Supplementary Information

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

Estimation of the spectral weight for the initial estimation of pure spectra

The exponential *p* of Eqn. (4) is inductively deduced as follows.

If *p* is set to two, equation (4) is transformed as follows.

$$\sum_{j=1}^{m}\sum_{k=1}^{q} {u_{j,k}}^2 \left\|\boldsymbol{x}_j - \boldsymbol{v}_k\right\|^2 \approx 0 \quad \text{(S-1)}$$

$$\sum_{j=1}^{m}\sum_{k=1}^{q} u_{j,k} \left\|\boldsymbol{x}_j - \boldsymbol{v}_k\right\| \approx 0 \quad \text{(S-2)}$$

Since $\sum_{k=1}^{q} u_{j,k} = 1$ at each *j*, $u_{j,k}$ can be regarded as the probabilities for the absolute components, and (S-2) represents the expectation of $\left\|\boldsymbol{x}_j - \boldsymbol{v}_k\right\|$ about *k* considering $\boldsymbol{x}_j$ is a given dataset and independent of *k*, and can be transformed to

$$\sum_{j=1}^{m}\left\|\boldsymbol{x}_j - \overline{\boldsymbol{v}}\right\| \approx 0 \quad \text{(S-3)}$$

$\overline{\boldsymbol{v}}$ is the average vector of the cluster center.

$$\boldsymbol{x}_j \approx \overline{\boldsymbol{v}} \cong \sum_{k=1}^{q} u_{j,k}\, \boldsymbol{v}_k \quad \text{(S-4)}$$

$$x_{j,i} \cong \sum_{k=1}^{q} u_{j,k}\, v_{k,i} \;\; (i = 1 \ldots n) \quad \text{(S-5)}$$

$$X \cong UV \quad \text{(S-6)}$$

These equations are general to the real number dataset. Let $x_{j,i}$ is the similarity matrix of $(cos\theta)_{j=j_1,j_2}$, for $j = j_1$, and $j = j_2$, the similarity map ($m \times m$), is represented as follows.

$$(cos\theta)_{j=j_1,j_2} \cong \sum_{k=1}^{q} \mu_{j_1,k}\, v_{k,j_2} \;\; (j = 1,2, \ldots m) \quad \text{(S-7)}$$

In Eqn. (S-7), *k* represents the cluster number, equal to the number of chemical species in the mixture. Eqn. (S-7) means that the similarity map is well approximated by the linear combination of the cluster center and its contribution to the whole data. Since the similarity map is originated from the mixture spectral data, both the information of the spectral shape and the concentration ratio of each species is included in the similarity map. Therefore, the result of the fuzzy c-means clustering could be utilized to estimate the shape of the pure spectra, since $\mu_{j_1,k}$ is the contribution of *k*-th chemical species to the

similarity map at wavelength j_1. However, further consideration is required for that calculation.

Based on Eqn. (S-7) and an analogy of the MCR matrix decomposition, $\mu_{j_1,k}$ and v_{k,j_2} can be regarded as a spectral shape and a concentration ratio profile, respectively. However, this is not correct because the dimension of V in Eqn. (S-6) is $(q \times m)$, and m is the number of parameters in spectrum measurement, meaning that V should have both the information on the spectral shape and concentration. Therefore, the spectral shape information included within v_{k,j_2} should be extracted as shown by Eqn. (S-8) and (S-9).

$$v_{k,j_2} \cong \sum_{l=1}^{q} y_{k,l}\, \mu_{l,j_2} \quad (l = 1,2,\dots q) \tag{S-8}$$

$$V \cong YU^T \tag{S-9}$$

Eqn. (S-9) is a matrix representation of Eqn. (S-8). Y is a square matrix that corresponds to the quasi-covariance matrix of each species, reflecting some parts of the concentration ratios since the spectral shape features U were extracted by Eqn. (S-8) and (S-9) from V. Based on Eqn. (S-6) and Eqn. (S-9), cosine similarity can be expressed as Eqn. (S-10),

$$cosine\ similarity \cong UYU^T \tag{S-10}$$

If there is no correlation between the concentration profile of each species, *cosine similarity* map represents only spectral feature, and quasi-covariance matrix Y can be an identity matrix I. Therefore, the similarity map representing only the spectral features can be transformed into Eqn. (S-11).

