Analyst

Initial estimation method by cosine similarity for multivariate curve resolution: Application to NMR spectra of chemical mixture

Journal:	*Analyst*
Manuscript ID	AN-ART-07-2019-001416.R1
Article Type:	Paper
Date Submitted by the Author:	02-Sep-2019
Complete List of Authors:	Nagai, Yuya; Chuo University, Department of Chemistry Sohn, Woon Yong; Chuo University, Department of Applied Chemistry Katayama, Kenji; Chuo University, Department of Chemistry; Kagaku Gijutsu Shinko Kiko, Precursory Research for Embryonic Science and Technology

Initial estimation method by cosine similarity for multivariate curve resolution: Application to NMR spectra of chemical mixture

Yuya Nagai[1] and Woon Yong Sohn,[1] Kenji Katayama[1,2*]

1 Department of Applied Chemistry, Chuo University, Tokyo 112-8551, Japan;

2 PRESTO, Japan Science and Technology Agency (JST), Saitama 332-0012, Japan

*Corresponding authors:

K. Katayama, Phone: +81-3-3817-1913, E-mail: kkata@kc.chuo-u.ac.jp

Abstract

Multivariate curve resolution (MCR) has been widely utilized to reveal the constituents of chemicals from the multiple spectral data of chemical mixtures. In the MCR calculation, the singular value decomposition (SVD) has been utilized to obtain the initial estimation of the spectra for pure chemicals and they are adjusted to obtain the best fit using the alternating least square (ALS) algorithm. However, wrong initial estimation by SVD frequently leads convergence at an incorrect local minimum of the least square error. To overcome this problem, we have developed a robust calculation technique, which utilizes a new initial estimation using cosine similarity, and the following optimization was performed by MCR. The calculation was applied for ^{1}H-NMR spectra of 4 different chemicals, and this methodology could recover the spectra of pure chemicals (>85 % consistency) and the concentration profile for each mixture within an accuracy of <10 %.

Keywords: multivariate curve resolution, cosine similarity, initial estimation, chemical mixture spectra, ^{1}H-NMR

Introduction

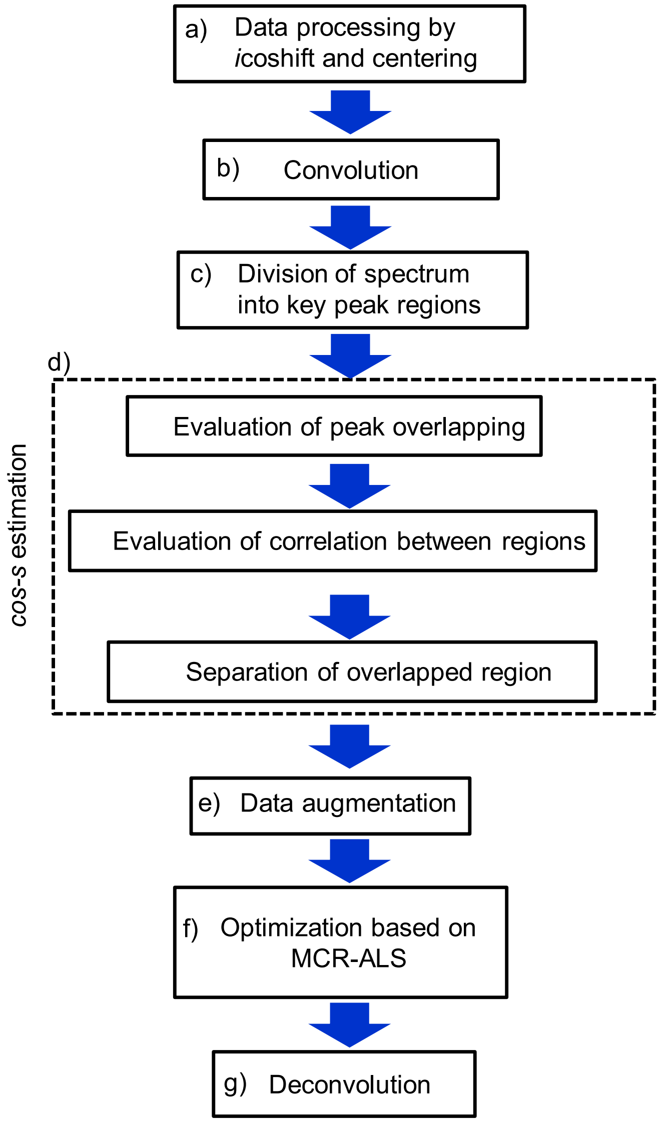
We often encounter various mixtures of chemicals in material researches. In chemical reactions, the reaction system includes not only the reactant and product species but also various by-products and intermediate species. For example, in biological processes in cells, many different proteins and lipids are involved and various chemicals are uptaken and discharged from the cell. To analyze the involved chemicals, we need to separate them into pure ones and to perform chemical analyses such as nuclear magnetic resonance (NMR), infrared absorption (IR), mass spectroscopy (MS), etc. In organic syntheses, usually, the reaction cannot be halted for the analyses of the reactants, products, and by-products by separating them during the reaction. In biological cells, much spectral information can be obtained but the spectrum is different at each position because the constituents are different depending on the local positions. As such, in many chemical processes, it is not always easy to extract pure chemicals, and we frequently encounter the situation where only information on a mixture of chemicals is obtained as an overlap of spectra.

Thus, it is beneficial if we could obtain spectra for pure chemicals and their concentrations from the overlapped spectra consisting of different ratio of chemicals for mixture samples. There are several methods to extract the information from the spectra of chemical mixtures in the field of chemoinformatics; the partial least square (PLS) regression has been widely used for this purpose[1] to identify important components from the multiple spectral data of chemical mixtures. In PLS, the data dimension is reduced much by extraction of the featured spectra using the principal component analysis (PCA) from the dataset. By ignoring the multicollinearity, a small number of components consisting of the spectra can be properly extracted. However, this process makes it difficult to interpret chemical information directly, because the loadings and scores obtained from the analyses are different from the pure spectra and the concentration profiles.

