Clustering analysis strategies for Electron Energy Loss Spectroscopy (EELS)

Pau Torruella*[1,2], Marta Estrader[3,#], Alberto López-Ortega[4], Maria Dolors Baró[5], Maria Varela[7], Francesca Peiró[1,2], Sònia Estradé[1,2]

[1] LENS-MIND, Departament d'Enginyeries: Electrònica, Universitat de Barcelona, 08028 Barcelona, Spain.
[2] Institute of Nanoscience and Nanotechnology (IN2UB), Universitat de Barcelona, 08028 Barcelona, Spain.
[3] Departament de Química Inorgànica i Orgànica, Universitat de Barcelona, Diagonal 645, 08028 Barcelona, Spain.
[#] Present Address: LPCNO, Université de Toulouse, CNRS, INSA, UPS, 135 avenue de Rangueil, 31077 Toulouse, France.
[4]CIC nanoGUNE, Tolosa Hiribidea, 76 E-20018 Donostia-San Sebastian, Spain.
[5] Departament de Física, Facultat de Ciències, Universitat Autònoma de Barcelona, E-08193 Bellaterra, Cerdanyola del Vallès, Barcelona, Spain
[7] Departamento de Física de Materiales, Instituto Pluridisciplinar and Instituto de Magnetismo Aplicado, Facultad de Ciencias Físicas, Universidad Complutense de Madrid, 28040 Madrid, Spain.

Abstract

In this work, the use of cluster analysis algorithms, widely applied in the field of big data, is proposed to explore and analyse electron energy loss spectroscopy (EELS) data sets. Three different data clustering approaches have been tested both with simulated and experimental data from Fe_3O_4/Mn_3O_4 core/shell nanoparticles. The first method consists on applying data clustering directly to the acquired spectra. A second approach is to analyse spectral variance with principal component analysis (PCA) within a given data cluster. Lastly, data clustering on PCA score maps is discussed. The advantages and requirements of each approach are studied. Results demonstrate how clustering is able to recover compositional and oxidation state information from EELS data with minimal user input, giving great prospects for its usage in EEL spectroscopy.

Introduction

The introduction of the first commercial electron energy loss spectrometer in 1986, model 666 by Gatan, set a huge milestone in the history of materials science. Since then, electron energy loss spectroscopy (EELS) has evolved from acquiring a single spectrum in several seconds[1] with a spatial resolution of tens of nanometers to atomic resolution spectrum images (SI) with thousands of pixels acquired in milliseconds[2]. Not only that, but the development of EELS tomography[3,4] at increasingly higher spatial and energy resolutions is starting to generate so-called EEL spectrum volumes of millions of spectra. In parallel, the development of automated electron tomography[5,6] and the foreseeable coalescence into automated EELS tomography might produce even larger EELS datasets.

EELS is, therefore, at a point where it faces new challenges and prospects of great developments. How to efficiently extract relevant information from enormous datasets is, nonetheless, a question that has been extensively addressed in the big data, data mining, machine learning and other data-science (DS) fields. Clearly, the microscopy community can benefit from the methods developed in these areas. A prominent example is the widespread use of principal component analysis (PCA) in EELS[3,7,8], either for reducing noise without loss of spatial resolution or for data exploration, i.e., the identification of different phases or compositions in a sample without thorough modelling of the spectra. This has enabled the mapping of energy loss near edge structure (ELNES) related properties from relatively noisy datasets[7] or even 3D reconstruction of oxidation states at the nanoscale[9].

However, PCA comes with its own problems, such as artefacts from ignoring too many components, obtaining components not related to physical properties of the sample or, simply, components that do not represent the data adequately. This stems from a fundamental flaw of applying PCA to EELS data. PCA solves the problem of finding an orthogonal coordinate system that maximizes the variance along its axes[10]. In this description it is implicitly assumed that the data treated follows a multivariate normal distribution[11]. However, this is not true in general for EEL SI.

For this reason, usually PCA is combined with other data processing techniques, such as Independent Component Analysis (ICA) [12–14]. Bayesian Linear Unmixing (BLU) must also be mentioned as an alternative method that has also been used in EELS[8,15]. All these spectral decomposition methods have yielded excellent results with EELS. However, they rely on the ability of the scientist to make a physical interpretation of the output components, something that is not always easy.

