PAPER

Novel spectral unmixing approach for electron energy-loss spectroscopy

Ren-Fong Cai\textsuperscript{1,2} \textsuperscript{a}, Mu-Tung Chang\textsuperscript{1}, Shen-Chuan Lo\textsuperscript{3} and Chien-Chun Chen\textsuperscript{1,3,4} \textsuperscript{a}

\textsuperscript{1} Department of Engineering and System Science, National Tsing Hua University, Hsinchu 30013, Taiwan
\textsuperscript{2} Material and Chemical Research Laboratories, Industrial Technology Research Institute, Hsinchu 31057, Taiwan
\textsuperscript{3} Instrument Technology Research Center, National Applied Research Laboratories, Hsinchu 30076, Taiwan
\textsuperscript{4} National Synchrotron Radiation Research Center, Hsinchu 30076, Taiwan

E-mail: ccchen0627@gmail.com

Keywords: electron energy-loss spectroscopy, scanning transmission electron microscopy, spectral unmixing

Abstract

Electron energy-loss spectroscopy is a well-established technique for characterizing low-Z elements in materials. Typically, a measured spectrum image is contributed from several materials when the composition of the specimen is sophisticated. Therefore, decomposing the distribution of each endmember is crucial to material scientists. In this article, we combined multiple linear least-squares fitting and $k$-means clustering to resolve the aforementioned issue. In addition, our method can nearly extract the true endmembers in materials unsupervisedly. Simulated and experimental data were employed to evaluate the performance and feasibility of our method.

1. Introduction

A spectrum image of a scanning transmission electron microscope (STEM) allows to collect high-resolution structural and chemical information simultaneously. The two most common spectroscopic techniques in STEM are electron energy-loss spectroscopy (EELS) and energy-dispersive x-ray spectroscopy (EDS), respectively. Although EDS is more efficient in detecting high-Z elements, EELS, on the contrary, is more sensitive to low-Z chemicals \cite{1–4}. A typical issue of EELS is that the measurement is a mixed spectrum; therefore, signal processing becomes essential when exploring a deep insight into materials, especially to extract the useful signal from the dataset \cite{5–10}.

Spectral unmixing is one of the tools to decompose the mixed spectrum into a collection of distinct spectra or endmembers—the absorption of individual elements or compounds at a specific range of energy \cite{11, 12}. Several algorithms for EDS/EELS data processing, such as principal component analysis (PCA), independent component analysis (ICA), multiple linear least-squares (MLLS) fitting, and non-negative matrix factorization (NMF), have been applied to extract the endmembers from a spectrum image.

PCA and ICA are used to reduce the dimension of the dataset. However, PCA is challenging to extract physically meaningful endmembers because the individual elements in the spectrum cannot be identified. Recently, ICA was applied to extract endmembers, but the ICA route requires denoising and is highly time-consuming \cite{13–17}.

MLLS fitting and NMF are both vector quantization techniques aiming to decompose the signal into a linear combination of principal components. It has been demonstrated that the MLLS fitting can solve the issue of edge overlapping in EELS by manually assigning principal spectra as \cite{9, 18, 19}. On the contrary, the NMF can automatically extract the meaningful components with the positivity constraints on the weighting of the components. Although NMF has been widely used in image processing, text mining, and spectral unmixing, great efforts are required to enhance the efficiency \cite{20–24}.

In order to extract physically meaningful endmembers, the cluster analysis, which groups the data with similarity, would be one of the solutions. Various algorithms have been developed for clusterings, such as agglomerative clustering, density-based spatial clustering of applications with noise, and $k$-means clustering.
K-means clustering is well-known as an unsupervised routine that groups the data only depending on the similarity and has been widely used for the data processing of electron microscopy. Most of the applications focus on image processing but rarely on spectroscopy. In EELS, the spectra are recorded by multiple channels corresponding to different electron energy at a single position. It can be considered as every single spectrum spanned in the multi-dimensional space, and the superposition forms the measured spectra. Therefore, the metric of similarity to each spectrum can be defined as the Euclidean distance in the energetic space. The centroids of each cluster can be represented as the reference spectra for MLLS routine. However, in most of the case, we cannot extract the true endmember by the k-means clustering.

