the chromatograms, respectively. The simulated spectra and the resolved spectra are shown in Figs. 3–5.

The first case is that of partially overlapping peaks in chromatograms and completely overlapping peaks in spectra, as shown in Fig. 3. The two original spectra have the same peak widths at half height and the degree of chromatographic resolution is 1. When the three improvements are implemented, NMF could give a narrow feasible region in which the resolution results are acceptable under certain experimental error, as the conventional resolution methods (such as EFA, HELP, etc.) do. But it exceeds them when the degree of chromato-

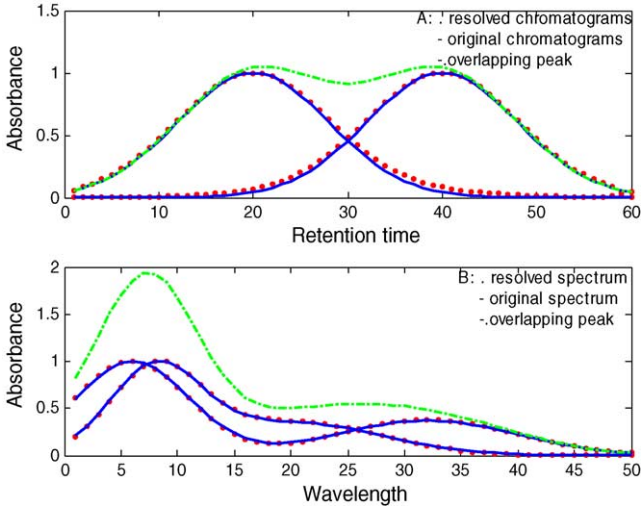

Fig. 3. Simulated and resolved chromatograms (A) and spectra (B) of the 2-D data. The solid lines denote the simulated chromatograms and spectra; the dot lines denote the overlapping curves, while the dashed lines represent the resolved chromatograms and spectra by the NMF.

graphic resolution is smaller than 0.5, NMF could still give reliable results, while the other methods mentioned above may not give correct answer.

A particular case is showed in Fig. 4. There are two peaks with the same peak width at half height, and they were severely overlapping ($R = 0.1$) in the chromatogram and in the spectrum. It appears when chromatographic profiles of two compounds are of similar peak-height and almost of the same retention time, and it is quite a troublesome issue for analysts. The solutions of NMF are reasonable and acceptable while the other methods mentioned above cannot give acceptable resolutions.

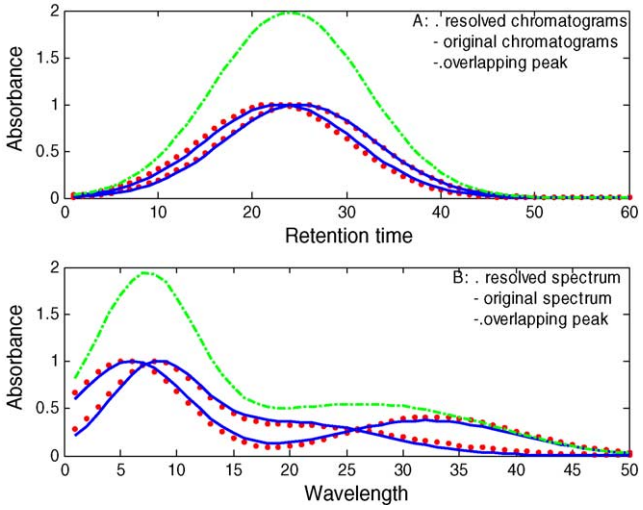

Fig. 4. Simulated and resolved chromatograms (A) and spectra (B) of the 2-D data. The solid lines denote the simulated chromatograms and spectra; the dot lines denote the overlapping curves, while the dashed lines represent the resolved chromatograms and spectra by the NMF.