It must be kept in mind that the goal of EEL spectrum imaging and the related data analysis techniques is mapping the spatial distribution of properties reflected in the shape of individual EEL spectra. In many cases, this implies finding a way to segment a given SI into different regions, where each corresponds to a different material or has a different property. This can be thought of as classifying spectra into groups with similar characteristics, such as spectra with a peak at the same energy position or spectra showing two edges with a similar intensity ratio. Interestingly, the task of grouping sets of objects by similarity is a well-known problem in DS, named cluster analysis or, simply, clustering[16–18].

In general terms, clustering aims to classify a number of objects, **n** , that have a number of attributes, **p**, in groups such that objects in the same group have similar attributes amongst them and as different as possible from those in other groups[19]. An EELS SI with **X**-by-**Y** pixels and **E** channels can be understood as a series of **n**=**X**·**Y** spectra (our objects), that have **p=E** attributes (the intensity value in each channel). It is straightforward to transform any SI into an **n**·**p** matrix, where each individual spectrum is a row. This matrix is a suitable input for most data clustering algorithms. This can also be understood as representing each spectrum as a point in a **p**-dimensional space. In this context, the concept of similarity between two spectra is simply the distance between the two corresponding **p**-dimensional points in any desired metric. Henceforth, the term “distance” will be used in this context.

In this frame, a clustering algorithm would label each of these points, or spectra, as pertaining to one cluster or another according to the distances between them. These labels can be assigned a colour and represented into shape of the original SI, resulting in an image where pixels of the same colour contain spectra from the same cluster. If applied successfully, the spectra clusters would segment the SI into regions with different chemical signatures, revealing the morphology and composition of the studied sample.

There are a wide variety of clustering algorithms available. K-means, density based methods and agglomerative clustering algorithms are widely used[17,19] and should be suitable for EELS. In this work, we limit the scope to hierarchical agglomerative clustering because of its speed and simplicity.

Because of its usage throughout the article a brief description of the hierarchical clustering algorithm is mandatory. Assume initially there are n points in a p-dimensional space to study. Let us denote the coordinates for a given point in that space by x_k (k = 1…p). In the first iteration of the algorithm, each of these points would be considered as a different cluster that contains one point. In successive iterations, the distances between each point are calculated as

$$\boldsymbol{d_{ij}} = \sqrt{\sum_{i}^{p} \left(x_k^i - x_k^j\right)^2} \tag{1}$$

, if the usual euclidean metric is used. Then the algorithm looks for the two points that belong to different clusters that are the closest. Let us name the distance between these points at iteration h by $\min(d_{ij})^h$. Once these points are found, the two clusters to which they belong are “linked” and in following iterations considered to be the same cluster. If n iterations are performed, all the points will be linked in one single cluster. The term “agglomerative clustering” describes this sequential linking of clusters, agglomerating points that are closer together, until there is only one cluster encompassing all the data[19]. But obtaining this final cluster does not provide any information or insight into the data. The way to identify significant clusters is to choose the iteration at which to stop the clustering process. A way to do this is to choose a distance threshold, so that the process is stopped at iteration h for which the condition $\min(d_{ij})^h > d_{threshold}$ is fulfilled. Therefore, if the last **l=n-h** links would occur across distances higher than the threshold, **l+1** clusters would be obtained. The way to choose an adequate value for $d_{threshold}$ is to consider how $\min(d_{ij})$ evolves with each iteration. Usually it will be small until a certain iteration from which it grows steeply. Defining the $\min(d_{ij})$ value before that significant increase will yield an adequate threshold.

The by considering the spectra in a SI an as a collection of **p**-dimensional (**p=E**) points the algorithm can be easily applied to EELS data. However, clustering performed this way is merely a segmentation method and some spectral variance might still be present in a given cluster. Nonetheless, there is no hindrance to perform PCA on each cluster, obtaining scores and factors that represent the data on that region. This would assure that all the information in a SI is being used.