Multivariate curve resolution (MCR) method paves the way to solve this problem, where multiple spectral data as a matrix is decomposed into a matrix of the spectra for pure chemicals (S) and a matrix of the concentration ratio (C) in the mixtures. In this calculation, S and C matrixes are updated alternatively to minimize the least square error between the spectra and concentrations for the mixtures and S and C are estimated. (MCR-ALS) with penalty terms such as non-negativity, restriction of the value range and the number of components, etc. The MCR-ALS has been utilized for various applications; chromatography data has been analyzed from the beginning of the MCR application in the field of analytical chemistry;[2–5] time-dependent spectral data were analyzed for the kinetic analysis of the protein folding,[6,7] the drug degradation[8] and the reaction of amino acid with a drug candidate;[9,10] electrochemical analysis;[11,11] mixture analysis of chemical blends using the near-infrared absorption spectra,[12,13] the UV/VIS absorption spectra;[14–16] the circular dichroism,[17] gas chromatography / liquid chromatography-mass spectrometry data[18,19] and x-ray absorption spectra;[20] the optical spectra, IR spectra and mass spectra obtained by scanning a sample surface was decomposed into pure spectra and concentration profiles;[21–24] the intermediate species were estimated from the temperature dependence of the near-infrared absorption;[25] metabolite profile analysis using the capillary-electrophoresis mass spectrometry and liquid chromatography-mass spectrometry data;[26–28] and ^{1}H-NMR data;[29] the separation of excitation-emission matrix into the components for different fluorophores;[30] polymer crystallinity at the side chain was estimated from the Raman spectra.[31,32] In recent years, spatial distribution of each chemical species or biological components was mapped out using Raman microscopy by collecting many spectra of mixtures from the different locations of biological cells,[33–37]

However, the calculation sometimes does not work well due to strong background,[38] unclear number of components,[39] rotational ambiguities in the matrix decomposition process,[40] and most severely affected by the initial estimation of the spectra of pure chemicals, which is conventionally obtained by the singular value decomposition (SVD). Once the least square error is in the local minimum, it is difficult to recover the correct spectra for pure chemicals.[41] To overcome this problem, we have developed the categorization of the spectral components by using the cosine similarity (hereafter called *cos-s*) of the peak intensity correlation in three steps. The *cos-s* estimation could provide a reasonable initial estimation and the following MCR process could refine the spectra and obtain the concentration profile with high reliability. In this paper, we demonstrated that the ^{1}H-NMR spectra of the mixtures consisted of 4 different chemicals were decomposed into the correct pure spectra and concentrations of the mixtures.

78 # Theory and method


a) Data processing by *i*coshift and centering
b) Convolution
c) Division of spectrum into key peak regions
d)
cos-s estimation
Evaluation of peak overlapping
Evaluation of correlation between regions
Separation of overlapped region
e) Data augmentation
f) Optimization based on MCR-ALS
g) Deconvolution

Scheme.1 The overall workflow of *cos*-s MCR is shown. The algorithm is divided into 7 sections: (a) data pre-processing, (b) spectral convolution, (c) division of the spectral regions, (d) cos-s estimation, (e) data augmentation, (f) MCR optimization, (g) deconvolution. The cos-s estimation is consisted of 3 steps; the peak overlap is evaluated in the first step, and the peak region correlation is examined in the second step, and the overlapped peak region is separated.

79

Scheme 1 represents the overall workflow of the MCR calculation we have developed. Because the original ^{1}H-NMR data had minor errors of chemical shifts for each measurement (<0.005 ppm), and a peak of ^{1}H-NMR typically is consisted of 10 data points, only 1-point shift gives a large error for the calculation. To remove the minor shifts, *i*coshift algorithm[42] was applied to adjust the true peak positions, (Scheme 1 (a)) which has been used for ^{1}H-NMR data to adjust the peak shifts of spectra.

Before the data processing, the spectrum data, $s_{i,j}$ for the sample number, *i* and the chemical shift, *j*, was centered to the average intensities of samples as:

$$s_{ij} = s_{ij} - \bar{s}_j \qquad (1)$$

This process ensures that the intensity variation for different samples is considered. (Scheme 1(a))

Even though the *i*coshift algorithm could adjust the chemical shifts, the original raw spectra have unknown shifts in peaks' positions and unexpected distortion or split. These biases of peaks did not satisfy the condition that a spectrum of the chemical mixture should be represented by a linear summation of the spectra of pure chemicals. To overcome this problem, each spectrum was convoluted by a Gaussian function to obscure tiny differences. (Scheme 1(b)) The used function had ~0.005 ppm of the full width of half maximum (FWHM), which was interactively determined. As shown in Eqn. (2), the convolution calculation was performed as

$$(f * g)(\delta) = \int f(\delta^*) g(\delta - \delta^*) d\delta^* \qquad (2),$$

where $g(\delta)$ is the Gaussian function and $f(\delta)$ is the original spectral data. Since this study focuses on the separation of major components in the mixture samples, the minimum molar fraction of each component was larger than 0.1, where this convolution procedure did not eliminate any signal peaks. Furthermore, any separated peaks were not merged into a single peak by the Gaussian function with a width of 0.005 ppm.

Then, the spectra were separated into each peak-region and non-peak regions were removed, which helped improve the calculation accuracy. (Scheme 1(c)) In ^{1}H-NMR spectra, many data points are near the baseline, and it is natural to pick up the peak regions and analyze them. To process this, the noise level was estimated from the standard deviation (σ) of the baseline. The signals/peaks were selected if the S/N ratio exceeded 2. The peaks were grouped under the criteria that the peaks are in the same group if the chemical shifts of a pair of peaks was less than 0.02 ppm. In practice, the incorrect grouping of peaks did not matter because the overlapped peak is separated and each peak are categorized again into each component corresponding to a single chemical species in the following cos-s estimation processes.

Generally, SVD has been utilized for the initial estimation for MCR-ALS, but we adopted the *cos*-s estimation as an alternative for the initial estimation. The *cos*-s procedure had three steps for the initial estimation of the pure spectra; the evaluation of the peak overlap in each peak region, the evaluation of the correlation between peak regions, and the separation of the overlapped peaks. The cosine similarity is defined as follows:

$$cos\theta = \frac{\boldsymbol{a} \cdot \boldsymbol{b}}{|\boldsymbol{a}||\boldsymbol{b}|} \qquad (3),$$

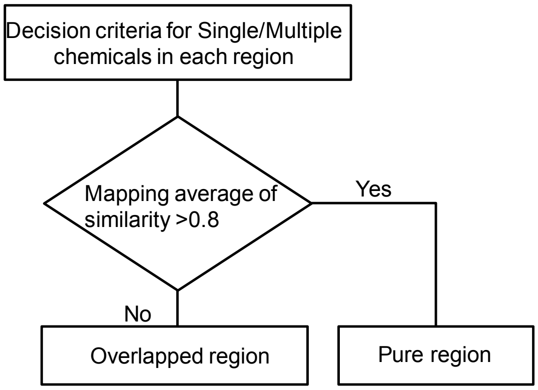
which provides the information about the similarity of the vector ***a*** and ***b***.