In this article, we propose a novel spectral unmixing approach named kMLLS clustering. By taking both advantages of the MLLS fitting and k-means clustering, the true individual endmembers can be extracted distinctly from a measured spectrum image.

The feasibility and robustness were examined by both simulated and experimental data.

2. Validation through simulation

The multiple linear least-squares (MLLS) fitting assumes that the nth spectrum in a spectrum image can be formulated as a linear combination of p reference spectra as follows

\[ S_n(E_{cm}) = \sum_{i=1}^{p} \beta_i R_i(E_{cm}) \]  \hspace{1cm} (1)

or

\[ S_n(E_{c1}, E_{c2},..,E_{cm}) = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} \begin{bmatrix} R_1(E_{c1}, E_{c2},..,E_{cm}) \\ \vdots \\ R_p(E_{c1}, E_{c2},..,E_{cm}) \end{bmatrix}, \]  \hspace{1cm} (2)

where \( S_n(E_{cm}) \) represents the nth energy channels selected to acquire the spectrum image. \( \beta \) refers to the fitting coefficient of each reference spectra. \( R_i(E_{cm}) \) is the endmember in a spectrum image. The main purpose of performing MLLS to a spectrum image is to solve \( \beta \). The solution can be expressed as

\[ \beta = S_n \times R^T \times (R \times R^T)^{-1}. \]  \hspace{1cm} (3)

The relative fractional abundance of the pth reference spectrum is represented by \( \beta_p \), so that we can obtain the spatial distribution of each reference spectrum [18, 19].
Usually, the $k$-means clustering algorithm, which groups data with the similarity defined by the Euclidean distance, is applied to extract the reference spectra or the endmember from a spectrum image. However, the EELS signal background decays exponentially, which may result in large differences of the similarity if we considered all the energy range during grouping. In order to avoid such situation, we cropped the energy region of interest before the analysis. After that, $k$ cluster centroids were randomly selected from the dataset. Then each data point was assigned to the specific group depending on the distance to the centroid. New centroids were computed as the center of mass of the data points in individual groups. The iterative process was not terminated until the centroids converged. To determine the proper value of $k$, one would need to compute the sum of squared errors (SSE) from a range of $k$ clusters and determine the $k$-value on the elbow of the SSE curve $[25, 30]$. The physical interpretation of $k$-means clustering can be realized as grouping the electron energy-loss spectra at specific positions into $k$ clusters. The centroid indicates the average spectra of each cluster.

In most cases, the accurate phase of endmember could not be extracted by $k$-means clustering. We hence combine $k$-means clustering with MLLS fitting to overcome the difficulty. The flow-chart of kMLLS is shown in Figure 2.

**Figure 2.** The simulated BN and C interface system. (a) The structure and the distribution of each endmember. (b) and (c) are simulated EELS spectrum of pure phases of BN and C at the position of 0 nm and 19.9 nm, respectively. 10% Poisson noise was imposed on the dataset.

**Figure 3.** The results of the elbow method and $k$-means clustering for the BN–C interface system. (a) The preliminary $k$-value determined by the SSE curve is three. (b)–(d) The centroids of three clusters from $k$-means clustering. The centroids 1 is the linear combination of centroids 2 and 3. (e), (f) The fraction determined by $k$-means and kMLLS clustering, respectively. (g), (h) The EELS spectra of extracted endmembers by kMLLS when the correct $k$-value is determined.

