Automated Unmixing of Comprehensive Two-Dimensional Chemical Separations with Mass Spectrometry[*]

Min Chen Stephen E. Reichenbach Jiazheng Shi
Computer Science and Engineering Department
University of Nebraska – Lincoln
Lincoln NE 68588-0115 USA

Abstract

This paper proposes techniques to automate unmixing of coeluted chemicals in data produced by comprehensive two-dimensional gas chromatography (GCxGC) coupled with mass spectrometry (MS). The approach consists of three steps: i) measure the pureness of a region of interest, ii) count and locate the peak points of underlying compounds in impure regions, and iii) unmix the region into pure compounds using parallel factor analysis (PARAFAC). This approach has parametric controls that allow tuning to balance demands for performance and computational efficiency. Experiments with real and simulated data demonstrate the approach is effective in automating the analysis of coelutions in GCxGC/MS.

1 Introduction

This paper proposes an approach to automating the unmixing of a region of interest in data produced by comprehensive two-dimensional gas chromatography (GCxGC) coupled with mass spectrometry (MS). GCxGC/MS is a powerful separation technique in which mass spectrometry provides additional chemical selectivity to the separations along two chromatographic columns of GCxGC systems [1]. The output of GCxGC/MS instruments is a three-way data cube. The first way (x-axis) represents the elapsed time for the first column separation; the second way (y-axis) represents the elapsed time for the second column separation; and the third way (z-axis) represents the mass spectrum. Details about GCxGC and GCxGC/MS can be found in [2].

Three-way chemometric algorithms, such as trilinear decomposition (TLD) and parallel factor analysis (PARAFAC), have been proposed for unmixing three-way trilinear chemical data [1, 3]. One common disadvantage is that they require *a priori* knowledge of the number of underlying compounds in the region. It is desirable to develop a method that can estimate the number of underlying compounds and thus automate the unmixing process for GCxGC/MS data.

This paper develops an automated approach to unmix a region of interest. Given a region, the first step is to measure its pureness. This paper proposes a statistical method for estimating the pureness of a selected region. Exploiting the additional selectivity provided by mass spectrometry, the paper also develops a method for locating and counting the peaks for the constituent compounds in the impure regions. PARAFAC is applied to unmix the impure regions after the number of underlying compounds is known. PARAFAC can provide a unique solution on conditions that the data have trilinear structure, a reasonable signal-to-noise ratio, and a known number of compounds in the mixture. When fitting GCxGC/MS data with the PARAFAC model, the solution has physical meaning, $i.e.$, the chromatographic profiles and mass spectra of underlying compounds [4, 1].

The rest of this paper is organized as follows. Section 2 describes the visualization and the mathematical model of the GCxGC/MS data. Section 3 presents the approach to automating the unmixing process. Section 4 presents the experimental results of simulation and real GCxGC/MS data. Finally, Section 5 gives conclusions and future work.

2 System modeling

GCxGC/MS data can be displayed as a two-dimensional Total Intensity Count (TIC) image, where each pixel value is the total of the corresponding mass spectrum (the third dimension or z-axis). The pixels are arranged so that the x-axis from left to right is the elapsed time for the first column separation and the y-axis from bottom to top is the elapsed time for the secondary column separation. Each chemical compound produces a two-dimensional chromatographic peak in the data, seen as a *blob* (or cluster) of adjacent pixels with values larger than the background. Visualization and analysis software, such as GC Image [5], can

[*]This work was supported by NSF Award IIS-0431119. Contact information: email: [mchen|reich|jshi]@cse.unl.edu, phone: 1.402.472.2401, fax: 1.402.472.7767.