First, the *cos*-s estimation was utilized for the evaluation of the peak overlap with multiple chemical species, (Scheme 1(d)) and the procedure is summarized in Scheme 2. The similarity in each peak region was evaluated if the peak region consists of multiple chemical species. The spectral intensities for each peak region and for the sample number are regarded as a matrix, whose component is represented as s_{ij} for the sample number, *i* and the chemical shift, *j*. For each chemical shift, *j*, *cos-s* was calculated as:

$$(cos\theta)_{j=j_1,j_2} = \frac{\boldsymbol{s}_{j_1} \cdot \boldsymbol{s}_{j_2}}{|\boldsymbol{s}_{j_1}||\boldsymbol{s}_{j_2}|} = \frac{\sum_{i=1}^{n} s_{ij_1} s_{ij_2}}{\sqrt{\sum_{i=1}^{n} {s_{ij_1}}^2}\sqrt{\sum_{i=1}^{n} {s_{ij_2}}^2}} \quad (i = 1 \ldots n) \qquad (4)$$

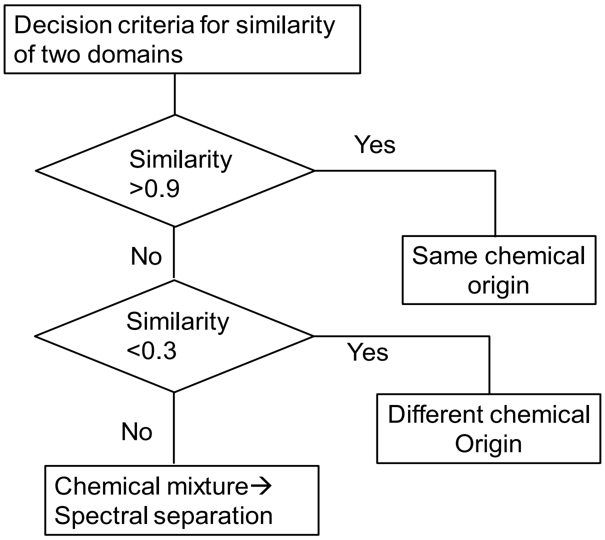
In Eqn. (4), *n* represents the total sample number. This calculation provides the correlation matrix indicating if the signal intensity variation in the sample number direction is correlated for the two chemical shifts, *i* and *j*.

Based on the mapping of the correlation matrix, $(cos\theta)_{j=j_1,j_2}$, it is determined if each peak region is composed of a single or multiple chemicals. In this study, if the average of the cosine similarity correlation matrix was larger than 0.8, we empirically evaluated that the peak region was dominated by a single chemical.


Decision criteria for Single/Multiple chemicals in each region
Mapping average of similarity >0.8
Yes
No
Overlapped region
Pure region

Scheme.2 The first step *cos-s* similarity procedure is shown. This process corresponds to the first process of Scheme 1(d). In each peak region, the similarity of the spectral intensity is examined. Based on the average similarity values in each peak region, the evaluation was made if there is an overlap in the region. The threshold of the similarity value was set to 0.8.

In the second step of the initial estimation, the correlation between the different peak regions was examined. In this procedure, the correlation of the peak areas were calculated and they were used for the cos similarity. The procedure is summarized in Scheme 3. At first, the peak area was calculated for each peak, which was represented as a matrix component a_{ij} for the sample number, i and the index of the peak regions, j. For each peak, *cos-s* was calculated by Eqn. (5).


Decision criteria for similarity of two domains
Similarity >0.9
Yes
Same chemical origin
No
Similarity <0.3
Yes
Different chemical Origin
No
Chemical mixture→ Spectral separation

Scheme.3 The second cos similarity procedure is shown. This process corresponds to the second process of Scheme 1(d). The similarity between the peak regions were evaluated. Based on the similarity values between the peak regions, it is evaluated if the peaks are derived from the same chemical.

$$(cos\theta)_{j=j_1,j_2} = \frac{\boldsymbol{a}_{j_1}\cdot\boldsymbol{a}_{j_2}}{|\boldsymbol{a}_{j_1}||\boldsymbol{a}_{j_2}|} = \frac{\sum_{i=1}^{n} a_{ij_1}a_{ij_2}}{\sqrt{\sum_{i=1}^{n} {a_{ij_1}}^{2}}\sqrt{\sum_{i=1}^{n} {a_{ij_2}}^{2}}} \quad (i = 1\ldots n) \qquad (5)$$

In Eqn. (5), n represents the total sample number. Considering that the spectral intensity for two peak regions derived from the same chemical origins should vary the intensities in a similar way in the sample number direction, i. It is obvious that highly correlated peak regions are derived from the same chemical species. Based on the correlation matrix, the similarity is utilized to evaluate the correlation of the peak regions and the independence of the peaks. From these two *cos-s* evaluation processes, it is evaluated how many components (chemicals) compose the spectra of chemical mixtures, and which peak regions are composed of multiple/single chemical species, and which peak regions are correlated each other.

A highly correlated peak region with an overlapped peak region is utilized to separate it into the peaks of pure chemicals, which is the third step of the *cos-s* procedure. Figure 1 illustrates the overview for the separation of peaks in the overlapped peak. When the overlapped peak region (red) is highly correlated with another peak region for a pure chemical (purple), the correlated component included in the overlapped peak region could be extracted using the cosine similarity and the residual signal component was regarded as a non-correlated component.

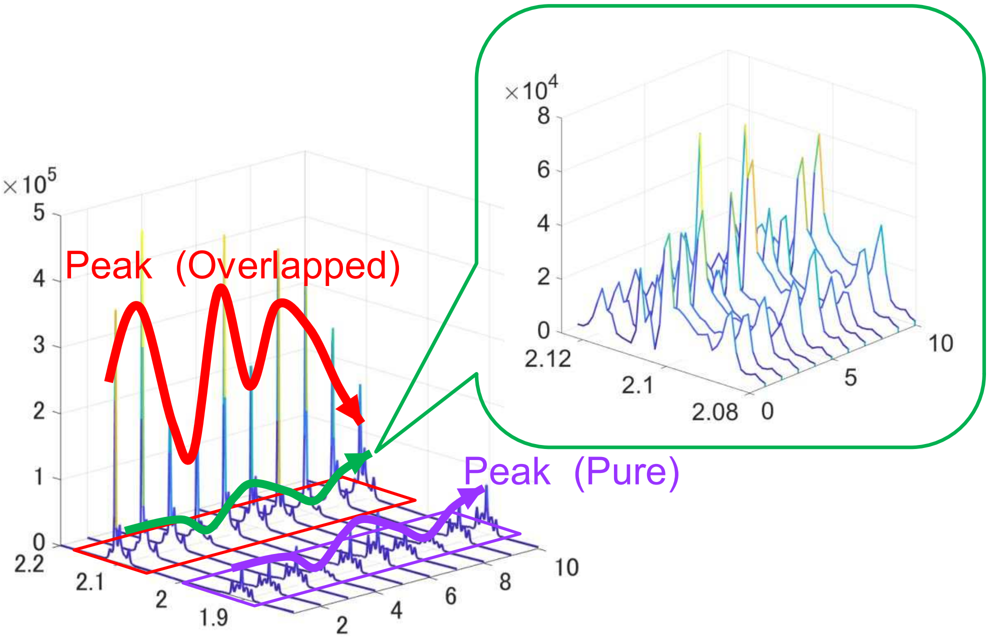


Figure 1 The schematic representation how to extract pure spectrum from the overlapped spectral region by using the correlation between the peak regions.