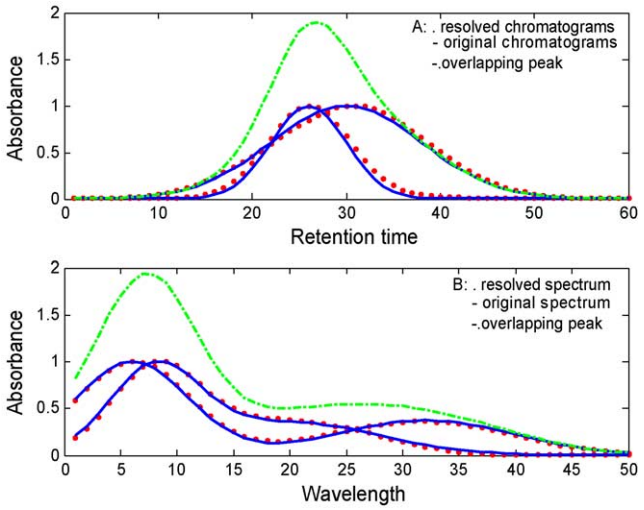

Fig. 5. Simulated and resolved chromatograms (A) and spectra (B) of the 2-D data. The solid lines denote the simulated chromatograms and spectra, the dot lines denote the overlapping curves, while the dashed lines represent the resolved chromatograms and spectra by the NMF.

The last case is completely overlapping peaks in chromatograms, also called embedded peaks or a peak in another peak, where the assumption of "first-in-first-out" is not satisfied [27]. No matter how advanced the analytical instruments and methods used are and how good the experience of the operator is, this possibly exists in the analytical data. It creates difficulty for analysts, and how to treat the particular elution pattern still remains a problem. One distinct embedded case is shown in Fig. 5. The maximum of two completely overlapping chromatograms does not appear at the same retention time. There are no appropriate methods at present to deal with the cases efficiently. NMF can succeed in getting reasonable and acceptable resolution results, which can be seen from Fig. 5.

As we have mentioned, NMF cannot give a unique resolution in most cases, and the results are in a feasible region in which all solutions satisfy the non-negative constraint. The feasible region is related to the overlapping degree of the chromatograms and the nature of the spectra.

We have performed NMF 100 times to resolve the same case where the chromatographic profiles are completely overlapped, and obtained acceptable resolutions in most cases, while occasionally we get unreliable results. The results are shown in Fig. 6. Though one does not get a unique solution, the obtained solutions are acceptable according to position and shape being very similar to the original ones and it was not difficult to identify the component.

4.3. Simulated MS data

The mass spectrum technique is applied more and more widely in analytical chemistry and other fields. There are many zero values in mass spectrum, which can be called sparse. Two-dimensional MS data were simulated

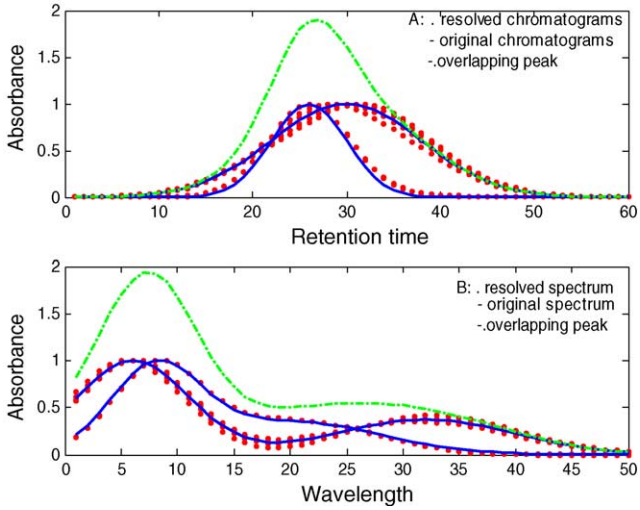

Fig. 6. The resolutions of NMF by carried out 100 times.

