Non-negative matrix factorization for mining big data obtained using four-dimensional scanning transmission electron microscopy

Fumihiko Uesugi[1,*], Shogo Koshiya[2], Jun Kikkawa[2], Takuro Nagai[2], Kazutaka Mitsuishi[1] and Koji Kimoto[2,*]

[1] National Institute for Materials Science, 1-2-1 Sengen, Tsukuba, Ibaraki 305-0047, Japan

[2] National Institute for Materials Science, 1-1 Namiki, Tsukuba, Ibaraki 305-0044, Japan

*Corresponding authors

Abstract

Scientific instruments for material characterization have recently been improved to yield big data. For instance, scanning transmission electron microscopy (STEM) allows us to acquire many diffraction patterns from a scanning area, which is referred to as four-dimensional (4D) STEM. Here we study a combination of 4D-STEM and a statistical technique called non-negative matrix factorization (NMF) to deduce sparse diffraction patterns from a 4D-STEM data consisting of 10,000 diffraction patterns. Titanium oxide nanosheets are analyzed using this combined technique, and we discriminate the two diffraction patterns from pristine TiO_2 and reduced Ti_2O_3 areas, where the latter is due to topotactic reduction induced by electron irradiation. The combination of NMF and 4D-STEM is expected to become a standard characterization technique for a wide range materials.

Keywords: Electron microscopy, four-dimensional scanning transmission electron microscopy, non-negative matrix factorization

---

1. Introduction

Scientific instruments for material characterization have recently been improved to yield big data [1–6]. For instance, advanced scanning transmission electron microscopy (STEM) allows us to acquire many diffraction patterns by varying the incident probe position, and this technique is called diffraction imaging, spatially resolved diffractometry or four-dimensional (4D) STEM [7–9]. A diffraction pattern obtained by electron microscopy is rich in crystallographic information (e.g., space group, crystal structure, strain, various ranges of atomic ordering), although conventional electron microscopy (e.g., selected area diffraction) cannot fully utilize the crystallographic information. Four-dimensional STEM enables crystallographic analyses to be performed with high spatial resolution of nanometer order. Since the data obtained by 4D-STEM is large in comparison with that obtained by conventional electron microscopy, statistical techniques are indispensable for deducing useful information.

There are various statistical techniques, and one of the standard techniques used in electron microscopy is principal component analysis (PCA). Principal component analysis has been successfully applied to two-dimensional (2D) spectroscopic analysis, i.e., spectrum imaging, using electron energy-loss spectroscopy or energy-dispersive X-ray spectroscopy with STEM [10], particularly for denoising. Because the components (i.e., spectra) resolved by primitive PCA include negative values, the components cannot be simply interpreted as spectra. Recently non-negative matrix factorization (NMF) based on alternating least-square multivariate curve resolution (ALS-MCR) has been demonstrated for spectrum-imaging data [11,12]. The resolved components and their spatial distributions have positive values, suggesting actual application for spectrum imaging. It has, however, been pointed out that there are two technical difficulties in NMF application [12]. First, in contrast to PCA, the number of components needs to be assumed in advance. Second, there is the possibility of convergence to a local minimum that is different from the global minimum of interest. Although both PCA and NMF can be applied to 4D-STEM data with appropriate data transformation procedures, there have been few applications of PCA [3] and no reports of NMF application to 4D-STEM so far.

In this study we apply NMF to 4D-STEM experimental data acquired from titanium oxide nanosheets overlapping each other. A titanium oxide nanosheet is an ideal specimen with homogeneous thickness and a known crystal structure that has been studied by STEM [13] and TEM [14,15]. The two above-mentioned difficulties of NMF are experimentally investigated in this study. The resolved components (i.e., diffraction patterns) and spatial distributions of the nanosheets show consistent results with the results of previous experiments [13–15], in which the pristine and topotactically reduced domains are included. In this paper, we demonstrate the validity of NMF to the big data obtained using 4D-STEM.

2. Methodological background

In the present study, 4D-STEM data is analyzed on the basis of linear combination model, and the analysis consists of a 4D data transformation and NMF procedure. In this section we outline the methodological background.

An experimental data is expressed as a linear combination of essential components (diffraction patterns) and their weights. In other words, many experimental diffraction patterns are resolved using sparse diffraction patterns and their distributions. To perform the ALS-MCR procedure for NMF, the experimental data **X**, essential diffraction patterns **S** and their spatial distribution **C** should be matrices, and the following equations must hold:

$$\mathbf{X} = \boldsymbol{SC} \qquad (1)$$

$$\mathbf{S} = (\mathbf{X}\mathbf{C}^{T})(\mathbf{C}\mathbf{C}^{T})^{-\mathbf{1}} \qquad (2)$$

Equation (2) is used to estimate a matrix **S**, and a matrix **C** can be calculated using the estimated matrix **S** as

$$\mathbf{C} = (\mathbf{S}^{T}\mathbf{S})^{-\mathbf{1}}(\mathbf{S}^{T}\mathbf{X}) \qquad (3)$$

The ALS-MCR procedure for NMF is described in detail later.

Four-dimensional STEM yields 4D experimental data **i**(*x*,*y*,*u*,*v*), where (*x*,*y*) is the incident probe position and (*u*,*v*) is the pixel of a diffraction pattern, as shown in Fig. 1a. An experimental 4D-STEM data **i**(*x*,*y*,*u*,*v*) should be transformed to a 2D matrix **X** for matrix calculations, and the flow of the transformation is shown in Figs. 1b-1d. First, each diffraction pattern, which is the 2D data of $m \times n$ pixels, is transformed to a one column × ($m \times n$) row vector, and the 4D data $\mathbf{i}(x,y,u,v)$ becomes the 3D data $\mathbf{i}(x,y,n_{uv})$. Then the 3D data is transformed to the 2D data $\mathbf{i}(n_{xy}, n_{uv})$. $\mathbf{X}$ becomes a matrix with $(m \times n) = n_{uv}$ rows and $(M \times N) = n_{xy}$ columns, where M and N are the numbers of probe positions along the x and y directions, respectively. We consider $\boldsymbol{X} \in \mathbb{R}_+^{n_{uv} \times n_{xy}}$, where $\mathbb{R}_+$ is the set of non-negative real numbers. If the number of essential components is k, matrices $\mathbf{S}$ and $\mathbf{C}$ are $\boldsymbol{S} \in \mathbb{R}_+^{n_{uv} \times k}$ and $\boldsymbol{C} \in \mathbb{R}_+^{k \times n_{xy}}$. In general, k is much smaller than n_{uv} and n_{xy}, resulting in thin matrices $\mathbf{S}$ and $\mathbf{C}$. As mentioned above, the number of components k must be assumed in advance for NMF.