The calculation process is described here. It is assumed that $s_{pure,\,i,\,j}$ and $s_{overlapped,\,i,\,j}$ as the spectral intensities for a pure region and an overlapped peak one, respectively. Since the pure component in the overlapped spectra should be varied in an intensity similar to the pure spectra, the following equations were utilized for separation of the peak intensities. Assuming k as the number of the overlapped components in the peak region, $r_{j,k}$ was regarded as the ratio representing how much the overlapped peak includes a pure component and it is expressed as:

$$r_{j,k} = \left(\frac{\boldsymbol{s_{pure}}_{j_0,k} \cdot \boldsymbol{s_{overlapped}}_{j}}{\left|\boldsymbol{s_{pure}}_{j_0,k}\right|\left|\boldsymbol{s_{overlapped}}_{j}\right|}\right)^2 = \left(\frac{\sum_{i=1}^{n} s_{pure_{i,j_0,k}} s_{overlapped\ i,j}}{\sqrt{\sum_{i=1}^{n} s_{pure_{i,j_0,k}}^{\ 2}} \sqrt{\sum_{i=1}^{n} s_{overlapped\ i,j}^{\ 2}}}\right)^2 \qquad (6)$$

$$s_{extracted,i,j,k} = r_{j,k} s_{i,j} \qquad (7)$$

$$s_{res_{i,j}} = (1 - \textstyle\sum_k r_{i,j,k}) s_{i,j} \qquad (8)$$

In Eqn.(6), $s_{pure_{i,j_0,k}}$ corresponds to the peak values in $s_{pure_{i,j,k}}$, representing a pure component. The similarity (6) was squared to make the value positive. A highly correlated component in the overlapped peak with another peak region is obtained by Eqn. (7). By subtracting all the correlated spectra, the residual spectrum $s_{res_{i,j,k}}$ can be obtained (Eqn. (8)). In most cases, the residual spectrum represents the baseline since it is not correlated with any peak regions. However, it could be a spectrum for a chemical with a single peak such as acetone, $CHCl_3$ or TMS. Thus, even if the spectrum does not have any correlation with other peak regions, it could be extracted as $s_{res_{i,j,k}}$.

Based on the analysis, the total number of chemical species is determined, and the spectrum intensity matrix was arranged as $s_{i,j,m}$ (*m*: number of species). The initial estimates for the pure spectra ($s_{est,m,j}$) was obtained by averaging it for the sample number, i. From these procedures, we can obtain a chemically meaningful initial estimation without using any prior information about the samples. Before the MCR calculation, the spectral data were augmented, (Scheme 1(e)) which indicates the procedure to increase the spectral data by mixing the original spectra with random ratios. In our calculation, the extended spectral number was set to 100 by compromising the solution stability and the calculation time.

For the MCR calculation, MCR-ALS GUI 2.0 [43] was utilized with some modification by setting the criteria of the program to keep the spectral intensity within the 50-150 % range from the initial *cos*-s estimation. The recovered spectra were deconvoluted to obtain the final pure spectra. (Scheme 1(g))

Experiment

Acetone, cyclopentanone, ethyl acetate, tetrahydrofuran (Wako) were utilized as purchased. These 4 chemicals were mixed in the molar fraction as shown in Table1. Each mixed sample was put into an NMR tube (OPTIMA), and ^{1}H-NMR (500 MHz, JEOL) spectra were measured at room temperature. A deuterated solvent, chloroform-*d* (Wako), was used as an internal standard. The reference spectra of these chemicals are shown in Figure 2(a) (The spectra around 2 ppm is expanded in Figure 2(b).)

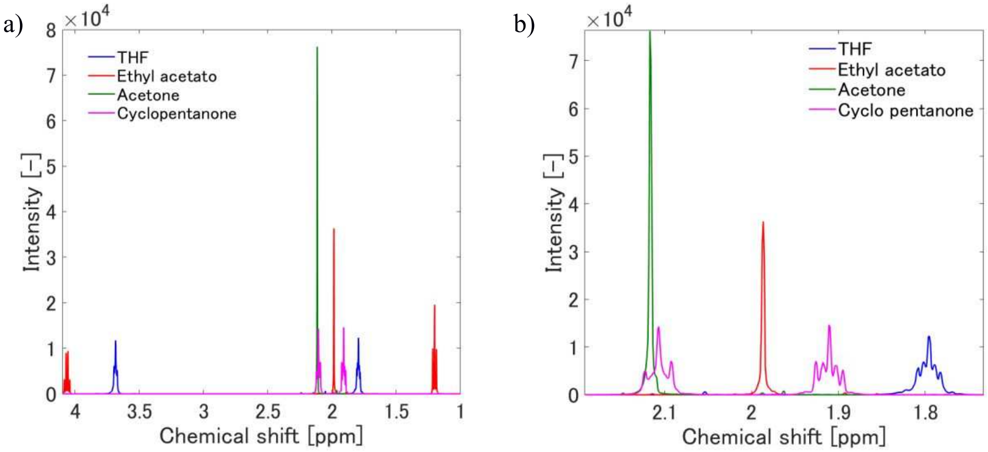


Figure 2 ^{1}H-NMR spectra of 4 chemicals (tetrahydrofuran, ethyl acetate, acetone, cyclopentanone) in the whole range (a) and enlarged around 2 ppm where the spectra are overlapped (b).

Table. 1 The molar fractions of the prepared samples.

Sample number	THF	Ethyl acetate	Acetone	Cyclopentanone
1	0.217	0.191	0.382	0.209
2	0.345	0.221	0.205	0.229
3	0.229	0.249	0.325	0.197
4	0.426	0.266	0.141	0.168
5	0.484	0.161	0.225	0.131
6	0.244	0.323	0.162	0.271
7	0.320	0.280	0.213	0.187
8	0.239	0.439	0.188	0.135
9	0.177	0.158	0.392	0.273
10	0.212	0.213	0.179	0.396
11	0.241	0.189	0.324	0.246
12	0.197	0.237	0.254	0.313
13	0.130	0.301	0.271	0.298
14	0.283	0.228	0.332	0.157
15	0.200	0.240	0.226	0.334
16	0.276	0.243	0.141	0.340

Result and discussions

^{1}H-NMR spectra for 16 different chemical mixtures are shown in Figure 3(a), and the overlapped peak region around 2 ppm is expanded in Figure 3(b). For comparison, the result by using the conventional initial estimation by SVD and the following MCR calculation is shown in Figure 4. In the SVD calculation, 4 components were selected because the number of compoents is known in advance in this case, and the following MCR calcuration was processed under the constraints of the number of components and the non-negativity of spectral intensities and concentrations. Figure 4(a) is the spectra obtained by the initial estimation by SVD, and Figure 4(b) shows the result after the MCR calculation.