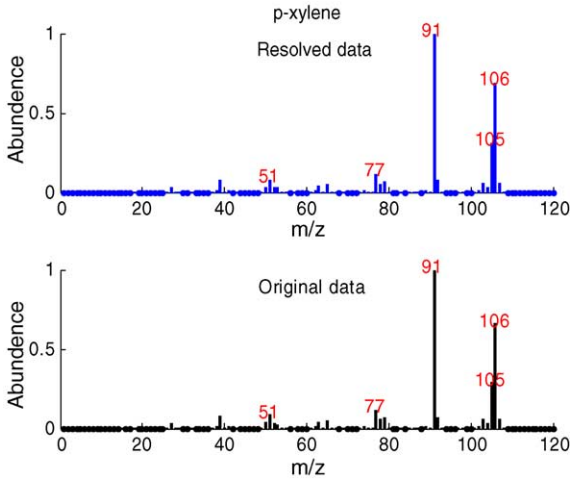

Fig. 8. Mass spectrum of *p*-xylene. The resolved spectrum is shown in the top and the original spectrum at the bottom.

to validate the application of NMF in sparse spectra data resolution.

The mass spectrum of ethylbenzene is quite similar to that of *p*-xylene, and it is difficult to identify them when ethylbenzene and *p*-xylene are mixed together at different proportions. NMF could succeed resolving the mixture of ethylbenzene and *p*-xylene, the resolution of the mass spectrum is acceptable. The matches of ethylbenzene and *p*-xylene are 98.1% and 97.2%, respectively, when searching automatically in the NIST MS database. The results are shown in Figs. 7 and 8, respectively.

4.4. Real GC–MS experimental data

Under the experimental condition mentioned in Section 3, the GC–MS chromatograms of GUJINGGONG liquor was obtained. Part data ranging 6.38–17.80 min were shown in Fig. 9.

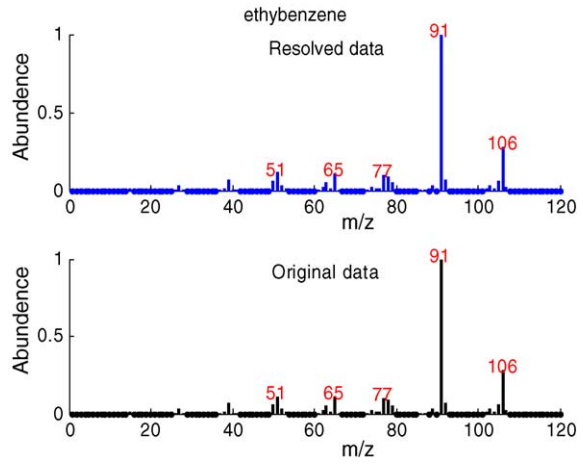

Fig. 7. Mass spectrum of ethylbenzene. The resolved spectrum is shown in the top and the original spectrum at the bottom.

There were many chromatograms in the data, which separated ideally and could be analyzed (qualitative analysis) directly by MSD data analysis software and NIST98 MS database. There were about 11 kinds of alcohols, 3 kinds of aldehydes and 22 kinds of esters, which could be identified. However, there were also some overlapping peaks in the data, 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 2-D 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 overlapping peak from 17.40 min to 17.55 min as an example. The GC–MS chromatogram was shown in Fig. 10. It seemed that there was only one component, but actually, it was a two-component peak.

Heuristic evolving latent projections [2,11] has been widely used to resolved 2-D chromatogram data, which is based on the chromatographic characters of filling in and eluting out to gain selective information of pure component chromatograms and spectra. We have used HELP to resolve the GC–MS data in this example, and the resolved chromatograms are shown in Fig. 11.

The resolution of the chromatogram is still a linear combination of pure components that contain negative and positive data, which contradicts reality. HELP could not give a reasonable resolution in this case.

NMF is successful in this case. The resolution of chromatograms and mass spectra are shown in Figs. 12 and 13, respectively. Resolved by NMF, the match of $C_{10}H_{20}O_2$ (butyl caproate) enhanced from 89.3% to 92.0% in the NIST MS database, and that of the isomeric compound (butyric acid, hexyl ester) enhanced from 86.3% to 95.0%.

