

## ARTICLE OPEN



# Unsupervised clustering for identifying spatial inhomogeneity on local electronic structures

Hideaki Iwasawa<sup>1,2,3</sup>✉, Tetsuro Ueno<sup>1,2</sup>, Takahiko Masui<sup>4</sup> and Setsuko Tajima<sup>5</sup>

Spatial inhomogeneity on the electronic structure is one of the vital keys to provide a better understanding of the emergent quantum phenomenon. Given the recent developments on spatially resolved ARPES (ARPES: angle-resolved photoemission spectroscopy), the information on the spatial inhomogeneity on the local electronic structure is now accessible. However, the next challenge becomes apparent as the conventional analysis encounters difficulty handling a large volume of a spatial mapping dataset, typically generated in the spatially resolved ARPES experiments. Here, we propose a machine-learning-based approach using unsupervised clustering algorithms (K-means and fuzzy-c-means) to examine the spatial mapping dataset. Our analysis methods enable automated categorization of the spatial mapping dataset with a much-reduced human intervention and workload, thereby allowing quick identification and visualization of the spatial inhomogeneity on the local electronic structures.

*npj Quantum Materials* (2022)7:24; <https://doi.org/10.1038/s41535-021-00407-5>

## INTRODUCTION

Not uncommonly, the emergence of the quantum phenomenon accompanies dominant states that are not spatially homogeneous<sup>1</sup>. This situation is generally originated from the cooperative interplay among internal degrees of freedom—charge, spin, orbital, and lattice. As ubiquitously observed in strongly correlated materials, such complexity leads to forming local structures with various sizes and scales (self-organization). For a fundamental understanding of the physical properties of the quantum materials, it is thus required to unravel the interconnection between the electronic and structural inhomogeneity.

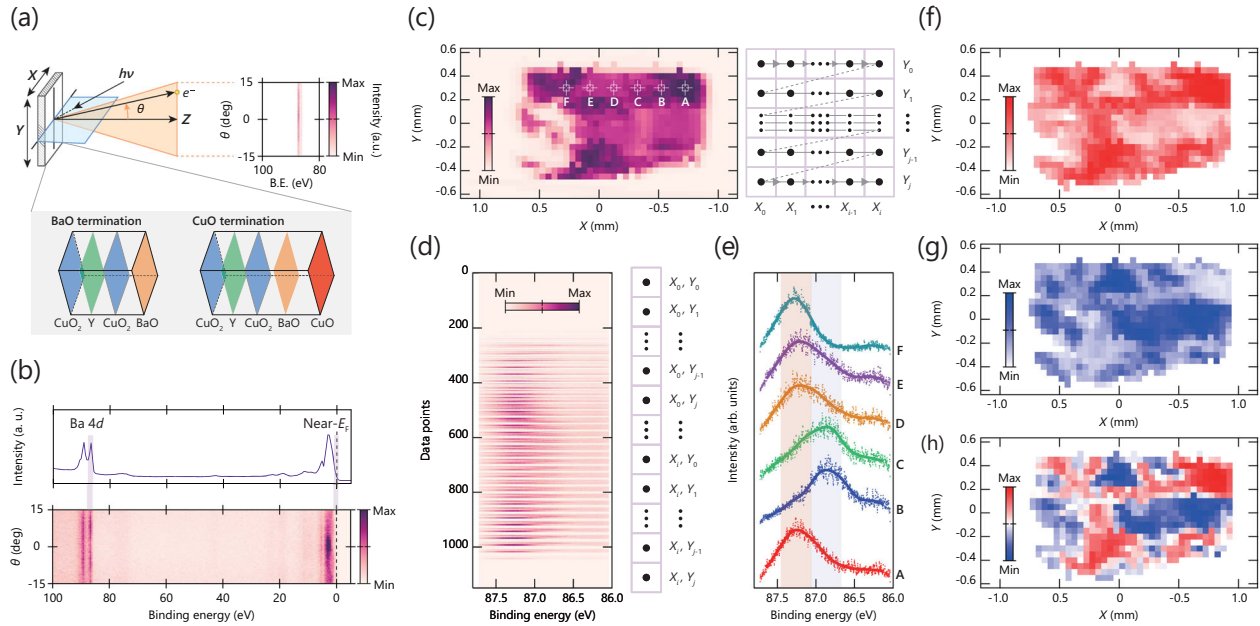
Angle-resolved photoemission spectroscopy (ARPES) is widely recognized as a representative tool for studying the electronic structures of quantum materials<sup>2–5</sup>. However, the local electronic structures have been elusive by ARPES. This is because the spatial resolution of ARPES has been relatively poor, typically in milli- to sub-milli-meter scale, as conventional ARPES systems pursued higher energy and momentum resolution. On the other hand, the situation has been changed these days with the development of spatially resolved ARPES, incorporating the advanced micro-/nano-focusing optics<sup>6–8</sup>. Indeed, the spatially resolved ARPES enabled probing the local electronic structures at micro-nano order length scales: the inhomogeneous phase transitions (metal-insulator transition) in transition metal oxides (vanadate<sup>9</sup>, manganite<sup>10</sup>), the termination-dependent electronic and chemical structures in Y-based high- $T_c$  cuprates<sup>11,12</sup>, the twinned domains picked from unstrained  $\text{BaFe}_2\text{As}_2$ <sup>13</sup>, the weak topological insulator states in Bismuth compounds<sup>14,15</sup>, and the heterostructures of a wide variety of two-dimensional materials<sup>16–19</sup>.

As in the typical and practical procedure of the spatially resolved ARPES experiments, these successful ARPES (=momentum-resolved) observations were typically performed only at limited position(s), selected based on the real-space mapping dataset prior. Subsequent ARPES data such as band-structures or Fermi surface are utilized as representatives of spatial electronic inhomogeneity for understanding the physical properties of the system. Therefore,

characterizing the spatial mapping dataset and selecting areas/points of interest is crucially important in spatially resolved ARPES experiments. In the conventional analysis of the spatial-mapping dataset, spectral feature extraction was mostly performed by integrating ARPES spectra with some energy and momentum windows<sup>11–13</sup> or fitting one-dimensional curves sliced from ARPES spectra<sup>20,21</sup>. These analyses indeed provided reasonable results because they were performed based on the researcher's experience and knowledge as well as the close check with eyes. Paradoxically, however, such conventional analyses inevitably require human intervention, thereby, arbitrariness and workload. In addition, conventional data analyses are getting more difficult because a spatial mapping dataset is usually generated in a vast volume, as supported by advances in automated instrumentation and data acquisition<sup>22</sup>. Hence, it is highly desired to develop a radically different approach that enables handling large and complex experimental data with minimal human intervention and workload. To this end, machine learning is the most promising approach, given the success of machine learning in materials science<sup>23–25</sup>.

Here, we propose the unsupervised clustering approach designed to recognize different types of local electronic structures in the spatial mapping dataset. The goals are to automatically categorize the spatial mapping dataset, visualize spatial evolutions of the local electronic structures, and identify each of the better locations representing each domain. Along this line, two representative clustering algorithms, K-means and fuzzy-c-means, were used as unsupervised learning. We first overview the spatially resolved ARPES experiment and data pre-processing. Subsequently, the conventional analysis, the K-means clustering, and the fuzzy-c-means clustering were applied to the simple spatial mapping dataset from a core level, mainly composed of a single peak. We demonstrate that both unsupervised clustering methods categorize the spatial mapping dataset into certain groups effectively and visualize the spatial evolution of the spatial inhomogeneity, with a much-reduced workload compared to the conventional analysis method. The advantages and disadvantages of the two types of

<sup>1</sup>Synchrotron Radiation Research Center, National Institutes for Quantum Science and Technology, Hyogo 679-5148, Japan. <sup>2</sup>QST Advanced Study Laboratory, National Institutes for Quantum Science and Technology, Chiba 263-8555, Japan. <sup>3</sup>Hiroshima Synchrotron Radiation Center, Hiroshima University, Hiroshima 739-0046, Japan. <sup>4</sup>Department of Physics, Kindai University, Higashi-Osaka 577-8502, Japan. <sup>5</sup>Department of Physics, Osaka University, Osaka 560-0043, Japan. ✉email: [iwasawa.hideaki@qst.go.jp](mailto:iwasawa.hideaki@qst.go.jp)



**Fig. 1** General overview, data-preprocessing, and conventional analysis in spatially resolved ARPES experiment. **a** Schematic drawings of the layout measuring a snapshot ARPES image and the two-types of surface terminations of  $\text{YBa}_2\text{Cu}_3\text{O}_{7-\delta}$ . **b** ARPES image and its angle-integrated EDC in wide-energy range. **c** Total ARPES intensity mapping in the real space and schematic illustration of acquisition flow. **d** Integrated EDCs after data pre-processing and its data indexes. **e** Exemplary EDCs extracted from several points, as labeled (A–F) in (c). **f, g** Partial ARPES intensity mapping obtained by a limited integration energy window as indicated by red and blue shaded area in (e), respectively. **h** Difference intensity mapping between (f) and (g), where each of pixel intensity is normalized by the averaged intensity of each map before taking the difference. Panels (b), (c) and (f–h) are based on ref. <sup>11</sup> with permission, copyright American Physical Society 2018.

unsupervised clustering and conventional analysis methods will also be described. Finally, we also demonstrate the versatility and extensibility of the clustering analysis methods by representing those applications to the more complex spatial mapping dataset from the electronic states near the Fermi level ( $E_F$ ), where several electronic bands disperse in energy and momentum.