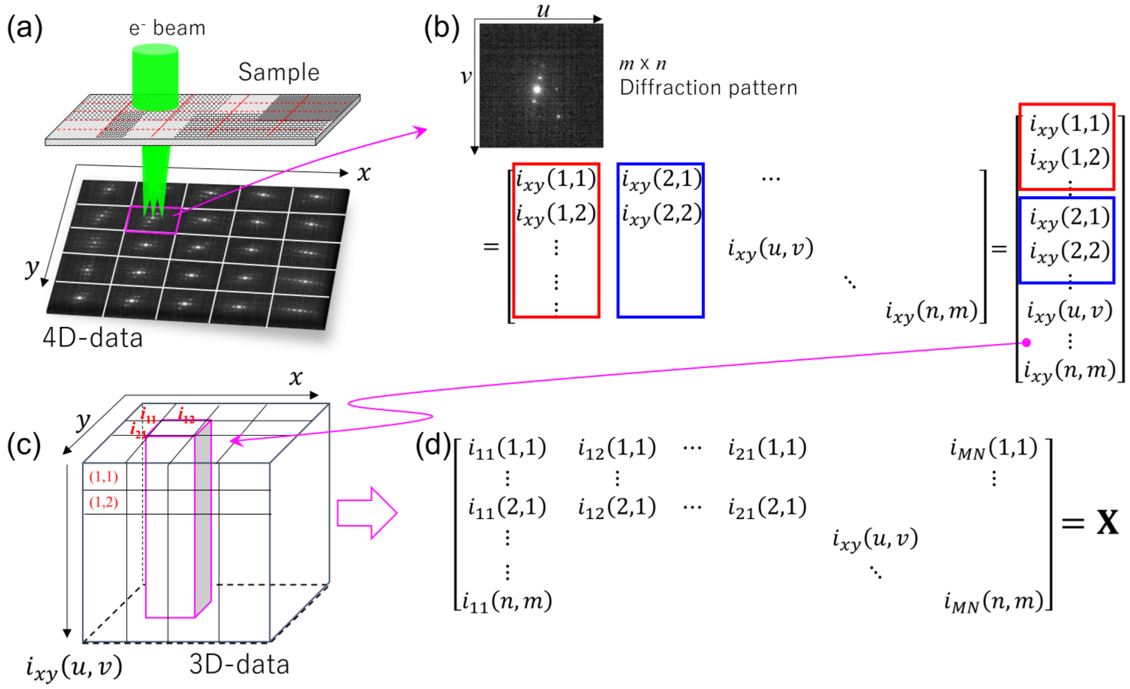


Fig. 1. Schematic diagram of data transformation from 4D to 2D matrix $\mathbf{X}$. (a) Schematic of 4D STEM. (b) Data transformation of diffraction pattern to a one-dimensional (1D) diffraction datum. (c) Schematic of 3D data consisting of a 1D diffraction datum. (d) 2D matrix transformed from 4D-STEM data.

We adopt an NMF scheme that has been demonstrated for spectrum imaging [12]. The NMF procedure of our study consists of the following nine steps:

1) The number k of essential components is assumed.

2) A matrix $\mathbf{C}$ is generated that consists of non-negative uniform-random numbers from zero to

one.

3) A matrix **S** is calculated using Eq. (2), $\mathbf{S} = (\mathbf{X}\mathbf{C}^T)(\mathbf{C}\mathbf{C}^T)^{-1}$.

4) Negative values in the matrix **S** are set to zero, and each column vector of **S** is normalized.

5) A matrix **C** is calculated using Eq. (3), $\mathbf{C} = (\mathbf{S}^T\mathbf{S})^{-1}(\mathbf{S}^T\mathbf{X})$.

6) Negative values of the matrix **C** are set to zero.

7) The mean square error (MSE) of the current estimation is calculated as $\overline{(\boldsymbol{X}-\boldsymbol{SC})^2} = \sum(\boldsymbol{X}-\boldsymbol{SC})^2\left(n_{xy}\times n_{uv}\right)^{-1}$, then its convergence is judged by comparison with the previous MSE. In the case of an enough lower MSE than the previous one, the iteration resumes from step 3). If not MSE, the iteration is stopped.

8) To survey the global minimum, the NMF procedure from step 2) to step 7) are performed multiple times (e.g., twenty in this study) and the minimum MSE and their matrices **S** and **C** are finally determined.

9) Because the number *k* of essential components is unknown, we perform the same NMF procedure from step 1) to step 8) for different values of *k* (e.g., from two to fifteen in this study).

With increasing *k*, the minimum MSE must decrease, because a larger number of components will be advantageous for reproducing experimental data. It should be noted that we can estimate a plausible value of *k* from the *k* dependence of the minimum MSE.

3. Method

3.1. Specimen preparation

Titanium oxide nanosheets were obtained from a single crystal of $K_{0.8}Ti_{1.73}Li_{0.27}O_4$ layered oxide by soft-chemical exfoliation [16]. K^+ and Li^+ ions of the layered oxide were removed, and a colloidal suspension comprising negatively charged $Ti_{0.87}O_2$ sheets surrounded by tetrabutylammonium (TBA) was formed. The colloidal suspension was added dropwise to a holey carbon film on a Cu grid, then was subjected to ultraviolet (UV) light irradiation in air before the STEM experiments. The UV light irradiation removed TBA ions via a photocatalytic reaction.

Because the thickness of the titanium oxide nanosheets is only one Ti atom or two O atoms, diffraction patterns can be analyzed on the basis of the kinematical approximation. The diffraction patterns from randomly overlapping nanosheets are considered to be a linear combination of individual diffraction patterns. Electron microscope observations of titanium oxide nanosheets have been reported elsewhere [13–15,17]. The crystal structures and kinematical diffraction calculations of the nanosheets are given in the Supplementary Note 1. A high-resolution STEM image of the nanosheet is shown in Supplementary Note 4.

