Four-Dimensional Big Data Analysis Using Two-Step Multivariate Curve Resolution Technique

Yutaka HOSHINA*, Shigeaki UEMURA, Haruka OKAMOTO, and Yugo KUBO

For research and development in material science, it is important to understand the three-dimensional (3D) distributions of chemical species in samples. The effective utilization of 4D big data which contain a lot of information about the 3D distributions is a key factor. This paper demonstrates a new 4D data analysis technique called "two-step multivariate curve resolution (MCR)". To obtain an intuitive expression of 4D data, we devised a process involving two iterations of MCR with digitization in between. The new technique was applied to the analysis of time-of-flight secondary ion mass spectrometry data derived from a thin-film sample to assist in the interpretation of complex three-dimensional local microstructures. Compared to conventional methods of data presentation, two-step MCR was found to greatly facilitate the clarification and understanding of the 4D analysis data.

Keywords: MCR, NMF, unsupervised machine learning, 4D/four-dimensional, ToF-SIMS

1. Introduction

According to advances in instrumental analysis technology, four-dimensional (4D) material data can be easily obtained today. For example, time-of-flight secondary ion mass spectrometry (ToF-SIMS)*[1] can provide 4D data when used in conjunction with sputtering techniques. As shown schematically in Fig. 1, the intensity of a secondary ion fragment possessing a given *m/z* varies as a function of the *X*, *Y*, and *Z* coordinates. Although such 4D data contain a wealth of valuable information about samples, these data are rarely utilized effectively because 4D data cannot be graphically shown on a 2D plane for intuitive interpretation.

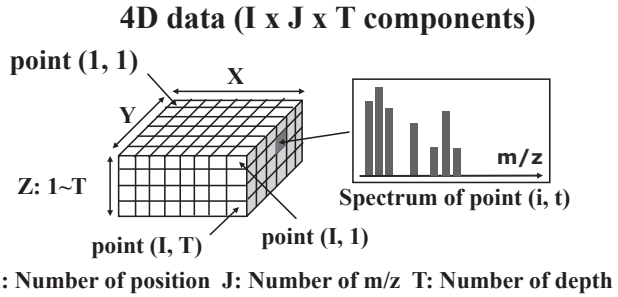

Fig. 1. Pictorial representation of ToF-SIMS 4D data

A description of 4D data on a 2D plane requires a means to summarize and represent the data. Three different data presentation methods are commonly used for this purpose. Method A is displaying only the depth profile (Fig. 2 (a)), Method B is a 3D plot of a given *m/z* signal (Fig. 2 (b)), and Method *C* is a display of the *m/z* intensity distribution at each pixel in a selected plane (Fig. 2 (c)).

In Method A, the 4D data are averaged over each *XY* plane. This process is therefore inappropriate for data of samples with an inhomogeneous *XY* plane. In Method B, it is difficult to capture the relationships between the fragment ions. In Method C, we observe only the specified planes. Thus, these techniques cannot present the entirety of the 4D information adequately.

To overcome this problem, we have conceived a new 4D data analysis technique based on unsupervised machine learning whereby important characteristics of the 4D data are extracted automatically without any arbitrary data editing. The remainder of the paper is organized as follows. An explanation of the mathematical procedures for the analysis is first given. Then, the proposed method is used to analyze 4D data obtained by ToF-SIMS measurement of a thin-film sample, demonstrating that the method provides