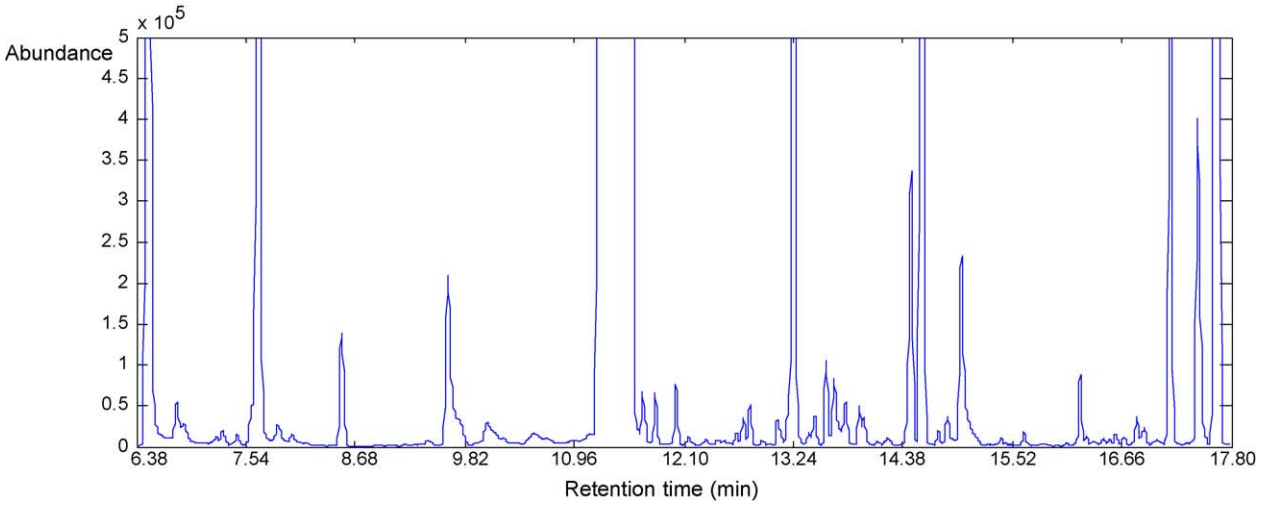

Fig. 9. The GC–MS chromatograms of liquor from 6.38 min to 17.80 min.

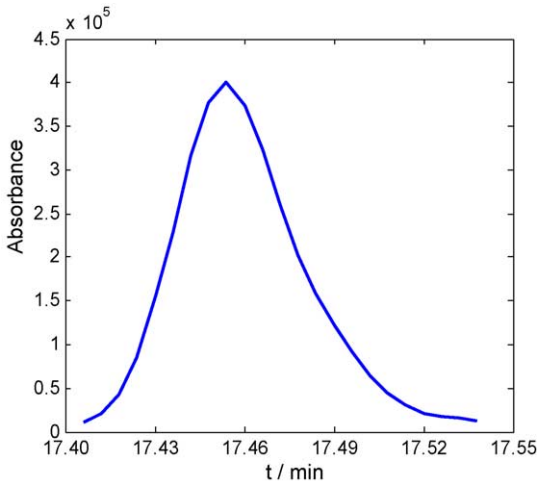

Fig. 10. The GC–MS chromatogram of liquor from 17.40 min to 17.55 min.

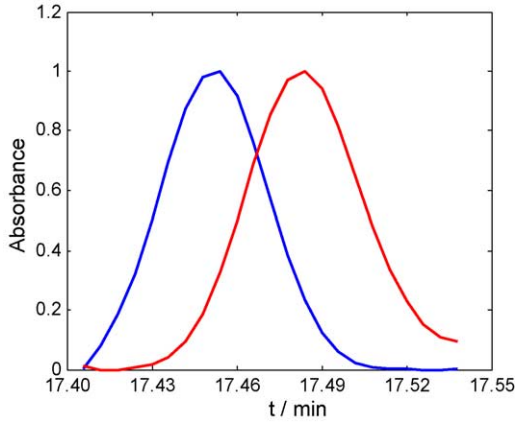

Fig. 12. The resolved GC–MS chromatograms from NMF.

