Analyst

Multivariate Curve Resolution Combined with Estimation by Cosine Similarity Mapping of Analytical Data

Journal:	*Analyst*
Manuscript ID	AN-ART-03-2021-000362.R5
Article Type:	Paper
Date Submitted by the Author:	20-Jun-2021
Complete List of Authors:	Nagai, Yuya; Chuo University, Department of Chemistry Katayama, Kenji; Chuo University, Department of Chemistry; Kagaku Gijutsu Shinko Kiko, Precursory Research for Embryonic Science and Technology

Multivariate Curve Resolution Combined with Estimation by Cosine Similarity Mapping of Analytical Data

Yuya Nagai[1] and 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

We developed a multivariate curve resolution (MCR) calculation combined with the mapping of cosine similarity (*cos-s*) for multiple mixture spectra of chemicals. The *cos-s map* was obtained by calculating the similarities of the variation of the signal intensities at each scanning parameter, such as the wavelength. The *cos-s map* was utilized for the initial estimation of the spectra for pure chemicals and also for the restriction of the iterative least-square calculation of MCR. These calculations were performed without arbitrary parameters by introducing the soft clustering to the *cos-s map*. The chemically meaningful initial estimation could prevent the convergence at an incorrect local minimum, which frequently happens for the wrong initial estimation of spectra far away from the real answer. Herein, we demonstrated the robustness of this calculation method by applying it for UV/Vis spectra and XRD patterns of multiple unknown chemical mixtures, whose shapes were totally different (broad overlapped peaks and multiple complicated peaks). Pure spectra/patterns were recovered as >84% consistency with the reference spectra, and <6% accuracy of the concentration ratios was demonstrated.

Keywords: multivariate curve resolution, initial estimation, cosine similarity, fuzzy c-means, non-parametric

Introduction

Multivariate curve resolution (MCR) has been extended for various applications of the separation of spectral data of chemical mixtures.[1] In analytical chemistry, much effort is devoted to the separation of chemicals to purify a single species for qualitative and quantitative determination. However, if we could obtain this information only from the spectra of chemical mixtures, tremendously chemical processes will be reduced for the analyses. This is the main reason why the MCR has been studied and developed in the field of analytical chemistry.

Two-dimensional correlation spectroscopy (2D-COS) is also an active research field to assign peaks and study their correlation in the spectra of chemical mixtures from 2D synchronous and asynchronous spectra obtained by applying a perturbation such as a temperature and concentration change.[2–4] The temperature dependent mixture spectra were used for demonstration of the identification of different protein moieties.[5] The correlation between the functional groups could be analyzed by applying temperature for polymer blend films.[6] 2D synchronous spectra obtained by changing temperature could reveal minor spectral differences for mixture of olive oils.[7] Recently, 2D asynchronous spectra can be used for the identification of the implicit isolated peaks from severely overlapped peaks and could successfully analyze bilinear data from mixture spectra.[8]

In MCR, spectral data of chemical mixtures as a number matrix is decomposed into a matrix of the spectra of pure chemicals (S) and a matrix of the concentration ratios (C) in the mixtures. This calculation accuracy has been drastically improved after the introduction of the alternating least square method (MCR-ALS).[9,10] The C and S matrixes are iteratively optimized sequentially, and also various restraints can be included in each iteration, such as non-negativity, limitation of concentrations, number of species,[11] area correlation,[12] etc. This chemically informative reconstruction of C and S is an advantage over the traditional methods such as the principal component analysis (PCA) and the partial least square (PLS).

In analytical chemistry, it was initially applied for the chromatography data,[13–15] and extended for the chromatographic spectral data,[12,16] and for many other analytical data such as UV/Vis,[17–19] NIR,[20,21] X-ray absorption.[22] Recently, an inhomogeneous sample surface was scanned combined with microscopic technique, and the mixture spectra were used for the component analysis, especially for biological applications using Raman spectroscopy.[23–26] For other utilities, it has been utilized for the data reduction for hyperspectral imaging data and for denoising for 2D-NMR data.[27] Furthermore, MCR was combined with deep learning and improved for the component analysis of the gas chromatography-mass spectrometry data.[28]

Still, there are several problems for MCR when the spectra include strong background,[29] several unidentified components,[30] rotational ambiguities,[31] and unprecedented initial estimations.[32] In the initial estimation, the spectra for pure chemicals are conventionally obtained by the singular

value decomposition (SVD). However, it is sometimes far from the real answer, and the iteration calculation often stops at the local minimum. To overcome this problem, there have been several efforts to improve the initial estimation, and one of the methods is called PURE used in SIMPLISMA, [33,34], where pure data points for each species are used for the prediction of the pure spectra. There are several other methods to solve the drawback of PURE that needs a pure component in the data, such as independent component analysis (ICA) [35] and orthogonal projection approach (OPA).[36] We also have developed the categorization of the spectral components by using the cosine similarity (*cos-s*) of the peak intensity correlation in the previous paper.[37] 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. This reasonable initial estimation can reduce the uncertainty of the matrix decomposition, which solved the problem of rotational ambiguity.[38] This method was applied for NMR spectra.

In principle, this calculation can be applied to any mixture spectra measured by different analytical methods without any prior information about the sample system. However, in practice, there are several difficult cases where this calculation cannot be applied as it is. One of the difficult cases is for the spectra with overlapped multiple broad peaks such as absorption spectra. In the initial estimation, the overlapped peaks must be separated at each wavelength for different chemical species, and the accuracy is lowered compared with the peak-isolated spectra like NMR. The other problem is the misclassification for the spectra with strong background or low signal-to-noise (SN) ratio. In this case, it is necessary to use multiple threshold parameters for better categorization, but these parameters are adjusted depending on analytical methods.

