Overlapped Chromatograms Separation Using Non-negative Matrix Factorization

S.Anbumalar, P.Rameshbabu, and R.Anandanatarajan

Abstract—In chromatogram analysis, overlapped chromatograms are difficult to analyze if they are not resolved. Experimental separation procedures are costlier and time consuming. Conventional multivariate curve resolution techniques have difficulties when the chromatograms are severely overlapped. Non-negative Matrix Factorization (NMF) is a tool generally used for image processing and data mining. The original NMF can also be applied for chemical analysis, after imposing some constraints. In this work, modified NMF with divergence objective algorithm (NMFdiv) has been proposed for the separation of severely overlapped chromatograms of acetone and acrolein mixture. The experimental chromatograms were obtained using Gas Chromatography –Flame Ionization Detector (GC-FID). The proposed NMFdiv algorithm effectively separates the overlapping chromatograms. Principal Component Analysis (PCA) was applied to determine the number of components in the mixtures taken. The correlation co-efficient is greater than 0.99 and resolved signal power is greater than 20dB always. The algorithm also effectively separates the simulated overlapped chromatograms even though the numbers of overlapped components are three or four. The simulated results are encouraging even for three and four components and the correlation coefficient is greater than 0.9 and resolved signal power is greater than 10 dB always.

Index Terms—Modified NMFdiv, Overlapping chromatograms, PCA, resolution.

I. INTRODUCTION

Gas chromatography (or) High Performance Liquid Chromatography is generally used for complex chemical mixture analysis. In that, there is a possibility to get overlapped chromatograms if the chemicals in the mixture have closer retention times. The analysis will be difficult if the overlaps are not resolved. The experimental separation procedures are time consuming and also require costly chemicals. Many resolution techniques are available to extract information from multivariate data. Evolving Factor Analysis (EFA) [1-2], Fixed Size Moving Window Evolving Factor Analysis (FSMW-EFA) [3], SIMPLS-to-use Interactive Self-modeling Mixture Analysis (SIMPLISMA) [4], Multivariate Curve Resolution (MCR) [5-7], Iterative Target Transformation Factor Analysis (ITTFA) [8-9] and Heuristic Evolving Latent Projections (HELP) [10-11] are some of the important techniques so far applied to many fields. All the above methods have their

limitations in separation of overlapped chromatograms when they are severely overlapped [12]. Many methods have been developed further to improve the resolution. However, it is still difficult to resolve severely overlapping peaks accurately.

Lee and Seung suggested NMF in 1990. So far, NMF has been used for image processing, text data mining, audio signal separation, biological signal separation and spectral resolution [13-25,27]. The original Lee and Seung's NMF has to be modified for chemical analysis. (i.e.) constraints have to be imposed based on the characteristics of chemical signal. Hong-Tao Gao et al., [20] applied modified NMF for the separation of two components chromatogram in which several constraints have been imposed based on sparseness and unimodality.

In this work, modified NMF divergence objective algorithm has been proposed for the separation of overlapped chromatograms. Acetone and acrolein are the chemicals found in the environment naturally and due to industrial release. An exposure of these chemicals beyond certain value, may lead to harmful effects. Hence the existence of these chemicals has to be identified. When the above mixture is analyzed using GC- FID, the resulting chromatogram is severely overlapping due to their closer retention times. It makes the analysis difficult. Hence, modified NMF with divergence objective algorithm (NMFdiv) has been applied on those chromatograms to resolve them. The proposed algorithm has also been applied on simulated two, three and four component chromatograms to resolve the overlap.

II. MATERIALS AND METHODS

A. Experiments

Instrumental condition:

Gas chromatography was used with Flame Ionization detector. Capillary Column (30 m × 0.25 mm ID, BP5) was used with a temperature of 100 ° C. The Temperature was programmed as 40 ° C for 2 min, 5° C/min, 68 ° C for 2min, 10 ° C/min, 100 ° C for 2min. Injector temperature was kept at 150 ° c. Nitrogen was used as a carrier gas at a flow rate of 90 ml/min.