$$cosine\ similarity_{reconst} \cong UIU^T = UU^T \tag{S-11}$$

$cosine\ similarity_{reconst}$ is the reconstructed similarity map only includes the spectral feature. From Eqn. (S-11), even if there is some correlation between the concentration ratio profile of each species, namely Y is not an identity matrix, the similarity map could be represented by UU^T as if there is no correlation between the concentration profile about each species. Then, $cosine\ similarity_{reconst}$ could be interpreted as the projection of U to the map space $(m \times m)$ to represent the spectral feature, and it corresponds to "square" of U as shown in Eqn. (S-11). Based on the interpretation above, one of the approximations for the projection should be an element-wise product of U, *i.e.* ${u_{j,k}}^2$, to the spectral space $(m \times q)$. Considering ${u_{j,k}}^2$ was utilized to evaluate the real contribution of the k-th species to the similarity map, the spectral information can be reasonably represented as ${u_{j,k}}^2$. Since the spectral intensities were centered, the differences from the averaged values over samples were obtained. The initial estimation of the spectra is calculated based on the averaged spectra as:

$$S_{est,j,k} = {u_{j,k}}^2 \bar{s}_j \tag{S-12}$$

$S_{est,j,k}$ 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.

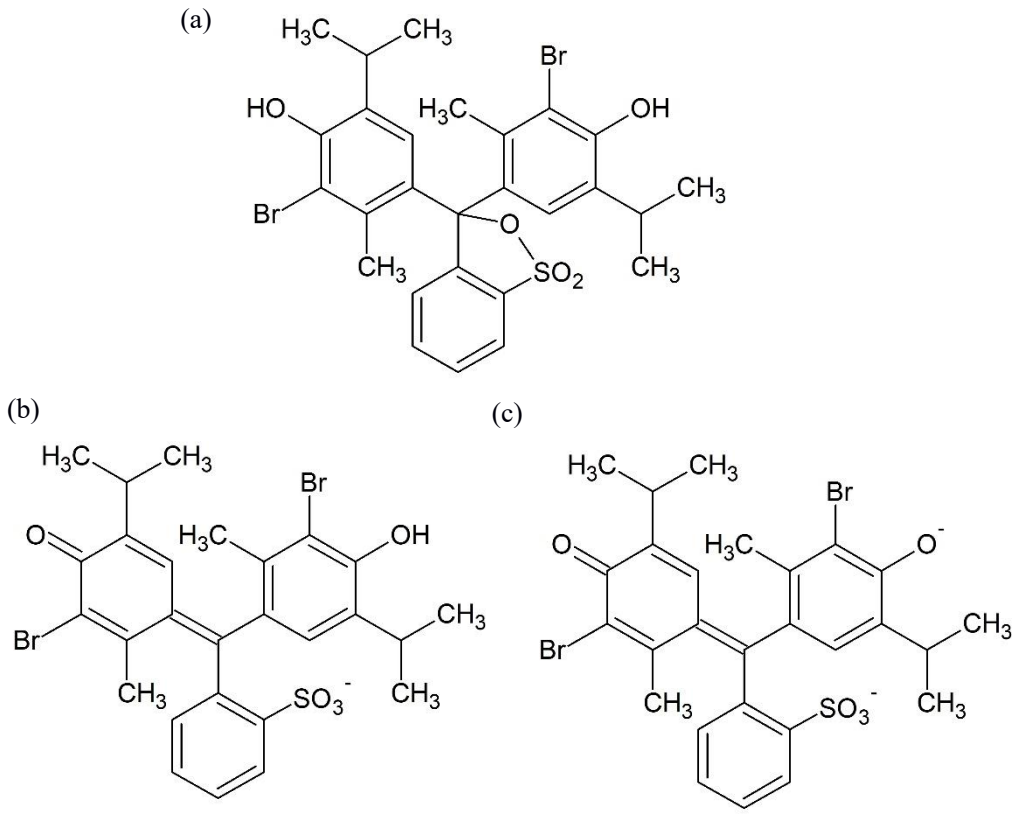

Fig. S1 (a) The molecular strucure of bromothymol blue (BTB) is shown. The proposed structures of BTB under acidic (b) and basic conditions (c) are shown, respectively.[1]

(1) Shimada, T.; Hasegawa, T. Determination of Equilibrium Structures of Bromothymol Blue Revealed by Using Quantum Chemistry with an Aid of Multivariate Analysis of Electronic Absorption Spectra. *Spectrochim. Acta. A. Mol. Biomol. Spectrosc.* **2017**, *185*, 104–110. https://doi.org/10.1016/j.saa.2017.05.040.