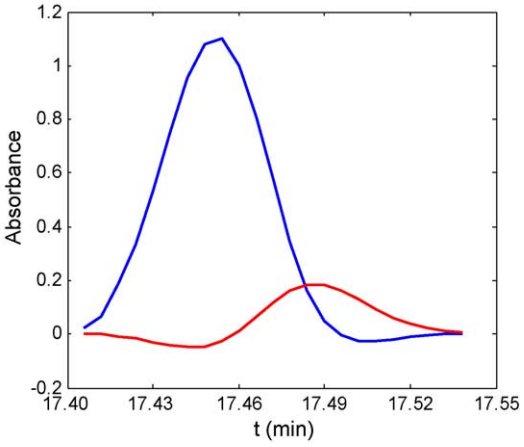

Fig. 11. The resolved GC–MS chromatograms from HELP.

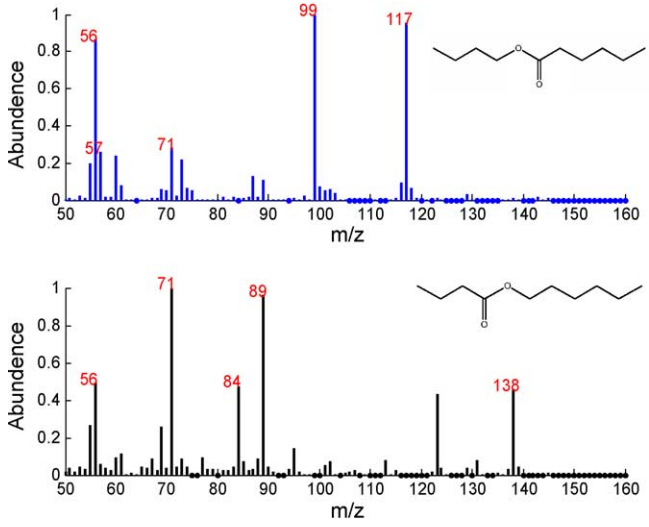

Fig. 13. The resolved mass spectra from NMF.

5. Conclusion

We have improved the algorithm of non-negative matrix factorization by undertaking the following steps. (1) Add curves smoothing; (2) impose unimodality constraint of chromatograms; and (3) set $\frac{V_{i\mu}}{(WH)_{i\mu}} = 0$ when $(WH)_{i\mu} = 0$ appears in the iterative process. The improved NMF has been used to resolve simulated chemical two-dimensional data and actual experimental GC–MS data. It has been shown that NMF is a powerful resolution method for overlapping chromatograms and mass spectra. When 2-D matrix contains overlapping chromatograms or mutually independent spectra, NMF can give reasonable and acceptable resolutions.

Both methods, PCA and NMF, are based on finding a projection matrix used to project new vectors. NMF implementation is based on element-to-element iterative calculations in which the basis functions are local, while PCA implementation is based on vector-to-vector decomposition in which basis functions are global and span the entire domain. The factorization matrices in NMF are non-orthogonal, while the factor matrices (score matrix and loading matrix) in PCA are orthogonal. The problem on PCA is that the data in the factor matrices have both positive and negative values, which contradicts physical reality. It must be solved in next steps to get linear combinations to represent similar "spectra". Unlike PCA, the resolutions of NMF are non-negative and can be easily interpreted and understood directly. The results are consistent with the properties of chemical spectra.

Modified by smoothness, unimodality and sparseness, non-negative matrix factorization can be applied to resolve chemical data. It is clear that NMF will be a promising resolution method for analysis of complex samples.