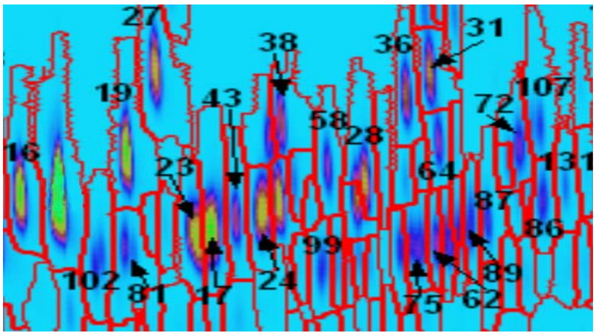


Figure 1: A portion of a TIC image of GCxGC/MS data. The red lines show the blob boundaries. The numbers are blob IDs.

accurately isolate, quantitate, and identify compounds that are well separated in the data, but coeluting compounds that produce blobs of multiple overlapping peaks pose a more difficult analysis problem. Figure 1 displays a TIC image where background removal and blob detection have been performed [6]. Some of the blobs are well separated peaks, but other blobs are composed of coeluting peaks.

In this paper, the individual elements of vectors, matrices, and three-way data are denoted by lowercase italics; vectors are denoted by lowercase bold characters; matrices are denoted by uppercase bold characters; and three-way data cubes are denoted by uppercase italics. Let a region of interest be an $I \times J \times K$ GCxGC/MS data cube *X*, with elements $x_{i,j,k}$. The vector $\mathbf{x}_{i,j}$ represents the mass spectrum of pixel (i, j) in the TIC image. An individual ion channel image for mass-charge-ratio k is $\mathbf{X}_{1..I,1..J,k}$.

Based on the properties of chemical compounds and GCxGC/MS instruments, the GCxGC/MS data can be formulated by a trilinear model:

$$\mathbf{X}^{I\times JK} = \sum_{f=1}^{F} \mathbf{a}_n(\mathbf{c}_n^T \otimes \mathbf{b}_n^T) + \mathbf{E}, \qquad (1)$$

where the $I \times JK$ matrix $\mathbf{X}$ is the GCxGC/MS data unfolded along the first mode; the operator $\otimes$ denotes the Kronecker product; T denotes matrix transpose; F is the number of chemical compounds; $\mathbf{E}$ is an $I \times JK$ matrix accounting for the noise and model error; and $\mathbf{a}_n, \mathbf{b}_n$, and $\mathbf{c}_n$ are the chromatographic profiles of first and second column and mass spectral profile for the n^{th} compound (see Figure 2) [1]. The pair of chromatographic profiles a_n and b_n determine the volume and the retention time of the n^{th} compound. The mass spectral profile c_n represents the mass spectrum of the n^{th} compound.

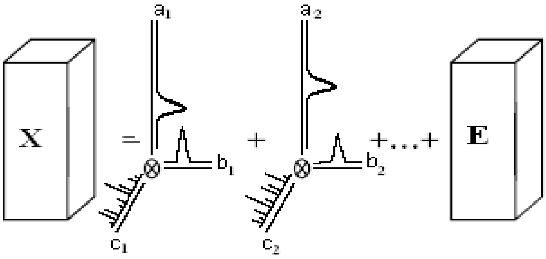


Figure 2: GCxGC/MS data formulated by a trilinear model [1].

3 Automated unmixing

Coeluting compounds produce data with overlapping peaks. The goals of data analysis are to identify and individually quantitate the peak for each constituent compound of interest. Implicit in this goal is the challenging problem of unmixing over-lapping peaks. Blob detection delineates peaked clusters of pixels, but, with coelution, a detected blob may account either for only a fraction of one chemical-related peak (rather than the whole peak) or for parts of multiple peaks (rather than a single peak). This paper proposes an effective approach to automating the unmixing of coeluted compounds in a region of interest. The approach has three steps: a pureness test, peak locating, and PARAFAC unmixing.

3.1 Pureness test

A region of the data may contain one or several blobs. The pureness test applies to each blob of the region. If there exists an impure blob, the region is designated as “impure”.