Experimental procedure:

Acetone and acrolein calibration standards were injected into GC-FID separately and chromatograms were obtained. The retention time of acetone and acrolein are 1.127 min and 1.187 min respectively. Then the acetone and acrolein were mixed in 4 different concentrations. The mixture was vigorously stirred separately under ice cold condition for half an hour and 0.2 µL of these mixtures were injected in GC-FID.

Manuscript received March 13, 2011; revised July 9, 2011. This work was funded by the Department of Science and Technology (DST), Pondicherry, India.

S.Anbumalar is with Sri Manakula Vinayagar Engineering College, Puducherry, India . e-mail: saravanan.anbumalar@gmail.com

P.Rameshbabu and R.Anandanatarajan are with Pondicherry Engineering College, Puducherry, India

The chromatograms were obtained using the software provided by Iris 32 chromatography provided by Chemito Technologies pvt ltd., running on PC with Intel (R) Pentium4 CPU 2.00GHz and 512MB RAM. The chromatographic data (i.e., retention time t and detector output) has been exported to an ASCII file through Iris 32 software and acquired through MATLAB R2008a software. The above instrumental condition produces overlapping chromatogram for the acetone and acrolein mixture. It is difficult to identify the individual components in the overlapping chromatograms.

B. Method

The modified NMFdiv algorithm has been applied on the overlapped chromatograms to resolve them. The performance of this algorithm is discussed in the Results section.

Proposed NMFdiv algorithm:

Non-negative Matrix Factorization is a method to decompose measured data matrix A of size $m \times n$, whose columns contain data samples, into two terms W and H of size $m \times k$ and $k \times n$ respectively, where $k < \min(m, n)$. The decomposition is performed so that the product WH should approximate the original data A . The columns of W are usually called basis vectors and the rows of H are called decomposition coefficients. Thus, the original data are represented as linear combinations of these basis vectors. NMF allows only additive operations to provide meaningful decomposition. Several different algorithms have been proposed for factorizing W and H under non-negativity constraints.

NMFdiv algorithm was suggested by Lee and Seung with the following objective function which has to be minimized:

$$\sum_{i=1}^m \sum_{j=1}^n [A_{ij} \log(A_{ij}/Y_{ij}) - A_{ij} + Y_{ij}] \quad (1)$$

where $Y = WH$.

The objective function is based on multiplicative update rules of W and H which are given by

$$H_{qj} = H_{qj} * (W^T A)_{qj} / (W^T W H + 1e^{-9})_{qj}; \text{ for } 1 \leq q \leq k \text{ and } 1 \leq j \leq n,$$

$$W_{iq} = W_{iq} * (A H^T)_{iq} / (W H H^T + 1e^{-9})_{iq}; \text{ for } 1 \leq i \leq m \text{ and } 1 \leq q \leq k.$$

Element wise operation has been carried out in the iterative steps.

When NMF is introduced to resolve the chemical data matrix, it must be modified according to the characteristics of signal (i.e.,) smoothness, unimodality of chromatograms and sparseness of data. Based on that, constraints have to be imposed into the iterative procedure to get a reasonable and acceptable resolution [20]. NMFdiv algorithm applied in this paper imposes non-negativity constraint to find the factors and additional constraint was imposed to deal with zero values in A and W matrices.

If A and W matrix contains zero value, the iteration breaks. Hence, some minimum threshold value ($1e^{-9}$) has been assigned for the zeros. The accuracy of the results are not affected due to the addition of a very small positive number which is almost equal to zero while compared with that of the data value taken in the matrices. The W and H matrices are normally initialized with random nonnegative

values. The unimodality and selectivity constraints were integrated into the iterative procedure additionally. The chromatograms taken for analysis are smooth and hence the curve smoothing constraint was not imposed. To speed up the convergence SVD-based initialization [26] was tested. But the results were not reliable. Then the random initialization was adopted for all cases. The maximum number of iterations has been specified. It takes only very few seconds to converge.