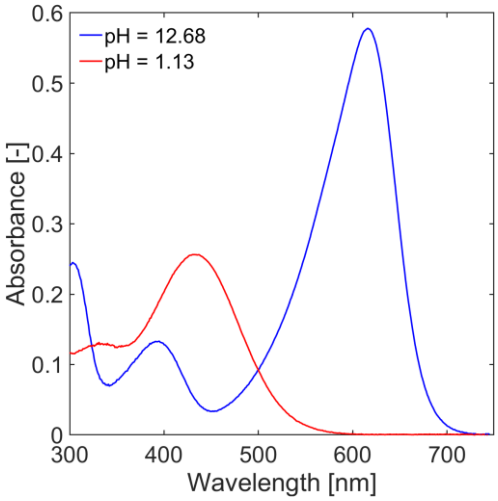

Fig. S2 The UV/Vis absorption spectra of the BTB solutions under highly acidic or basic conditions (pH = 1.13 and 12.68) are shown.

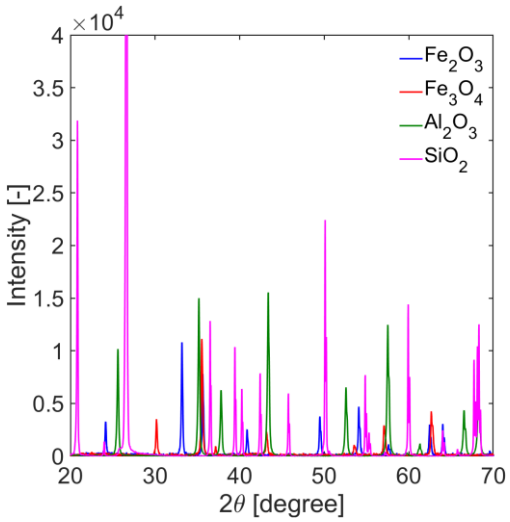

Fig. S3 The reference XRD patterns for four chemicals (Fe_2O_3, Fe_3O_4, Al_2O_3, and SiO_2) are shown.