Suppose a blob contains only one chemical compound. The mass spectrum of every pixel in the blob should be very close to that of the blob peak. If there exist two or more overlapping compounds in a blob, however, the mass spectra of some pixels could be quite different from that of the peak. With this observation, the paper proposes a statistical function to measure the impurity of a given blob:

$$P = \sum_{i=1,j=1}^{I,J} \left\| \frac{\mathbf{x}_{i,j}}{\sum_{k=1}^{K} x_{i,j,k}} - \frac{\mathbf{x}_{p,q}}{\sum_{k=1}^{K} x_{p,q,k}} \right\|^2 \frac{\sum_{k=1}^{K} x_{i,j,k}}{\sum_{k=1}^{K} x_{p,q,k}}, \qquad (2)$$

where (p, q) is the blob peak, $\mathbf{x}_{i,j}$ and $\mathbf{x}_{p,q}$ are the pixel and peak mass spectra. The sum of the squared differences between the normalized pixel and peak mass spectra is weighted by the relative TIC of the pixel (i, j). The assumption is that the coeluted compounds have different mass spectra. The expected impurity P should be very small if the blob is pure; otherwise, it should be large. A blob is designated as impure if P is greater than a threshold.

3.2 Peak locating

If a region is determined to be impure, the next step is to locate and count the peaks in the region. In general, overlapping peaks can be distinguished only if there is some separation in time between the largest values in each peak. If the separation is relatively large, distinct peaks can be located even if the mass spectra are nearly identical. If the separation is relatively small, distinct peaks can be located only if the mass spectra of the compounds are different to some degree.

The peak locating process searches for local maxima in an ion image by sliding a small window. The center of the sliding window is considered as a peak candidate if it is the local maximum for the window. A practical concern is that noise may generate false peak candidates. Therefore, two additional criteria are used to verify peak candidates. The candidate also must have: i) ion intensity that is much greater than noise level and ii) neighboring values that are reasonably large.

Performing the peak locating process and the subsequent step of PARAFAC unmixing on all ion images is computationally expensive because GCxGC/MS data typical have hundreds of ion channels. These computations can be reduced by operating only on the ion channels that provide the greatest discrimination. This paper proposes a diamond algorithm to identify the ion channels used for these steps. As noted previously, underlying this algorithm is the requirement that there is some separation between peaks. Using an impure region with two compounds c_1 and c_2 as an example, Figure 3 shows peak c_1 at the center and a diamond with eight general offset directions for peak c_2. The size of the diamond is regulated by the bounding box of the blob which contains all pixels with TIC at least $w\%$ of the TIC at the blob peak. The diamond has apexes p_1, p_3, p_5 and p_7, defined from the blob bounding box (w) as:

$$\begin{aligned}
\big[x(p_1), y(p_1)\big] &= \big[s_1(w), \quad (s_2(w)+e_2(w))/2\big],\\
\big[x(p_3), y(p_3)\big] &= \big[(s_1(w)+e_1(w))/2, \quad e_2(w)\big],\\
\big[x(p_5), y(p_5)\big] &= \big[e_1(w), \quad (s_2(w)+e_2(w))/2\big],\\
\big[x(p_7), y(p_7)\big] &= \big[(s_1(w)+e_1(w))/2, \quad s_2(w)\big],
\end{aligned}$$

where s_1 and s_2 are the starting points of the bounding box along first and second columns and e_1 and e_2 are the ending points. Another four points p_2, p_4, p_6 and p_8 are defined:

$$\begin{aligned}
\big[x(p_2), y(p_2)\big] &= \big[(x(p_1)+x(p_3))/2, (y(p_1)+y(p_3))/2\big],\\
\big[x(p_4), y(p_4)\big] &= \big[(x(p_3)+x(p_5))/2, (y(p_3)+y(p_5))/2\big],\\
\big[x(p_6), y(p_6)\big] &= \big[(x(p_5)+x(p_7))/2, (y(p_5)+y(p_7))/2\big],\\
\big[x(p_8), y(p_8)\big] &= \big[(x(p_7)+x(p_1))/2, (y(p_7)+y(p_1))/2\big].
\end{aligned}$$

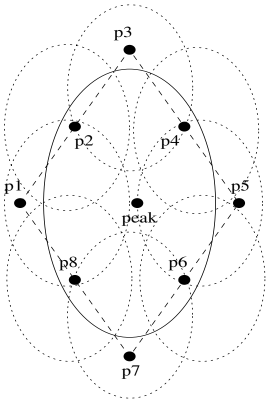


Figure 3: Diamond algorithm. The solid oval represents c_1. The dotted ovals are possible positions for c_2 around c_1.