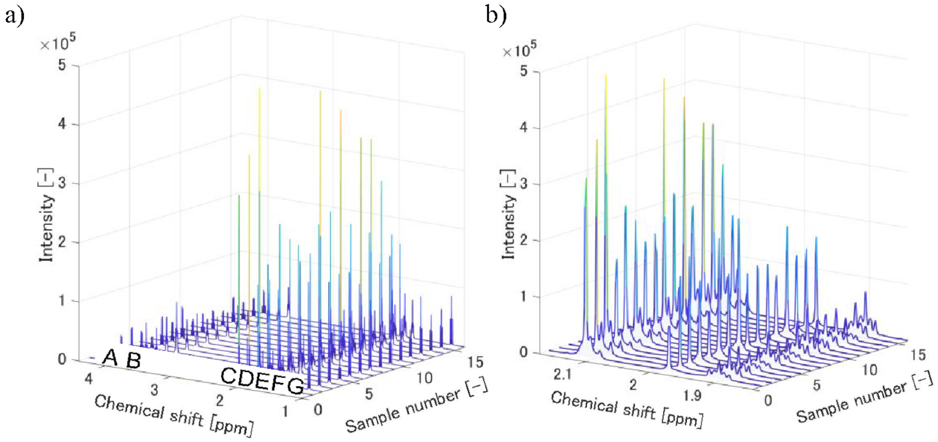


Figure 3 ^{1}H-NMR spectra of chemical mixtures of 4 chemicals (THF, ethyl acetate, acetone, cyclopentanone) (a) in the whole range and (b) enlarged around 2 ppm.

The comparison between Figure 2(b) and 4(b) clearly showed an inconsistency between them. From the similarity between the initial guess by SVD and the obtained result by MCR (Figure 4(a) and (b)), the final result by MCR is greatly affected by the initial estimation, and a reliable initial estimation is needed for the MCR calculation.

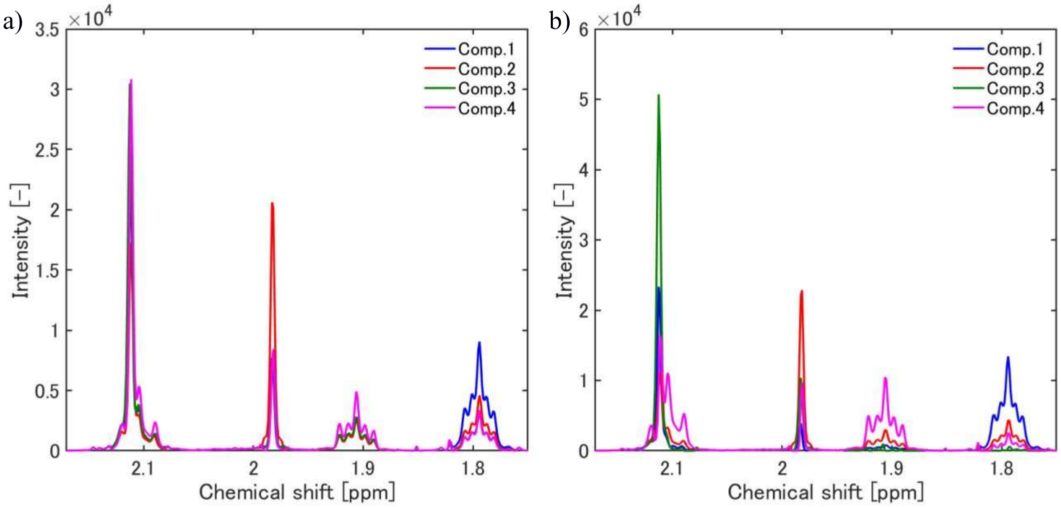


Figure 4 The estimated ^{1}H-NMR spectra are shown; (a) the initial estimation by singular value decomposition, and (b) the optimized spectra of (a) after the MCR calculation.

Then, the initial estimation described in the theory was utilized. At first, the *i*coshift calculation was processed for all the spectrum data, which adjusted the minor peak shifts, and the spectra were centered to the average spectrum to have a variation of the peak intensity from the average spectrum for the samples. (Scheme 1(a)) Next, the spectra were convoluted with a gaussian function to blur the peaks to some extent (Scheme 1(b)), and separated into 7 peak regions (Scheme 1(c)).

To recognize if each peak region has an overlapped region, the first *cos-s* calculation (Scheme 1(d)) for each region was processed based on Eqn. (4). Figure 5 represents a correlation mapping at each region and decided if there is an overlap of different chemicals based on the criteria as described in Scheme 2. The averaged similarity values are also presented in each map. In the six out of seven regions, the average values were larger than 0.8; meaning a single chemical constitutes the peak, except for the peak region C, which showed lower similarity values in the correlation mapping, and it was evaluated as an overlapped peak region.

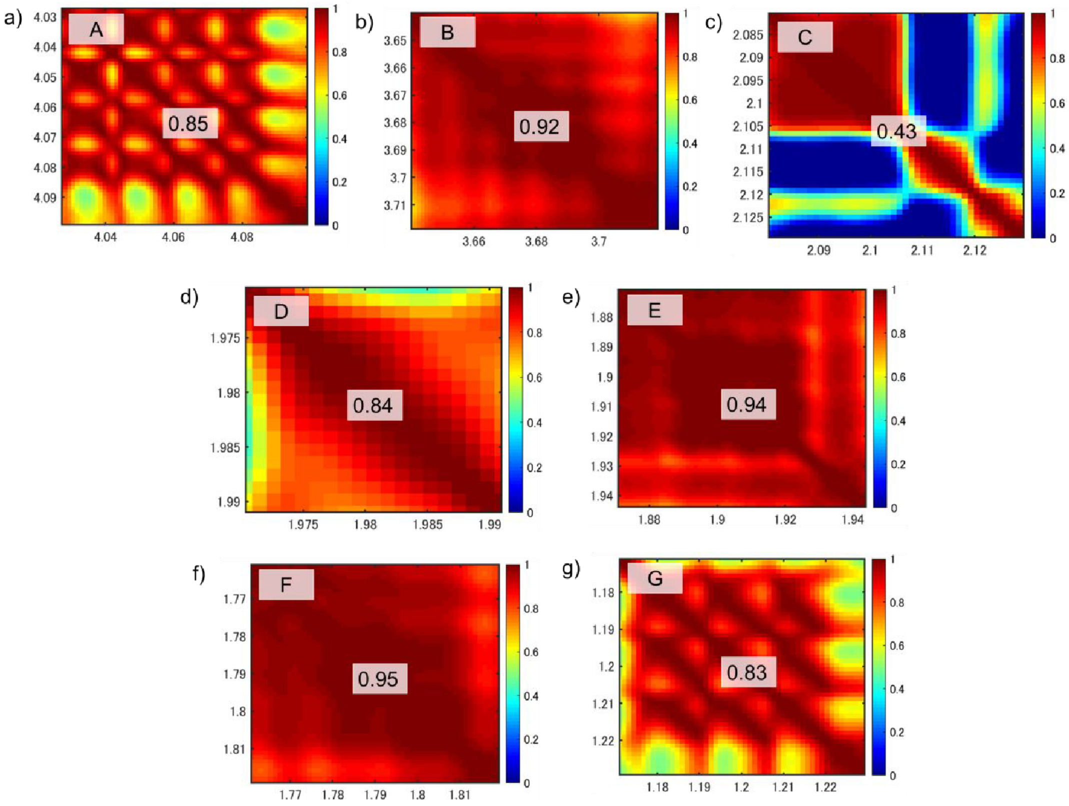


Figure 5 The correlation mapping in each peak region is shown, obtained by the first *cos-s* calculation. The alphabets correspond to the peak regions labelled in Figure 3. The average similarity values are shown in the center.