Acknowledgements

The authors thank the National Nature Science Foundation of China for financial support (Grant no. 20275026).

References

[1] D.D. Lee, H.S. Seung, Nature 401 (1999) 788.
[2] P.V. van Zomeren, H. Darwinkel, P.M.J. Coenegracht, G.J. de Jong, Anal. Chim. Acta 487 (2003) 155.
[3] H. Gampp, M. Maeder, C.J. Meyer, A.D. Zuberbuehler, Talanta 32 (1985) 1133.
[4] P.J. Gemperline, J. Chem. Inf. Comput. Sci. 24 (1984) 206.
[5] P.J. Gemperline, Anal. Chem. 67 (1989) 2240.
[6] Z.L. Zhu, W.Z. Cheng, Y. Zhao, Chemom. Intell. Lab. Syts. 64 (2002) 157.
[7] E. Sanchez, B.R. Kowalski, Anal. Chem. 58 (1986) 496.
[8] G.F. Carlos, A.B. Carsten, E.S. Robert, Anal. Chem. 73 (2001) 675.
[9] Q.S. Xu, Y.Z. Liang, Chemom. Intell. Lab. Syst. 45 (1999) 335.
[10] Z.P. Chen, J.H. Jiang, Y. Li, H.L. Shen, Y.Z. Liang, R.Q. Yu, Anal. Chim. Acta 381 (1999) 233.
[11] O.M. Kvalheiml, Y.Z. Liang, Anal. Chem. 64 (1992) 936.
[12] P. Paatero, U. Tapper, Environmetrics 5 (1994) 111.
[13] F.A. Garrido, G.M. Martínez, J.L. Vidal, D.L. Massart, et al., Anal. Chim. Acta 411 (2000) 145.
[14] H.T. Gao, T.H. Li, K. Chen, X. Bi, Chem. J. Anal. Chem. 32 (2004) 993.
[15] D.D. Lee, H.S. Seung, Adv. Neural Inf. Proc. Syst. 9 (1997) 515.
[16] D.D. Lee, H.S. Seung, Adv. Neural Inform. Process. Syst. (2000) 556.
[17] B. Gershon, B. Orin, Vis. Res. (2002) 42.
[18] D. Guillamet, J. Vitriz, Pattern Recognit. Lett. 24 (2003) 1599.
[19] D. Guillamet, J. Vitria, Proceedings of 16th International Conference on Pattern Recognition, 2, 2002, p. 128.
[20] D. Guillamet, B. Schiele, J. Vitria, Proceedings of 16th International Conference on Pattern Recognition, 2, 2002, p. 116.
[21] T. Feng, S.Z. Li, H.Y. Shum, H.J. Zhang, Proceedings on Development and Learning, 2, 2002, p. 178.
[22] D. Guillamet, J. Vitri, B. Schiele, Pattern Recognit. Lett. 24 (2003) 2447.
[23] D. David, S. Victoris, http://www-stat.stanford.edu/~donoho/Reports/2003/NMFCDP.pdf.
[24] M. Bogusz, M. Erkens, J. Chromatogr. A 674 (1994) 97.
[25] H.R. Keller, D.L. Massart, Y.Z. Liang, O.M. Kvalheim, Anal. Chim. Acta 263 (1992) 125.
[26] A.K.M. Leung, F. Gong, Y.Z. Liang, F.T. Chau, Anal. Lett. 33 (2000) 3195.
[27] F. Gong, Y.Z. Liang, Q.S. Xu, F.T. Chau, A.K.M. Leung, J. Chromatogr. A 905 (2001) 193.
[28] F. Gong, Y.Z. Liang, H. Cui, F.T. Chau, B.T.P. Chan, J. Chromatogr. A 909 (2001) 237.
[29] F. Gong, Y.G. Peng, H. Cui, Y.Z. Liang, A.K.M. Leung, F.T. Chau, Chem. J. Chin. Univ. 20 (1999) 199.