## RESULTS

### Spatially resolved ARPES: overview and data pre-processing

The experimental setup of the spatially resolved ARPES experiment is essentially the same as the scanning photoemission microscopy (SPEM)<sup>8</sup>, as illustrated in Fig. 1a. Two-dimensional ARPES image  $I(E_k, \theta)$  was measured as a function of kinetic energy ( $E_k$ ) and emission angle ( $\theta$ ) at a local area irradiated by focused incident light. Note that the energy unit is converted to the binding energy  $E_B = hv - \phi - E_k$ , while the angle unit is not converted to momentum in this paper. The data acquisition was made by a snapshot (fixed) mode, which measures a detector image for a short dwell time ( $\sim$ sub-sec) and is advantageous for the spatially resolved ARPES experiment. In this work, we analyzed two different types of spatial mapping datasets from Y-based high- $T_c$  cuprate  $\text{YBa}_2\text{Cu}_3\text{O}_{7-\delta}$  (see Methods). One data type focuses on the Ba  $4d_{5/2}$  core level composed of simpler lineshape, while another one is composed of more complex electronic states near the Fermi level (near- $E_F$ ). These energy levels are indicated by the shaded area in the angle-integrated energy distribution curve (EDC) (solid line in Fig. 1b). We will first focus on the Ba  $4d_{5/2}$  core level, which is sensitive to exposed surface terminations due to BaO- and CuO-layer, as schematically shown in Fig. 1a<sup>11</sup>, and its simpler lineshape is more suitable to verify analysis methods.

As shown in the right panel in Fig. 1c, the spatially resolved ARPES mapping dataset  $I_{\text{ARPES}}(E_B, \theta, x_i, y_j)$  can be obtained by sequentially measuring the series of a 2D ARPES image  $I_{\text{ARPES}}(E_B, \theta)$  as functions of the spatial coordinates  $x_i$  and  $y_j$  along the  $X$  and  $Y$  axes, respectively. Here,  $i(j)$  is the integer ranging from 1 to  $n_x(n_y)$ ,

corresponding to the number of acquisition points along the  $X$  ( $Y$ ) axis. A most simple way to visualize the spatial distribution of ARPES images is to convert a 2D ARPES image to an intensity at each point by taking a full integration over energy and angle dimensions, namely,  $I(x_i, y_j) = \sum_{E_B, \theta} I_{\text{ARPES}}(E_B, \theta, x_i, y_j)$ . As shown in Fig. 1c, the resulting spatial image captures intensity inhomogeneity reflecting electronic modulations on the surface (see the optical microscope image in Supplementary Fig. 1). However, the limited information on the integrated intensity is not enough to identify the surface terminations. To extend information while keeping ease of analysis and data handling, we simplified the 2D ARPES image into an integrated-EDC (iEDC) obtained by summing up the whole points in the angle dimension [ $I_{\text{iEDC}}(E_B) = \sum_{\theta} I(E_B, \theta)$ ]. Consequently, the spatial mapping dataset  $I_{\text{ARPES}}(E_B, \theta, x_i, y_j)$  was simplified into iEDCs,  $I_{\text{iEDC}}(E_B, p)$  (Fig. 1d), where the 2D spatial coordinates  $(x_i, y_j)$  are flattened into the 1D array with the integer index of  $p$  ranging from 1 to  $n$  ( $=n_x \times n_y$ ). Note that all the clustering analyses were performed utilizing iEDCs throughout the paper, while the integrated angle distribution curves (iADCs) can also be used (see Supplementary Fig. 2 for data slicing). We found that similar analyses on iADCs lead to the essentially same conclusions, as confirmed by Supplementary Figs. 6 and 7.

### Conventional analysis

The electronic modulations on the surface are exemplarily shown in Fig. 1e, where we picked up several raw EDCs at spatial coordinates (labeled as A–F in Fig. 1c). At first sight, one can roughly categorize them into two groups in terms of peak energy. The spatial distribution of each group can then be visualized by employing the narrow integration windows (red and blue shaded area in Fig. 1e) indicating almost opposite distributions (Fig. 1f, g). Further, by taking the intensity difference between these two, one can see the spatial distribution of two groups on the cleaved surface, as shown in Fig. 1h. Note that the deeper and shallower peak position indicates that those measured positions are dominantly terminated by a CuO and BaO layer, respectively<sup>11</sup>.

The analysis described here can be regarded as a kind of supervised classification. Namely, it is performed based on the assumption of the presence of three classes ( $C_{0-2}$ ) in advance: we labeled CuO- and BaO-terminated surface ( $C_1$  and  $C_2$ ) at which the red and blue integration window returns the highest intensity, and outside the sample ( $C_0$ ) for negligible intensity. Such analysis based on the spectrum integration has been usually used and worked effectively to characterize the surface morphology in spatially resolved ARPES experiments<sup>6-8</sup>. Peak fitting analysis is another commonly used data-categorization technique. It has an advantage to the spectral integration that it can handle more spectral information such as peak amplitude, center, and width<sup>20</sup>. However, the peak fitting typically requires heavier time-cost and workloads, and thereby, it is inferior to the spectral integration in terms of versatility. Hence, we employed the spectral integration as a conventional analysis in this work. For reference, we also performed the peak fitting analysis for quantitative comparison between different spatial-mapping data analysis techniques, as discussed in Supplementary Note 5 and Supplementary Fig. 5. Indeed, the peak fitting analysis showed the quantitative agreement between the conventional and clustering analysis methods, indicating the strength of the clustering analysis having much-reduced workloads than the conventional analysis.

### Unsupervised clustering

In the previous section, we reviewed how the conventional analysis enabled investigating the spatial inhomogeneity on the local electronic structures. However, such analysis requires visual and manual data confirmation for extracting spectral features. In general, manual feature extraction becomes more difficult and time-consuming for the larger dataset and thus not practical for distinguishing minor differences in the spectral features. More problematically, the goodness of analysis highly depends on one's knowledge and/or experience in the conventional analysis. As already mentioned, the procedure of conventional analysis can be regarded as a kind of data classification in supervised learning. In other words, the output data is labeled and classified by following the pre-labeled features extracted manually. However, no matter what results are brought by such a conventional analysis, any analysis methods accompanying human intervention more or less generate arbitrariness and workload inevitably. It is, therefore, necessary to develop an unsupervised analysis method to classify the large dataset and understand data characteristics without prior knowledge and/or experience. Unsupervised clustering is suitable to achieve this aim and is divided into two types, hard- and soft-clustering. In the following, we apply K-means<sup>26</sup> and fuzzy-c-means<sup>27</sup> clustering methods, which are symbolic hard- and soft-clustering algorithms, respectively. Note that the general descriptions of unsupervised learning, including two clustering methods, are given in "Methods".

**K-means clustering.** Next, we present the clustering analysis to categorize the dataset into designated groups using the K-means method, which is a computationally light method and one of the most commonly used unsupervised learning algorithms<sup>26</sup>. The K-means algorithm finds a group where the data points have high similarity between them against the maximum number of clusters ( $n_k$ ); the  $n_k$  is an input hyperparameter and should not be determined beforehand. To estimate the optimal value of  $n_k$ , we examined three evaluation approaches (the elbow method, the Silhouette score, and the gap statistic), whose explanations are described in "Methods". As seen from these results shown in Fig. 2a–c, their estimations ( $n_k = 3, 2,$  and  $5$ ), indicated by the red line, are different, meaning that the absolute determination of  $n_k$  is quite difficult (see Methods). We thus do not go into detail about this issue in this work. While the obtained gap statistics suggests that  $n_k^{\text{opt}} = 5$ , we alternatively adopted a higher value of  $n_k = 8$  (blue dashed line in Fig. 2c) for further analyses, to examine