Table 2 The table of the cosine similarity values between different peak regions. They were calculated for the peak area in each region.

	A	B	C	D	E	F	G
A		-0.031	-0.554	0.998	-0.173	-0.032	1.000
B			-0.635	-0.081	-0.498	1.000	-0.033
C				-0.513	0.338	-0.635	-0.553
D					-0.135	-0.081	0.998
E						-0.493	-0.170
F							-0.033
G							

The second step of the *cos-s* calculation was processed to examine if multiple peak regions are originated from an identical chemical species, based on Scheme 1(d) and Eqn. (5). The correlation table between the peak regions is tabulated in Table 2. The peak region, A had a correlation with D and G, from which these three regions are originated from the identical chemicals. Similarly, B and F were due to the same chemical. However, the correlation between C and E showed medium values for any other peak regions, and further analyses were necessary.

Based on the second *cos-s* estimation, the peak regions, C and E showed a weak correlation. Also, the first *cos-s* estimation indicated the spectral overlap in the peak region C. Considering these results, it is assumed that the peak region C includes the overlap region which is attributed to the same chemical origin as the peak region E. Thus, by applying Eqn. (6) to the peak region C as $\boldsymbol{s_{overlapped}}$

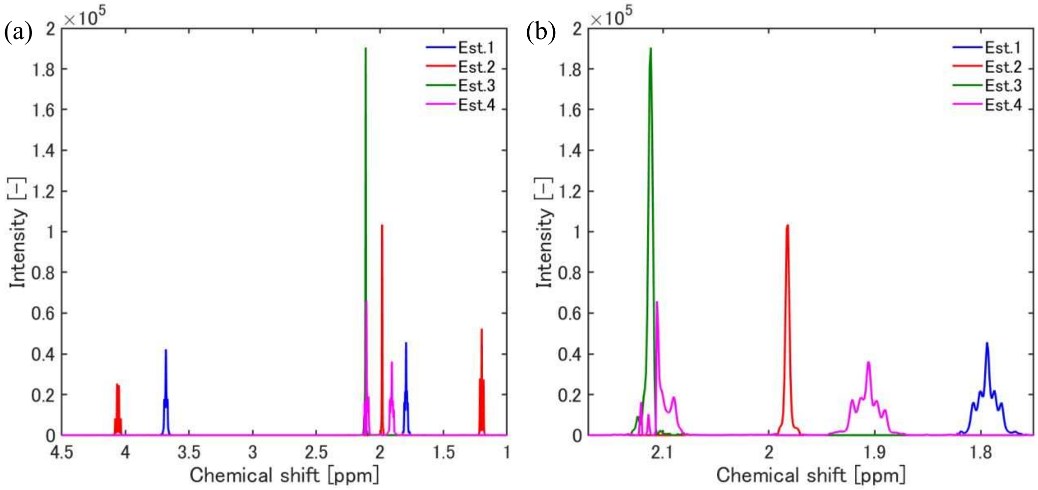


Figure 6 (a) The initial estimated spectra by using the cosine similarity estimations.in the whole spectrum region. and (b) an enlarged figure around 2 ppm region.

and the peak region E as $\boldsymbol{s_{pure}}$, the corresponding pure spectrum from the overlapped region C (Eqn. (7)) was extracted. At the same time, the non-correlated spectrum was calculated by Eqn. (8). The initial estimation of the spectra is calculated by Eqn. (9), (10) and shown in Figure 6. ^{1}H-NMR spectra for 4 chemicals were estimated without any prior information.

After this new initial estimation, the MCR calculation was processed for the spectrum/sample data matrix to refine the spectra for pure chemicals and also to obtain the concentrations of the chemicals in the samples. The result is shown in Figure 7 with the reference spectra and the experimentally prepared concentration ratio. From the comparison between the initially estimated spectra (Figure 6(b)) and the MCR-optimized spectra (Figure 7(d)), it is obvious that the spectral shape was optimized. More detaild performance of the calculated resutls are tablated in the Table. 3 and Table. 4 with the relative errors of the calcualted concentrations from the orignal sample concentrations and the correaltion coefficients between the pure reference spectra and the caluclated spectra. Therefore, even if the initial estimation would have a minor error, the MCR process could