3.2. 4D-STEM experiment

An aberration-corrected scanning transmission electron microscope (Thermo Fisher Scientific, Titan cubed) was used at an acceleration voltage of 80 kV. The convergence semiangle α was 1 mrad. The diffraction limit and the defocus of the objective lens regulate the incident probe shape, and the probe size was roughly estimated to be 3 nm from λ/α, where λ is the wavelength (4.2 pm) of an electron at 80 kV [18]. Four-dimensional STEM data was acquired from an area of 500 nm × 500 nm at 100 × 100 probe positions. Diffraction patterns (128 × 128 pixels) were acquired using a CCD camera (Gatan, Inc., UltraScan) of a post-column energy filter (Gatan, Inc., Quantum ERS) with an exposure time of 0.05 s, and the total acquisition time was about 12 min. During the exposure for acquiring a diffraction pattern the incident probe was scanned over 16×16 sub-positions within each probe position (5 nm × 5 nm) to avoid discrete sampling from the specimen. The 4D-STEM data size was 640 MB.

3.3. PCA and NMF analyses

We devised a few programs coded in DigitalMicrograph (Gatan Inc.) script for the data transformation and NMF calculation. A PCA calculation was carried out using a commercial package (HREM Research Inc. MSA plug-in for DigitalMicrograph). Since the direct beam spot in each diffraction pattern was stronger than the other diffraction spots, the direct spots of the 4D-STEM data were masked before NMF and PCA calculations. Non-negative matrix factorization results for a different number of components (k=11) are given in the Supplementary Note 3. The NMF calculation was performed using a standard personal computer. The number of NMF iterations required for convergence was less than 100 and the calculation time was less than 20 min. Examples of the convergence are given in the Supplementary Note 5.

4. Results

4.1. Experimental result of 4D-STEM

The 4D-STEM experimental data is shown in Fig. 2. Figure 2a shows an annular dark-field (ADF) image simultaneously acquired in the 4D-STEM experiment. The white dots with number in Fig. 2a indicate the corresponding probe positions of the diffraction patterns shown in Figs. 2b. Several domains of the observed area show brighter ADF contrast, suggesting the overlap of a few layers at positions 3, 4 and 5. The upper row of Fig.2b shows diffraction patterns extracted from the 4D-STEM data and the lower row shows virtual dark field (VDF) images created selecting diffraction spots indicated by yellow arrow heads in upper row's diffraction pattern. All the diffraction patterns show streaks from the center, which are due to so-called smearing of the charge-coupled device (CCD) camera. The two diffraction patterns taken on the same nanosheet (probe position 1 and 2 in Fig. 2a) show similar diffraction spots except for some extra spots as indicated by arrows in Fig.2b-1. These extra spots have already been identified [13], and one of the authors reported that a titanium oxide nanosheet showed a topotactic transformation; a pristine $Ti_{0.87}O_2$ nanosheet was partly reduced to a Ti_2O_3 structure during TEM observation without changing its nanosheet features [14,15]. Our previous analysis elucidated that the Ti_2O_3 structure showed extra reflection spots, corresponding to the forbidden reflections of the $Ti_{0.87}O_2$ structure (see the Supplementary Information and our previous work [14]). Although the VDF images can be used to visualize each layer, it is difficult to estimate the diffraction pattern of each layer without overlapping. Furthermore, in case that Ti_2O_3 and $Ti_{0.87}O_2$ coexist, their distribution could not be determined. In the present study, we try to discriminate these structures and to determine distribution using NMF and 4D-STEM.

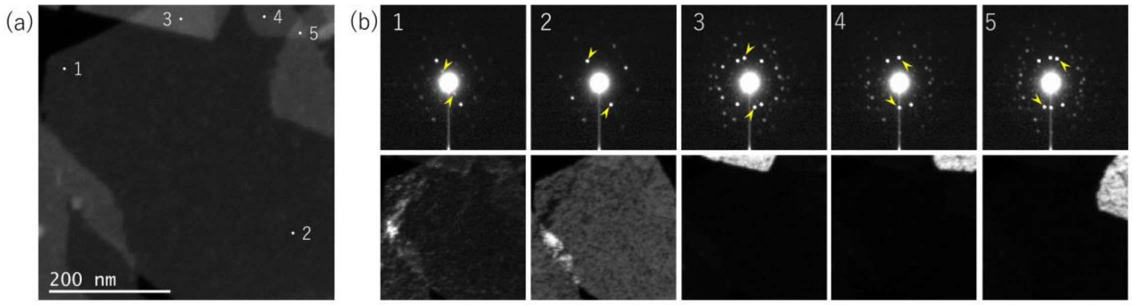


Fig. 2. (a) ADF image and diffraction patterns (upper row of (b)) extracted from 4D-STEM data. The white dots labelled (1-5) in (a) show the incident probe positions corresponding to the diffraction patterns (1-5). Lower row of (b) shows virtual dark field images created by selecting two diffraction spots indicated by yellow arrows of each diffraction pattern.

4.2. PCA and NMF for 4D-STEM

We first apply PCA to the 4D-STEM data as shown in Fig. 3. A scree plot, which indicates the logarithmic eigenvalue of each component, is often used to estimate the number of essential components. Although the number of components could be estimated (about twenty according to Fig. 3a), it appears to be larger than the number of nanosheets in the observed area. Figure 3b shows the principal components and their distributions, whose component indices are shown by the numbers (1-8) in Figs. 3a and 3b. Red and blue colors of the components and distributions represent positive and negative values, respectively. The first component appears to be an averaged diffraction pattern of the entire the observed area, and the component and the distribution have positive values. By contrast, other components and distributions include negative values, and are difficult to interpret as diffraction patterns or spatial distributions. Overlapping diffraction patterns are not resolved. Consequently, diffraction information and the number of components could not be estimated based on the PCA. It should also suggest that the PCA is not effective for noise reduction.