Wherever c_2 lies, at least one of these eight points has a mass spectrum which is more similar to the mass spectrum of the compound for c_2 than is the mass spectrum at the blob peak. The diamond can be constricted by increasing w or enlarged it by decreasing w. Normalizing the mass spectra at the eight diamond points and computing their differences with the peak yields eight arrays. The ion channels that contain the largest mass spectral differences are used for peak locating.

3.3 PARAFAC unmixing

PARAFAC is a method to decompose trilinear data. It yields three physically meaningful component matrices **A**, **B**, and **C** that contain $\mathbf{a}_n, \mathbf{b}_n$, and $\mathbf{c}_n$ in Equation 1.

Initializing and refining the profiles are two major steps in the implementation of PARAFAC unmixing. Any initialization leads to the same result if the data are strictly trilinear because PARAFAC has the property of uniqueness under this condition [7, 4]. However the GCxGC/MS data are not strictly trilinear, so different initializations may produce different results. A good starting point can not only speed up the convergence dramatically, but also helps to avoid local minima. In this paper, an enhanced trilinear decomposition (TLD) is used to initialize PARAFAC because it is a non-iterative algorithm and its solution is close to real profiles. The enhanced TLD is initialized by Tucker3 and adjusted by generalized rank annihilation method (GRAM) so that the solution gets closer to the final profiles than does TLD. Details about TLD, Tucker3, and GRAM can be found in [3, 4, 8]. Alternating least squares (ALS) technique is used to refine the component matrices during the PARAFAC fitting. In the course of ALS, constraints of unimodality and nonnegativity [9] are imposed to make the component matrices more physically meaningful. Figure 4 presents the algorithm and the mathematical derivations for

```
$[\mathbf{A}, \mathbf{B}, \mathbf{C}]$ = PARAFAC($X$, F)
  • Initialized by the enhanced TLD:
    $[\mathbf{A}, \mathbf{B}, \mathbf{C}]$ = eTLD($X$, F)
      – Initialized by Tucker3
          1. $[\mathbf{A}_{Tuck}, \mathbf{B}_{Tuck}, \mathbf{C}_{Tuck}, G]$ = Tucker3($X$, F)
      – Adjusted by GRAM
          1. $[\mathbf{B}_{GRAM}, \mathbf{C}_{GRAM}]$ = GRAM($G$, F)
          2. $\mathbf{B} = \mathbf{B}_{Tuck}\mathbf{B}_{GRAM}$, $\mathbf{C} = \mathbf{C}_{Tuck}\mathbf{C}_{GRAM}$,
             $\mathbf{A} = \mathbf{X}^{I \times JK}((\mathbf{C}|\otimes|\mathbf{B})^{-1})^{T}$
  • Refining component matrices with ALS
      1. Repeat if relative change between two consecutive iterations is greater than threshold
      2. Reshape $X$ into $\mathbf{X}^{I \times JK}$
      3. Improve $\mathbf{A}$: $\mathbf{Z} = \mathbf{C}|\otimes|\mathbf{B}$, $\mathbf{A} = \mathbf{X}\mathbf{Z}(\mathbf{Z}^T\mathbf{Z})^{-1}$
      4. Reshape $X$ into $\mathbf{X}^{J \times IK}$
      5. Improve $\mathbf{B}$: $\mathbf{Z} = \mathbf{C}|\otimes|\mathbf{A}$, $\mathbf{B} = \mathbf{X}\mathbf{Z}(\mathbf{Z}^T\mathbf{Z})^{-1}$
      6. Reshape $X$ into $\mathbf{X}^{K \times IJ}$
      7. Improve $\mathbf{C}$: $\mathbf{Z} = \mathbf{B}|\otimes|\mathbf{A}$, $\mathbf{C} = \mathbf{X}\mathbf{Z}(\mathbf{Z}^T\mathbf{Z})^{-1}$
      8. Impose constraints of unimodality and nonnegativity
  • Postprocessing
```

Figure 4: PARAFAC unmixing algorithm.