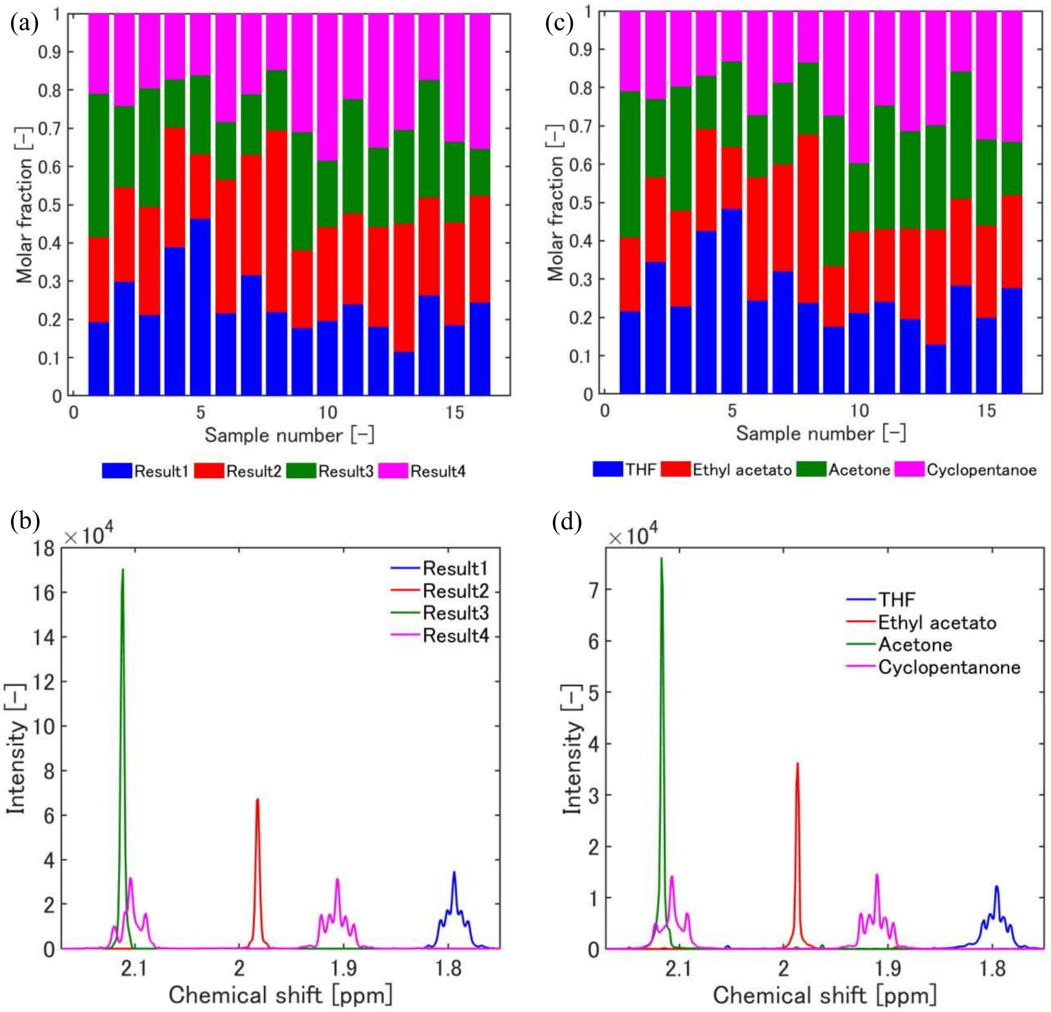


Figure 7 The concentration profile (a) and the pure spectra (b) recovered by the newly developed *cos-s* MCR is shown. For comparison, the actual ratio of the prepared samples (c), and the reference spectra of pure chemicals (d) are shown.

adjust it. From this method, even though we do not have any prior knowledge such as the number and the species of chemicals and the mixture ratio, we could recover the spectra for pure chemicals with more than 85 % consistency and the concentration profile for each mixture within an accuracy of less than 10 % error on average. There was some minor systematic error in the concentrations, and it is now under research.

Table. 3 Relative error (%) of the calculated concentrations from the prepared ones.

Sample number	THF	Ethyl acetate	Acetone	Cyclopentanone
1	11.30	-16.86	1.94	0.09
2	13.36	-11.47	-4.41	-5.16
3	7.72	-13.51	4.20	1.25
4	8.52	-17.86	10.5	-2.07
5	4.15	-5.46	8.42	-23.00
6	11.35	-8.19	6.36	-4.26
7	1.31	-12.41	25.71	-13.02
8	7.83	-8.42	16.04	-8.84
9	-0.23	-28.63	20.91	-13.24
10	7.64	-15.75	2.73	3.15
11	0.15	-24.79	7.03	9.68
12	8.22	-10.17	17.75	-11.87
13	10.92	-11.70	9.93	-2.05
14	7.08	-13.02	7.47	-9.64
15	7.23	-11.97	6.64	-0.24
16	11.5	-14.93	12.14	-3.74

Table. 4 Correlation coefficients between the calculated spectra and the reference ones.

Chemicals	Correlation coefficients
THF	0.9731
Ethyl acetate	0.9606
Acetone	0.8738
Cyclopentanone	0.9645

Although the peak splitting due to the spin-spin coupling and the integrated area of the recovered spectra almost matched with the pure ones, some peak positions did not perfectly match, which caused a minor error. It is supposed that this is caused by a small shift of the spectra due to the intermolecular interaction in chemical mixtures.[44] In this application, the linear combination of the multiple pure spectra with different coefficients was utilized from the viewpoint of qualitative and quantitative chemical analyses, and this minor difference of the chemical shift was ignored. With the obtained accuracy, it is safely stated that the pure spectra can be recovered by using the MCR method with the new initial estimation using cosine similarity. On the other hand, it is a useful application if the chemical environemnt were varied, leading to the spectral change for species, due to molecular interaction,[45] intermediate or aggregate formation, etc, and the monitoring of the chemical state change is another important application of MCR,[7,46] and our method will be extended for such applications, too.

Conclusion

We have developed a new estimation technique from the multiple spectral information of unknown chemical mixtures to extract the pure spectra and the concentration profiles of them. We utilized a combination methodology of the multivariate curve resolution and the cosine similarity as an initial estimation instead of using the conventional singular value decomposition. By applying this method to ^{1}H-NMR spectral data of chemical mixtures, we could obtain the spectra for pure chemicals and the concentration profiles in the chemical mixtures with high accuracy. By using this robust initial estimation procedures, the process can be completed without using reference spectra, therefore it can be applied to many other spectral data to extract the quantitative and qualitative information.

Acknowledgments

The research was financially supported by the Institute of Science and Engineering, Chuo University, JST PRESTO (#JPMJPR1675), The Science Research Promotion Fund from the Promotion and Mutual Aid Corporation for Private Schools of Japan.

References

1 Y. Uwadaira, Y. Sekiyama and A. Ikehata, *Heliyon*, 2018, **4**, e00531.

2 H. Parastar and R. Tauler, *Anal. Chem.*, 2014, **86**, 286–297.

3 D. W. Osten and B. R. Kowalski, *Anal. Chem.*, 1984, **56**, 991–995.

4 E. Bezemer and S. Rutan, *Anal. Chem.*, 2001, **73**, 4403–4409.

5 J. C. Nicholson, J. J. Meister, D. R. Patil and L. R. Field, *Anal. Chem.*, 1984, **56**, 2447–2451.

6 S. Navea, A. de Juan and R. Tauler, *Anal. Chem.*, 2002, **74**, 6031–6039.

7 A. Domínguez-Vidal, M. P. Saenz-Navajas, M. J. Ayora-Cañada and B. Lendl, *Anal. Chem.*, 2006, **78**, 3257–3264.

8 R. M. Bianchini and T. S. Kaufman, *Int. J. Chem. Kinet.*, 2013, **45**, 734–743.

9 J. Jaumot, V. Marchán, R. Gargallo, A. Grandas and R. Tauler, *Anal. Chem.*, 2004, **76**, 7094–7101.

10 J. Saurina, S. Hernández-Cassou and R. Tauler, *Anal. Chem.*, 1997, **69**, 2329–2336.

11 M. C. Antunes, J. E. J. Simão, A. C. Duarte and R. Tauler, *Analyst*, 2002, **127**, 809–817.

12 J. Jaumot, B. Igne, C. A. Anderson, J. K. Drennen and A. de Juan, *Talanta*, 2013, **117**, 492–504.

13 R. R. de Oliveira, K. M. G. de Lima, R. Tauler and A. de Juan, *Talanta*, 2014, **125**, 233–241.

14 M. A. Hegazy, N. S. Abdelwahab and A. S. Fayed, *Spectrochim. Acta. A. Mol. Biomol. Spectrosc.*, 2015, **140**, 524–533.

15 S. Nigam, A. de Juan, R. J. Stubbs and S. C. Rutan, *Anal. Chem.*, 2000, **72**, 1956–1963.

16 S. Nigam, A. de Juan, V. Cui and S. C. Rutan, *Anal. Chem.*, 1999, **71**, 5225–5234.

17 R. Gusmão, S. Cavanillas, C. Ariño, J. M. Díaz-Cruz and M. Esteban, *Anal. Chem.*, 2010, **82**, 9006–9013.

18 D. W. Cook and S. C. Rutan, *Anal. Chem.*, 2017, **89**, 8405–8412.

19 H. Parastar, J. R. Radović, M. Jalali-Heravi, S. Diez, J. M. Bayona and R. Tauler, *Anal. Chem.*, 2011, **83**, 9289–9297.