To overcome these problems, we have developed a new, improved initial estimation and optimization method with reasonable assignments of overlapped peaks without adjustable parameters by employing the similarity map and soft clustering. The similarity map visualizes the internal correlation of the whole region of spectral data, and this map is basically same as the 2D synchronous spectra used in 2D-COS except that the spectral intensities are normalized in the cosine similarity, and can provide the coincidental intensity changes of signal pairs in the sampling.[39] And each peak region can be automatically assigned without manual assignments, meaning that no parameter tuning was necessary by utilizing the clustering for the similarity map. The highly-trustable initial estimation was also utilized for the restraints in the iterative MCR optimization. Hereinafter we call the *cos-s map* estimation and the following MCR calculation as *cos-s map* MCR. This calculation could extract the pure spectra and their concentration profiles with high accuracy without adjustable parameters by users and does not need any prior information. We selected two general problems of chemical mixtures, which can be solved by other initial estimation methods; UV/Vis absorption spectra and X-ray diffraction (XRD) patterns. In this paper, we applied our calculation techiniqe to two distinct general problemsand showed the robustness and versatility of

this calculation technique for the mixture spectra.

Theory and method

Assuming that multiple spectra for mixture samples with different compositions are obtained, the spectrum for the ith mixture is represented as S_{ij} (j: parameter for the spectrum such as wavelength). We processed the spectral data matrix in Scheme 1; (a): data pre-processing and initial estimation of pure spectra by *cos-s* mapping, (b): iterative calculation of pure spectra and concentration ratios by MCR.

In Scheme 1(a), the background was removed by a built-in function of MATLAB (R2020a) named msbackadj, which regresses the background to the multiple shifted windows using spline approximation. It is used only when the background level was similar to the signal level. When spectra have isosbestic points, observed for UV/Vis absorption spectra, random numbers (0.5-1.5) were multiplied to the mixture spectrum to provide the dispersion at the isosbestic points intentionally. The signal variation is necessary for the peak intensity assignment for the overlapped peaks. Next, the peak shifts due to the equipment and sampling errors were removed by the *i*coshift algorithm.[40] These pre-processing calculations for mixture spectra were general methods for spectral analyses, and they are effective for the MCR optimization because the intensities of the original data are directly proportional to the finally obtained concentrations and spectra. Even after application of *i*coshift, peak distortion or peak splitting could cause non-proportionality of the peak intensity.[37] To solve this, each spectrum was convoluted by a Gaussian function with a width less than 5% of the total window size. When no-signal regions of the spectra exceed 50 %, such as NMR and XRD patterns, no-signal regions were removed to reduce the data points.

The pre-processed spectra were centered by their average spectra for all the mixture samples ($\bar{\boldsymbol{s}}_j$) to extract the variation for different samples as shown in Eqn. (1)

$$s_{ij} = S_{ij} - \bar{\boldsymbol{s}}_j \quad (1),$$

where s_{ij} reflects the spectral intensity variation for each mixture spectrum.

From the similarity of the spectral intensity trend for different mixture samples, the spectral regions for pure chemicals were decided, and those for the overlapped regions were separated, as shown in Scheme 1(a). *Cosine similarity* was used to understand the spectral intensity tendency quantitatively and was calculated as:

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

where n and m represent the total sample number and the number of discrete data-points in a spectrum. This calculation provides the correlation matrix indicating the correlation between the spectral intensities for the two wavelengths, j_1 and j_2 in the direction of the sample number. The size of the correlation matrix is $m \times m$, converted into a heat map, and this two-dimensional correlation map is called a similarity map. This is the normalized version of the 2D synchronous spectrum in

2D-COS, providing information on the coincidental intensity changes of pairs of peaks.[39] From the analysis of the similarity map, the spectral regions for the same chemical species were decided. Furthermore, the overlapped regions for multiple chemical species could be recognized and separated for each chemical species based on the similarity values as described in the previous paper.[37] However, the separation criteria depended on the threshold value of the similarity, determined by analysts. Fuzzy c-means (FCM) clustering[41] has been introduced for the automatic feature extraction of the similarity map to avoid the baseless decision. Clustering is one of the unsupervised learning methods for data classification based on the distance function. The distance function evaluates the internal pattern of the data, and the data is divided into groups according to the similarity of the recognized patterns. In the calculation of FCM, similarity vectors, $\boldsymbol{x}_j$ (j = 1,2,…m) ($m \times 1$), which were the row vectors in the similarity map, were classified into q clusters, and the center of each cluster ($\boldsymbol{v}_k$ ($m \times 1$): central vector of the k-th cluster), and its weight to all clusters ($u_{j,k}$: contribution of the j-th feature value to the cluster k) were obtained by solving this equation.

$$\arg \min_{\boldsymbol{u},\boldsymbol{v}} \left(\sum_{j=1}^{m} \sum_{k=1}^{q} {u_{j,k}}^{p} \|\boldsymbol{x}_j - \boldsymbol{v}_k\|^2 \right) \quad (3),$$

where p is the exponent for u. The number of species q was iteratively determined by assessing the finally obtained mean square error.

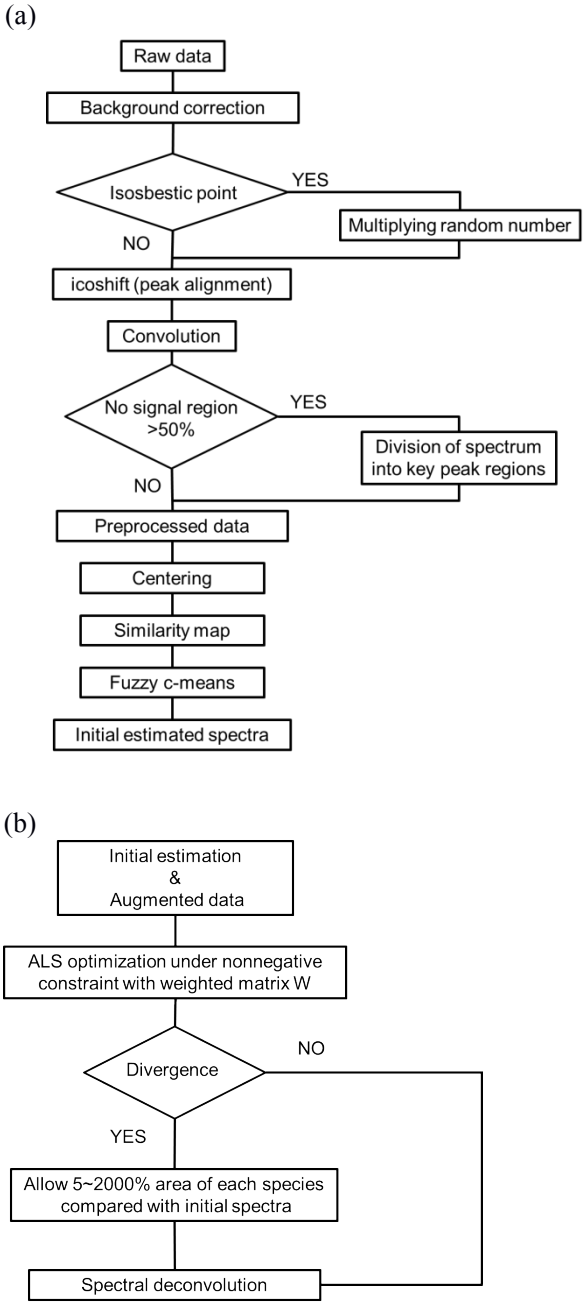
From the similarity matrix of $X = (cos\theta)_{j=j_1,j_2}$, ($m \times m$), $\boldsymbol{v}_k$ and its weight $u_{j,k}$ are estimated. The result of FCM clustering can be directly utilized to estimate the shape of the pure spectra because $u_{j,k}$ is regarded as the contribution of the k-th chemical species to the similarity map, and p was regarded as 2. (Appendix in Supporting Information (SI)). Since the spectral intensities were centered, the differences from the averaged values over samples were obtained. The initial estimation of the spectra was calculated based on the averaged spectra as:

$$S_{est,j,k} = {u_{j,k}}^2 \bar{s}_j \quad (4)$$

$S_{est,j,k}$ represents the initial estimation of the spectrum of the k-th component at wavelength j. $\bar{s}_j$ is the average spectral intensity in the whole mixture samples.

Scheme 1 (b) represents the updated alternating least square (ALS) optimization process under several constraints after the initial estimation. Before the ALS optimization, the number of spectral data for the calculation was augmented about 100-500 by mixing the measured mixture spectra with arbitrary ratios to improve the calculation accuracy. During the ALS optimization, the weight matrix was multiplied to the estimated spectra in every iteration to reflect the initial estimation to the final result. This procedure can avoid unrealistic peaks during the iteration. The weight matrix was obtained in Scheme 2, as described below. The ALS algorithm was performed until the convergence was reached for both the concentration and spectrum profiles. The calculation was converged when the difference between the lack of fit of each iteration was below 0.1%. The lack of fit is the ratio of

the data matrix and the residual after the fitting.[42] If the calculation diverges, additional constraints were also used in the optimization process by restricting the spectral area from 5% to 2000% to the spectra obtained by the initial estimation. Finally, the optimized spectra were deconvoluted by the Gaussian function used for the convolution in Scheme 1.



Scheme.1 The analysis flow charts for the cosine-similarity map multivariate curve resolution (*cos-s map* MCR) are shown. (a) Data pre-processing of *cos*-s *map* MCR and initial estimation process of *cos*-s *map* MCR are shown. (b) corresponds to the alternating least square (ALS) optimization based on the initial estimation with constraints.

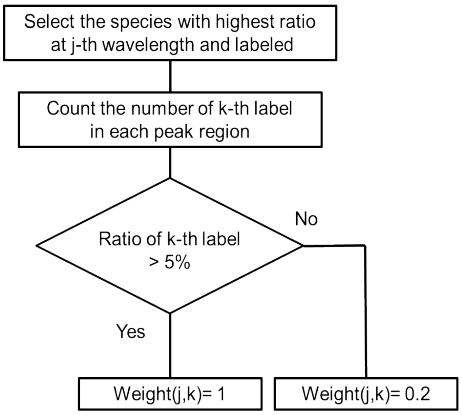
Scheme 2 represents the calculation process for the weight matrix $W=w_{j,k}$, at wavelength j and k-th species. The weight matrix was determined within each peak region defined in Scheme 1(a) and controls how much the initial estimation of the spectra is reflected by multiplying it to the estimated spectra in all iterations. This weight matrix was introduced to maintain the spectral assignment to species in the initial estimation because the peak shape was not accurate in the initial estimation, though the peak assignment for species was correct. First, count $L_{j,k}$, a label for the species with the largest contribution to the spectral intensity at j by evaluating $u_{j,k}$.

$$L_{j,k} = \begin{cases} 1 & if\ k = \arg\max_{k'} (u_{j,k'}) \\ 0 & otherwise \end{cases} \quad (6)$$

By summation of the number of labels ($N_{r,k}$) for the k-th species in each peak region defined in Scheme 1(a). The ratio of the k-th species contribution to the r-th peak region ($p_{r,k}$) was calculated as:

$$p_{r,k} = \frac{N_{r,k}}{\sum_{k=1}^{N} N_{r,k}} = \frac{N_{r,k}}{M_r} \quad (7),$$

where M_r is the wavelength size of each peak region. If the ratio $p_{r,k} < 0.05$, the weight matrix $W=w_{j,k}$ was set to 0.2 to suppress the unnecessary peak evolution during the iteration. Otherwise, it was set to 1 and do nothing in the iteration. This threshold was set to improve the robustness for the peak assignment about the pure peak region. Since a pure peak was typically consisted of at least 20 data points, the peak was properly classified even if it includes a different species component, and this process was helpful for the assignment around the boundary of peak regions. This improved ALS optimization can refine the spectral shape of each species while keeping the peak assignment of the initial estimation.



Scheme 2 Determination of weight matrix based on the initial estimation for convergence in

the ALS optimization.