Usually, the $k$-means clustering algorithm, which groups data with the similarity defined by the Euclidean distance, is applied to extract the reference spectra or the endmember from a spectrum image. However, the EELS signal background decays exponentially, which may result in large differences of the similarity if we considered all the energy range during grouping. In order to avoid such situation, we cropped the energy region of interest before the analysis. After that, $k$ cluster centroids were randomly selected from the dataset. Then each data point was assigned to the specific group depending on the distance to the centroid. New centroids were computed as the center of mass of the data points in individual groups. The iterative process was not terminated until the centroids converged. To determine the proper value of $k$, one would need to compute the sum of squared errors (SSE) from a range of $k$ clusters and determine the $k$-value on the elbow of the SSE curve $[25, 30]$. The physical interpretation of $k$-means clustering can be realized as grouping the electron energy-loss spectra at specific positions into $k$ clusters. The centroid indicates the average spectra of each cluster.

In most cases, the accurate phase of endmember could not be extracted by $k$-means clustering. We hence combine $k$-means clustering with MLLS fitting to overcome the difficulty. The flow-chart of kMLLS is shown in...
Figure 1. First, we determined the preliminary $k$-value of the system and then performed the $k$-means clustering to obtain the candidates for the MLLS. Second, a brute-force algorithm enumerates all the combinations of centroids and then examines whether a centroid is composed of other two or more centroids. If the centroid is a linear combination (with positive coefficients) of other centroids, it will be removed from the reference spectra for further MLLS calculation. Third, we performed the MLLS routine pixel-by-pixel to obtain the coefficient of each reference. Fourth, the coefficients were employed to calculate the new reference. Finally, we repeated the procedures until no coefficients were changed beyond the tolerance of each reference. The endmembers would be obtained when the references converged.

Simulated spectra with different compositions of endmembers were generated by the EELS Advisor package in the GATAN DigitalMicrograph software to demonstrate our kMLLS clustering algorithm [35]. In this study, the experimental parameters of JEOL JEM-ARM200F were used to simulate the spectra (primary energy: 200 keV;...
convergence angle: 44.2 mrad; collection angle of the EEL spectra: 29.6 mrad. To be more realistic, 10% Poisson noise was imposed on each spectrum. Although neither background subtraction nor denoising were needed in advance, the combination of the endmembers in the mixed region was assumed to be linear. All the data processing in this paper were implemented by the Numpy, Scikit-learn, and HyperSpy packages in Python [36, 37].

The first example is a two-endmember system. The pure BN and C are located at two endpoints, and the abundant profiles were shown in figure 2(a). Figures 2(b), (c) shows the simulated noisy EELS spectra of BN and C, respectively. The elbow method was applied to find out the optimal k-value for clustering. The result of the elbow route presented in figure 3(a) indicates that the number of clusters is three, which is different from the
given two-endmember model. The $k$-means clustering was performed, and the centroids of each cluster are shown in figures 3(b)–(d). It should be noted that all the centroids are the mixture of BN and C. The components in each centroid examined by the MLLS route are 0.49 BN + 0.51 C, 0.95 BN + 0.05 C, and 0.04 BN + 0.96 C, respectively. Obviously, figure 3(b) is the linear combination of figures 3(c) and (d). The correct $k$-value is then determined as two. The MLLS result from figures 3(c) and (d) was shown in figure 3(e). Because of the mixture of BN and C, the average fractions in the pure BN and C regions were 105.52% and 104.73%. To improve accuracy, we performed the $k$-MLLS clustering, and the result was shown in figures 3(f)–(h). The average fractions in the pure BN and C regions were 100/70% and 100.59%, respectively.

It is notable that if we pick the mixture phase (figure 3(b)) as one of the reference spectra, the $k$-means clustering always trap at the incorrect fractions shown in figures 4(a) and (e). By contrast, the $k$MLLS clustering still converges to nearly pure phases. For example, if figure 3(c) (nearly pure BN) and figure 3(b) (mixture BN and C) are selected as the reference spectra, the figure 3(b) in kMLLS clustering converges to the nearly pure C phase. A similar result was obtained when we pick figure 3(d) (nearly pure C) and figure 3(b) as the reference spectra. The results were shown in figures 4(b) and (f). This indicates no matter which references are selected for kMLLS, the true endmembers can be extracted if the $k$-value is correctly determined. It is notable that the $k$-values are typically larger than the number of endmembers, and can be corrected by the brute-force algorithm. If the $k$-value is underestimated, it may result in negative fractions for some endmembers. In such a case, we can then manually determine the proper $k$-value for kMLLS clustering.