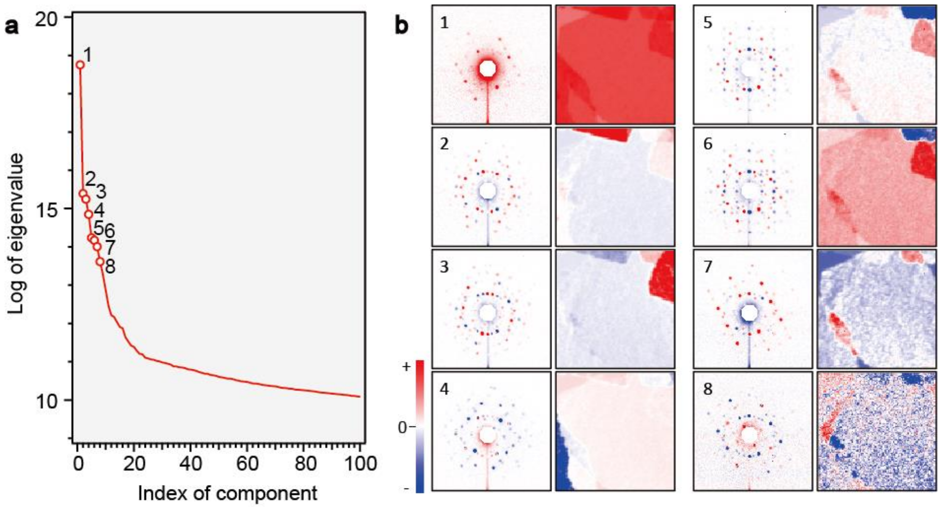


Fig. 3. Principal component analysis result for 4D-STEM data. (a) Scree plot, which shows logarithmic eigenvalues as a function of the index of the component. (b) Principal components (left) and their spatial distributions (right), whose component indices are given as numbers. Red and blue colors correspond to positive and negative values, respectively.

Next, we apply NMF to the same 4D-STEM data. Since the number of components *k* is unknown, we perform NMF with the number *k* varied from two to fifteen. To survey the global minimum, we repeat the NMF procedure twenty times at each *k* value, and calculate the MSE, $\overline{(\boldsymbol{X}-\boldsymbol{SC})^2}$. Figure 4a shows the variation of MSEs as a function of the number of components *k*. The minimum MSE at each *k* decreases with increasing *k*, although the scattered points (indicated by gray rectangles in Fig. 4a) represent convergences to various local minima. Also note that the averaged MSE and the minimum MSE diverge with increasing *k*. This suggests that the probability of convergence to a local minimum increases with increasing number of components. The minimum MSE at each *k* appears to be close to saturation at *k* of around ten. This graph can be practically used to estimate the number of components similarly to a scree plot in PCA. The MSE of NMF instantaneously decays in comparison with the scree plot of PCA (Fig. 3a), suggesting that the validity of NMF procedure. We will discuss the number of components in the Discussion section.

Although the number of components is still arbitrary in Fig. 4a, here we see the NMF result for *k*=7 for instance. Figure 4b shows the components and their distributions that exhibit the minimum MSE for *k*=7 as indicated by an arrow in Fig. 4a. The seven components and their distributions are in descending order of the integrated intensity of each distribution. As a result of the

NMF procedure, all the components and distributions have non-negative values, although positive values are not a sufficient condition to be able to interpret each component as a diffraction pattern. The first component corresponds to the background in the diffraction patterns and its distribution covers almost all the observed area. A small amount of contamination on the surface, quantum noise, dark currents and the smearing of the CCD camera are possible sources of this component. This suggests that NMF can be used to deduce a spatially independent background and noise. The second component corresponds to the diffraction patterns of a pristine $Ti_{0.87}O_2$ nanosheet without topotactic reduction. Note that the observed spots of the all resolved component can be interpreted as an ordinary diffraction pattern (see Supplementary Information). The sixth component is similar to the second but it has extra spots corresponding to the diffraction pattern of Ti_2O_3 nanosheets.

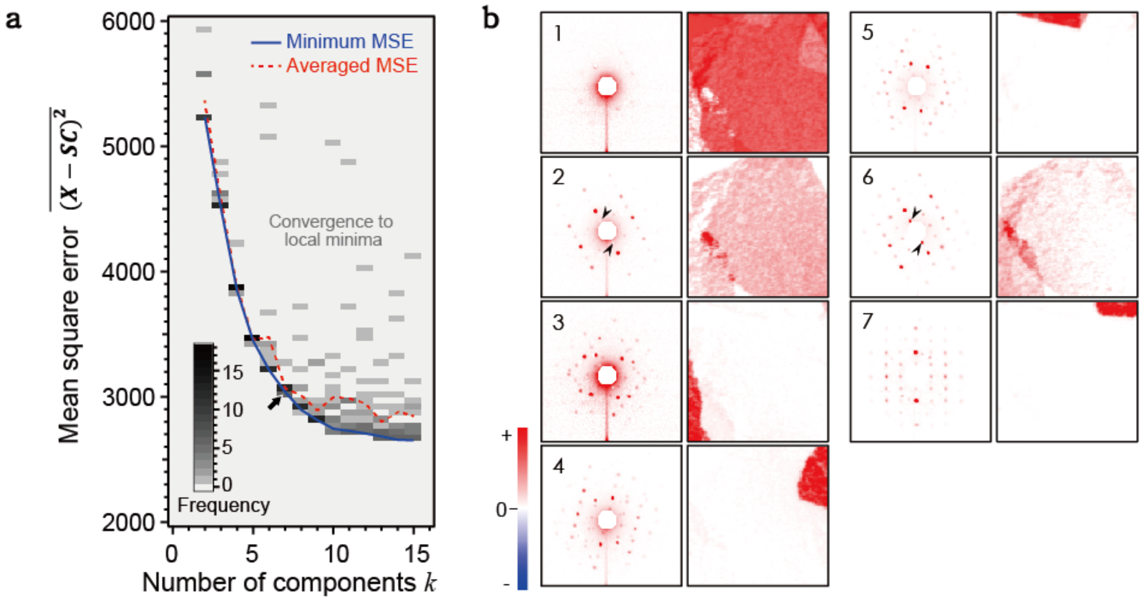


Fig. 4. Non-negative matrix factorization result for 4D-STEM data. (a) Mean square error as a function of number of components. Gray squares denote the frequency of the convergences, as shown by the inset brightness bar. The blue solid line and red broken line respectively indicate the minimum and averaged MSEs for each number of components. (b) Components (left) and their spatial distributions (right) of the minimum MSE in the case of k=7, as shown by the arrow in Fig. 4a.

Non-negative matrix factorization successfully resolves the overlapping patterns of $Ti_{0.87}O_2$ and Ti_2O_3. The third, fourth, fifth and seventh components represent other individual nanosheets. The third and fifth components do not correspond to a single-layer diffraction pattern; the components are considered to correspond to folded areas because their domains exhibit higher ADF contrast as shown in Fig. 1a. Other NMF results for a different number of components are given in the Supplementary Note 3.