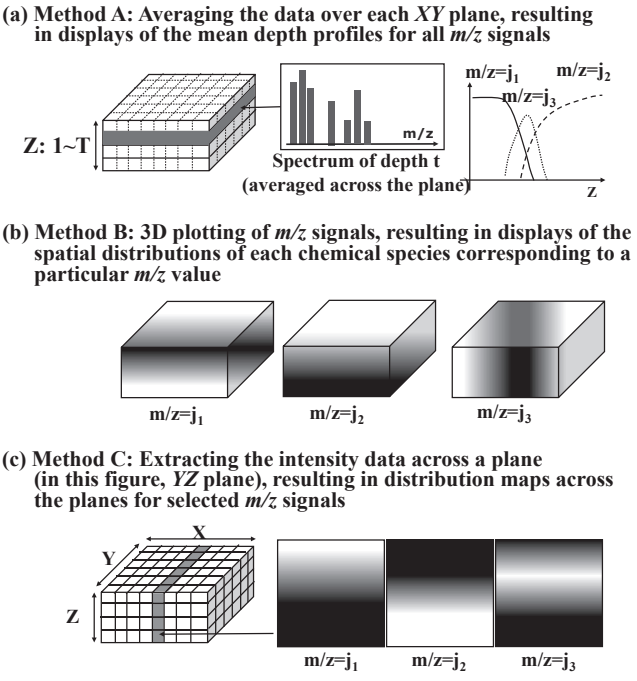

Fig. 2. Three conventional data presentation methods

a clarified and intuitive display of the spatial distributions of chemical species in the sample.

2. The Mathematical Procedures of the New Analysis Method

2-1 Framework of two-step MCR

There are many mathematical techniques to compress enormous amounts of spectra data without critical information loss.[1] Among them, the most popular one is multivariate curve resolution (MCR)*2.[1),(2)] The new analysis method demonstrated in this paper is referred to as "two-step MCR" because it involves the iteration of MCR a second time after an appropriate intermediate treatment step.

It is important to show the matrix expression of the 4D data for the two-step MCR process since MCR is based on a matrix factorization technique. The 4D data for ToF-SIMS measurement can be expressed as

$$D = \begin{pmatrix} d_{111} & \cdots & d_{1J1}d_{112} & \cdots & d_{1J2} & d_{11T} & \cdots & d_{1JT} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \cdots & \vdots & \ddots & \vdots \\ d_{I11} & \cdots & d_{IJ1}d_{I12} & \cdots & d_{IJ2} & d_{I1T} & \cdots & d_{IJT} \end{pmatrix} \cdots (1).$$

In this expression, the matrix components are specified by three indices: the lateral position (X and Y: from 1 to I), the mass (m/z: from 1 to J), and the depth (Z: from 1 to T). The matrix component d_{ijt} is the data that corresponds to the XY position i, the nominal mass j, and the depth t. Generally, the XY planes are usually transferred into one-dimensional rows in MCR process. In this study, the original data were preprocessed using Poisson scaling under the assumption that ToF-SIMS signals are likely to be governed by Poisson statistics. In the example in this paper, J is set to 500 because unit mass binning of the peaks in the mass spectra was performed for a mass range of 1 to 500.

The two-step MCR approximates the raw 4D data in Eq. (1) as follows.

$$d_{ijt} \cong \sum_{k=1}^{K}\sum_{l=1}^{L} c_{ik}s_{jl}f_{tl}^{(k)} \quad \cdots\cdots (2)$$

Here, c_{ik}, s_{jl}, and $f_{tl}^{(k)}$ correspond to the lateral distribution of the "unit" k, the mass spectrum of the chemical species l, and the depth profile of the chemical species l in the unit k, respectively. The units correspond to the combinations of the lateral concentration distributions and the depth profiles (Z vs. m/z fragment intensity). Here, K and L are the numbers of units and chemical species, respectively.

These two decomposition parameters K and L are characteristic of the two-step MCR, which are not used in parallel factor analysis (PARAFAC),[3] which is formalistically similar to the two-step MCR. PARAFAC is a tensor decomposition method that decomposes the data matrix (1) into three matrices as follows.

$$d_{ijt} \cong \sum_{l=1}^{L} c_{il}s_{jl}f_{tl} \quad \cdots\cdots (3)$$

Here, c_{il}, s_{jl}, and f_{tl} correspond to the lateral distribution, the mass spectrum, and the depth profile of the chemical species l, respectively. The l in Eq. (3) can be regarded as the index of the chemical species in the sample, and corresponds to the l in Eq. (2). It is worth noting that the two-step MCR has higher degrees of freedom with which to express a system compared to PARAFAC, since it has two independent parameters (K and L). The two-step MCR has therefore an advantage when, for example, expressing the complicated material distribution of a sample.