Another important factor to take into account is noise in the spectra. If it is too severe, different clusters may overlap, hampering the clustering process. Additionally, the computational time of the clustering algorithm would be proportional, among other factors, to the number of energy channels of the spectra. These issues can be simultaneously addressed by following a different clustering strategy. If PCA is initially carried out, series components and their corresponding score maps will be obtained. Let us assume that the first **c** components explain all the relevant variance of the SI and the rest are noise. Now, the SI can be represented with only **c** values for each pixel, each one corresponding to the weight of a component in that pixel. This can also be understood in the clustering nomenclature as having **n=X·Y** objects each with **c** attributes (the scores of each PCA component). This great reduction in the number of attributes means that the clustering algorithm will now deal with a much smaller, mostly noise-free dataset, improving the computational time and accuracy in the cluster formation.

In this work, the three proposed strategies to apply clustering to EEL SI are considered: performing data clustering on the raw data, performing PCA within data clusters, and finally, performing clustering on the score maps of an initial PCA decomposition. The different methods will be tested out both with simulated and experimental EELS data from Fe_3O_4/Mn_3O_4 core/shell nanoparticles, to study how they perform in different scenarios and what pre-processing is necessary in each case.

Results and Discussion

Spectrum image simulation.

To demonstrate the application of data clustering to EELS datasets an artificial EELS SI was generated. The 128x128 pixels SI contain a 1024 channel spectrum in each one, which reflect the composition of the model shown in Figure 1A. It was chosen to be a particle with four main regions: a FeO core of constant composition, a FeCoO region (as if it was a precipitate), an internal void and, finally, a shell with varying Fe/O ratio. The intensity of the iron edge in this shell increases linearly from the edge of the particle to the core region, while the oxygen edge intensity decreases in the same proportion. For each of these regions, the spectrum in each pixel was built as a linear model comprised of several factors. First, a power-law background was used. The corresponding hydrogenic cross-sections for each element were added, with intensity proportional to the chosen concentration of the element. Two gaussian curves were used to model the white lines for Fe (at 710 eV and 724eV) and Co (at 781 eV and 796 eV). Finally, gaussian and poissonian noise with a signal-to-noise ratio (SNR) of 10:1 were added, resulting in the spectra in figure 1C for the positions highlighted in figure 1B.

In Figure 1D the intensity of the channel corresponding to an energy loss of 781 eV is plotted against the intensity of channel 710 eV, where each simulated spectrum is represented by a single point. The latter panel serves to discuss the spectral variance distribution within an SI. It is clear that the data form groups related to the different compositions present in the sample, even with the presence of noise, demonstrating that the EELS data does not follow a normal, PCA-suitable distribution. With this visualization, the adequacy of data clustering in EELS can be appreciated.

Raw data clustering

The simulated SI was transformed into a matrix of 128x128=16384 rows and 1024 columns according to the previously discussed data interpretation. The simulated data was then fed to the hierarchical agglomerative clustering algorithm. The code, implemented mainly with the Scipy[20] and Hyperspy[21] python libraries, is available as Supplementary information.

An interesting way to visualize the clustering process is to plot the so called linkage tree of the clustering process. The clusters are laid along the x-axis of the plot, sorted by an arbitrary id number. The y-axis represents distance as defined by (1). A vertical line arises from each cluster. These vertical lines are linked by horizontal lines at a height which corresponds to the min(d_{ij}) at which they were merged together in the clustering process. The result is a tree-like plot, denominated “dendogram plot”, which can be seen in figure 2, that gives an idea of the clustering evolution.

In figure 2, it is evident that the last three links occur across distances at least an order of magnitude greater than the previous ones and so they are discarded, yielding four clusters. The output of the algorithm is a vector of length 16384 which contains the labels of the cluster to which each spectrum pertains. This vector can be reshaped into the original image as shown in figure 1E, where pixels that contain spectra from a given cluster are displayed with the same colour. Here they are chosen to match those used in the model, green for the shell cluster, blue for the core cluster, red for the FeCoO region and black for the pixel cluster that represents the background. Clearly, the obtained clusters give an excellent representation of the data, recovering the initial composition of the simulated SI. Moreover, it is now possible to obtain the average spectrum of each cluster (Figure 1F), showing the chemical composition from each of the image segmented regions.