5. Discussion

As shown in Fig.4, NMF successfully factorizes sparse diffraction patterns from big data obtained using 4D-STEM. Here, we directly compare the NMF results with PCA results from the viewpoint of sparse modeling. Figure 5 shows the MSEs in PCA and NMF analyses as functions of the number of components. Here we can quantitatively compare the validity of sparse modeling by PCA and NMF. The PCA analysis has higher MSEs up to k=9, indicating that NMF is suitable for finding a small number of essential components for 4D-STEM. The MSE of PCA becomes smaller than the minimum MSE of NMF for k>9. This suggests that the negative components effectively reduce the MSE in PCA. This number of components (k=9) is considered to be the critical number required to factorize the present experimental result using non-negative components. We think that the quantitative difference of MSEs between PCA and NMF may provide a clue when estimating an unknown number of components.

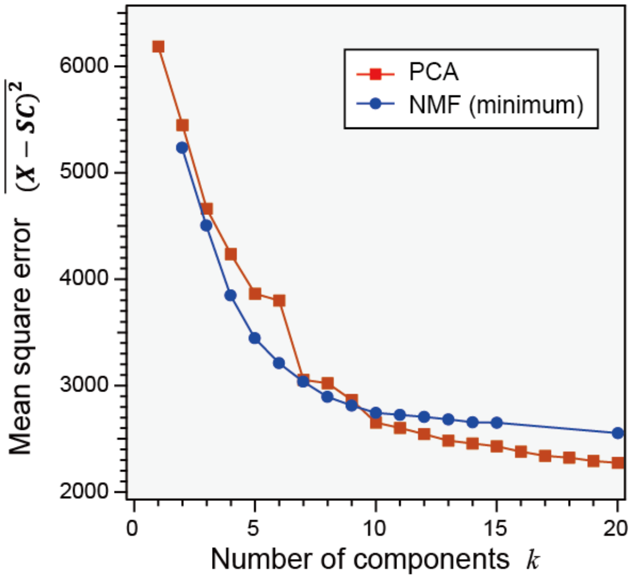


Fig. 5. Mean square errors in PCA and NMF analyses for various numbers/indices of components.

Lastly, we elucidate the NMF results at different number of components (k=7 and 11, see Fig. S3 of Supplementary Note 3). It is found that there are similar distributions estimated for k=7 and 11. The NMF results for large number of components (k=11) resolve variation of bending of nanosheet. Although the number of components could not be conclusively determined by using the MSE plot, the plateau in the MSE could be a guideline to estimate the number of components in NMF.

6. Summary

To mine useful crystallographic information from big data obtained using 4D-STEM, we have applied PCA and NMF. PCA is often applied for denoising in spectrum imaging; however, a resolved component cannot be interpreted as a diffraction pattern. We have applied NMF to 4D-STEM, which successfully resolves the components that can be interpreted as diffraction patterns. Two diffraction patterns of $Ti_{0.87}O_2$ and Ti_2O_3 nanosheets, in which many diffraction spots overlap, are successfully factorized. The k dependence of the minimum MSE has been used to estimate the number of components and to survey the global minimum of interest. A quantitative comparison of MSEs between NMF and PCA elucidates the validity of NMF and is informative for estimating the number of essential components. The combination of NMF and 4D-STEM is expected to a standard characterization technique for a wide range of materials.

Acknowledgments

The authors would like to express their thanks Dr. Sasaki and Dr. Ebina for providing the titanium oxide nanosheets. They also thank Dr. Ishizuka and Dr. Shiga for invaluable discussions. This study was partly supported by a Grant-in-Aid for Scientific Research on Innovative Area "Mixed Anion Compounds" (Project JP16H06440) of JSPS. K.K. also thanks MEXT Elements Science and Technology Project and Elements Strategy Initiative to Form Core Research Center.

References

[1] A. Hirata, S. Kohara, T. Asada, M. Arao, C. Yogi, H. Imai, Y. Tan, T. Fujita, M. Chen, Atomic-scale disproportionation in amorphous silicon monoxide, Nat. Commun. 7 (2016) 58–59. https://doi.org/10.1038/ncomms11591.

[2] S. V. Kalinin, E. Strelcov, A. Belianinov, S. Somnath, R.K. Vasudevan, E.J. Lingerfelt, R.K. Archibald, C. Chen, R. Proksch, N. Laanait, S. Jesse, Big, Deep, and Smart Data in Scanning Probe Microscopy, ACS Nano. 10 (2016) 9068–9086. https://doi.org/10.1021/acsnano.6b04212.

[3] S. Jesse, M. Chi, A. Belianinov, C. Beekman, S. V. Kalinin, A.Y. Borisevich, A.R. Lupini, Big Data Analytics for Scanning Transmission Electron Microscopy Ptychography, Sci. Rep. 6 (2016) 1–8. https://doi.org/10.1038/srep26348.

[4] S. V Kalinin, B.G. Sumpter, R.K. Archibald, Big-deep-smart data in imaging for guiding materials design., Nat. Mater. 14 (2015) 973–80. https://doi.org/10.1038/nmat4395.

[5] W. Xu, J.M. Lebeau, Ultramicroscopy A deep convolutional neural network to analyze position averaged convergent beam electron diffraction patterns, 188 (2018) 59–69. https://doi.org/10.1016/j.ultramic.2018.03.004.

[6] P. Torruella, M. Estrader, A. López-ortega, M. Dolors, M. Varela, F. Peiró, S. Estradé, Ultramicroscopy Clustering analysis strategies for electron energy loss spectroscopy (EELS), 185 (2018) 42–48. https://doi.org/10.1016/j.ultramic.2017.11.010.

[7] J. Tao, D. Niebieskikwiat, M. Varela, W. Luo, M.A. Schofield, Y. Zhu, M.B. Salamon, J.M. Zuo, S.T. Pantelides, S.J. Pennycook, Direct imaging of nanoscale phase separation in La(0.55)Ca(0.45)MnO(3): relationship to colossal magnetoresistance., Phys. Rev. Lett. 103 (2009) 097202. https://doi.org/10.1103/PhysRevLett.103.097202.

[8] F. Uesugi, A. Hokazono, S. Takeno, Evaluation of two-dimensional strain distribution by STEM/NBD, Ultramicroscopy. 111 (2011). https://doi.org/10.1016/j.ultramic.2011.01.035.

[9] K. Kimoto, K. Ishizuka, Ultramicroscopy Spatially resolved diffractometry with atomic-column resolution, Ultramicroscopy. 111 (2011) 1111–1116. https://doi.org/10.1016/j.ultramic.2011.01.029.

[10] M. Watanabe, E. Okunishi, K. Ishizuka, Analysis of Spectrum-Imaging Datasets in Atomic-Resolution Electron Microscopy, Microsc. Anal. 23 (2009) 5–7.