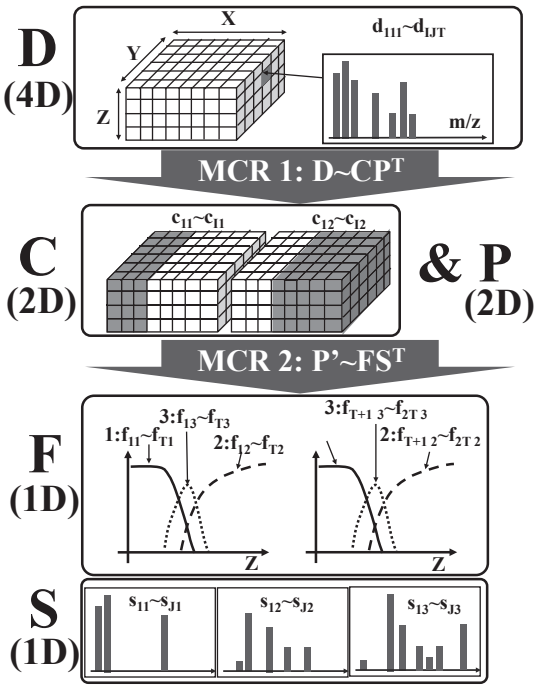

Fig. 3. Schematic summary of the two-step MCR process

An overview of the two-step MCR is shown in Fig. 3. The matrix components such as "c_{I1}," "$f_{T+1\ 3}$," and "s_{J2}" correspond to those given in after-mentioned Eqs. (4) and (6). Here, the special case where $K = 2$ (in Eq. (4)) and $L = 3$ (in Eq (6)) is shown for simplicity. Summarized data matrices C, F, and S, are sequentially obtained from the original data matrix D by two iterations of MCR. The decomposition is realized using an alternating least squares (ALS) method.[2),(3)] First, the Poisson-scaled data matrix D in Eq. (1) is decomposed into the product of two low-dimensional matrices as given by

$$D \sim CP^T$$

$$C = \begin{pmatrix} c_{11} & \cdots & c_{1K} \\ \vdots & \ddots & \vdots \\ c_{I1} & \cdots & c_{IK} \end{pmatrix}$$

$$P^T = \begin{pmatrix} p_{1\,1} & \cdots & p_{J\,1}p_{J+1\,1} & \cdots & p_{J\times 2\,1} & p_{J\times(T-1)+1\,1} & \cdots & p_{J\times T\,1} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \cdots & \vdots & \ddots & \vdots \\ p_{1\,K} & \cdots & p_{J\,K}p_{J+1\,K} & \cdots & p_{J\times 2\,K} & p_{J\times(T-1)+1\,K} & \cdots & p_{J\times T\,K} \end{pmatrix} \cdots (4).$$

Here, the matrix C is the lateral concentration distribution of units. The matrix P corresponds to the depth

profiles of all *m/z* signals in all the units. After the decomposition of Eq. (4), *C* is digitized and *P* is correspondingly modified. The significance and the specific procedures of digitization are explained in the next subsection.

The calculation of Eq. (4) is generally time- and memory-consuming so some simplification processes such as data decimation are sometimes necessary.

P can be further decomposed by an additional MCR process. Before the decomposition, *P* is transformed as follows in order to derive the chemical components from the data by the additional MCR.

$$P \to P' = \begin{pmatrix} p_{1\,1} & \cdots & p_{J\,1} \\ p_{J+1\,1} & \cdots & p_{J\times 2\,1} \\ \vdots & & \vdots \\ p_{J\times(T-1)+1\,1} & \cdots & p_{J\times T\,1} \\ & \vdots & \\ p_{1\,K} & \cdots & p_{J\,K} \\ p_{J+1\,K} & \cdots & p_{J\times 2\,K} \\ \vdots & & \vdots \\ p_{J\times(T-1)+1\,K} & \cdots & p_{J\times T\,K} \end{pmatrix} \quad \cdots\cdots\cdots\cdots (5)$$