Nevertheless, according to figure 1E, the procedure seems to have failed in identifying the existence of the composition gradient in the shell. This exemplifies a case where there are relevant differences within a spectra cluster. This exemple illustrates the relevance of an appropriated choice of the free parameter $d_{threshold}$ to unveil all the significant information contained in the spectra.. Figure 3 shows the segmentation results with three different threshold values, revealing several clusters within the shell related to the linear variation of Fe and O. Still, discrete segmentation of a property that varies continuously is a suboptimal representation and an inherent limitation of this “only clustering” approach. Nonetheless, this limitation can be overcome by performing PCA for each cluster.

Principal component analysis within a cluster.

Instead of searching for an adequate threshold value that reveals the spectral variation in the outer shell of the particle, it is possible to apply PCA to the spectra that comprise this cluster.

This is equivalent to performing PCA while masking the spectra for all pixels not pertaining to the cluster.

The results of PCA applied to the $Fe_{1-x}O_{1+x}$ shell of the simulated data (green cluster in Figure 3A) are shown in Figure 4. The algorithm resolves that two main components explain the data, a mean iron oxide spectrum (Figure 4A) and a spectrum of positive oxygen signal and negative iron signal that varies in intensity along the shell (Figure 4B). The second factor demonstrates a varying Fe/O ratio in the shell, which is the origin of the problems in the segmentation of the shell seen in Figure 3. Whit this secondary analysis, the usage of all the information in the SI is assured, and no compositional changes are missed.

Clustering PCA scores

As promising as these results are, some additional considerations can be made. If the SNR of the simulated spectra decreased, the point groups in figure 1D would be more spread, and different clusters could overlap. This would obviously hinder the clustering process and is one of the reasons to perform clustering on PCA score maps rather than on the raw data.

PCA of the simulated SI demonstrates that the data can be explained with only three components (Figure S1). This means that now the SI information can be represented by only three values per pixel, and that the clustering input matrix will have 16384 rows and three columns.

The clustering result on these PCA scores is essentially the same as on the raw data, as seen in figure 5. However, now, less pixels in the core of the particle as mislabelled as shell, and the segmentation into two shells occurs at a slightly higher distance, meaning that the different clusters are more clearly segmented (dendogram tree in figure S2). Note that a segmentation of the shell into several clusters is a better representation of the composition variation. Otherwise, the (wrong) idea that the shell is uniform might be apparent. Additionally, links of the dendogram tree at higher distances, more spaced, mean that the algorithm is able to distinguish the differences between clusters more easily. These results demonstrate a significant improvement with respect to the raw spectra clustering, and the benefits may be even more noticeable in SI with lower SNR.

These are promising results, allowing identification of regions with different chemical composition and demand further testing with experimental data.

Experimental Spectrum image.

The three described strategies have also been tested in an experimental SI. The sample studied consisted in a Fe_3O_4/Mn_3O_4 core/shell nanoparticles. The nanoparticles were obtained using the seeded-growth method[22,23] where the Mn oxide layer was grown on Fe_3O_4 11 nm seeds by hot injection[24]. An EEL SI of the nanoparticles was acquired in an aberration-corrected Nion UltraSTEM200 operated at 200 kV and equipped with a Gatan Enfina spectrometer (Gatan). The simultaneous scanning TEM dark field image of the SI can is shown in Figure 6A. Previous analysis of the nanoparticles was performed through ELNES modelling[25]. The obtained ELNES parameter maps revealed that the core of the nanoparticle was Fe_3O_4. Mn L_3/L_2 ratio maps revealed the surrounding manganese oxide to be Mn_3O_4 nanocrystals and in an external shell of MnO.

Raw data clustering

Even though the simulated example demonstrates how clustering can correctly identify regions of an SI with different compositions, some factors must be taken into account before moving on to experimental data. Most TEM samples have changes in thickness that would hinder clustering in terms of chemical composition. Regions with the same composition but different thicknesses would be at large distances because, even if the “shape” of the spectra is the same, the total intensity is very different. This effect is shown in figure 6. The clustering analysis of the raw data (figure 6B) yielded three different clusters, which fail to adequately represent the chemical composition of the nanoparticle[24,25].