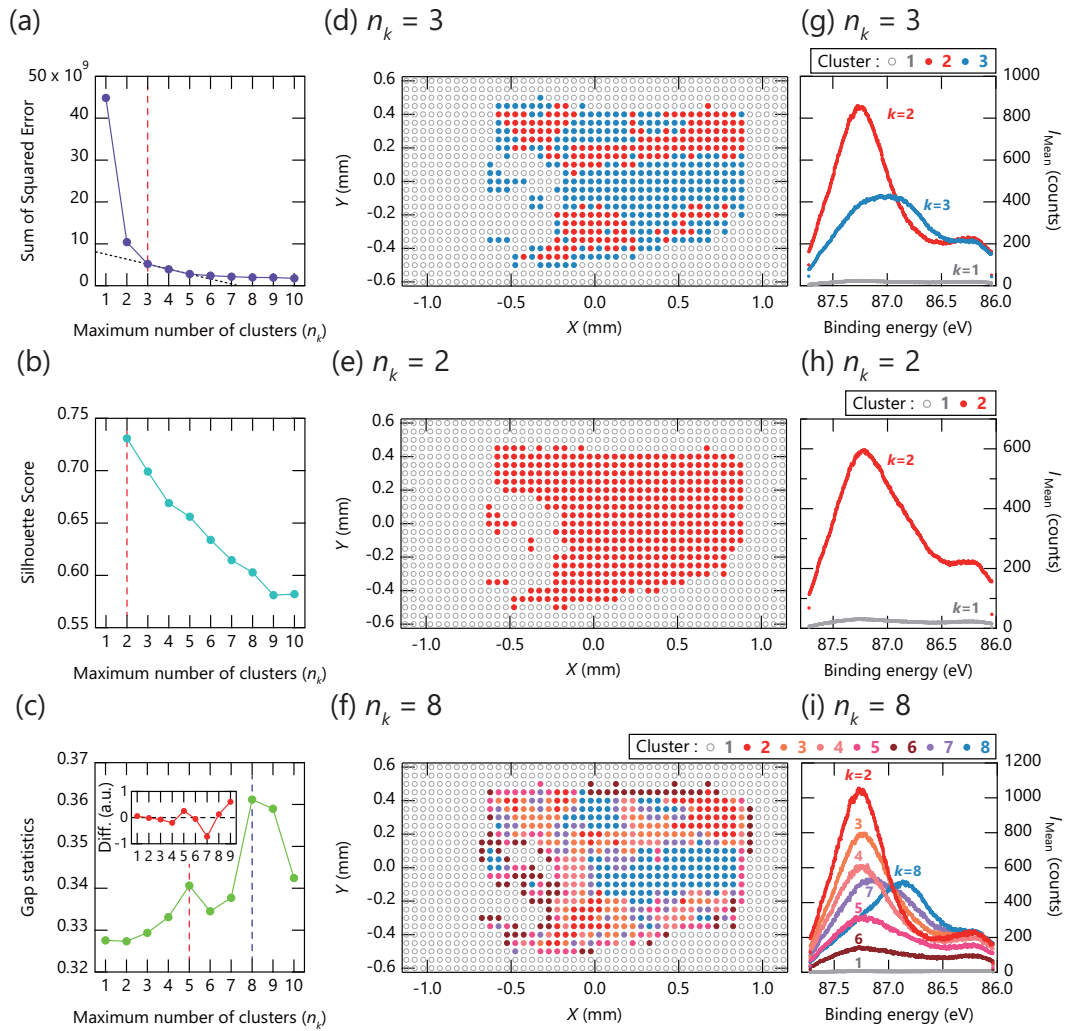
the  $n_k$ -dependence clearly. In the following, we thus examine the  $n_k$ -dependence on the clustering results for  $n_k = 3, 2,$  and  $8$ , which help understand the characteristics of the spatial dataset.

Figure 2d–i visualizes the results of K-means clustering for different  $n_k$ -values. Figure 2d–f and (g–i) shows the spatial distribution of each cluster and the mean-EDC averaging all the belonging members in a cluster, respectively. First, we found that the absolute value of  $n_k$  should be greater than two by comparing the results between  $n_k = 2$  and  $3$  (Fig. 2h, g) as the mean-EDC assigned for the cluster 3 appears, differently from the existing mean-EDCs for  $n_k = 2$ . In addition, the spatial distribution of clusters for  $n_k = 3$  is similar to conventional analysis, and the peak position of mean-EDC is different between clusters 2 (red) and 3 (blue). However, the peak width for cluster 2 (Fig. 2g) is much broader compared with the EDC extracted from the pixel (B and C in Fig. 1e), indicating the mixing of two clusters (2 and 3). As seen in the mean-EDCs in Fig. 2i, such mixture is, of course, reduced by increasing  $n_k$ . Indeed, the peak positions of the mean-EDCs ( $k = 2, 8$ ) are quantitatively consistent with the conventional analysis, as discussed in Supplementary Note 5 and Supplementary Fig. 5.

Also, the eight mean-EDCs can be classified into four groups by similarity. Two clusters ( $k = 1$  and  $8$ ) constitute an independent cluster, representing the outside the sample ( $C_0$ ) and one domain ( $C_1$ ). Whereas, five clusters ( $k = 2-6$ ) show the similar shape of the mean-EDCs with almost the same peak energy, confirming the presence of the second domain ( $C_2$ ). The remaining cluster ( $k = 7$ ) seems to be the admixture of two domains  $C_1$  and  $C_2$ . Consequently, while we ignored the absolute determination of the  $n_k$ , the  $n_k$ -dependence of the K-means results can lead to a reasonable estimation of  $n_k$ ; three classes in this case. Note that this estimation is also supported by a different approach using principal component analysis, as discussed in Supplementary Note 3 and Supplementary Fig. 3. On the other hand, however, the K-means clustering analysis still left arbitrariness in determining a better position from the cluster members in each cluster.

**Fuzzy-c-means clustering.** The K-means method is a famous hard clustering algorithm whereby the data items are classified into K clusters such that each item only belongs to one cluster. In other words, each item's attribution probability  $\rho_i$  at each data point ( $i = 1, \dots, n$ ) is given only 1 or 0 in the K-means method. Due to this property, it is difficult to judge which item (position) is most representing the cluster. Also, the K-means easily and inevitably lead to admixture between different clusters. In contrast, we will present that these disadvantages can be overcome by using the soft clustering method, in which the attribution probability ( $\rho_i^k$ ) ranges from 0 to 1, including decimal numbers, for each cluster ( $k = 1, \dots, n_k$ ). In this work, as the representative of the soft clustering algorithm, we employ the fuzzy-c-means method<sup>27</sup> that requires two hyperparameters: the fuzzifier  $m$  besides the maximum number of clusters  $n_k$ . Although the goodness of the fuzzy-c-means clustering is often evaluated by a fuzzy partition coefficient (FPC), we did not find reasonable  $n_k$ - and  $m$ -dependence on the FPC (see Supplementary Fig. 4 in Supplementary Information). We thus employ  $n_k = 3$ , as suggested above by the K-means clustering, in the fuzzy-c-means clustering analysis. On the other hand,  $m = 1.5$  was chosen as a moderate value from  $m$ -dependence on the clustering results, as shown below.

Figure 3a, b shows the results of the fuzzy-c-means clustering with  $n_k = 3$  and  $m = 1.5$ . The probability density  $\rho_i^k$  for each cluster ( $k = 1, \dots, n_k$ ) and the data points ( $i = 1, \dots, n$ ) is plotted in Fig. 3a, where one can confirm that  $\rho_i^k$  takes 0 to 1 and  $\sum_{k=1}^{n_k} \rho_i^k = 1$  and  $\sum_{i=1}^n \sum_{k=1}^{n_k} \rho_i^k = n$ . Since the attribution probability expresses the strength of the belonging of the item into the cluster, it is possible to further classify the cluster members by  $\rho_i^k$  in the case of the soft clustering. This is well demonstrated by the mean-EDC within each cluster (Fig. 3b), where the