Note that columns *j* in matrix *P′* correspond to the nominal mass *j*. This transformed matrix *P′* is decomposed into the product of two low-dimensional matrices as

$$P' \sim \begin{pmatrix} f_{1\,1} & \cdots & f_{1\,L} \\ f_{2\,1} & \cdots & f_{2\,L} \\ \vdots & \vdots & \vdots \\ f_{T\,1} & \cdots & f_{T\,L} \\ & \vdots & \\ f_{T\times(K-1)+1\,1} & \cdots & f_{T\times(K-1)+1\,L} \\ f_{T\times(K-1)+2\,1} & \cdots & f_{T\times(K-1)+2\,L} \\ \vdots & \vdots & \vdots \\ f_{T\times K\,1} & \cdots & f_{T\times K\,L} \end{pmatrix} \begin{pmatrix} s_{11} & \cdots & s_{J1} \\ \vdots & \ddots & \vdots \\ s_{1L} & \cdots & s_{JL} \end{pmatrix} = FS^T \cdots\cdots (6)$$

where the matrix *S* corresponds to the mass spectra of the compounds, matrix *F* gives the depth profiles of the chemical compounds, and *L* is the number of chemical compounds that are necessary for expressing the essential characteristics of the sample.

The matrix *F* includes the depth profiles in all *K* units. Therefore it can be expressed as

$$F = \begin{pmatrix} F^{(1)} \\ F^{(2)} \\ \vdots \\ F^{(K)} \end{pmatrix} \quad F^{(k)} = \begin{pmatrix} f_{11}^{(k)} & \cdots & f_{1L}^{(k)} \\ f_{21}^{(k)} & \cdots & f_{1L}^{(k)} \\ \vdots & \vdots & \vdots \\ f_{T1}^{(k)} & \cdots & f_{TL}^{(k)} \end{pmatrix} \quad \cdots\cdots\cdots\cdots (7)$$

where sub-matrix $F^{(k)}$ corresponds to the depth profile in the unit *k*.

As a result, the original 4D data are expressed approximately by three matrices, namely, the lateral concentration distributions of the units, the depth profiles of the compounds, and the mass spectra of the compounds. The necessary numbers of regions and compounds (*K* and *L* in Eqs. (4) and (6), respectively) are typically less than ten, as shown in the following example. These small numbers permit the entire 4D dataset to be expressed within the limited space of a 2D plane.

In this paper, the MCR algorithm has been used in a twofold decomposition process. Other algorithms such as the non-negative matrix factorization (NMF)[1] are also suitable for this purpose and offer similar results since both MCR and NMF are based on the minimization of the sum of squares of the error between *D* and CP^T.

2-2 Digitization of the matrix C

In MCR, the matrix *C* generally has a number of nonzero values in each row, which means that the lateral distributions of different units overlap with each other. In 3D data analysis, this overlap simply corresponds to a mixture of compounds and does not introduce any difficulty in interpreting the data intuitively. In contrast, the overlapping of the components of *C* in the two-step MCR complicates the intuitive interpretation of the data.

To improve the clarity of the expression, the matrix *C* is digitized such that the overlapping of units is prohibited. In other words, for each row in *C*, there is only a single "1" value, and all other components are set to "0." *C* can be digitized using some clustering methods such as *k*-means clustering. The corresponding *P* is then modified by the simple least squares process. The distributions of the *C* are expressed using only "1" or "0" as shown in aftermentioned Fig. 5.

MCR is classified as a "soft clustering" method that permits the overlap of the components. On the other hand, the combination of the first MCR and the above digitization is considered "hard clustering," as it prohibits overlaps. In the two-step MCR, intuitive expressions of 4D data are realized by doing hard and soft clustering about the lateral and vertical distributions of the materials in the sample, respectively.