To circumvent this problem, the normalization of each spectrum by the integrated intensity of all energy channels is proposed. After normalization, the three data clusters that are found (figure 6C) and their corresponding mean spectra (figure 6D) clearly segment the SI into manganese oxide (red), iron oxide (blue) and supporting grid (green). This shows that automatic segmentation of real data into chemically distinct phases is possibly through spectra clustering.

As promising as this result is, two things should be noted. Firstly, the average spectrum of the nanoparticle core has some manganese signal. Its origin is the three dimensional nature of the nanoparticle and that there is some manganese oxide above the iron core. The type of image segmentation that clustering provides cannot unmix these signals. Additionally, this first method fails to “see” different Mn oxidation states (the chosen clustering threshold is shown in Figure S3).

Principal component analysis within a cluster

If PCA is performed on the manganese oxide related data cluster, labelled in Figure 6C, the components shown in Figure 7 are obtained. Clearly, the PCA factors (Figure 7C) have Mn L edges with very different L_3/L_2 intensity ratios. The value found for the blue component is $L_3/L_2 \approx 2.2$ and for the red component $L_3/L_2 \approx 4.3$. Actually, the components with this ELNES features[26–28] directly map the Mn_3O_4 (Figure 7A) and MnO (Figure 7B) concentrations. It must be pointed out that such clear results are very hard to obtain straight away from PCA. In fact, PCA of the whole raw data set was performed (figure S4) and no physically meaningful components, with straightforward interpretation were obtained, demonstrating the virtues of the clustering plus PCA scheme.

Clustering PCA scores

Last but not least, the clustering on PCA scores strategy was tried out. From Figure S4A it follows that the first 6 components are enough to explain the data of the SI. Again normalization must be undertaken in order to get results relevant to chemical composition rather than just related to thickness variation. In this case, each score map was divided by its maximum value. Following by hierarchical clustering (final clustering tree in Figure S5), four spectra clusters were found. As can be seen in Figure 6, they correspond to the support membrane, iron oxide, and two different manganese oxides. The average spectra for each of these regions show manganese oxide spectra with different Mn L_3/L_2 intensity ratios, demonstrating the already known Mn_3O_4 – MnO structure of the crystals. Therefore, the fact that the PCA representation cuts a lot of the noise was critical to allow the segmentation of both oxidation states.

Discussion

The presented results show that data clustering is a suitable analysis tool for EELS data sets. It is able to identify regions of different composition without any prior assumption of the data. This is certainly an advantage over the mentioned spectra modelling, which is limited to well-understood spectral features such as white lines. This advantage is shared with PCA and ICA.. ICA in particular is expected to yield physically meaningful components, but that is not always the case. In the present case ICA was applied to the data for completion, yielding the results in figure S6. Although some components can berelated to the different oxidation states of manganese and to the iron oxide core, a direct interpretation is not possible since all of them have non-physical features (negative edges, negative background, detector intensity steps…). Contrarily, clustering is bound to yield easily interpretable results since the averaged spectra from a cluster will always be physically meaningful.

However, as explained in the different sections, data clustering should be rather seen as a complementary tool. Spectra modelling could certainly be benefited by having an SI segmented and allowing different models to be used in different clusters. On the other hand, performing spectral decomposition techniques on segmented images has already been demonstrated to give good results[29], but this segmentation is usually performed manually. Clustering can therefore be used as an automatic segmentation to be implemented in more challenging data sets or when consistency and automation is required. Lastly, the benefits of working with PCA plus clustering have already been demonstrated in Figures 4-5 and 7-8, namely, the obtaining of physically meaningful components, noise reduction and lack of missing information.

Conclusions

By means of data clustering, SI segmentation according to chemical composition is obtained for EEL spectra, even recovering ELNES information. The authors believe that the three different strategies presented here will enable the use of data clustering in a wide range of problems, allowing easy and fast EELS data exploration. Importantly, it should be noted that this method is not limited to SI and could also be potentially used in EELS tomography in order to provide straightforward volume segmentation. The obtained results clearly show the usefulness and robustness of data clustering methods to deal with large EELS datasets.