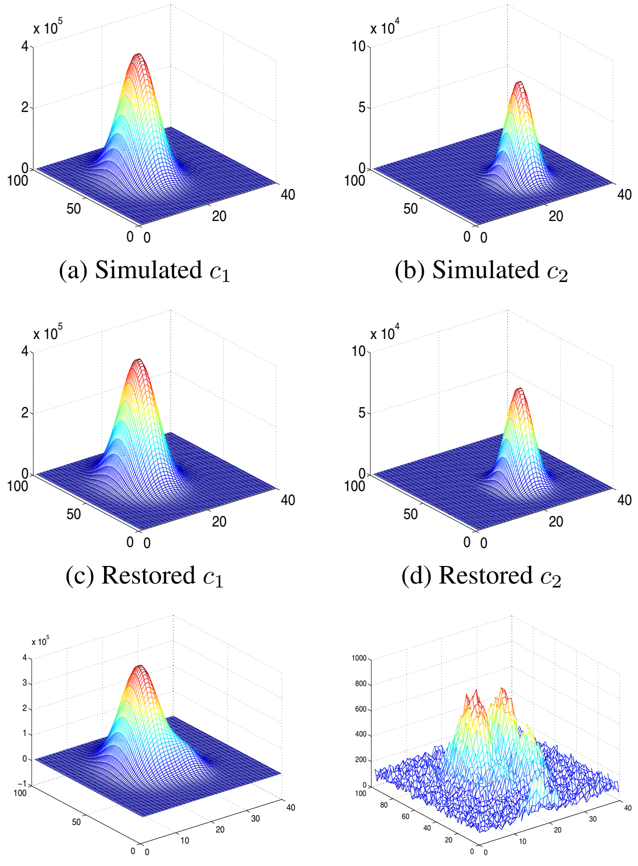
a three-way data cube X that has F components. The operator $|\otimes|$ in Figure 4 stands for khatri_Rao product [4].

4 Experimental results

4.1 Simulation

The simulation data is synthesized according to Equation 1. Two Gaussian-shaped components are created with different standard deviations along abscissa and ordinate as well as two different mass spectra. White noise is added so that the signal to noise ratio (SNR) is consistent with real data. The three-way data cube has dimensions $I = 40$, $J = 100$, and $K = 10$ for three modes respectively. Figure 5(a) and (b) show the TIC images of the simulated components c_1 and c_2, with peaks at $(16, 56)$ and $(23, 37)$, respectively. Their volumes (summation of all the elements) are 1.2×10^8 and 1.2×10^7, respectively. Figure 6 illustrates the simulated mass spectra of the two components. Figure 5(e) illustrates the synthesized data, the summation of component c_1, component c_2, and noise.

Figure 5(c) and (d) demonstrate the TIC images of the components restored from the outer product of chromatographic profiles extracted by PARAFAC. The peaks of restored components are located exactly at the same places as the simulated components c_1 and c_2. The volumes of the restored components are 1.2034×10^8 and 1.1752×10^7, respectively. Figure 7 illustrates the mass spectral profiles for the two restored components. For clearer visualization, Figure 8 shows the unmixing results in two-dimension.



(a) Simulated c_1

(b) Simulated c_2

(c) Restored c_1

(d) Restored c_2

(e) TIC image

(f) Residual after PARAFAC

Figure 5: Simulated data and PARAFAC unmixing results.

Let the original three-way data be X and the restored three-way data from the resulting profiles be $\hat{X}$. The residual is computed by:

$$\mathbf{E}(i, j) = \sum_k |x_{ijk} - \hat{x}_{ijk}|. \tag{3}$$

The ratio r of residual to original data is computed by:

$$r = \sum_{ij} \sum_k |x_{ijk} - \hat{x}_{ijk}| \Big/ \sum_{ij} \sum_k x_{ijk}. \tag{4}$$

Figure 5(f) shows the residual, which has residual ratio 0.5%.