3. ToF-SIMS Measurement and Two-Step MCR Analysis of BZY Thin Film

The analysis of $BaZr_{1-x}Y_xO_{3-\delta}$ (BZY) is demonstrated as an example of the two-step MCR. BZY is a promising candidate as an electrolyte in protonic ceramic fuel cells. The ToF-SIMS test sample used in this study was comprised of two BZY layers (BZY-A (100 nm) and BZY-B (>1 μm)) on a NiO+BZY substrate as shown in Fig. 4 (a). In this case, it was particularly important to evaluate the diffusion of Ni into the BZY layers.

A PHI nano TOF II ToF-SIMS instrument (ULVAC-PHI Inc., Kanagawa, Japan) was used for measurements. A 30 kV Bi^{3++} beam was adopted for the primary ions. A 2 kV Cs^+ ion beam was used as the sputtering source. The sputtering times were converted to depth from the surface using the sputtering rate for SiO_2. The lateral area analyzed was 100 μm × 100 μm. The positive ion data were analyzed. The signals of *m/z* = 133 (Cs), 266 (Cs_2), and 399 (Cs_3) have been omitted from the raw data because the sensitivity of Cs is extremely high in positive ion ToF-SIMS measurement, which disrupts the two-step MCR process and interpretation of the data.

First, we examine the results using Method A in Fig. 2 (a), as shown in Fig. 4 (b). Ni (*m/z* = 58) was detected in the BZY-B layer. However, details of the spatial distribution cannot be well understood from the depth profile alone.

The same data processed using the two-step MCR are presented in Fig. 5. Three distinct units and three chemical

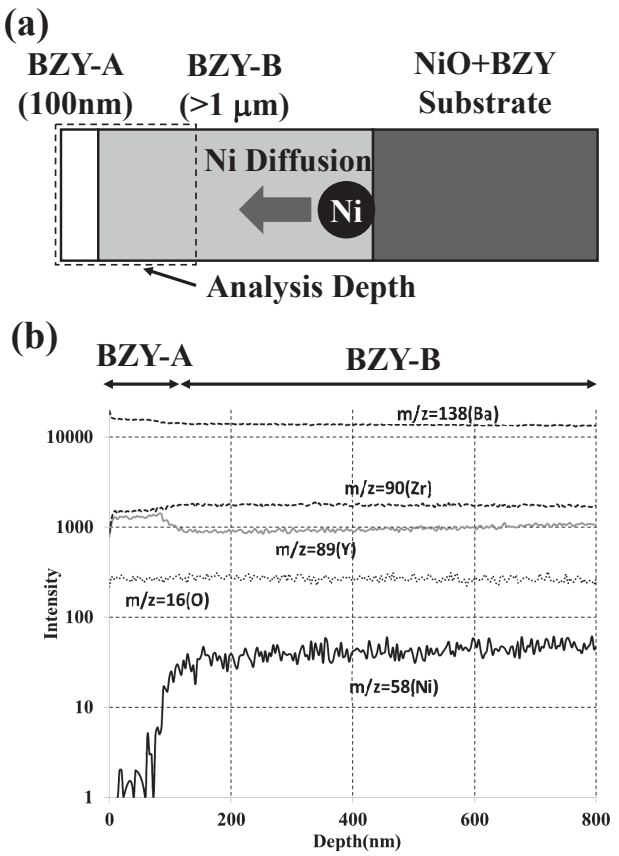

Fig. 4. (a) The cross-section structure of the test sample for ToF-SIMS measurement. (b) Depth profiles of typical *m/z* signals, which are averaged over the whole analyzed *XY* plane