Experiment

For the demonstration purpose of this calculation, bromothymol blue (BTB) (Wako) was dissolved into 1 M HCl or 1 M NaOH aqueous solutions (Wako), and the pH of each solution was adjusted by deionized (DI) water and phosphate buffer (Wako) (4.5~9). The molecular structure of BTB is provided in Fig. S1 in Supporting Information (SI), together with two moclecular structures in acidic and basic solutions, where the absorption spectrum changes due to the electrolytic dissociation. BTB solutions with eight different pH were prepared at the same final concentration of BTB. The UV/Vis absorption spectra were measured at room temperature (Shimadzu, UV-3600) and shown in Fig. 1. For the reference spectra, BTB solutions of pH = 1.13 and 12.68 were used as strongly acidic or basic conditions. (Fig. S2 in SI). The pKa value was 7.1.[43]

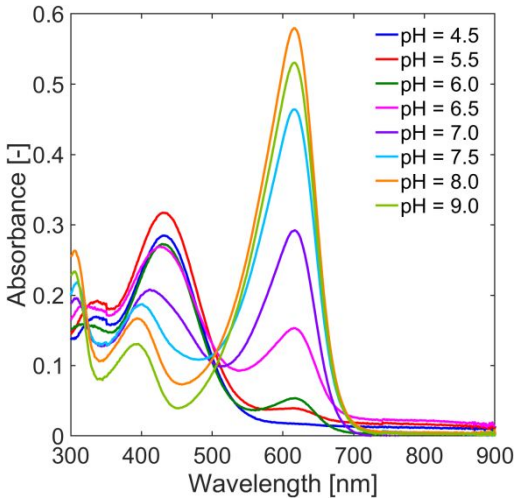


Fig. 1 Absorption spectra of BTB solutions for different pHs from 4.5 to 9 are shown.

With regard to XRD data, iron(III) oxide (Fe_2O_3), iron(II, III) oxide (Fe_3O_4), aluminum (III) oxide (Al_2O_3), and silicon dioxide (SiO_2) (Wako) were utilized. These four powders of chemicals were scaled as the ratios described in Table 1, and mixed by a mortar for five minutes The total mass of each mixutre was 300 mg. The XRD patterns were measured by an X-ray diffractometer (Rigaku, Ultima IV) using Cu-Kα radiation (40 kV and 30 mA). The diffraction patterns were recorded from 20° to 70° (2θ) with a step size of 0.02° (2θ) and a scan rate was 1° per minute. The XRD patterns of 12 samples with different mixture ratios are shown in Fig 2. The reference spectra of the pure

powders are shown in Fig. S3 in SI.

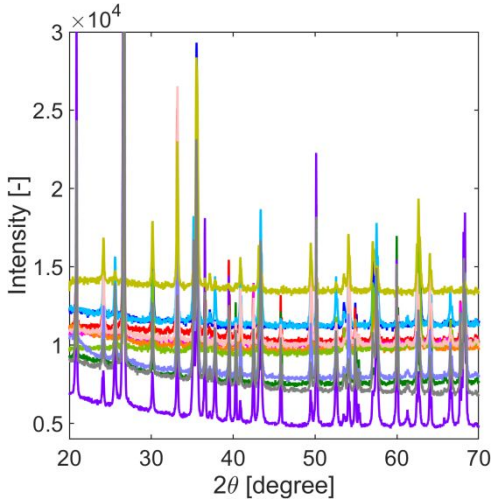


Fig. 2 12 XRD pattern of chemical mixtures of 4 chemicals (Fe_2O_3, Fe_3O_4, Al_2O_3, and SiO_2) are shown.

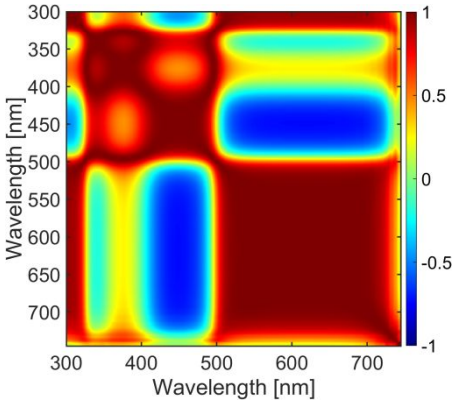
Sample number	Fe_2O_3	Fe_3O_4	Al_2O_3	SiO_2
1	21	61	5.6	12
2	35	25	11	30
3	30	6	14	50
4	33	26	14	27
5	6.6	9.3	36	48
6	29	29	35	7.0
7	32	28	17	23
8	7.8	52	29	11
9	42	6.2	25	27
10	20	17	38	26
11	25	60	8.4	6.1
12	7.1	19	34	40

Table. 1 The mass fractions of the inorganic mixture samples for XRD measurements. (wt%)

Results and discussions

The demonstration for the spectra, including broad overlapped peaks, is shown as a first example. The absorption spectra of the BTB solutions for different pHs were used, as was demonstrated by Shimada et al.[43] In the procedures as described in theory, the data was pre-processed as Scheme 1(a). Since there were merely minor baseline shifts among each spectrum, the

msbackadj was not used, and they were removed by simple subtraction using the average intensity from 745 nm to 900 nm.. Then, the background correction and peak separation were skipped because they had no strong background nor isolated peaks. Before the convolution, random numbers were multiplied to each mixture spectrum since there were two isosbestic points in the spectra, as shown in Fig. 1. In Scheme 1(a), the initial estimation was conducted based on the similarity map. The similarity values were obtained from any two wavelengths j_1 and j_2 by Eqn. (2), which evaluates the



spectral intensity variation in the sample number direction. The similarity values were converted into a heat map as shown in Fig. 3. The map showed a specific pattern corresponding to spectral features of pure chemicals.

Fig. 3 The cosine similarity map of the UV/Vis spectra of the BTB solutions with different pHs obtained as Scheme 1(a) (*cos-s map* estimation).

Based on the fuzzy c-means clustering of the similarity map and the following calculation in Eqn (4), the initial estimation of the spectra was obtained as shown in Fig. 4(a). Two components were extracted and it was consistent with the previous research for BTB acid base equilibrium using spectroscopy.[43] Compared with the reference spectra in Fig. 4(b), the positions of the initial estimation of the spectra were reasonable; however, the spectral shape needs to be optimized, as described in the theory section.