[11] S. Muto, T. Yoshida, K. Tatsumi, Diagnostic Nano-Analysis of Materials Properties by Multivariate Curve Resolution Applied to Spectrum Images by S/TEM-EELS, Mater. Trans. 50 (2009) 964–969. https://doi.org/10.2320/matertrans.MC200805.

[12] M. Shiga, K. Tatsumi, S. Muto, K. Tsuda, Y. Yamamoto, Ultramicroscopy Sparse modeling of EELS and EDX spectral imaging data by nonnegative matrix factorization, Ultramicroscopy. 170 (2016) 43–59. https://doi.org/10.1016/j.ultramic.2016.08.006.

[13] S. Koshiya, S. Yamashita, K. Kimoto, Microscopic observation of dye molecules for solar cells on a titania surface, Sci. Rep. 6 (2016) 2–7. https://doi.org/10.1038/srep24616.

[14] M. Ohwada, K. Kimoto, K. Suenaga, Y. Sato, Y. Ebina, T. Sasaki, Synthesis and Atomic Characterization of a Ti_2O_3 Nanosheet, J. Phys. Chem. Lett. 2 (2011) 1820–1823. https://doi.org/10.1021/jz200781u.

[15] M. Ohwada, K. Kimoto, T. Mizoguchi, Y. Ebina, T. Sasaki, Atomic structure of titania nanosheet with vacancies, Sci. Rep. 3 (2013). https://doi.org/10.1038/srep02801.

[16] T. Sasaki, M. Watanabe, H. Hashizume, H. Yamada, H. Nakazawa, Macromolecule-like aspects for a colloidal suspension of an exfoliated titanate. Pairwise association of nanosheets and dynamic reassembling process initiated from it, J. Am. Chem. Soc. 118 (1996) 8329–8335. https://doi.org/10.1021/ja960073b.

[17] H.S. Al Qahtani, K. Kimoto, T. Bennett, J.F. Alvino, G.G. Andersson, G.F. Metha, V.B. Golovko, T. Sasaki, T. Nakayama, Atomically resolved structure of ligand-protected Au9 clusters on TiO2 nanosheets using aberration-corrected STEM., J. Chem. Phys. 144 (2016) 114703. https://doi.org/10.1063/1.4943203.

[18] S.J. Pennycook, P.D. Nellist, Scaninng transmission electron microscopy, Springer, 2011.

Supplementary Information

Non-negative matrix factorization for mining big data obtained using four-dimensional scanning transmission electron microscopy

Fumihiko Uesugi[1], Shogo Koshiya[2], Jun Kikkawa[2], Takuro Nagai[2], Kazutaka Mitsuishi[1] and Koji Kimoto[2]

[1] National Institute for Materials Science, 1-2-1 Sengen, Tsukuba, Ibaraki 305-0047, Japan
[2] National Institute for Materials Science, 1-1 Namiki, Tsukuba, Ibaraki 305-0044, Japan

Contents

Supplementary note 1: Crystal structures and diffraction patterns of titanium oxide nanosheets

Supplementary note 2: Comparison between deduced components and kinematical diffraction calculations

Supplementary note 3: Examples of non-negative matrix factorization (NMF) with different number of components

Supplementary note 4: High-resolution annular dark-field (ADF) image of titanium oxide nanosheet

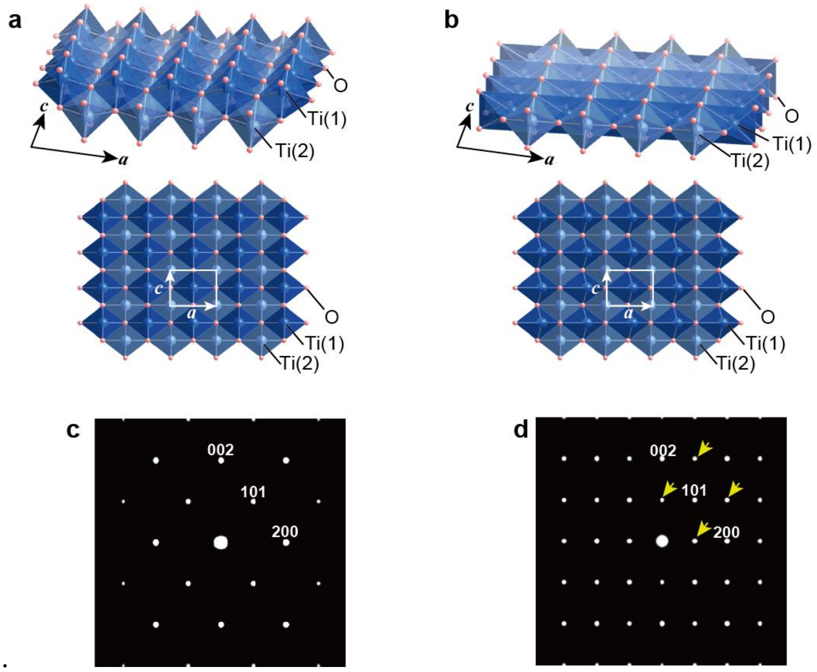
Supplementary note 5: Examples of convergence in NMF procedure

Supplementary note 6: DigitalMicrograph script for NMF procedure

Supplementary note 1: Crystal structures and diffraction patterns of titanium oxide nanosheets

The crystal structures of titanium oxide nanosheets are shown in Supplementary Fig. 1. Although an actual titanium oxide nanosheet includes titanium vacancies[1], we consider an ideal structure without vacancies for the diffraction calculation[2]. A pristine TiO_2 nanosheet (Supplementary Fig. S1a) consists of TiO_6 octahedrons with adjacent octahedrons having shared edges. A pristine TiO_2 nanosheet is topotactically reduced by electron irradiation to a Ti_2O_3 nanosheet, in which TiO_6 octahedrons share faces (Supplementary Fig. S1b). The variation in these crystal structures is known as Magneli phase. Half of the octahedrons of the Ti_2O_3 nanosheet structure are highly distorted and half of the titanium atoms (Ti(1) in Supplementary Fig. S1b) are slightly shifted along the *a* axis.