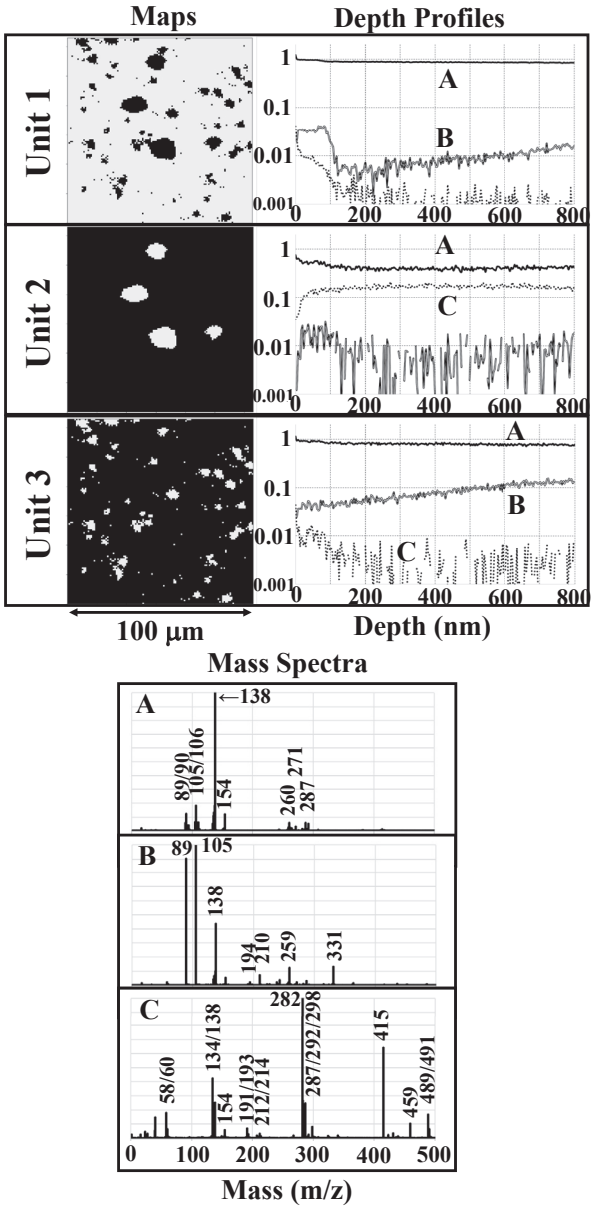

Fig. 5. Results for two-step MCR analysis of the BZY data

species were automatically extracted. Three maps correspond to the digitalized matrix *C*. White points in the maps show each units. In Fig. 5, the *XY* plane is divided into three regions. It is contrast to the expression of Fig. 4, where the 4D data are averaged over each *XY* plane. The depth profiles correspond to the matrix $F^{(k)}$ in Eq. (7), which show the distributions of material A, B, and C along *Z* axis. The mass spectra correspond to the matrix *S*, which show the components of material A, B, and C.

In Unit 1, the signal intensity is almost constant with depth for material A. Material A mainly includes the components of *m/z* = 89 (Y), 90 (Zr), 105 (YO), 138 (Ba), and 154 (BaO), which indicates that Unit 1 correspond to the main structure of BZY.

The other two units express the local structure of the sample. In Unit 2, the ratio of material C is higher than that in Unit 1. Material C includes the components of *m/z* = 58 (Ni), 60 (Ni), 191 (CsNi), 193 (CsNi), 212 (BaNiO), 214 (BaNiO), 282 (Cs_2O), 415 (Cs_3O), 489 (Cs_3NiO_2), and 491 (Cs_3NiO_2). This indicates that the Ni atoms diffuse into the BZY layer along local paths.

In Unit 3, the ratio of material B is higher than that in Unit 1. Material B includes the components of *m/z* = 89 (Y), 105 (YO), 194 (Y_2O), 210 (Y_2O_2), 259 ($BaYO_2$), and 331 (Y_3O_4). Unit 3 thus indicates that the Y atoms aggregate locally in the BZY-B layer.

The results shown in Fig. 5 can also be interpreted from a different perspective. Figure 6 shows 3D intensity plots for representative *m/z* signals. The distributions of *m/z* = 89 and 105 resemble each other, which corresponds to material B in Unit 3 in Fig. 5. It is noteworthy that the distribution of *m/z* = 58 (Ni), which is relatively unclear compared to other five 3D maps due to the order of magnitude lower intensity of the *m/z* = 58 signal as shown in Fig. 4 (b), can be indirectly understood by the distribution of *m/z* = 282 (Cs_2O) because they are found to be similar by the two-step MCR as shown in Fig. 5.