4.2 Real data

This subsection presents experimental results for real GCxGC/MS data, whose TIC image is shown in Figure 1.

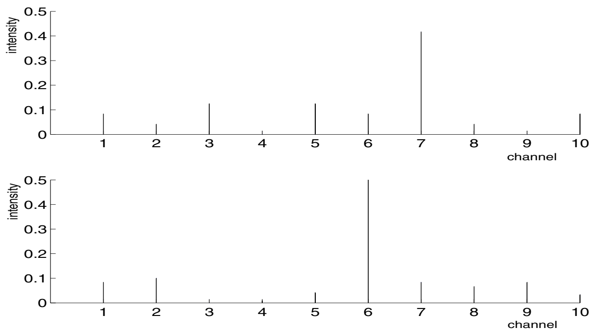


Figure 6: Simulated mass spectra.

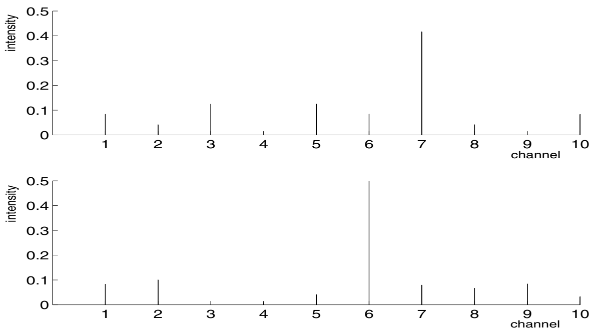


Figure 7: Mass spectra extracted by PARAFAC.

In terms of Equation 2, the impurities P of most blobs in Figure 1 are very small. For example, the impurities P of blob 19 and blob 27 are 0.10 and 0.08, indicating they are relatively pure. No further analysis is needed in these cases. However, the impurity P of the blob 102 is 5.37, indicating that this blob is impure.

Figure 9(a) illustrates the results of the peak locating process for blob 102. The peak finding process locates two peaks, noted by black crosses, c_1 at $(321, 279)$ and c_2 at $(324, 286)$. Figure 9(b) and (c) display the two peaks extracted by PARAFAC unmixing. The peaks of the two compounds, marked by black crosses at $(321, 279)$ and $(324, 286)$, are consistent with the positions found by the peak locating process. Figure 9(d) shows the residual.

The chromatographic profiles along the first column, pictured in Figure 10(a), carry the volume information, because the chromatographic profiles along the second column, pictured in Figure 10(b), as well as mass spectral profiles are normalized. The outer product of chromatographic profiles generates two pure peaks (Figure 9(b) and (c)). The two extracted peaks are well-shaped for quantitation and their locations provide the retention times of the corresponding chemical compounds. Figure 10 (c) and (d) illustrate the integrated profiles along first and second columns of the extracted model $\mathbf{X}$ and the residual $\mathbf{E}$. The

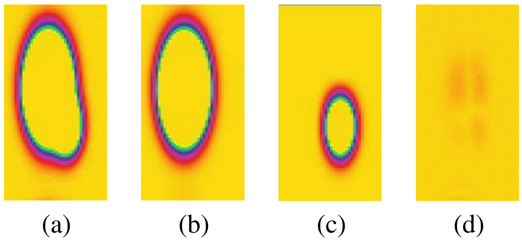


Figure 8: 2D view of the result of PARAFAC unmixing on the simulated data. (a) is the TIC image of the synthesized data. (b) and (c) are two component TIC images restored by the resulting profiles. (d) is residual.