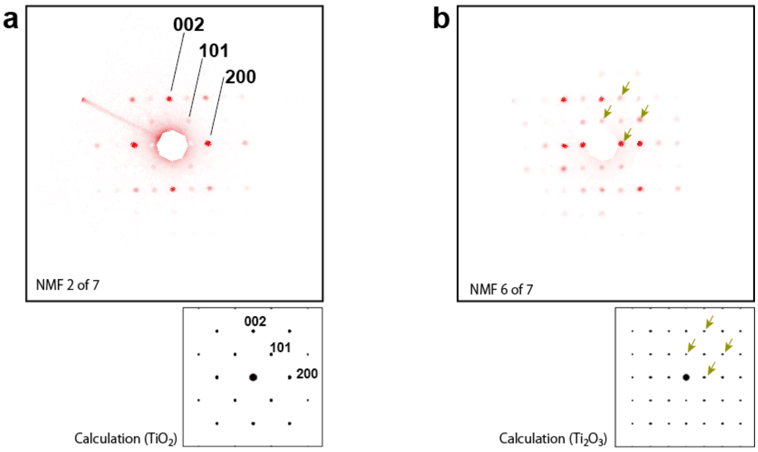
Owing to this crystallographic modification, the forbidden diffraction spots of a pristine TiO_2 structure become observable. Supplementary Figs. S1c and S1d are respectively the diffraction patterns of TiO_2 and Ti_2O_3 structures obtained by kinematical calculations. The forbidden diffraction spots, such as 100, 001 and 201, become evident in the Ti_2O_3 diffraction pattern as shown by arrows in Supplementary Fig. S1d.



Supplementary Fig. S1 Crystal structures and diffraction patterns of TiO_2 and Ti_2O_3 nanosheets. **a**, **b** Crystal structures of TiO_2 and Ti_2O_3 nanosheets. **c**, **d** Kinematical diffraction patterns of TiO_2 and Ti_2O_3 nanosheets.

Supplementary note 2: Comparison between deduced components and kinematical diffraction calculations

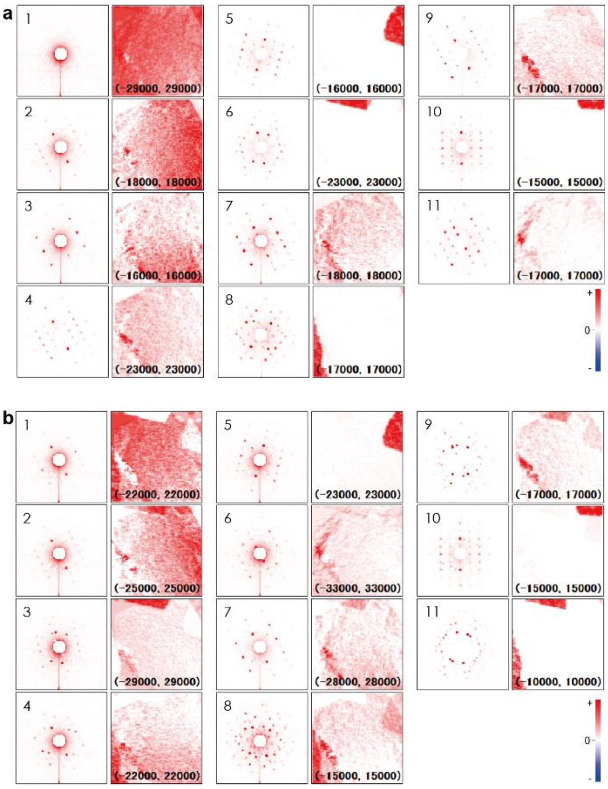
Here we compare the components deduced by non-negative matrix factorization (NMF) (Fig. 4b) with the kinematical diffraction calculations (Supplementary Fig. S1). Supplementary Figs S2a and S2b show the two deduced components in the case of k=7 (see Fig. 4). The components are rotated to fit the direction of the ***a*** axis in the kinematical calculations. The second (Supplementary Fig. S2a) and sixth components(Supplementary Figs. S2a and S2b) are similar to the TiO_2 and Ti_2O_3 diffraction patterns, respectively.



Supplementary Fig. S2 Comparison between deduced components and kinematical diffraction calculations. **a** The upper figure is the second component in Fig. 4b. The lower figure shows the kinematical calculation of the TiO_2 nanosheet. **b** The upper figure is the sixth component in Fig. 4b. The lower figure shows the kinematical calculation of the Ti_2O_3 nanosheet.

Supplementary note 3: Examples of non-negative matrix factorization (NMF) with different number of components

In the main text and Fig. 4 we gave the result of NMF in the case of k=7, where k is the number of components. Here we show NMF results for k=11. As described in the main text, NMF was performed twenty times with different starting values for each number of components, and we selected the result with the minimum mean square error (MSE) from the twenty trials. Supplementary Fig. S3a shows the NMF result whose MSE is the minimum of twenty calculations for k=11. The components (left) and their distributions (right) are in descending order of the integrated intensity of the distributions. We compare this NMF result for k=11 (Supplementary Fig. S3a) with that for k=7 (Fig. 4b). The first components of both NMF results have very similar diffraction patterns and distributions. The fifth, sixth, eighth and tenth components in Supplementary Fig. S3a (k=11) correspond to the fourth, fifth, third and seventh components in Fig. 4b (k=7), respectively. The other (second, third, fourth, seventh, ninth, eleventh) components in Supplementary Fig. S3a correspond to the second and sixth components in Fig. 4b. The distributions of the components (second, third, fourth, seventh) of Supplementary Fig. S3a show a gradation, suggesting the bending of the nanosheet.

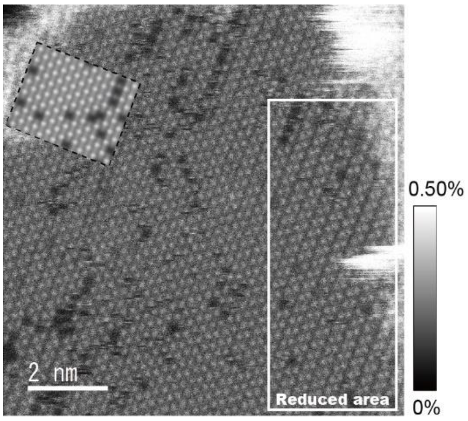


Supplementary Fig. S3 Non-negative matrix factorization results obtained for k=11. Results with the **a** minimum MSE and **b** largest MSE among twenty calculations for k=11. The inset numbers are the display range of each distribution.

As mentioned in the main text, the twenty trials have a variety of MSEs, representing conversions to local minima. Supplementary Fig. S3b shows an example with a large MSE. It is clear that the NMF does not successfully factorize the diffraction patterns. For instance, the eighth, ninth and eleventh distributions in Supplementary Fig. S3b have intensities on plural domains, and their deduced components are not simple diffraction patterns.