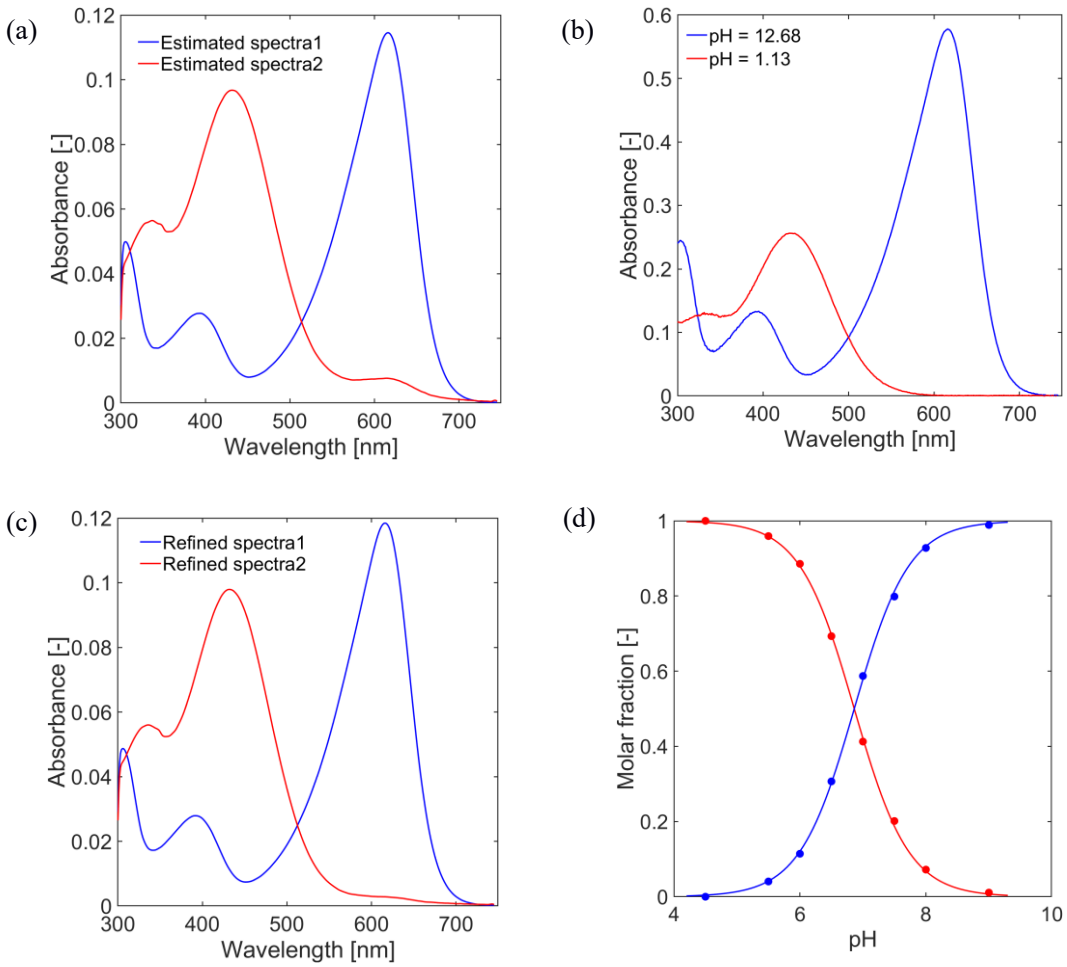

Fig. S4 (a) The initial estimation of the spectra is shown, which were obtained from the mixture spectra of the BTB solutions by using PURE in SIMPLISMA. (b) The UV/Vis absorption spectra of the BTB solutions under highly acidic or basic conditions (pH = 1.13 and 12.68) is shown. (c) The absorption spectra 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.

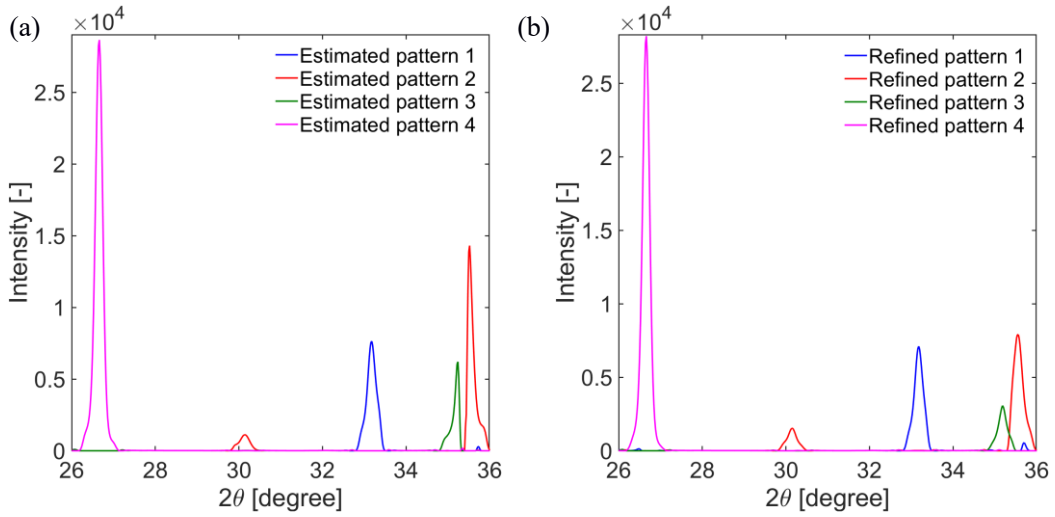

Fig. S5 (a) The initial estimation of the spectra by using the *cos-s map* estimations in an enlarged figure around 30 degrees. (b) The final estimation of the pure spectra obtained by the *cos-s map* MCR is shown.