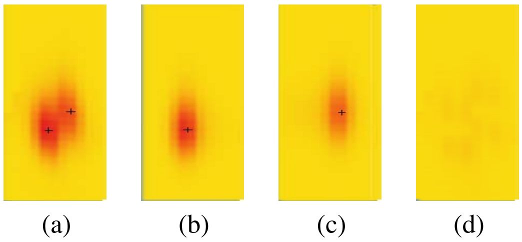
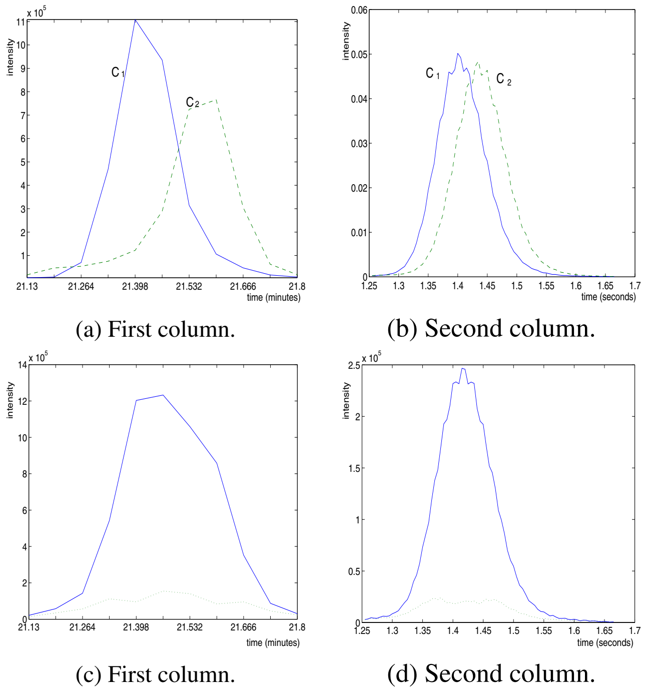


Figure 9: The PARAFAC unmixing result on blob 102. (a) shows TIC image of blob 102 and the result of peak locating. (b) and (c) show the TIC images of two restored compounds c_1 and c_2, respectively, as well as the peaks of c_1 and c_2. (d) illustrates the residual.



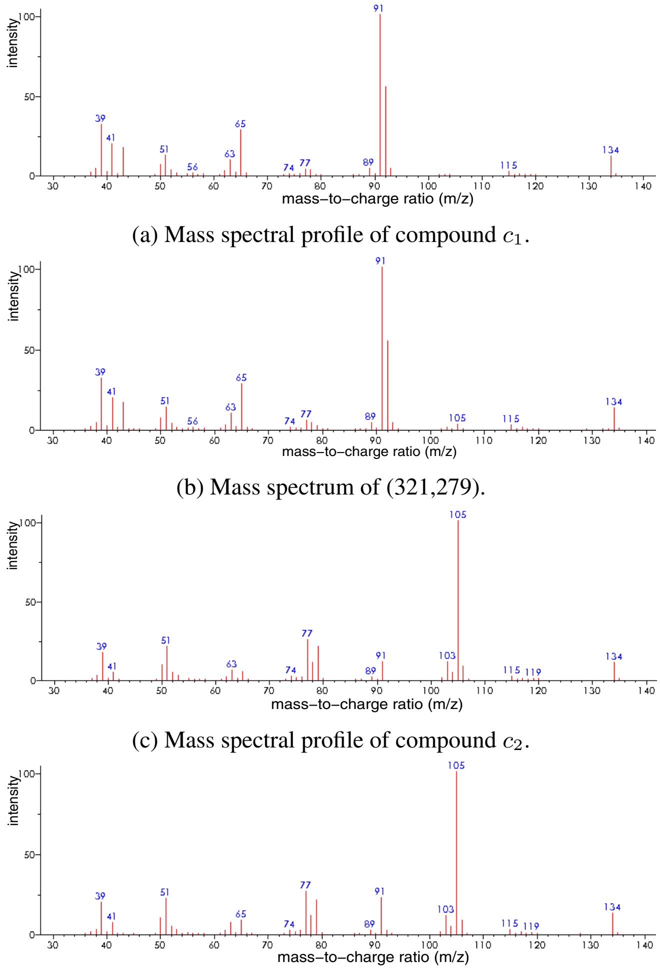
(a) First column. (b) Second column.