Acknowledgements

First and foremost, the authors ackowledge Pedro Delicado from the department of Statiscs of the Universitat Politècnica de Catalunya for fruitful discussions and valuable advice. We also acknowlege Josep Nogués from the catalan institue of nanoscience and technology (ICN2) for his colaboration in the sinthesis of the studied nanoparticles. EELS data from the sample of Figures 6-8 was acquired at the Oak Ridge National Laboratory. The authors acknowledge the financial support from the Spanish Ministry of Economy, Industry and Competitiveness through the projects MAT2016-79455-P and MAT2015-66888-C3-3-R with support of FEDER funds.

Bibliography

(1) Egerton, R. F.; Yang, Y.-Y.; Cheng, S. C. Characterization and Use of the Gatan 666 Parallel-Recording Electron Energy-Loss Spectrometer. *Ultramicroscopy* **1993**, *48*, 239–250.

(2) Muller, D. a; Kourkoutis, L. F.; Murfitt, M.; Song, J. H.; Hwang, H. Y.; Silcox, J.; Dellby, N.; Krivanek, O. L. Atomic-Scale Chemical Imaging of Composition and Bonding by Aberration-Corrected Microscopy. *Science (80-.).* **2008**, *319*, 1073–1076.

(3) Yedra, L.; Eljarrat, A.; Arenal, R.; Pellicer, E.; Cabo, M.; López-Ortega, A.; Estrader, M.; Sort, J.; Baró, M. D.; Estradé, S.; *et al.* EEL Spectroscopic Tomography: Towards a New Dimension in Nanomaterials Analysis. *Ultramicroscopy* **2012**, *122*, 12–18.

(4) Jarausch, K.; Thomas, P.; Leonard, D. N.; Twesten, R.; Booth, C. R. Four-Dimensional STEM-EELS: Enabling Nano-Scale Chemical Tomography. *Ultramicroscopy* **2009**, *109*, 326–337.

(5) Koster, A. J.; Chen, H.; Sedat, J. W.; Agard, D. A. Automated Microscopy for Electron Tomography. *Ultramicroscopy* **1992**, *46*, 207–227.

(6) Mastronarde, D. N. Automated Electron Microscope Tomography Using Robust Prediction of Specimen Movements. *J. Struct. Biol.* **2005**, *152*, 36–51.

(7) De la Peña, F.; Berger, M.-H.; Hochepied, J.-F.; Dynys, F.; Stephan, O.; Walls, M. Mapping Titanium and Tin Oxide Phases Using EELS: An Application of Independent Component Analysis. *Ultramicroscopy* **2011**, *111*, 169–176.

(8) Yedra, L.; Eljarrat, A.; Rebled, J. M.; López-Conesa, L.; Dix, N.; Sánchez, F.; Estradé, S.; Peiró, F. EELS Tomography in Multiferroic Nanocomposites: From Spectrum Images to the Spectrum Volume. *Nanoscale* **2014**, *6*, 6646–6650.

(9) Torruella, P.; Arenal, R.; de la Peña, F.; Saghi, Z.; Yedra, L.; Eljarrat, A.; López-Conesa, L.; Estrader, M.; López-Ortega, A.; Salazar-Alvarez, G.; *et al.* 3D Visualization of the Iron Oxidation State in FeO/Fe_3O_4 Core−Shell Nanocubes from Electron Energy Loss Tomography. *Nano Lett.* **2016**, *16*, 5068–5073.

(10) Jolliffe, I. *Principal Component Analysis*; Balakrishnan, N.; Colton, T.; Everitt, B.; Piegorsch, W.; Ruggeri, F.; Teugels, J. L., Eds.; John Wiley & Sons, Ltd: Chichester, UK, 2014.

(11) O'Rourke, N.; Hatcher, L.; Stepanski, E. J.; Hatcher, L. *A Step-by-Step Approach to Using SAS for Univariate & Multivariate Statistics*; Wiley-Interscience, 2005.

(12) Hyvärinen, A.; Karhunen, J.; Oja, E. *Independent Component Analysis*; John Wiley & Sons, 2004.

(13) Bonnet, N.; Nuzillard, D. Independent Component Analysis: A New Possibility for Analysing Series of Electron Energy Loss Spectra. *Ultramicroscopy* **2005**, *102*, 327–337.