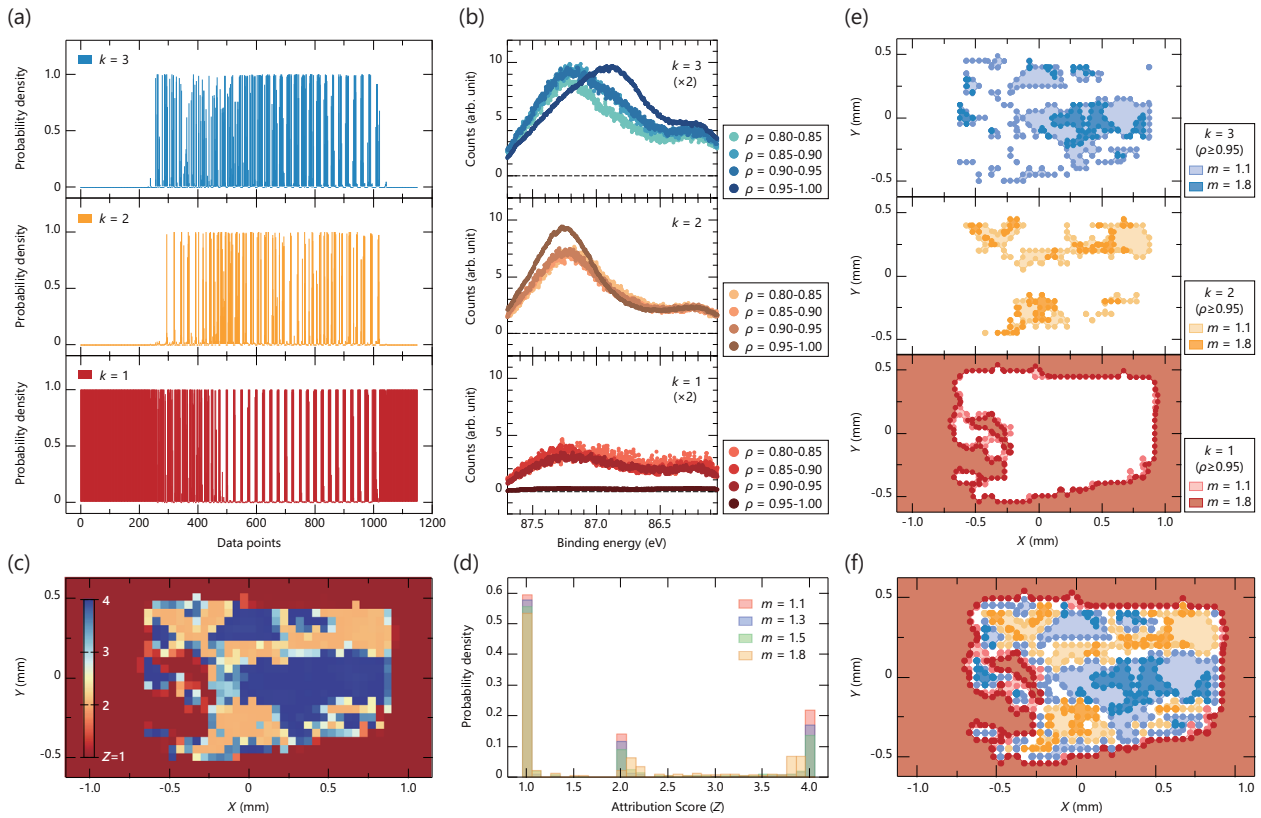
**Fig. 2 Unsupervised clustering by the K-means algorithm.** **a–c** The maximum number of clusters ( $n_k$ ) dependence on **(a)** the sum of squared error, **b** silhouette score, and **(c)** gap statistic, where red dashed line indicates the estimated optimum number of clusters  $n_k^{\text{opt}}$ . The inset of **(c)** shows the gap difference (Diff.), where red and blue dashed lines represent the estimated  $n_k^{\text{opt}}$  and the adopted trial value ( $n_k = 8$ ), respectively. **d–f** K-means clustering results for  $n_k = 3$ , 2, and 8, respectively: **d–f** Spatial distribution of each of clusters for different number of clusters, **g–i** mean-EDCs obtained by averaging the cluster members in each cluster.

averaging of objects is divided into four levels based on the attribution probability;  $\rho_i^k = 0.80–0.85$ ,  $0.85–0.90$ ,  $0.90–0.95$ , and  $0.95–1.00$ . Such spectral variation seen in the mean-EDCs is similar to what is observed by the K-means clustering with the higher maximum number of clusters ( $n_k = 8$ ), and the peak positions of two clusters ( $k = 2, 3$ ) for high  $\rho_i^k$  show quantitative agreement (see Supplementary Fig. 5). In this way, the fuzzy-c-means clustering allows us to find more intrinsic spectra consisted of a single cluster.

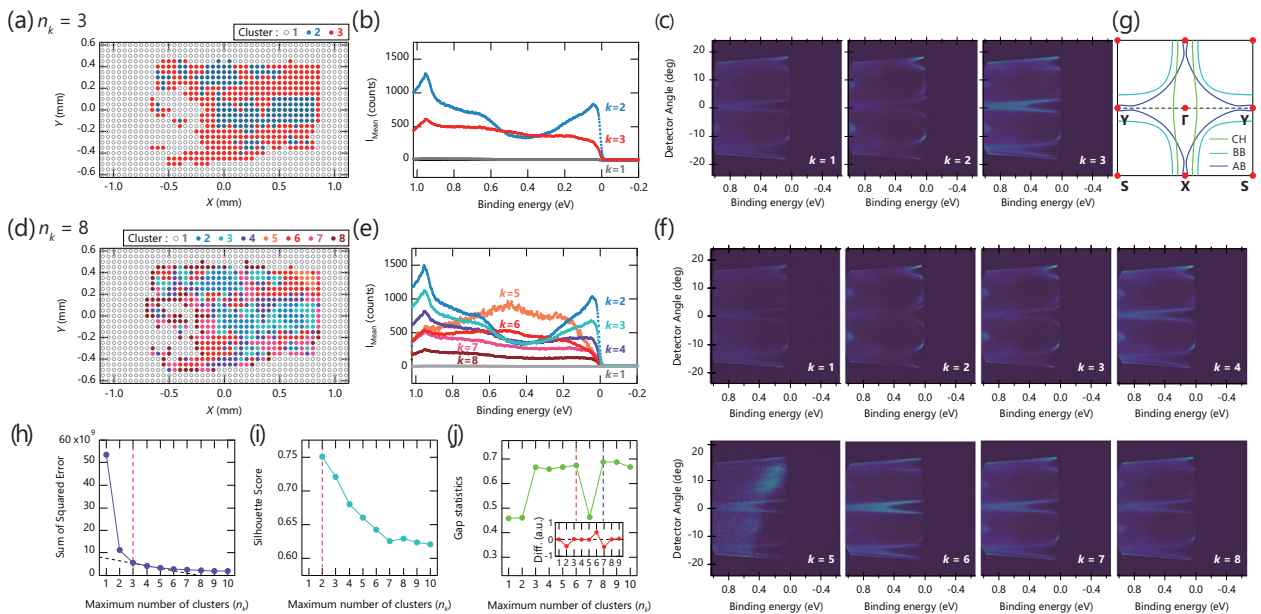
On the other hand, the spatial visualization of a representing cluster at each point is not straightforward compared with the K-means clustering because each point has finite probability density against multiple clusters in the fuzzy-c-mean clustering. For the spatial visualization of the fuzzy-c-means results, we developed an attribution score ( $Z$ ) calculated by the binary-coded decimal ( $Z_i = \sum_{k=1}^{n_k} \rho_i^k \cdot 2^{k-1}$ ) and also employed a contour plot. In the present case, the  $Z$  score was found to be not greater than 4. As indicated in Table 1, the attribution score of 1, 2, and 4 consists of the single cluster, while  $Z = 3$  and the other non-integers indicate the admixture of clusters. The spatial  $Z$ -distribution shown in Fig. 3c is similar to that is obtained by the K-means clustering with  $n_k = 3$ , while the image contrast illustrates a higher-purity region in each

cluster. Moreover, the probability distribution of  $Z$  is helpful to understand the influence of the fuzzifier ( $m$ ) on the clustering results. As shown in the histogram of  $Z$  (Fig. 3d), main components by three scores ( $Z = 1, 2$ , and 4) decrease as  $m$  increases. Conversely, the remaining cluster members, even for higher  $m$ , represent the higher-purity region in the cluster. This trend can be more clearly visualized in the contour plots shown in Fig. 3e, f, where the cluster region is considerably shrunk with  $m = 1.8$ . This means that the boundary between different clusters becomes more unclear. Again, the remaining small area, even in higher  $m$ -value, should be manifesting the higher-purity region. These results indicate that the fuzzy-c-means clustering is an effective tool to find suitable areas among spatial-mapping data for further ARPES experiments in the  $k$ -space.