(c) First column. (d) Second column.

Figure 10: (a) and (b) are chromatographic profiles along first and second column extracted by PARAFAC unmixing. The profiles in (a) carry the volume information; the chromatographic profiles in (b) are normalized so that the summation is 1.0. Solid lines are for compound c_1 and dotted lines are for compound c_2. (c) and (d) show the combined chromatographic profiles $\mathbf{X}$ as solid lines and the residual $\mathbf{E}$ as dotted lines.

residual ratio (Equation 4) is 15%. The reason for the larger residual is that the real GCxGC/MS data deviates slightly from the trilinear model.

Figure 11 displays mass spectral profiles extracted by PARAFAC unmixing and the mass spectra of the data at the two peak points. The spectra in Figure 11 (a) and (b) are very similar and so are the spectra in Figure 11 (c) and (d).



(a) Mass spectral profile of compound c_1.

(b) Mass spectrum of (321,279).

(c) Mass spectral profile of compound c_2.

(d) Mass spectrum of (324,286).

Figure 11: Mass spectral profiles of two pure compounds extracted by PARAFAC unmixing and mass spectra of two points in blob 102.

5 Conclusion

This paper proposes an approach for automated unmixing of coelutions in comprehensive two-dimensional chemical separations with mass spectrometry. The approach has three steps: a pureness test, peak locating, and PARAFAC unmixing. The pureness test evaluates whether the mass spectrum of a region is relatively uniform, indicating if it has more than one compound. Any impure region is subjected to further analysis. Peak locating determines the number of the underlying compounds as well as their location. PARAFAC unmixing with constraints of unmodality and nonnegativity decomposes the region into several pure peaks described by their individual chromatographic profiles and mass spectra.

The approach developed in this paper is demonstrated experimentally to be effective. In both simulated and real data, the coeluting compounds are unmixed and quantified successfully. Ongoing work is investigating models for GCxGC/MS data, the limits of PARAFAC unmixing, and the performance of the approach for various chemical mixtures.

References

[1] A. E. Sinha, C. G. Fraga, B. J. Prazen, and R. E. Synovec, “Trilinear chemometric analysis of two-dimensional comprehensive gas chromatography–time-of-flight mass spectrometry data,” *Journal of Chromatography A*, vol. 1027, pp. 269–277, 2004.

[2] J. Dalluge, J. Beens, and U. A. Brinkman, “Comprehensive two-dimensional gas chromatography: a powerful and versatile analytical tool,” *Journal of Chromatography A*, vol. 1000, pp. 69–108, 2003.

[3] Z. Lin, K. S. Booksh, L. W. Burgess, and B. R. Kowalski, “Second-order fiber optic heavy metal sensor employing second-order tensorial calibration,” *Analytical Chemistry*, vol. 66, pp. 2552–2560, 1994.

[4] R. Bro, “Multi-way analysis in the food industry: Models, algorithms, and applications,” Ph.D. dissertation, Department of Dairy and Food Science, Royal Veterinary and Agricultural Univeristy, Denmark, 1998.

[5] S. E. Reichenbach, “GC Image, LLC, University of Nebraska-Lincoln. http://www.gcimage.com/ (2004).”

[6] S. E. Reichenbach, M. Ni, D. Zhang, and E. B. J. Ledford, “Image background removal in comprehensive two-dimensional gas chromatography,” *Journal of Chromatography A*, vol. 985, pp. 47–56, 2003.

[7] J. F. t. Berge, “Partial uniqueness in candecomp/parafac,” *Journal of Chemometrics*, vol. 18, pp. 12–16, 2004.

[8] E. Sanchez and B. Kowalski, “Generalized rank annihilation factor analysis,” *Analytical Chemistry*, vol. 58, pp. 496–499, 1986.

[9] R. Bro and N. D. Sidiropoulos, “Least squares algorithms under unimodality and non-negativity constraints,” *Journal of Chemometrics*, vol. 12, pp. 223–247, 1998.

-7803-9232-9/05/$20.00 2005 IEEE.