These six *m/z* distributions in Fig. 6 have been identified by simply looking at three mass spectra in Fig. 5. Normally, to determine these *m/z* groups that coexist in the sample, all (1~500) *m/z* 3D distributions must be plotted, as in Fig. 1 (c), which is extremely troublesome. In contrast, the two-step MCR enables the automatic extraction and intuitive display of important characteristic 3D units and

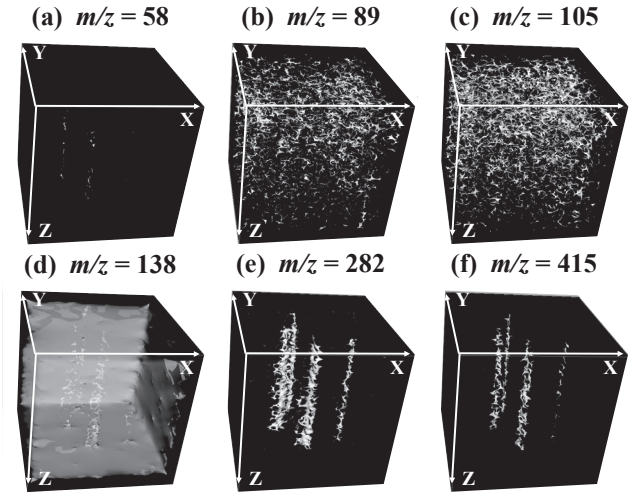

Fig. 6. Distribution of six representative *m/z* signals in the BZY film. Among them, only (d) is a translucent plot showing the deficiency regions of *m/z* = 138 (Ba), which corresponds to Unit 2 in Fig. 5. The *X*, *Y*, and *Z* axes range from 0 to 100 µm, 0 to 100 µm, and 0 to 800 nm, respectively

materials from raw big 4D data, which powerfully assists the understanding and development of the material.

4. Conclusion

A new data analysis technique called "two-step MCR" for expressing 4D material data has been proposed. The proposed method was applied to a 4D dataset of ToF-SIMS measurements of a BZY thin-film sample to demonstrate the simplification of displays that are used to identify the local distributions of Ni and Y in the BZY layer. Such information is difficult to determine by using conventional data presentation techniques.

In principle, the two-step MCR technique can be applied not only to ToF-SIMS data but also to other varieties of 4D data sets including energy dispersive X-ray spectrometry with focused ion beam processing (*X*, *Y*, *Z*, and *hv*) and X-ray computed tomography (*X*, *Y*, *Z*, and *T*), for example. We anticipate in the future that the two-step MCR will provide solutions to a wide range of data analysis problems.

Technical Terms

∗1 Time-of-flight secondary ion mass spectrometry (ToF-SIMS): A surface analysis technique. Ion beam (primary ion) irradiation is applied to a sample surface and mass separation of the ions emitted from the surface (secondary ions) is performed using the difference in time-of-flight (time-of-flight is proportional to the square root of the weight).

∗2 Multivariate curve resolution (MCR): A mathematical processing technique based on unsupervised machine learning, by which pure component spectra in a sample are extracted from raw 3D (e.g., *X*, *Y*, and *m/z*) data.

References

(1) N. Verbeeck, R. M. Caprioli, and R. Van de Plas, Mass Spectrometry Reviews, 2019 00, 1–47 (2019)
(2) S. Muto, T. Yoshida, and K. Tatsumi, Materials Transactions, 50 (5) 964 (2009)
(3) R. Bro, Chemometrics and Intelligent Laboratory Systems, 38 (2) 149-171 (1997)

Contributors

The lead author is indicated by an asterisk (∗).

Y. HOSHINA*
• Ph.D.
Assistant Manager, Analysis Technology Research Center

S. UEMURA
• Ph.D.
Assistant General Manager, Analysis Technology Research Center

H. OKAMOTO
• Analysis Technology Research Center

Y. KUBO
• Ph.D.
Assistant General Manager, Analysis Technology Research Center