Supplementary note 4: High-resolution annular dark-field (ADF) image of titanium oxide nanosheet

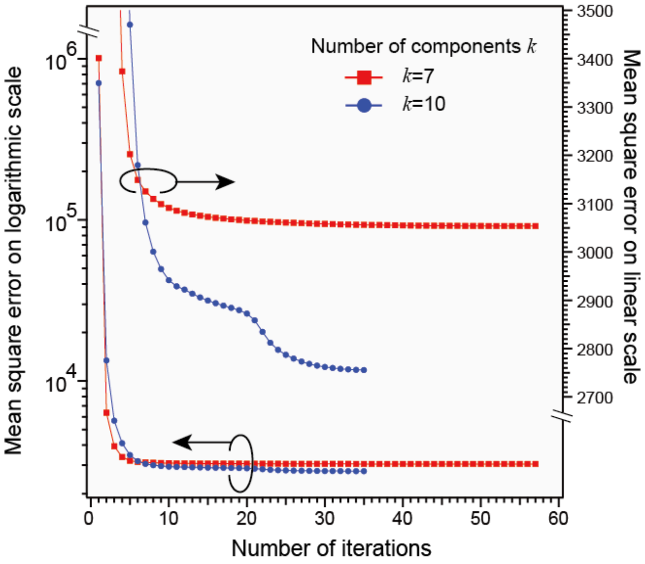
In our previous reports we analyzed the microstructure and topotactic reduction of titanium oxide nanosheets using electron diffraction and high-resolution TEM imaging[1,2]. We also observed the atomic structure using annular dark-field (ADF) imaging[3]. Here we show an ADF image of a titanium oxide nanosheet that includes both pristine and reduced titanium oxide domains. Supplementary Fig. S4 shows an ADF image of a titanium oxide nanosheet observed using an aberration-corrected STEM instrument (Thermo Fisher Scientific, Titan^3) at 80 kV. A liquid-nitrogen cooling holder (Gatan, Inc., UHRTR3500) was used at −180 °C to reduce the contamination and irradiation damage of the specimen. The ADF contrast was quantified as the scattering probability[4,5], and bright dots of the image correspond to titanium atoms. The inset in the upper left (the rectangle with a broken line) shows the multislice simulation (HREM Research Inc., xHREM) result for a pristine titanium oxide nanosheet with titanium vacancies. The rectangular area on the right corresponds to a reduced area, i.e., this ADF image shows the different atomic arrangements of titanium.



Supplementary Fig. S4 ADF image of titanium oxide nanosheet observed using aberration-corrected STEM.

Supplementary note 5: Examples of convergence in NMF procedure

Non-negative matrix factorization requires iterative calculations to find the minimum MSE. The iteration time required for convergence depends on the number randomly generated in the initial step (see step 3 in the main text) of the iteration. Supplementary Fig. S5 shows examples of convergence, in which the MSEs are plotted on both logarithmic (left) and linear (right) intensity scales as a function of the iteration number; the two examples for the cases of k = 7 and 10 are shown as squares and circles, respectively. The typical calculation time required for convergence is less than 20 min using a desktop personal computer and a custom script of DigitalMicrograph (Gatan, Inc.), although it could be reduced by preparing DLL files for DigitalMicrograph. Thus, the calculation time for NMF is practical.



Supplementary Fig. S5 Examples of convergence in NMF procedure. Mean square errors are plotted on both logarithmic (left) and linear (right) scales as a function of the number of iterations. The two examples of k = 7 and 10 are shown as squares and circles, respectively.

Supplementary note 6: DigitalMicrograph script for NMF procedure

We prepared a few DigitalMicrograph scripts for this study. DigitalMicrgraph software can be downloaded through the website of the manufacturer (Gatan, Inc.)[6]. The NMF algorithm that consists of nine steps in this study was fully described in the main text. The core steps for NMF can be written using DigitalMicrograph functions as follows:

```
// step 2
     C = UniformRandom()
// step 3
     CT=MatrixTranspose(C)
     CCT = MatrixMultiply(C, CT)
     ICCT = MatrixInverse(CCT)
     XCT = MatrixMultiply(X, CT)
     S = MatrixMultiply(XCT, ICCT)
// step 4
     S = tert(S<0, 0, S)
// step 5
     ST=MatrixTranspose(S)
     STS = MatrixMultiply(ST,S)
     ISTS = MatrixInverse(STS)
     STX = MatrixMultiply(ST, X)
     C = MatrixMultiply(ISTS, STX)
// step 6
     C = tert(C<0, 0, C)
// step 7
     MSE = MeanSquare(X - MatrixMutiply(S, C))
```

Supplementary References

[1] M. Ohwada, K. Kimoto, T. Mizoguchi, Y. Ebina, T. Sasaki, Atomic structure of titania nanosheet with vacancies, Sci. Rep. 3 (2013). https://doi.org/10.1038/srep02801.

[2] M. Ohwada, K. Kimoto, K. Suenaga, Y. Sato, Y. Ebina, T. Sasaki, Synthesis and Atomic Characterization of a Ti_2O_3 Nanosheet, J. Phys. Chem. Lett. 2 (2011) 1820–1823. https://doi.org/10.1021/jz200781u.

[3] S. Koshiya, S. Yamashita, K. Kimoto, Microscopic observation of dye molecules for solar cells on a titania surface, Sci. Rep. 6 (2016) 2–7. https://doi.org/10.1038/srep24616.

[4] S. Yamashita, S. Koshiya, T. Nagai, J. Kikkawa, K. Ishizuka, K. Kimoto, Quantitative annular dark-field imaging of single-layer graphene—II: Atomic-resolution image contrast, Microscopy. 64 (2015) 409–418. https://doi.org/10.1093/jmicro/dfv053.

[5] S. Yamashita, S. Koshiya, K. Ishizuka, K. Kimoto, Quantitative annular dark-field imaging of single-layer grapheme, Microscopy. 64 (2015) 143–150. https://doi.org/10.1093/jmicro/dfu115.

[6] Gatan, Inc., Gatan Microscopy Suite Software, (2018). https://www.gatan.com/products/tem-analysis/gatan-microscopy-suite-software.

Highlights

The Non-negative matrix factorization (NMF) is applied to analyze 4D-STEM data.

NMF can resolve 4D-STEM data to diffraction patterns and corresponding distributions.

We demonstrated the present method to 4D-STEM data from titanium oxide nanosheets.