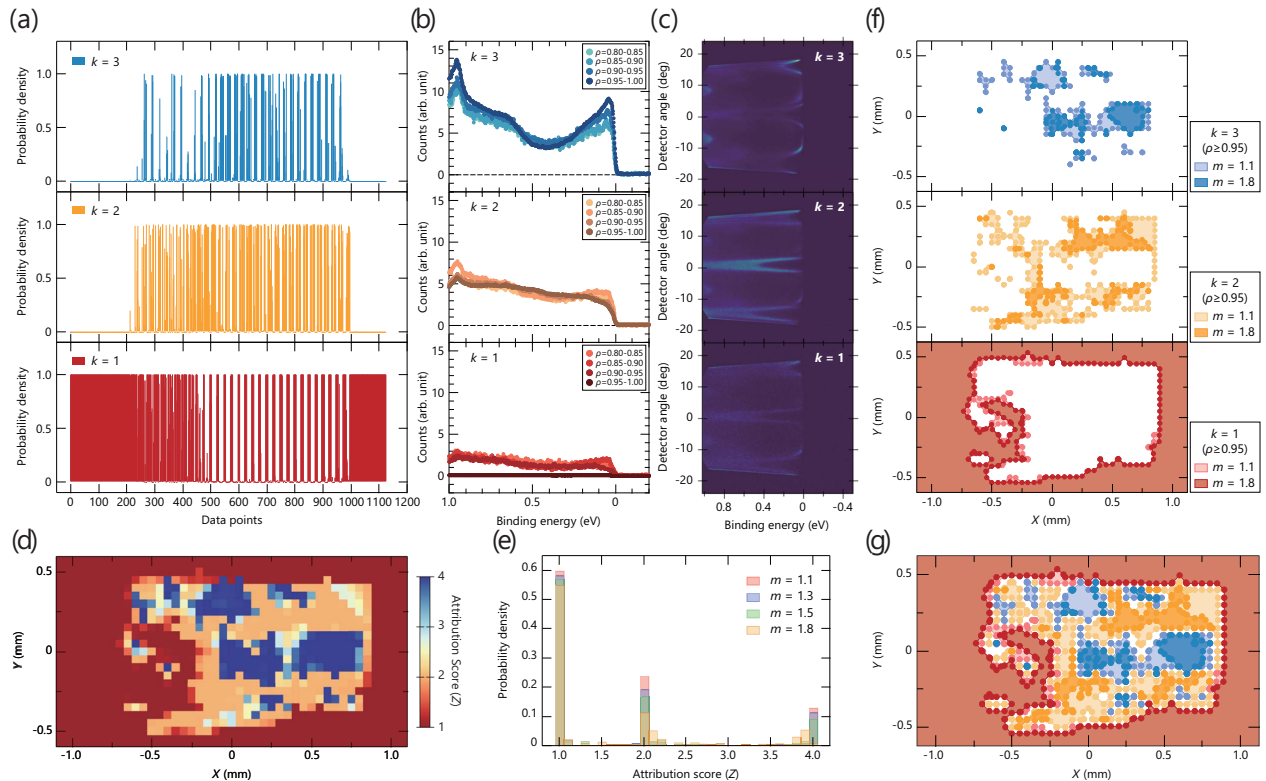
**Versatility and extensibility of unsupervised clustering.** So far, we have demonstrated that the clustering analysis is useful to categorize the spatial mapping dataset. However, there exists a concern whether the present analysis is applicable to the more complex dataset as the core level spectra are relatively simple. We thus applied the clustering analysis to a more complex spatial-mapping dataset from the near- $E_F$  electronic states, where energy bands are more dispersive in energy and momentum.



**Fig. 3 Unsupervised clustering by the fuzzy-c-means algorithm.** **a** Probability density ( $\rho^k$ ) distribution as a function of the acquisition point for each cluster. **b** The mean-EDCs within each cluster, where the averaging range is limited by a probability window as indicated in the figure annotation. **c** Spatial distribution of the attribution score ( $Z$ ). **d** Probability density distribution as a function of the  $Z$  for several  $m$ 's. **e, f** Contour plots of the spatial distribution of high probability density region ( $\rho > 0.95$ ) for  $m = 1.1$  and  $1.8$ , displayed by **(e)** each cluster and **(f)** all clusters.



**Fig. 4 K-means clustering applied for complex near- $E_F$  dataset.** **a-f** The K-means clustering results for  $n_k = 3$  and  $8$ , respectively: **a, d** spatial distribution of each of clusters for different number of clusters, **b, e** the EDCs obtained by averaging all the belonging cluster members in each cluster, and **c, f** the ARPES images obtained by averaging the cluster members in each cluster. The measured  $k$ -location is near the  $Y$ - $T$ - $Y$  line as indicated in the schematic Fermi surface in **(g)**. **h-j** The  $n_k$ -dependence on **(h)** the sum of squared error, **(i)** silhouette score, and **(j)** gap statistic, where the red line indicates the estimated optimum number of clusters  $n_k^{\text{opt}}$ . In **(j)**, the blue line represents the adopted trial value ( $n_k = 8$ ), and the inset shows the gap difference (Diff.).



**Fig. 5 Fuzzy-c-means clustering applied for complex near- $E_F$  dataset.** **a** Probability density ( $\rho_i^k$ ) as a function of the acquisition point for each cluster. **b, c** The mean of EDCs and ARPES images within each cluster, respectively. The averaging range is limited by a probability window as indicated in the figure annotation for the mean-EDC while the ARPES image is averaged within  $0.95 \leq \rho_i^k \leq 1.00$ . **d** Spatial distribution of the attribution score ( $Z$ ). **e** Histogram of the probability density distribution as a function of the  $Z$  for several  $m$ 's. **f, g** Contour plots of the spatial distribution of high probability density region ( $\rho > 0.95$ ) for different  $m$ 's, displayed for **(f)** each cluster and **(g)** all clusters.

Figure 4 shows the K-means clustering results with  $n_k = 3$  and 8, which were selected for the direct comparison to the clustering results on the core-level mapping dataset, shown in Fig. 2. Note that we excluded  $n_k = 2$  as it is evident that the clustering with  $n_k = 2$  just categorizes inside and outside the sample. Just for reference, the evaluations on  $n_k$ -values are shown in Fig. 4h–j, where (h) the sum of squared error (SSE), (i) silhouette score, and (j) gap statistic suggested the optimum  $n_k$ -value as 3, 2, and 6, respectively.

By comparing the clustering results for  $n_k = 3$  between the near- $E_F$  and core-level datasets shown in Figs. 4a–c and 2d, respectively, one can notice that the clustering analysis worked regardless of the complexity of the dataset. As was the case for the core-level dataset, it is easy to identify the main components of each cluster: outside the sample ( $k = 1$ ), BaO-termination ( $k = 2$ ), and CuO-termination ( $k = 3$ ). Looking more closely at the spatial distribution of cluster 2 (Fig. 4a), its spatial region is much smaller than one obtained from the clustering analysis for the core-level dataset with  $n_k = 3$  (Fig. 2d). Furthermore, it is somewhat similar to the clustering results for the core-level dataset with  $n_k = 8$  in Fig. 2f. Indeed, the admixture between clusters 2 and 3 is not apparent in the mean-EDCs ( $k = 2$  and 3) in Fig. 4b, compared with the broader mean-EDC ( $k = 2$ ) obtained by the core-level analysis in Fig. 2d. These results would indicate that the K-means analysis on the near- $E_F$  dataset brought favorable clustering results, probably due to the presence of more spectral features.

Then, let us focus on the K-means clustering results from the near- $E_F$  spatial-mapping dataset using the higher  $n_k$ -value ( $n_k = 8$ ). As seen in Fig. 4e, the mean-EDC can be roughly categorized into three types: outside the sample ( $k = 1$ ), the BaO-termination showing higher intensity with a peak near the Fermi level ( $k = 2$ –4), and the CuO-termination showing a hump structure around 0.5 eV ( $k = 5$ –8). The degree of admixture between different clusters is hardly distinguishable from the mean-EDC but can be somehow conjectured from

spectral features recognized in mean-ARPES images in Fig. 4f. For instance, a characteristic V-shape band disperses in a narrow energy region ( $\sim 0.2$  eV) centered around  $\theta = +15^\circ$ . This band is most clearly seen in the BaO-termination ( $k = 2$ ) then becomes weaker for  $k = 3$  and 4, while it is faintly but surely present for  $k = 7$  and 8. These observations are thus indicating that the CuO-termination-dominant clusters ( $k = 7$  and 8) contain signals from the BaO-termination. Similarly, the CuO-chain bands representing the CuO-termination were observed, centered around  $\theta = 0, \pm 15^\circ$  with wide energy dispersion ( $\sim 0.8$  eV). These bands are most clearly observed for  $k = 6$ , then become weaker for  $k = 7$  and 8, and extra background is merged in the case of  $k = 5$ . Note that the background is apparent in the mean-ARPES image, though it is hardly expected from the mean-EDC. These results consistently suggest that more spectral features help to expect the essential number of clusters and find higher-purity areas in the spatial-mapping dataset.