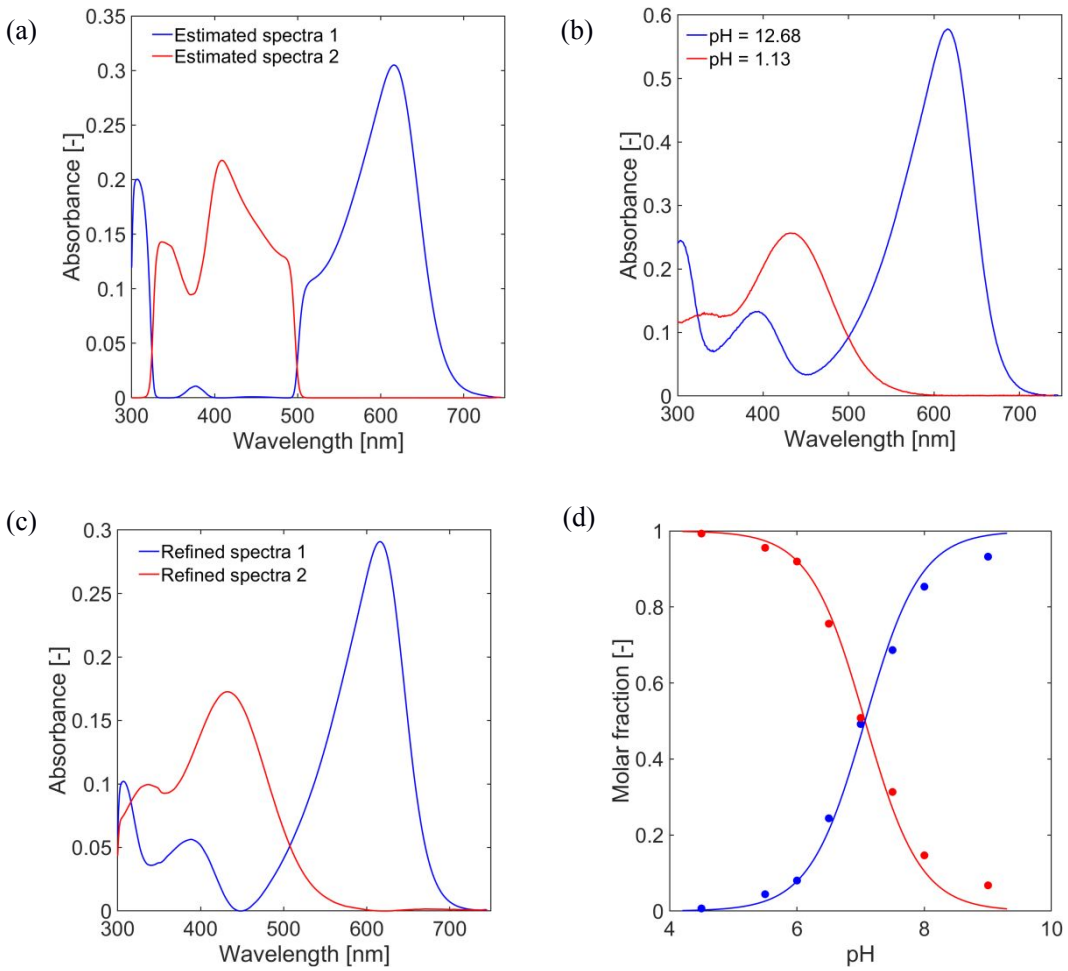


Fig. 4 (a) The initial estimation of the spectra is shown, which were obtained from the mixture spectra of the BTB solutions by using the *cos-s map* estimation. (b) The UV/Vis absorption spectra of the BTB solutions under highly acidic or basic conditions (pH = 1.13 and 12.68) are shown. (c) The absorption refined by the MCR calculation from the initial estimation of the spectra is shown. (d) The concentration ratios of the two components were obtained in the MCR calculation.

Then, the updated ALS optimization was followed as in Scheme 1(b). The recovered spectra were shown in Fig. 4(c), and they matched well with the reference spectra in Fig. 4(b). The spectral shape was refined from the initial estimation of the spectra. The correlation coefficients of the two predicted spectra and the reference spectra under the acidic and basic conditions had a consistency of >99 %.

The concentration profile corresponding for each pure spectrum was obtained in Fig. 4(d). The plots correspond to the concentration ratio of each species. The calculation results were fitted by the least square minimization based on Eqns. (8) from the acid-base equilibrium of BTB.

$$[HA] = \frac{[H^+]}{[H^+] + K_a}, [A^-] = \frac{K_a}{[H^+] + K_a} \quad (8),$$

where K_a is the equilibrium constant. The concentration profiles were well fitted with these equations as a parameter of pKa, and the estimated pK_a was 7.07, in agreement with the literature value.[36] This result indicates that this new calculation technique can be applied for the spectra including broad overlapping peaks.

We would like to mention that this problem can be also solved by the general procedure using PURE in SIMPLISMA and the following MCR calculation. The results were summarized in Fig.S4 in SI. The correlation coefficients between the predicted and the reference spectra was 0.99 on average, and the prediceted pKa was 6.86, which is 3% smaller than reported value. For this simple example, our method had a similar accuracy compared with the results by the PURE-based initial estimation from the comparison between Fig. 4 and Fig. S4.

Next, the *cos*-s *map* MCR was applied for the XRD patterns. In the pre-processing of the data, the procedures in Scheme 1 were performed except the random number multiplication because the XRD patterns had no isosbestic points. The peak intensity trend at every diffraction angle in the sample number direction were converted to cosine similarty values by Eqn. (2). All the similairty values were converted into a heat map as shown in Fig. 5(a), and it represents the internal correlation of the peak intensity at each diffraction angle. The FCM clustering was applied and, the initial estimation of the XRD patterns was calculated. We could confirm four species from the map (Fig. 5(b)). Finally, the updated ALS optimization was applied to refine the pure patterns and to obtain the concentration profiles. For the calculation, weight matrix *W* was used in the calculation (Scheme 2). The recovered patterns and concentrations profile is shown in Figure. 6.