Figure 8. The comparisons of $k$-means and kMLLS clustering using experimental measurements. (a) ADF image (above) and the corresponding spectrum image (below). (b)–(d) Three centroids for clusters. The fraction determined by the $k$-means and kMLLS clustering as well as the extracted endmembers by kMLLS algorithm from (e)–(h) centroids 1 and 2, (i)–(l) centroids 2 and 3, (m)–(p) centroids 1 and 3.
An overlapped multiple-edge system, Ti–N–TiO₂, is designed to demonstrate the feasibility of our method in advance. As shown in figure 5, the O K edge is located between Ti L₃,2 and Ti L₁ edges. The k-value determined by the SSE curve shown in figure 6(a) is three. The corresponding centroids of k-means clustering are shown in figures 6(b)–(d). Although the k-value is correctly determined, figures 6(c) and (d) are still mixed spectra. After we implemented kMLLS clustering, the weighting distribution and the endmembers are both correctly retrieved. The comparisons of k-means and kMLLS clustering are shown in figure 7.

3. Demonstrate the feasibility using experimental data

Finally, we examined the feasibility of kMLLS clustering through experimental data. A line-scanned spectrum image of oxide–nitride–oxide (ONO) multilayer thin film was acquired using a JEOL JEM-ARM200F Cs-Corrected STEM with a GATAN Quantum 965 EELS camera at 200 keV. Since the EELS signal decays exponentially, inaccurate Euclidian distances might result from poor background subtraction. In reality, the background subtraction is usually problematic when multiple elements exist in the material. Therefore, we only select the region of absorption edge for kMLLS clustering. The advantage is that the features in the spectra dominate the value of Euclidian distances (i.e., similarity), thus alleviating the bias of clustering due to the background. Figure 8(a) showed the stack of the annular dark-field image and the corresponding line-scanned spectrum image. Three centroids determined from the elbow test are shown in figures 8(b), (c). Apparently, figure 8(d) is the linear combination of figures 8(b) and (c). We then performed the kMLLS clustering with arbitrary two references from the three centroids. Figures 8(e)–(h), (i)–(l), and (m)–(p) were the comparisons of k-means and kMLLS clustering from the centroids 1 and 2, centroids 2 and 3, and centroids 3 and 1, respectively. Our kMLLS clustering showed great consistency of extracting endmembers of N and O with correct fractions no matter which references were selected.

4. Conclusion

In this paper, we have demonstrated that a novel algorithm, the kMLLS clustering, can successfully extract the endmembers in a spectrum image from both simulated and experimental data. In comparison with k-means clustering, no pre-knowledge or pre-selection of references is needed; therefore, the endmember extraction can be conducted unsupervisedly and lead to nearly true endmembers. We believe that the kMLLS clustering has great potential to in-line investigations and provides a significant insight into materials.

Acknowledgments

This work was supported by the Taiwan Ministry of Science and Technology (106-2112-M-007-029-MY3).

ORCID iDs

Ren-Fong Cai  https://orcid.org/0000-0003-1205-0024
Chien-Chun Chen  https://orcid.org/0000-0002-9141-1088