We next present the application of the fuzzy-c-means clustering to the near- $E_F$  spatial-mapping dataset. Figure 5a–d is the results of the fuzzy-c-means clustering with  $n_k = 3$  and  $m = 1.5$ . Similar to the application to the core-level dataset, the categorization of cluster members by the probability density yielded similar classifications by the K-means with the higher  $n_k$ -value. Indeed, obtained mean-EDCs for  $\rho_i^k = 0.95$ –1.00 (Fig. 5b) and mean-ARPES images (Fig. 5c) are similar to those obtained by the K-means clustering (Fig. 4e, f). Here,  $k = 1$ –3 in the fuzzy-c-means corresponds to  $k = 1, 2,$  and 6 in the K-means. As for the spatial distribution of the attribution score  $Z$  and  $m$ -dependence on the histogram of the  $Z$ -score, comparable results have been obtained in the fuzzy-c-means clustering on core-level (Fig. 3c, d) and near- $E_F$  (Fig. 5d, e) datasets. This tendency is the same for the spatial distribution of high- $\rho_i^k$  region in the contour plots in Fig. 5f, g, which are comparable to Fig. 3e, f. On the other hand, one might recognize the slight difference in the spatial distributions of

probability density between them. This dissimilarity is probably originated in differences in spectral features of core-level and near- $E_F$  datasets. Since two types of datasets were measured from the same surface, the clustering accuracy may be improved by clustering based on both datasets (mutual information), which we leave for future work.

Finally, we point out that finding a clustering algorithm suitable for spatial-mapping ARPES dataset still needs to be pursued, as this work treated representative K-means and fuzzy-c-means clustering algorithms introductory. Indeed, there is a suite of choices for the clustering algorithms: for instance, derived algorithms such as K-means++<sup>28</sup> or generalized fuzzy-c-means<sup>29</sup>, and algorithms with different concepts such as agglomerative clustering<sup>30</sup> or density-based spatial clustering of applications with noise (DBSCAN)<sup>31</sup>. Incorporating such an algorithm or developing existing algorithms may improve the clustering accuracy. Despite that, it should also be noted that the purpose of the analysis is likely more important than trying various algorithms. The results for principal features must be rather robust, whichever algorithm is used. In contrast, the clustering accuracy should be more sensitive to the properties of the employed algorithm when features of interest are weak or merged, as the boundary of clusters.

## DISCUSSION

Recent progress on the spatial resolution in the ARPES experiment enabled us to investigate spatial inhomogeneity on the local electronic structures, resulting in successful observations on various quantum materials<sup>6–8</sup>. But meanwhile, the advantage of spatially resolved ARPES, namely, significantly increased spatial degrees of freedom, also brought us arbitrariness in identifying the spatial evolution of the electronic structures as well as selecting an appropriate measurement position from a large volume of spatial mapping dataset. In this regard, we applied machine learning in order to establish an effective analysis method for the spatial mapping dataset.

We used two unsupervised clustering algorithms, K-means<sup>26</sup> and fuzzy-c-means<sup>27</sup> methods, as the representative hard- and soft-clustering, respectively. Both algorithms require the input hyperparameter(s), the number of clusters  $n_k$  for K-means, and the  $n_k$  and fuzzifier  $m$  for fuzzy-c-means, which should be given appropriately to obtain proper clustering results. Apart from the difficulty in the absolute determination of these parameters, we showed that the parameter dependence on the clustering results enables reasonable estimations on these hyperparameters. In the first step, we used the  $n_k$ -dependence in the K-means to infer the  $n_k$ , which is also supported by the principal component analysis. Subsequently, by using the estimated  $n_k$ , the  $m$ -dependence in the fuzzy-c-means enabled visualizing the high-purity region for each cluster. We remark that this series of analyses can grasp and classify the overall characteristics of the input spatial-mapping dataset. We also demonstrate that the present clustering analysis essentially works on either simple or complex datasets. We believe that our analysis procedure presented here puts forward a novel and effective analysis methodology for spatially resolved ARPES experiments. Furthermore, the applications of present clustering analysis can be expected to provide benefits on the categorization and visualization of any multidimensional ARPES datasets without prior knowledge.

## METHODS

### Materials and spatially resolved ARPES experiments

High-quality single crystals of optimally doped YBCO ( $\delta = 0.1$ ,  $T_c = 93$  K) were grown by the crystal pulling technique and detwinned by annealing under uniaxial pressure<sup>32</sup>. Micro-ARPES experiments were performed at beamline I05 of the Diamond Light Source<sup>33</sup> using a photon energy of 150 eV at  $\sim 7$  K. All the data were measured by a high-resolution hemispherical electron analyzer (R4000, Scienta) after cleaving the samples in situ in ultrahigh

vacuum better than  $2 \times 10^{-10}$  mbar at  $\sim 8$  K. The energy, angular, and spatial resolution were set to be 8 meV,  $0.2^\circ$ , and  $\sim 60$   $\mu\text{m}$ , respectively.

### Unsupervised clustering: hierarchic and non-hierarchic clusterings

Clustering is unsupervised learning to find the series of grouping in the dataset that maximizes or minimizes a given criterion, evaluating the similarity or dissimilarity of the data points within the same cluster. The dataset consists of  $n$  observables, that is,  $n (=n_x \times n_y)$  ARPES spectra  $I_i = (I_1, I_2, \dots, I_n)$ , measured as functions of  $X$  and  $Y$  coordinates, where  $n_x(n_y)$  is the number of  $X$  ( $Y$ ) measurement points in this work. The clustering model can be divided into two types: hierarchic and non-hierarchic clusterings<sup>34,35</sup>. The hierarchic clustering starts from  $n$  clusters, where each cluster contains only one observable. Then, it iteratively combines the clusters having the highest similarity until all the clusters are merged, thus creating a hierarchical structure. Several linkage criteria have been proposed, such as the nearest-neighbor (single linkage) method<sup>30</sup>, furthest-neighbor (complete linkage)<sup>36</sup>, and Ward's method<sup>37</sup>. However, the hierarchic clusterings are in principle unsuitable for handling a large dataset because the increase of computational cost and more complex hierarchical structure are expected in that case. In contrast, the non-hierarchic clusterings generally require much less computational cost, thus, suitable for handling a large dataset. The non-hierarchic clustering aims to categorize  $n$  observables into the preset value for the number of clusters ( $n_k$ ). As the output, an attribution probability to a cluster ( $\rho$ ) is obtained for each observable. Then, the non-hierarchic clustering can be further divided into two types, hard- and soft-clusterings, depending on  $\rho$ -value permitted: the  $\rho$  only takes 0 or 1 in the hard clustering, while it ranges between 0 and 1, including decimals, in the case of the soft clustering. There are various algorithms proposed for both clustering methods: for instance, partition around medoids (PAM)<sup>38</sup>, K-means<sup>26</sup>, spectral clustering<sup>39</sup>, and EM (expectation-maximization) algorithm based Gaussian mixture model (GMM)<sup>40</sup> for the hard-clustering, and fuzzy-c-means<sup>27</sup>, GMM<sup>40</sup>, probabilistic latent semantic analysis<sup>41</sup>, and non-negative matrix factorization (NMF)<sup>42</sup> for the soft-clustering.

In this work, we used the K-means and fuzzy-c-means algorithms, as they are representative hard- and soft-clustering methods, respectively. All the codes were developed by Python using the scikit-learn package<sup>43</sup> for K-means and the scikit-fuzzy package for fuzzy-c-means clustering, all of which can be found at <https://github.com/h-iwasawa/arpes-clustering>. In addition, a platform using Igor Pro for the K-means clustering is also available. In the following, we will give a brief explanation of these algorithms.

**K-means clustering.** The K-means clustering categorizes the  $n$  observables,  $I_i = (I_1, I_2, \dots, I_n)$ , into the  $n_k$  clusters  $C_k = (C_1, C_2, \dots, C_{n_k})$ , where  $n_k$  is the preset value, and each member must belong to a single cluster. In the K-means algorithm, the grouping is iteratively performed by determining the centroids of the clusters  $C_k = (c_1, c_2, \dots, c_{n_k})$  and by assigning the belonging cluster of each member. The centroids are also regarded as the centers of the gravity of the clusters. The K-means algorithm has various implementations, though we used a standard iterative refinement approach as follows.

Step 1. Specify  $n_k$  and randomly assign  $c_k$ .

Step 2. Calculate the distances between a member and all the centroids of the clusters, and then assign a belonging cluster  $C_k$  for each member, whose centroid provides the nearest neighbor for a member.

Step 3. Calculate the new centroids  $c_k$  based on the updated members in Step 2.

Step 4. Repeat Steps 2 and 3 until the assignment in Step 2 no longer changes.

This algorithm aims at minimizing the evaluating criterion, or objective function ( $W$ ), the sum of squared distances (Euclidean distance) used in this work, given by

$$W(I_i, c_k) = \sum_{k=1}^{n_k} \sum_{i \in C_k} \|I_i - c_k\|^2. \quad (1)$$

though different distance metrics  $d(I_i, c_k)$  can also be applicable, and Eq. (1) can then be more generalized as

$$W(I_i, c_k) = \sum_{k=1}^{n_k} \sum_{i \in C_k} d(I_i, c_k). \quad (2)$$

It should be noted that the results of the K-means depend on the initial assignment of  $c_k$  and the preset value of  $n_k$ . For the estimation of the optimal number of cluster ( $n_k^{\text{opt}}$ ), the elbow method<sup>44</sup>, silhouette analysis<sup>45</sup>, and the gap statistic method<sup>46</sup> were used in this work, as shown in Fig. 2a–c, respectively. Those explanations are briefly given in the following.