Procedure

The input matrix A was formed by assigning the detector outputs obtained in various experiments as rows of A and the various experiments as columns. The number of components K in the mixture was determined by principal component analysis (PCA). Based on the size of A matrix and k value, W and H matrix have been initialized with random non negative and non zero values and were updated with multiplicative step (Lee and Seung). After resolved, the basis vectors W denotes the individual chemical's chromatogram, while H is the coefficient matrix. The proposed NMFdiv algorithm of Equation (1) has been performed to deconvolute the data matrix A . Initially, one dimensional, GC-FID data matrices of two, three and four component chromatograms were simulated using cross product multiplication of Gaussian functions. Then the proposed NMFdiv algorithm was tested on the simulated chromatograms of two cases, i.e., severely overlapped and embedded peaks. Finally, the proposed NMFdiv algorithm was applied on the experimental overlapped chromatograms of acetone and acrolein mixtures. After decomposition, each column of W matrix gives the chromatogram of individual chemical present in the mixture.

The resolved components are compared with standard chromatograms and the correlation coefficients are calculated. The resolved signal power is calculated using the formula given in equation 2.

$$\text{Pr(dB)} = 10 \log_{10} \left[\frac{\sum_i (S_i / \text{std}(S_i))^2}{\sum_i ((S_i / \text{std}(S_i)) - (R_i / \text{std}(R_i)))^2} \right] \quad (2)$$

where S_i - standard signal,

R_i - resolved signal

Std - standard deviation

Pr (dB) - closeness of resolved signal power with that of standard signal

$i = 1, 2, \dots, n$; where n is the number of components in the mixture.

III. RESULTS AND DISCUSSION

A. Simulated GC-FID data matrix

Two components system:

At first, the proposed NMFdiv algorithm was applied on the simulated dataset of severely overlapped chromatograms (two components). The simulated overlapped chromatogram and the resolved chromatograms using proposed NMFdiv algorithm are shown in Figure.1. The resolved components were compared with the standard chromatograms. The correlation coefficients are 0.9993, 0.9997 and resolved signal powers are 28.8040 dB, 32.0223 dB for the first and second components respectively.

Secondly, the proposed NMFdiv algorithm was applied

on the simulated embedded peaks (two components). The simulated embedded chromatograms and the resolved chromatograms are shown in Fig.2. The resolved components were compared with the standard chromatograms. The correlation coefficients are 0.9999, 0.9986 and the resolved signal powers are 37.9929 dB, 25.4354dB for the first and second components respectively.

Fig.1. Simulated severely overlapped chromatogram of two components mixture and resolved c.hromatograms obtained using proposed NMFdiv.

Fig.2. Simulated embedded chromatogram of two components mixture and resolved chromatograms obtained using proposed NMFdiv.

Three and Four component systems:

The simulated three components system was also tested with proposed NMFdiv algorithm. The results are encouraging. The correlation coefficient is greater than 0.99 and the resolved signal power is greater than 20 dB always. The NMFmse algorithm was then applied on the simulated severely overlapped four component chromatograms. The simulated overlapped chromatogram and resolved chromatograms using proposed NMFdiv are shown in Fig.3. The resolved components were compared with the simulated standard chromatograms and the correlation coefficients are 0.9452, 0.9997, 0.9993 and 0.9333 for the first, second, third and fourth components respectively. The resolved signal powers are 9.6813dB, 32.0223 dB, 28.8040 dB and 8.9043 dB for the first, second, third and fourth components respectively. The proposed NMFdiv was then applied on the embedded peak containing four components. The simulated embedded chromatogram and the resolved chromatograms using NMFdiv are shown in Fig.4. The

resolved results were compared with the simulated standard chromatograms and its correlation coefficients are 0.9854, 0.9715, 0.9879 and 0.9993 for the first, second, third and fourth components respectively. The resolved signal powers are 15.9691 dB, 12.4932 dB, 16.4645 dB and 28.4323 dB for the first, second, third and fourth components respectively.