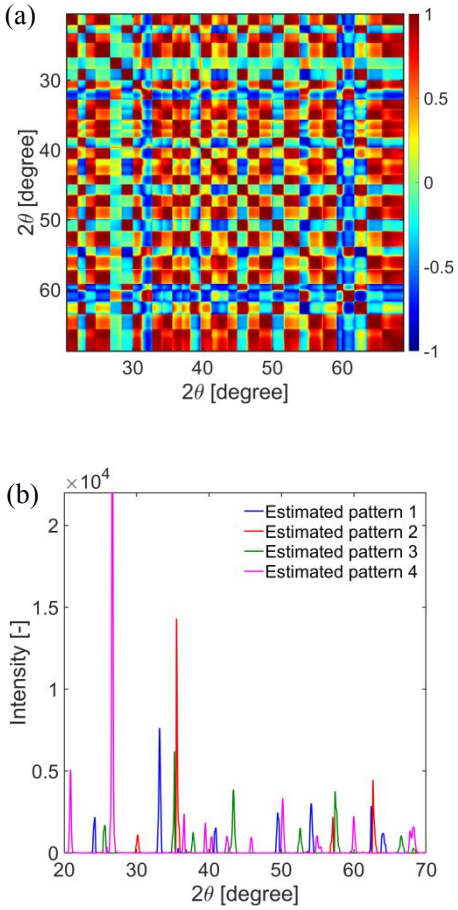


Fig. 5 (a) The *cos-s* map for the XRD patterns obtained in the process of *cos-s map* estimation is shown. (b) The initial estimation of the XRD patterns by using the *cos-s map* estimations in the whole angle region.

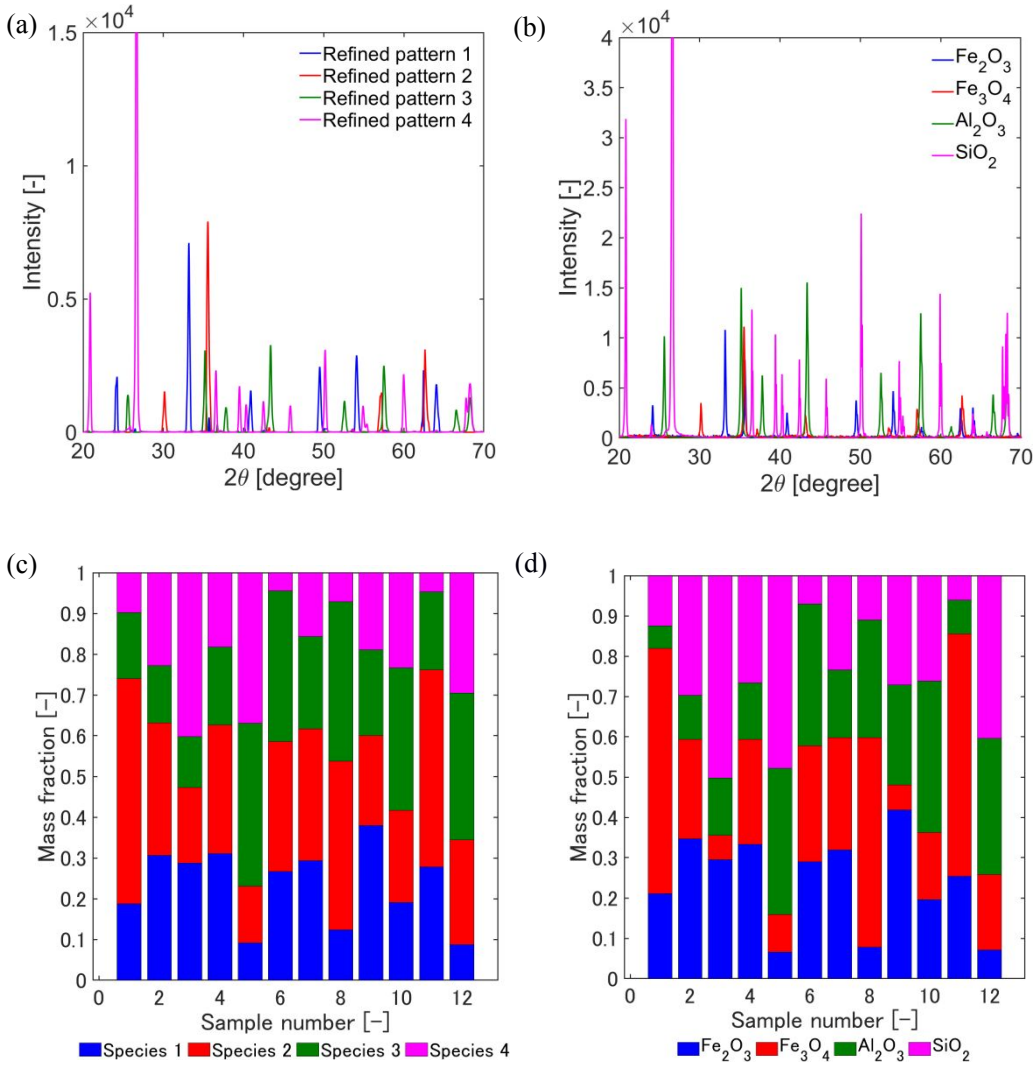


Fig. 6 (a) The final estimation of pure XRD patterns obtained by the *cos-s map* MCR and (b) The reference XRD patterns for four chemicals (Fe_2O_3, Fe_3O_4, Al_2O_3, and SiO_2) are shown. (c) The concentration profile obtained by the *cos-s map* MCR and (d) the prepared concentrations are shown.

From the comparison between the initial estimation of the XRD patterns (Fig. S5(a) in SI) and the optimized XRD pattern (Fig. S5(b) in SI), the spectral shape and intensities were optimized. Compared with the reference XRD pattern (Fig.6 (b)) and the optimized spectra (Fig. 6(a)), the peak positions of each calculated spectra mostly matched with those of the reference patterns. The detailed comparison between the calculated patterns and pure patterns was shown in Fig. 7. The correlation coefficients were 0.90 on average (Table 3), and they were sufficiently accurate for the

assignment of chemical species.