20 P. Conti, S. Zamponi, M. Giorgetti, M. Berrettoni and W. H. Smyrl, *Anal. Chem.*, 2010, **82**, 3629–3635.

21 M. B. Mamián-López and R. J. Poppi, *Microchem. J.*, 2015, **123**, 243–251.

22 C. J. G. Colares, T. C. M. Pastore, V. T. R. Coradin, L. F. Marques, A. C. O. Moreira, G. L. Alexandrino, R. J. Poppi and J. W. B. Braga, *Microchem. J.*, 2016, **124**, 356–363.

23 J. Jaumot and R. Tauler, *Analyst*, 2015, **140**, 837–846.

24 S.-T. Tan, K. Chen, S. Ong and W. Chew, *Analyst*, 2008, **133**, 1395–1408.

25 M. R. Alcaráz, A. Schwaighofer, H. Goicoechea and B. Lendl, *Spectrochim. Acta. A. Mol. Biomol. Spectrosc.*, 2017, **185**, 304–309.

26 E. Ortiz-Villanueva, F. Benavente, B. Piña, V. Sanz-Nebot, R. Tauler and J. Jaumot, *Anal. Chim. Acta*, 2017, **978**, 10–23.

27 M. M. Sinanian, D. W. Cook, S. C. Rutan and D. S. Wijesinghe, *Anal. Chem.*, 2016, **88**, 11092–11099.

28 M. Navarro-Reig, J. Jaumot, A. Baglai, G. Vivó-Truyols, P. J. Schoenmakers and R. Tauler, *Anal. Chem.*, 2017, **89**, 7675–7683.

29 F. Puig-Castellví, I. Alfonso and R. Tauler, *Anal. Chim. Acta*, 2017, **964**, 55–66.

30 Y. Casamayou-Boucau and A. G. Ryder, *Anal. Chim. Acta*, 2018, **1000**, 132–143.

31 A. Z. Samuel, M. Zhou, M. Ando, R. Mueller, T. Liebert, T. Heinze and H. Hamaguchi, *Anal. Chem.*, 2016, **88**, 4644–4650.

32 A. Z. Samuel, B.-H. Lai, S.-T. Lan, M. Ando, C.-L. Wang and H. Hamaguchi, *Anal. Chem.*, 2017, **89**, 3043–3050.

33 M. Ando and H. Hamaguchi, *J. Biomed. Opt.*, 2013, **19**, 011016.

34 J. P. Smith, F. C. Smith and K. S. Booksh, *Analyst*, 2017, **142**, 3140–3156.

35 D. Zhang, P. Wang, M. N. Slipchenko, D. Ben-Amotz, A. M. Weiner and J.-X. Cheng, *Anal. Chem.*, 2013, **85**, 98–106.

36 H. Noothalapati and S. Shigeto, *Anal. Chem.*, 2014, **86**, 7828–7834.

37 L. Zhang, T. Cambron, Y. Niu, Z. Xu, N. Su, H. Zheng, K. Wei and P. Ray, *Anal. Chem.*, 2019, **91**, 2784–2790.

38 J. Kuligowski, G. Quintás, R. Tauler, B. Lendl and M. de la Guardia, *Anal. Chem.*, 2011, **83**, 4855–4862.

39 C. Fauteux-Lefebvre, F. Lavoie and R. Gosselin, *Anal. Chem.*, 2018, **90**, 13118–13125.

40 R. B. Pellegrino Vidal, A. C. Olivieri and R. Tauler, *Anal. Chem.*, 2018, **90**, 7040–7047.

41 J. A. Johnson, J. H. Gray, N. T. Rodeberg and R. M. Wightman, *Anal. Chem.*, 2017, **89**, 10547–10555.

42 F. Savorani, G. Tomasi and S. B. Engelsen, *J. Magn. Reson.*, 2010, **202**, 190–202.

43 J. Jaumot, A. de Juan and R. Tauler, *Chemom. Intell. Lab. Syst.*, 2015, **140**, 1–12.

44 J. V. Hatton and R. E. Richards, *Trans. Faraday Soc.*, 1961, **57**, 28–33.

45 M. Besemer, R. Bloemenkamp, F. Ariese and H.-J. van Manen, *J. Phys. Chem. A*, 2016, **120**, 709–714.

46 S. Navea, R. Tauler and A. de Juan, *Anal. Chem.*, 2006, **78**, 4768–4778.

Figure captions

Scheme.1 The overall workflow of *cos*-s MCR is shown. The algorithm is divided into 7 sections: (a) data pre-processing, (b) spectral convolution, (c) division of the spectral regions, (d) cos-s estimation, (e) data augmentation, (f) MCR optimization, (g) deconvolution. The cos-s estimation is consisted of 3 steps; the peak overlap is evaluated in the first step, and the peak region correlation is examined in the second step, and the overlapped peak region is separated.

Scheme.2 The first step *cos-s* similarity procedure is shown. This process corresponds to the first process of Scheme 1(d). In each peak region, the similarity of the spectral intensity is examined. Based on the average similarity values in each peak region, the evaluation was made if there is an overlap in the region. The threshold of the similarity value was set to 0.8.

Scheme.3 The second cos similarity procedure is shown. This process corresponds to the second process of Scheme 1(d). The similarity between the peak regions were evaluated. Based on the similarity values between the peak regions, it is evaluated if the peaks are derived from the same chemical.

Figure 1 The schematic representation how to extract pure spectrum from the overlapped spectral region by using the correlation between the peak regions.

Figure 2 ^{1}H-NMR spectra of 4 chemicals (tetrahydrofuran, ethyl acetate, acetone, cyclopentanone) in the whole range (a) and enlarged around 2 ppm where the spectra are overlapped (b).

Figure 3 ^{1}H-NMR spectra of chemical mixtures of 4 chemicals (THF, ethyl acetate, acetone, cyclopentanone) (a) in the whole range and (b) enlarged around 2 ppm.

Figure 4 The estimated ^{1}H-NMR spectra are shown; (a) the initial estimation by singular value decomposition, and (b) the optimized spectra of (a) after the MCR calculation.

Figure 5 The correlation mapping in each peak region is shown, obtained by the first *cos-s* calculation. The alphabets correspond to the peak regions labelled in Figure 3. The average similarity values are shown in the center.

Figure 6 (a) The initial estimated spectra by using the cosine similarity estimations.in the whole spectrum region. and (b) an enlarged figure around 2 ppm region.

Figure 7 The concentration profile (a) and the pure spectra (b) recovered by the newly developed *cos-s* MCR is shown. For comparison, the actual ratio of the prepared samples (c), and the reference spectra of pure chemicals (d) are shown.

Table. 1 The molar fractions of the prepared samples.

Table 2 The table of the cosine similarity values between different peak regions. They were calculated for the peak area in each region.

Table. 3 Relative error (%) of the calculated concentrations from the prepared ones.

Table. 4 Correlation coefficients between the calculated spectra and the reference ones.