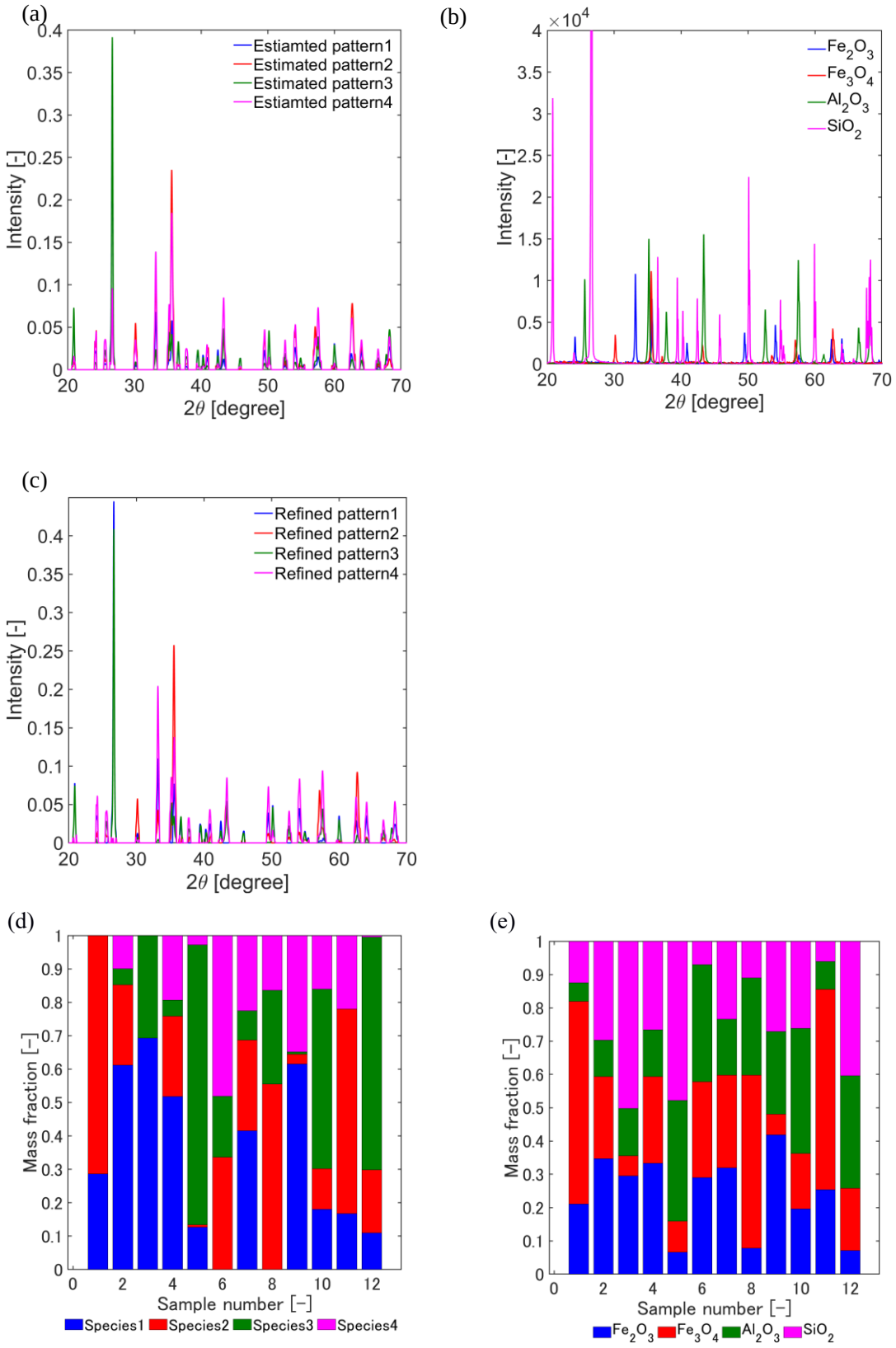

Fig. S6 (a) The final estimation of pure XRD patterns obtained by the PURE (SIMPLISMA)−

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 XRD patterns refined by the MCR calculation from the initial estimation of the patterns are shown. (d) The concentration profile obtained by the PURE (SIMPLISMA)−MCR and (e) the prepared concentrations are shown.