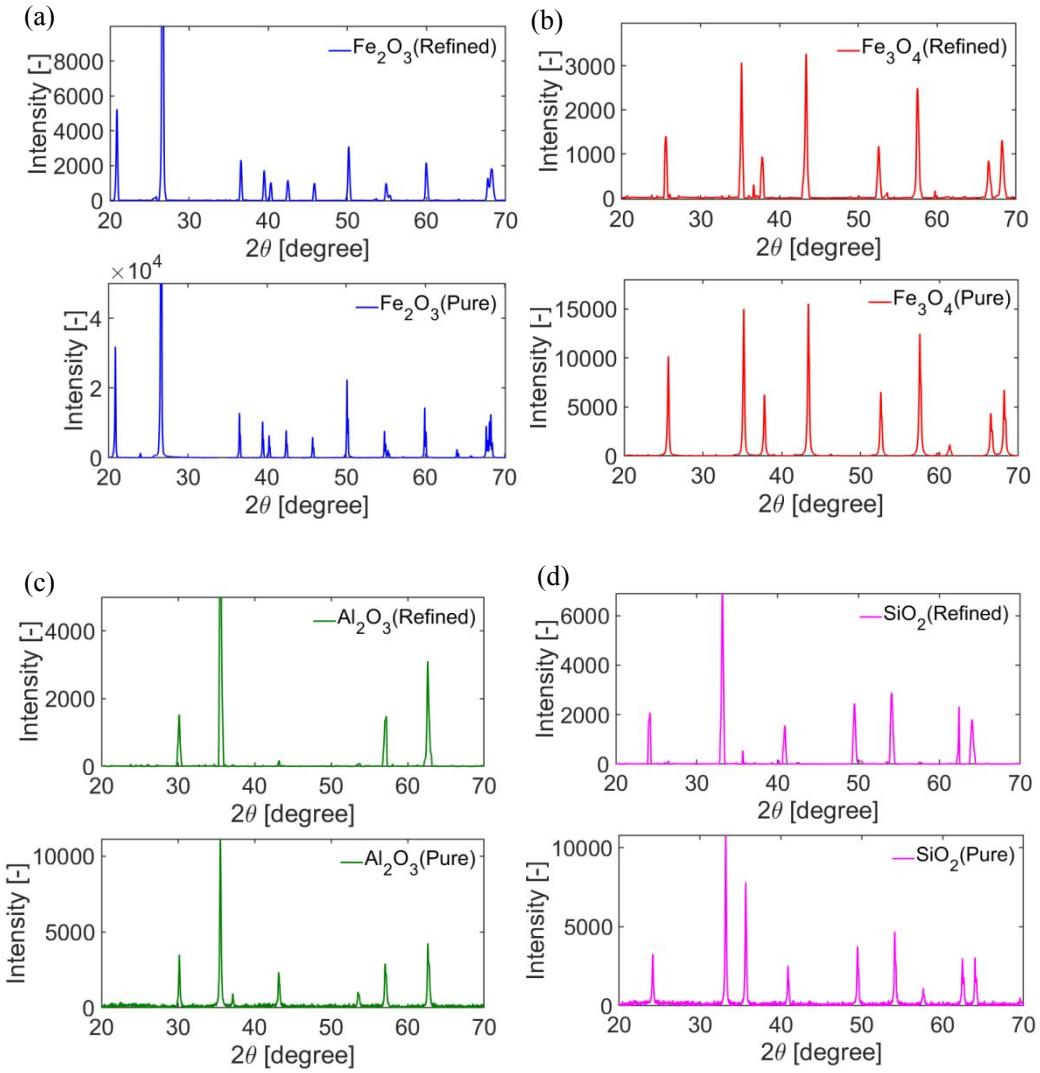


Fig. 7 The comparison between the calculated patterns and the corresponding pure patterns of (a) Fe_2O_3 (b) Fe_3O_4 (c) Al_2O_3 (d) SiO_2 is shown.

Table 2 The correlation coefficients of agreement between the predicted and the reference spectra.

Chemicals	Correlation coefficients
Fe_2O_3	0.84
Fe_3O_4	0.93
Al_2O_3	0.98
SiO_2	0.87

The absolute errors for the concentration ratios were tabulated in Table 3. From the comparison of the concentration profiles between the predicted ratios and the actual preparation, the absolute error was less than 6% on average. The errors of the prediction were suppressed both for the large and small portions of chemicals. The prediction error was in the same range as the preparation errors.

Table 3 The absolute errors (%) of the predicted concentrations from the prepared conditions.

Sample number	Fe_2O_3	Fe_3O_4	Al_2O_3	SiO_2
1	-2.3	-5.6	11	-2.7
2	-4.1	7.9	3.1	-6.9
3	-0.78	13	-1.8	-10
4	-2.2	5.6	5.0	-8.3
5	2.6	4.6	3.6	-11
6	-2.3	3.0	1.9	-2.6
7	-2.6	4.5	5.9	-7.8
8	4.6	-11	9.9	-3.9
9	-3.8	16	-3.8	-8.2
10	-0.50	5.9	-2.6	-2.8
11	2.4	-12	11	-1.4
12	1.7	6.9	2.1	-11

Finally, we compared the calculation results by our method and PURE (SIMPLISMA) initial estimation-MCR. The pure XRD patterns and concentration profiles are shown in Fig. S6. The correlation coefficients between the predicted and the reference spectra was 0.38 on average, and the absolute error of the concentrations was 14% on average. It is obvious that the errors were much worse than the ones obtained by our method. The reason for this discrepancy is not fully understood, however, the initial estimation of the spectra was much worse than the one obtained by our method. It can be said that the proposed initial estimation and the updated optimization scheme was effective and necessary for the extraction of the pure XRD patterns and the concentration ratios with high accuracy.

Based on these two demonstrations, our calculation method could be applied to various types of spectral data composed of sharp and broad peaks with overlapped regions. Both of the mixture spectra, whose shapes were totally different, were analyzed by the same schemes, and we could obtain accurate results without using any prior information about pure spectra and the concentration profile in the sample system. Since the reasonable initinal estimation is used and the following MCR optimiazation is performed with the constraints using the initial estimation in this calculation technique, the process can reduce the uncertainty of the matrix decomposition and solve the problem of rotational ambiguity.[38]

Conclusion

We could develop a parameter-less multivariate curve resolution (MCR) method by the improved initial estimation and its integration into the MCR optimization and demonstrated the extraction of pure spectra and the estimation of the concentration ratios for totally different types of analytical spectroscopic data of unknown chemical mixtures. We could develop the robust initial estimation of pure spectra by a combination of cosine similarity mapping and soft clustering. The initial estimation was integrated into the MCR optimization calculation as a new constraint for the alternating least square algorithm. By applying this method for the UV/Vis spectra and XRD patterns of unknown chemical mixtures, we could recover the pure spectra/patterns and concentration ratios with high accuracy in both cases. Although general problems were solved in this paper, more difficult problems with featureless and large-overlapped peaks must be solved and compared with other calculation techniques, and we are now in progress on it. Since this method was applied to various types of spectral data (broad overlapped peaks multiple complicated peaks) obtained by different analytical equipment, it will be a general spectral analysis method to quantify and qualify chemical species in mixture samples.