Even though the degree of overlapping is very high, proposed NMFdiv algorithm resolves the overlap efficiently with a correlation coefficient greater than 0.93 and resolved signal power greater than 8dB. The chromatographic situation of the cases taken will be difficult for an analyst to get a reasonable resolution. But the solutions of proposed NMFdiv are reasonable and acceptable. If the W matrix was randomly initialized, the algorithm converges to a stationary point for most of the runs in two, three and four components (severely overlapped) chromatogram. But, for the complex chromatograms (i.e., four components embedded and more components chromatograms), the algorithm converges to a stationary point only for few runs. But all the above said results were in a narrow feasible region that is reasonable and acceptable.

Fig.3. Simulated severely overlapped chromatogram of four components mixture and resolved chromatograms obtained using proposed NMFdiv.

Fig.4. Simulated embedded chromatogram of four components mixture and resolved chromatograms obtained using proposed NMFdiv

B. Real GC-FID experimental data

The overlapping chromatograms were obtained by GC-FID for acetone and acrolein mixtures. The experiments were conducted under condition mentioned in Section 2.1.

Chromatograms obtained are shown in fig.5 and fig.6. It can be analyzed qualitatively with the help of standard chromatograms. But, the obtained chromatograms are overlapping due to almost similar retention times of acetone and acrolein (i.e., 1.127 min and 1.187min respectively). Hence, during peak identification, only one peak has been identified instead of two in the overlapped region. If these overlapped peaks were not resolved, the chromatograms of mixtures obtained could not get a good match with that of a pure component in the database. Hence, proposed NMFdiv algorithm was applied on the chromatograms obtained to resolve acetone and acrolein. Before applying NMFdiv, the number of components k in the data A was determined using PCA, which identified two components. The standard chromatograms of acetone and acrolein, overlapped chromatogram of acetone and acrolein mixture and the resolved chromatograms are shown in Fig.7.

The correlation coefficients were obtained by comparing the resolved results with the standard chromatograms and those coefficients are 0.9967 and 0.9994 for acetone and acrolein respectively. The resolved signal powers for the resolved acetone and acrolein are 22.9044 dB and 30.1120 dB respectively. The algorithm works well with random initial matrices of W and H . The resolving power of proposed NMFdiv was tested with different size of A matrix by varying n . Initially, the algorithm was tested only with 2 experiments. i.e., n was taken as 2. The result shows that there was reliable resolution even with 2 datasets.

Fig.5. Real experimental severely overlapped chromatogram of acetone and acrolein mixture of 2 μ l acrolein and 8 μ l acetone. Experimental condition: Gas chromatography-Flame Ionization Detector; Carrier gas: Nitrogen (90ml/min); Detector temperature: 150° C; Retention time of acetone 1.127min; Retention time of acrolein 1.187min.

Fig.6a. Real experimental chromatogram of acetone standard. Experimental condition: Gas chromatography-Flame Ionization Detector; Carrier gas: Nitrogen (90ml/min); Detector temperature: 150° C; Retention time of acetone 1.127min.

Fig.6b. Real experimental chromatogram of acrolein standard. Experimental condition: Gas chromatography-Flame Ionization Detector; Carrier gas: Nitrogen (90ml/min); Detector temperature: 150° C; Retention time of acrolein 1.187 min.

Fig.7. Experimental severely overlapped chromatogram of acetone and acrolein mixture and resolved chromatograms using proposed NMFdiv (applied on 4 dataset). Experimental condition: Gas chromatography-Flame Ionization Detector; Carrier gas: Nitrogen (90ml/min); Detector temperature: 150° C; Retention time of acetone 1.127min; Retention time of acrolein 1.187 min. Standards used to prepare a mixture are also given at the top.