**Table 1.** Attribution score ( $Z$ ) when the attribution probability ( $\rho^k$ ) takes binary value for each cluster ( $k = 1-3$ ) with the case of  $n_k=3$ .

$Z$	$\rho^{k=3}$	$\rho^{k=2}$	$\rho^{k=1}$
0	0	0	0
1	0	0	1
2	0	1	0
3	0	1	1
4	1	0	0
5	1	0	1
6	1	1	0
7	1	1	1

**Elbow method.** The elbow method is a well-known and naive approach. It calculates the SSE, which measures the cohesion of the clusters. As seen in Fig. 2a, the SSE decreases rapidly with increasing  $n_k$  for  $n_k < n_k^{\text{opt}}$ , while it also decreases for  $n_k \geq n_k^{\text{opt}}$  but the reduction rate should be rather gradual. Thus, the SSE curve shows an arm-like behavior with an elbow as a function of  $n_k$ , where the elbow point is considered to provide  $n_k^{\text{opt}}$  empirically. In this work, the elbow point can be clearly seen at  $k = 3$  as shown in Fig. 2a. However, it should be noted that the elbow method does not always work well, especially if the data are not very cohesive.

**Silhouette method.** The silhouette method evaluates the cohesion ( $a^{(i)}$ ) and separation ( $b^{(i)}$ ) of clusters as

$$a^{(i)} = \frac{1}{|C_{\text{in}} - 1|} \sum_{x^{(j)} \in C_{\text{in}}} x^{(i)} - x^{(j)}, \quad (3)$$

$$b^{(i)} = \frac{1}{|C_{\text{near}}|} \sum_{x^{(j)} \in C_{\text{near}}} x^{(i)} - x^{(j)}, \quad (4)$$

where  $a^{(i)}$  is the average distance between  $x^{(i)}$  and other points in the same cluster  $C_{\text{in}}$ , while  $b^{(i)}$  is the average distance between  $x^{(i)}$  and all points in  $C_{\text{near}}$  (the nearest-neighbor cluster of  $x^{(i)}$ ). As a measure of the cohesion and separation of clusters, the silhouette score is then given as

$$s^{(i)} = \frac{b^{(i)} - a^{(i)}}{\max(a^{(i)}, b^{(i)})}, \quad (5)$$

where  $s^{(i)}$  ranges  $[-1, 1]$ . Thus, the best clustering is given when  $s^{(i)} = 1$ , the assignment of data belonging would be incorrect when  $s^{(i)} < 0$ , and the clusters overlap when  $s^{(i)} = 0$ . As shown in Fig. 2b, the highest silhouette score in the dataset is given at  $k = 2$ , while it was obviously not appropriate. This stems from the fact that the silhouette method is not good at treating the dataset that includes adjacent or overlapping clusters.

**Gap statistic method.** The gap statistic method is based on the statistical testing methods. It aims to standardize the cohesion measure  $W_k$  for the dataset  $I_i = (I_1, I_2, \dots, I_n)$  by using  $W_{k,b}^*$  for the null reference distribution of data (random dataset)  $I_b = (I_1, I_2, \dots, I_B)$ . The  $B$  is the number of bootstrap samples. The optimal number of clusters  $n_k^{\text{opt}}$  will be given at which the gap statistic becomes maximum. Let  $C_r$  denotes the indices of observations in cluster  $r$ ,  $n_r$  is the number of points in the cluster  $C_r$ , and  $I_i$  and  $I_j$  are a data point of the dataset  $I_i$ . Then,  $W_k$  is defined as

$$W_k = \sum_{r=1}^{n_k} \frac{1}{2n_r} \sum_{i,j \in C_r} \|I_i - I_j\|. \quad (6)$$

Since  $\log W_k$  falls the farthest when the value of  $k$  takes the optimal number of clusters, the gap statistics can be estimated as

$$\text{Gap}(k) = \frac{1}{B} \sum_{b=1}^B \log W_{k,b}^* - \log W_k. \quad (7)$$

In general, the gap becomes unchanged, and the results become precise for  $B \geq 500$ . The optimum number of clusters should provide a higher gap value and is also given as the minimum value of  $k$  satisfying

$$\text{Gap}(k) \geq \text{Gap}(k+1) - s(k+1), \quad (8)$$

where

$$s(k) = \text{sd}_k \sqrt{1 + \frac{1}{B}} \quad (9)$$

and  $\text{sd}_k$  denotes the standard deviation of  $\{\log W_k\}_{b=1}^B$ . Thus, the gap difference (Diff.) can be used as the criterion as

$$\text{Diff.} = \text{Gap}(k) - \text{Gap}(k+1) + s(k+1) \geq 0. \quad (10)$$

The obtained gap statistics for the core-level dataset shown in Fig. 2c indicates the local maximum at  $k = 8$ , while  $k = 5$  seems more appropriate judging from the gap difference shown in the inset of Fig. 2c. Similarly, in the case of the near- $E_F$  dataset shown in Fig. 4f, the local maximum is given at  $k = 8$ , while  $k = 6$  should be appropriate judging from the gap difference shown in the inset of Fig. 4f. However, it should also be noted that the results of gap statistics fluctuate each time, even using a much higher number of bootstrap ( $B = 1000$  and  $2000$ ). Thus, we regarded the gap statistics just as a reference and adapted  $B = 500$  to reduce the computational cost. Indeed, we adopted  $k = 8$  as a trial and higher  $n_k$ -value in this work, to see the  $n_k$ -dependence on the K-means clustering results.

**Fuzzy-c-means clustering.** The fuzzy-c-means clustering is very similar to the K-means clustering, though it assigns the attribution probability  $u_{ik} \in [0, 1]$  ( $i = 1, \dots, n$ ,  $k = 1, \dots, n_k$ ) to all the clusters  $C_k = (C_1, C_2, \dots, C_{n_k})$  for each data point of the  $n$  observables,  $I_i = (I_1, I_2, \dots, I_n)$ . Thus, each data point may have the attribution probability across multiple clusters in the fuzzy-c-means clustering (soft-clustering). This is in contrast to the K-means clustering, in which each data point belongs to only a single cluster (hard-clustering). Then, the fuzzy-c-means algorithm aims to minimize the objective function ( $W$ ) given by

$$W(u_{ik}, c_k) = \sum_{i=1}^n \sum_{k=1}^{n_k} (u_{ik})^m d(I_i, c_k), \quad (11)$$

where  $m$  is a hyperparameter called a fuzzifier that defines the maximum fuzziness or noise in the dataset, and  $d(I_i, c_k) = d_{ik} = \|I_i - c_k\|^2$  in the case of Euclidean distance. The attribution probability  $u_{ki}$  and the centroids of clusters  $c_k$ , minimizing  $W(u_{ki}, c_k)$ , are respectively given as follows:

$$u_{ki} = \frac{(1/d_{ik})^{\frac{1}{m-1}}}{\sum_{i=1}^n (1/d_{ik})^{\frac{1}{m-1}}}, \quad (12)$$

and

$$c_k = \frac{\sum_{i=1}^n (u_{ik})^m I_i}{\sum_{i=1}^n (u_{ik})^m}. \quad (13)$$

The fuzzy-c-means algorithm aims to find  $u_{ik}$  and  $c_k$ , yielding the minimum of objective function  $W(u_{ik}, c_k)$  by an alternate optimization as follows.

- Step 1. Specify  $n_k$  and  $m$ , and initialize  $u_{ik}$  randomly.
- Step 2. Calculate  $c_k$ .
- Step 3. Calculate  $u_{ik}$ .
- Step 4. Repeat Step 2 and 3 until  $W(u_{ik}, c_k)$  is minimized or  $\|u_{ik}^{(t+1)} - u_{ik}^{(t)}\| < \epsilon$ .

Here,  $t$  is the iteration number, and  $\epsilon$  is a predefined convergence value. The performance of the fuzzy-c-means clustering is often evaluated by using the fuzzy partition coefficient (FPC). The FPC index is calculated by

$$\text{FPC} = \frac{1}{n_k} \sum_{k=1}^{n_k} \sum_{i=1}^n u_{ik}^2. \quad (14)$$

The FPC ranges between  $[0, 1]$ , and the higher value is generally expected to result in better clustering performance. As shown in Supplementary Fig. 3, however, the FPC became higher with smaller  $n_k$  and  $m$  without showing any indication of the optimum number of these hyperparameters. Thus, these hyperparameters were not determined intuitively in the present case.

## DATA AVAILABILITY

The datasets and codes that support the findings of this study can be found at <https://github.com/h-iwasawa/arpes-clustering>.