References

[1] Jeaquailmaurice C and Colliex C 1989 Ultramicroscopy 28 252–7
[2] Hunt J A and Williams D B 1991 Ultramicroscopy 38 47–73
[3] Egerton R 2011 Electron Energy-Loss Spectroscopy in the Electron Microscope (Boston, MA: Springer)
[4] Pennycook S J and Nellist P D 2011 Scanning Transmission Electron Microscopy (New York: Springer)
[5] Stephan O, Ajayan P M, Colliex C, Redlich P, Lambert J M, Bernier P and Lefin P 1994 Science 266 1683–5
[6] Schmid H K 1995 Micro. Microanal. Microstruct. 6 99–111
[7] Kimoto K, Asaka T, Nagai T, Saito M, Matsu Y and Ishiaka K 2007 Nature 450 702–4
[8] Bosman M, Kars T J, Garcia-Muñoz J L, D’Alfonso A J, Findlay S D and Allen L J 2007 Ultramicroscopy 109 32–8
[9] Arenal R, de la Peña F, Stéphan O, Walls M, Tencé M, Loiseau A and Colliex C 2008 Ultramicroscopy 109 113–24
[10] Angadi V C, Abhayaratne C and Walther T 2016 J. Microsc. 262 157–66
[11] Keshava N and Mustard J 2002 IEEE Signal Process. Mag. 19 44–57
[12] Miao L and Qi H 2007 IEEE Trans. Geosci. Remote Sens. 45 765–77
[13] Titchmarsh J M and Dumbill S 1996 J. Microsc. 184 195–207
[14] de la Peña F, Berger M H, Hochepied J F, Dynys F, Stéphan O and Walls M 2011 Ultramicroscopy 111 169–76
[15] Cueva P, Hovden R, Mundy J A, Xin H L and Muller D A 2012 Microsc. Microanal. 18 667–75
[16] Rossouw D, Burdet P, de la Peña F, Ducati C, Knappett B R, Wheatley A E H and Midgley P A 2015 Nano Lett. 15 2716–20
[17] Potapov P and Lubk A 2019 Adv. Struct. Chem. Imaging 5 4
[18] Shuman H and Somlyo A P 1987 Ultramicroscopy 21 23–32
[19] Leapman R D and Svity C R 1988 Ultramicroscopy 26 393–403
[20] Lee D D and Seung H S 1999 Nature 401 788–91
[21] Muñoz-Barrutia A, García-Muñoz J, Ucar B, Fernández-García I and Ortiz-de Solorzano C 2007 Conf. Proc. : Annual Int. Conf. of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conf. pp 6248–51
[22] Arngren M, Schmidt M N and Larsen J 2011 J. Signal Process. Syst. 65 479–96
[23] Wang N, Du B and Zhang L 2013 IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 6 534–69
[24] Gillis N 2014 The Why and How of Nonnegative Matrix Factorization arXiv:1401.5226v2
[25] Raschka S and Mirjalili V 2017 Python Machine Learning 2nd edn (Birmingham: Packt Publishing Ltd.)
[26] Voyles P M 2017 Curr. Opin. Solid State Mater. Sci. 21 141–58
[27] Torruella P, Estrader M, López-Ortega A, Baró M D, Varela M, Peiró F and Estradé S 2018 Ultramicroscopy 185 42–8
[28] Hall R 2004 J. Struct. Biol. 145 19–28
[29] Kumar V 2004 J. Struct. Biol. 145 41–51
[30] Jain A K 2010 Pattern Recognit. Lett. 31 651–66
[31] Yang Z, Fang J, Chitturlu J, Asturias F J and Penczek P A 2012 Structure 20 237–47
[32] Alizade Nacini A, Jamshidzadeh A, Saadatseresht M and Homayouni S 2014 ISPRS—Int. Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences vol XL-2/W3, pp 35–9
[33] Shapiro D A et al 2014 Nat. Photon. 8 765–9
[34] Jesse S, Chi M, Belianinov A, Beekman C, Kalinin S V, Borisevich A Y and Lupini A R 2016 Sci. Rep. 6 26348
[35] Menon N K and Krivanek O L 2002 Microsc. Microanal. 8 203–15
[36] Pedregosa F et al 2011 J. Mach. Learn. Res. 12 2825–30
[37] de la Peña F et al 2019 Hyperspy v1.5.2: https://doi.org/10.5281/zenodo.3396791