The proposed NMFdiv algorithm for chromatographic resolution is effective. The correlation coefficient is greater than 0.99 and resolved signal power is greater than 20 dB. When chromatographic profiles of two compounds mixture of similar peak-height and almost of the same retention time are analysed, it is quite a troublesome issue for analysts. But the solutions of proposed algorithm are reasonable and acceptable. Fig.8 shows the minimization of objective function Vs number of iterations for proposed NMFdiv. It shows that the objective function has come to a minimum error in less than 40 iterations. For proposed NMFdiv, additional constraint was imposed to deal with zero values in A and W matrices.

IV. CONCLUSION

In this paper, proposed NMFdiv algorithm was applied to resolve the overlapped chromatograms. The algorithm was

tested on simulated data of two, three and four component systems of severely overlapped as well as embedded peaks and the results are encouraging. The algorithm was also applied on real experimental overlapped chromatograms of acetone and acrolein mixtures and it is proved that it resolves overlapping chromatograms efficiently. The efficiency is checked from the correlation coefficient of the resolved result with the standard which is always greater than 0.9 and the resolved signal power which is always greater than 8 dB. Principal Component Analysis (PCA) was applied to determine the number of components in the mixtures taken.

Fig.8. Objective function Vs number of iterations for proposed NMFdiv implementation.

ACKNOWLEDGMENT

This work was funded by the Department of Science and Technology (DST), Puducherry, India.

REFERENCES

- [1] H. Gampp, M. Maeder, C.J. Meyer and A.D. Zuberbühler, "Calculation of equilibrium constants from multiwavelength spectroscopic data. III. Model-free analysis of spectrophotometric and ESR titrations", *Talanta*, vol.32, pp. 1133-1139, 1985.
- [2] M. Maeder, "Evolving Factor Analysis for the resolution of overlapping chromatographic peaks", *Anal. Chem.*, vol.59, pp. 527-530, 1987.
- [3] H.R. Keller and D.L. Massart, "Peak purity control in liquid chromatography with photodiode array detection by fixed size moving window evolving factor analysis", *Anal. Chim. Acta*, vol.246, pp. 379-390, 1991.
- [4] W. Windig and J. Guilment, "Interactive self-modeling mixture analysis", *Anal. Chem.*, vol.63, pp.1425-1432, 1991.
- [5] R. Tauler, A.K. Smilde and B.R. Kowalski, "Selectivity, local rank, three-way data analysis and ambiguity in multivariate curve resolution", *J. Chemometr.* vol.9, pp. 31-58, 1995.
- [6] R. Tauler, "Multivariate curve resolution applied to second order data", *Chemom. Intell. Lab. Sys.*, vol.30, pp.133-146, 1995.
- [7] R. Tauler, S. Lacorte and D. Barcelo, "Application of multivariate curve self-modeling curve resolution for the quantitation of trace levels of organophosphorous pesticides in natural waters from interlaboratory studies", *J. of Chromatogr. A*, vol.730, pp.177-183, 1996.
- [8] P.J. Gemperline, "A priori estimates of the elution profiles of the pure components in overlapped liquid chromatography peaks using target