Code availability

The codes included in this paper is available via GitHub (https://github.com/Katayama-ChuoU/cos_s_map_mcr).

Author Contribution Statement

YN made the experiments and analyses, and YN and KK discussed the results. YN and KK wrote the manuscript, and both of them reviewed it.

Conflict of interests

I declare that the authors have no competing interests or other interests that might be perceived to influence the results and/or discussion reported in this article.

Acknowledgments

The research was financially supported by JST PRESTO (#JPMJPR1675), KIOXIA corporation, and the Institute of Science and Engineering, Chuo University.

References

1 A. de Juan and R. Tauler, *Anal. Chim. Acta*, , DOI:10.1016/j.aca.2020.10.051.

2 P. Lasch and I. Noda, *Appl. Spectrosc.*, 2019, **73**, 359–379.

3 Y. Park, I. Noda and Y. M. Jung, *Front. Chem.*, , DOI:10.3389/fchem.2015.00014.

4 Y. Park, S. Jin, I. Noda and Y. M. Jung, *J. Mol. Struct.*, 2020, **1217**, 128405.

5 I. Noda, *Spectrochim. Acta. A. Mol. Biomol. Spectrosc.*, 2017, **187**, 119–129.

6 Y. Park, S. Jin, E. Park, M. Hwang, I. Noda, B. Chea and Y. M. Jung, *J. Mol. Struct.*, 2020, **1216**, 128344.

7 W. Sohng, Y. Park, D. Jang, K. Cha, Y. M. Jung and H. Chung, *Talanta*, 2020, **212**, 120748.

8 R. Guo, X. Zhang, A.-Q. He, F. Zhang, Q.-B. Li, Z.-Y. Zhang, R. Tauler, Z.-Q. Yu, S. Morita, Y.-Z. Xu, I. Noda, Y. Ozaki and J.-G. Wu, *Spectrochim. Acta. A. Mol. Biomol. Spectrosc.*, 2019, **220**, 117103.

9 B. G. M. Vandeginste, W. Derks and G. Kateman, *Anal. Chim. Acta*, 1985, **173**, 253–264.

10 P. J. Gemperline, *J. Chem. Inf. Comput. Sci.*, 1984, **24**, 206–212.

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

12 M. Ghaffari and H. Abdollahi, *Chemom. Intell. Lab. Syst.*, 2019, **189**, 121–129.

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

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

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

16 M. B. Anzardi, J. A. Arancibia and A. C. Olivieri, *J. Chromatogr. A*, 2019, **1604**, 460502.

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

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

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

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

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

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

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

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

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

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

27 F. Bruno, R. Francischello, G. Bellomo, L. Gigli, A. Flori, L. Menichetti, L. Tenori, C. Luchinat and E. Ravera, *Anal. Chem.*, 2020, **92**, 4451–4458.

28 X. Fan, P. Ma, M. Hou, Y. Ni, Z. Fang, H. Lu and Z. Zhang, *J. Chromatogr. A*, 2020, 461713.

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

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

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

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

33 P. De B. Harrington, E. S. Reese, P. J. Rauch, L. Hu and D. M. Davis, *Appl. Spectrosc.*, 1997, **51**, 808–816.

34 W. Windig, A. Bogomolov and S. Kucheryavskiy, in *Comprehensive Chemometrics (Second Edition)*, eds. S. Brown, R. Tauler and B. Walczak, Elsevier, Oxford, 2020, pp. 107–136.

35 L. Valderrama, R. P. Gonçalves, P. H. Março, D. N. Rutledge and P. Valderrama, *J. Adv. Res.*, 2016, **7**, 795–802.

36 F. C. Sánchez, J. Toft, B. van den Bogaert and D. L. Massart, *Anal. Chem.*, 1996, **68**, 79–85.

37 Y. Nagai, W. Y. Sohn and K. Katayama, *Analyst*, 2019, **144**, 5986–5995.

38 H. Abdollahi and R. Tauler, *Chemom. Intell. Lab. Syst.*, 2011, **108**, 100–111.

39 I. Noda, *J. Mol. Struct.*, 2020, **1211**, 128068.

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

41 J. C. Bezdek, R. Ehrlich and W. Full, *Comput. Geosci.*, 1984, **10**, 191–203.

42 A. de Juan, J. Jaumot and R. Tauler, *Anal Methods*, 2014, **6**, 4964–4976.

43 T. Shimada and T. Hasegawa, *Spectrochim. Acta. A. Mol. Biomol. Spectrosc.*, 2017, **185**, 104–110.

Figure captions

Fig. 1 Absorption spectra of BTB solutions for different pHs from 4.5 to 9 are shown.

Fig. 2 12 XRD pattern of chemical mixtures of 4 chemicals (Fe_2O_3, Fe_3O_4, Al_2O_3, and SiO_2) are shown.

Fig. 3 The cosine similarity map of the UV/Vis spectra of the BTB solutions with different pHs obtained as Scheme 1(a) (*cos-s map* estimation).

Fig. 4 (a) The initial estimation of the spectra is shown, which were obtained from the mixture spectra of the BTB solutions by using the *cos-s map* estimation. (b) The absorption refined by the MCR calculation from the initial estimation of the spectra is shown. (c) The concentration ratios of the two components were obtained in the MCR calculation.

Fig. 5 (a) The similarity map for the XRD patterns obtained in the process of *cos-s map* estimation is shown. (b) The initial estimation of the XRD patterns by using the *cos-s map* estimations in the whole angle region.

Fig. 6 (a) The final estimated pure XRD patterns obtained by the *cos-s map* MCR. (b) The concentration profile obtained by the *cos-s map* MCR and (c) the prepared concentrations are shown.

Fig. 7 The comparison between the calculated patterns and the corresponding pure patterns of (a) Fe_2O_3 (b) Fe_3O_4 (c) Al_2O_3 (d) SiO_2 is shown.

1