Received: 7 July 2021; Accepted: 1 December 2021;  
Published online: 18 February 2022



## REFERENCES

- Dagotto, E. Complexity in strongly correlated electronic systems. *Science* **309**, 257–262 (2005).
- Damascelli, A., Hussain, Z. & Shen, Z.-X. Angle-resolved photoemission studies of the cuprate superconductors. *Rev. Mod. Phys.* **75**, 473 (2003).
- Yang, H. et al. Visualizing electronic structures of quantum materials by angle-resolved photoemission spectroscopy. *Nat. Rev. Mater.* **3**, 341–353 (2018).
- Lv, B., Qian, T. & Ding, H. Angle-resolved photoemission spectroscopy and its application to topological materials. *Nat. Rev. Phys.* **1**, 609–626 (2019).
- Sobota, J. A., He, Y. & Shen, Z.-X. Angle-resolved photoemission studies of quantum materials. *Rev. Mod. Phys.* **93**, 025006 (2021).
- Rotenberg, E. & Bostwick, A. microARPES and nanoARPES at diffraction-limited light sources: opportunities and performance gains. *J. Synchrotron Radiat.* **21**, 1048–1056 (2014).
- Cattelan, M. & Fox, N. A. A perspective on the application of spatially resolved ARPES for 2D materials. *Nanomaterials* **8**, 284 (2018).
- Iwasawa, H. High-resolution angle-resolved photoemission spectroscopy and microscopy. *Electron. Struct.* **2**, 043001 (2020).
- Lupi, S. et al. A microscopic view on the Mott transition in chromium-doped  $V_2O_3$ . *Nat. Commun.* **1**, 105 (2010).
- Massee, F. et al. Bilayer manganites reveal polarons in the midst of a metallic breakdown. *Nat. Phys.* **7**, 978–982 (2011).
- Iwasawa, H. et al. Surface termination and electronic reconstruction in  $YBa_2Cu_3O_{7-x}$ . *Phys. Rev. B* **98**, 081112(R) (2018).
- Iwasawa, H. et al. Buried double  $CuO$  chains in  $YBa_2Cu_4O_8$  uncovered by nano-ARPES. *Phys. Rev. B* **99**, 140510(R) (2019).
- Watson, M. D. et al. Probing the reconstructed Fermi surface of antiferromagnetic  $BaFe_2As_2$  in one domain. *npj Quantum Mater.* **4**, 36 (2019).
- Noguchi, R. et al. A weak topological insulator state in quasi-one-dimensional bismuth iodide. *Nature* **566**, 518–522 (2019).
- Lee, K. et al. Discovery of a weak topological insulating state and Van Hove singularity in triclinic  $RhBi_2$ . *Nat. Commun.* **12**, 1855 (2021).
- Nguyen, P. V. et al. Visualizing electrostatic gating effects in two-dimensional heterostructures. *Nature* **572**, 220–223 (2019).
- Joucken, F. et al. Visualizing the effect of an electrostatic gate with angle-resolved photoemission spectroscopy. *Nano Lett.* **19**, 2682–2687 (2019).
- Jones, A. J. H. et al. Observation of electrically tunable Van Hove singularities in twisted bilayer graphene from nanoARPES. *Adv. Mater.* **32**, 2001656 (2020).
- Lisi, S. et al. Observation of flat bands in twisted bilayer graphene. *Nat. Phys.* **17**, 189–193 (2021).
- Iwasawa, H. et al. Development of laser-based scanning  $\mu$ -ARPES system with ultimate energy and momentum resolutions. *Ultramicroscopy* **182**, 85–91 (2017).
- Kastl, C. et al. Effects of defects on band structure and excitons in  $WS_2$  revealed by nanoscale photoemission spectroscopy. *ACS Nano* **13**, 1284–1291 (2019).
- Iwasawa, H. et al. Accurate and efficient data acquisition methods for high-resolution angle-resolved photoemission microscopy. *Sci. Rep.* **8**, 17431 (2018).
- Rosenbrock, C. W., Homer, E. R., Csányi, G. & Hart, G. L. W. Discovering the building blocks of atomic systems using machine learning: application to grain boundaries. *npj Comput. Mater.* **3**, 29 (2017).
- Stanev, V. et al. Machine learning modeling of superconducting critical temperature. *npj Comput. Mater.* **4**, 29 (2018).
- Zhang, Y. et al. Machine learning in electronic-quantum-matter imaging experiments. *Nature* **570**, 484–490 (2019).
- MacQueen, J. Some methods for classification and analysis of multivariate observations. In *Proc. of 5th Berkeley Symposium on Math. Sta. and Prob.* (eds Le Cam, L. M. & Neyman, J.) 281–297 (University of California Press, 1967).
- Bezdek, J. C. *Pattern Recognition with Fuzzy Objective Function Algorithms* (Springer, 1981).
- Arthur, D. & Vassilvitskii, S. k-means++: the advantages of careful seeding. In *Proc. Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms* (eds Bansal, N., Pruess, K. R. & Stein, C.) 1027–1035 (Association for Computing Machinery, 2007).
- Miyamoto, S., Ichihashi, H. & Honda, K. *Algorithms for Fuzzy Clustering* (Springer, 2008).
- Sibson, R. SLINK: an optimally efficient algorithm for the single-link cluster method. *Comput. J.* **16**, 30–34 (1973).
- Ester, M., Kriegel, H.-P., Sander, J. & Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proc. 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96)* (eds Simoudis, E., Han, J. & Fayyad, U.) 226–231 (AAAI Press, 1996).
- Yamada, Y. & Shiohara, Y. Continuous crystal growth of  $YBa_2Cu_3O_{7-x}$  by the modified top-seeded crystal pulling method. *Phys. C* **217**, 182–188 (1993).
- Hoesch, M. et al. A facility for the analysis of the electronic structures of solids and their surfaces by synchrotron radiation photoelectron spectroscopy. *Rev. Sci. Instrum.* **88**, 013106 (2017).

- Rokach, L. & Maimon, O. Clustering Methods. In *Data Mining and Knowledge Discovery Handbook* (eds Maimon, O. & Rokach, L.) 321–352 (Springer, 2005).
- Giordan, P., Ferraro, M. B. & Martella, F. Non-Hierarchical Clustering. In *An Introduction to Clustering with R*, 75–109 (Springer, 2020).
- Defays, D. An efficient algorithm for a complete-link method. *Comput. J.* **20**, 364–366 (1977).
- Ward, J. H. Jr Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.* **58**, 236–244 (1963).
- Kaufman, L. & Rousseeuw, P. J. *Finding Groups in Data: an Introduction to Cluster Analysis* (John Wiley & Sons, Inc., 1990).
- Luxburg, U. V. A tutorial on spectral clustering. *Stat. Comput.* **17**, 395–416 (2007).
- Bishop, C. M. *Pattern Recognition and Machine Learning* (Springer, 2006).
- Hofmann, T. Probabilistic latent semantic analysis. In *UAI'99: Proc. Conference on Uncertainty in Artificial Intelligence* (eds Lasky, K. B. & Prade, H.) 289–296 (Morgan Kaufmann Publishers Inc., 1999).
- Dhillon, I. S. & Sra, S. Generalized nonnegative matrix approximations with Bregman divergences. In *NIPS'05 Proc. 18th International Conference on Neural Information Processing Systems* (eds Weiss, Y., Schölkopf, B. & Platt, J. C.) 283–290 (MIT Press, 2005).
- Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
- Thorndike, R. L. Who belongs in the family? *Psychometrika* **18**, 267–276 (1953).
- Rousseeuw, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20**, 53–65 (1987).
- Tibshirani, R., Walther, G. & Hastie, T. Estimating the number of clusters in a data set via the gap statistic. *J. R. Stat. Soc. B* **63**, 411–423 (2001).

## ACKNOWLEDGEMENTS

We thank Niels B. M. Schröter, Timur K. Kim, and Moritz Hoesch for their supports on the ARPES experiments. We thank Diamond Light Source for access to beamline I05 (Proposals No. NT16871 and No. NT17192) that contributed to the results presented here. This work was supported by Grant-in-Aid for Scientific Research (C) JSPS KAKENHI Grant Number 19K03749, JSPS Bilateral Program Grant Number JPJSBP120209941, and QST President's Strategic Grant (QST Advanced Study Laboratory).

## AUTHOR CONTRIBUTIONS

H.I. supervised the project, performed spatially resolved ARPES experiments, and wrote the manuscript. H.I. developed the analytical codes and analyzed the data with input from T.U. High-quality samples were grown by T.M. and S.T. All authors discussed the results and commented on the manuscript.

## COMPETING INTERESTS

The authors declare no competing interests.

## ADDITIONAL INFORMATION

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41535-021-00407-5>.

**Correspondence** and requests for materials should be addressed to Hideaki Iwasawa.

**Reprints and permission information** is available at <http://www.nature.com/reprints>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.