- factor analysis", *J. Chem. Inf. Comput. Sci.*, vol.24, pp.206-212, 1984.
- [9] B.G.M. Vandeginste, W. Derks and G. Kateman, "Multicomponent self-modelling curve resolution in high-performance liquid chromatography by iterative target transformation analysis", *Anal. Chim. Acta*, vol.173, pp.253-264, 1985.
- [10] O.M. Kvalheim and Y.Z. Liang, "Heuristic evolving latent projections: resolving two-way multicomponent data. 1. Selectivity, latent-projective graph, datascope, local rank, and unique resolution", *Anal. Chem.*, vol.64, pp.936-946, 1992.
- [11] Y.Z. Liang, O.M. Kvalheim, H.R. Keller, D.L. Massart, P. Kiechle and F. Erni, "Heuristic evolving latent projections: resolving two-way multicomponent data. 2. Detection and resolution of minor constituents", *Anal. Chem.*, vol.64, pp.946-953, 1992.
- [12] P. V. van Zomeren, H. Darwinkel, P. M. J. Coenegracht and G. J. de Jong, "Comparison of several curve resolution methods for drug impurity profiling using high-performance liquid chromatography with diode array detection", *Anal. Chim. Acta*, vol.487, pp.155-170, 2003.
- [13] D.D. Lee and H. Seung, "Learning the parts of objects by non-negative matrix factorization", *Nature*, vol.401, pp.788-791, 1999.
- [14] D.D. Lee and H. Seung, "Algorithms for non-negative matrix factorization", *Adv. Neural Inf. Process. Syst.* vol.13, pp.556-562, 2001.
- [15] P. O. Hoyer, "Non-negative Matrix Factorization with Sparseness Constraints", *Journal of Machine Learning Research*, vol.5, pp.1457-1469, 2004.
- [16] S. Z. Li, X. W. Hou, H. J. Zhang, and Q. S. Cheng, "Learning Spatially Localized, Parts-based Representation", *International Conference on Computer Vision and Pattern Recognition*, pp.207-212, 2001.
- [17] S. Zafeiriou, A. Tefas, I. Buciu, and I. Pitas, "Exploiting Discriminant Information in Nonnegative Matrix Factorization with Application to Frontal Face Verification", *IEEE Trans. on Neural Networks*, vol.17(3), pp. 683-695, 2006.
- [18] I. Buciu and I. Pitas, "A New Sparse Image Representation Algorithm Applied to Facial Expression Recognition", *IEEE Workshop on Machine Learning for Signal Processing*, pp.539-548, 2004.
- [19] M. N. Schmidt and M. Morup, "Nonnegative Matrix Factor 2- D deconvolution for blind single channel source separation," Independent Component Analysis and Blind Signal Separation Lecture Notes in Computer Science, 3889/2006, pp.700-707, 2006.
- [20] Hong-Tao Gao, Tong-Hua Li, Kai Chen, Wei-Guang Li and Xian Bi, "Overlapping spectra resolution using non-negative matrix factorization" *Talanta*, vol. 66, pp.65-73, 2005.
- [21] Liu Mingyu, Ji Hongbing and Zhao Chunhong, "Non negative Matrix Factorization and Its Application in EEG Signal Processing", *IEEE Xplore*, 978-1-4244-1748-3/08.
- [22] H. Li, T. Adali, W. Wang, D. Emge and A. Cichocki, "Non-negative matrix factorization with orthogonality constraints and its application to Raman spectroscopy," *J. of VLSI Signal Processing*, 48, 1-2, pp.83-97, 2007.
- [23] Z. Yuan and E. Oja, "Projective nonnegative matrix factorization for image compression and feature extraction," *14th Scandinavian Conference on Image Analysis*, pp.333-342, 2005.
- [24] M. Morup, K.H. Madsen and L.K. Hansen, "Shifted Non-negative Matrix Factorization", *IEEE Workshop on Machine Learning for Signal Processing*, 2007.
- [25] S.S. Bucak, B. Günsel and O. Gürsoy, "Incremental non-negative matrix factorization for dynamic background modelling," *ICEIS 8th International Workshop on Pattern Recognition in Information Systems*, 2007.
- [26] C. Boutsidis and E. Gallopoulos, "On SVD-based initialization for nonnegative Matrix factorization," Technical Report HPCLAB-SCG-6/08-05, University of Patras, Patras, Greece., 2005.
- [27] Francesca Guimet, Ricard Boque and Joan Ferré "Application of non-negative matrix factorization combined with Fisher's linear discriminant analysis for classification of olive oil excitation-emission fluorescence spectra" *Chemom. Intell. Lab. Sys.*, vol.81, pp 94 - 106, 2006.