(14) Rossouw, D.; Langelier, B.; Scullion, A.; Danaie, M.; Botton, G. A. Multivariate-Aided Mapping of Rare-Earth Partitioning in a Wrought Magnesium Alloy. *Scr. Mater.* **2016**, *124*, 174–178.

(15) Dobigeon, N.; Brun, N. Spectral Mixture Analysis of EELS Spectrum-Images. *Ultramicroscopy* **2012**, *120*, 25–34.

(16) Oikonomakou, N.; Vazirgiannis, M. A Review of Web Document Clustering Approaches. In *Data Mining and Knowledge Discovery Handbook*; Springer US: Boston, MA, 2009; pp. 931–948.

(17) Rokach, L. A Survey of Clustering Algorithms. In *Data Mining and Knowledge Discovery Handbook*; Springer US: Boston, MA, 2009; pp. 269–298.

(18) Maimon, O.; Rokach, L. *Data Mining and Knowledge Discovery Handbook*; Springer, 2010.

(19) Kaufman, L.; Rousseeuw, P. J. *Finding Groups in Data: An Introduction to Cluster Analysis*; Wiley Series in Probability and Statistics; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 1990.

(20) Jones, E.; Oliphant, T.; Peterson, P.; others. {SciPy}: Open Source Scientific Tools for {Python}, 2001.

(21) De la Peña, F.; Burdet, P.; Ostasevicius, T.; Sarahan, M.; Nord, M.; Fauske, V. T.; Taillon, J.; Eljarrat, A.; Mazzucco, S.; Donval, G.; *et al.* Hyperspy: HyperSpy 0.8.2, 2015.

(22) Salazar-Alvarez, G.; Lidbaum, H.; López-Ortega, A.; Estrader, M.; Leifer, K.; Sort, J.; Suriñach, S.; Baró, M. D.; Nogués, J. Two-, Three-, and Four-Component Magnetic Multilayer Onion Nanoparticles Based on Iron Oxides and Manganese Oxides. *J. Am. Chem. Soc.* **2011**, *133*, 16738–16741.

(23) López-Ortega, A.; Estrader, M.; Salazar-Alvarez, G.; Estradé, S.; Golosovsky, I. V.; Dumas, R. K.; Keavney, D. J.; Vasilakaki, M.; Trohidou, K. N.; Sort, J.; *et al.* Strongly Exchange Coupled Inverse Ferrimagnetic Soft/hard, MnxFe3−xO4/FexMn3−xO4, Core/shell Heterostructured Nanoparticles. *Nanoscale* **2012**, *4*, 5138.

(24) Estrader, M.; López-Ortega, A.; Estradé, S.; Golosovsky, I. V; Salazar-Alvarez, G.; Vasilakaki, M.; Trohidou, K. N.; Varela, M.; Stanley, D. C.; Sinko, M.; *et al.* Robust Antiferromagnetic Coupling in Hard-Soft Bi-Magnetic Core/shell Nanoparticles. *Nat. Commun.* **2013**, *4*, 2960.

(25) Yedra, L.; Xuriguera, E.; Estrader, M.; López-ortega, A.; Baró, M. D.; Nogués, J.; Roldan, M.; Varela, M.; Estradé, S.; Peiró, F. Oxide Wizard : An EELS Application to Characterize the White Lines of Transition Metal Edges. *Microsc. Microanal.* **2014**, *20*, 698–705.

(26) Schmid, H. K.; Mader, W. Oxidation States of Mn and Fe in Various Compound Oxide Systems. *Micron* **2006**, *37*, 426–432.

(27) Colliex, C.; Manoubi, T.; Ortiz, C. Electron-Energy-Loss-Spectroscopy near-Edge Fine Structures in the Iron-Oxygen System. *Phys. Rev. B* **1991**, *44*, 11402–11411.

(28) Varela, M.; Oxley, M. P.; Luo, W.; Tao, J.; Watanabe, M.; Lupini, a. R.; Pantelides, S. T.; Pennycook, S. J. Atomic-Resolution Imaging of Oxidation States in Manganites. *Phys. Rev. B - Condens. Matter Mater. Phys.* **2009**, *79*, 1–14.

(29) Eljarrat, A.; López-Conesa, L.; López-Vidrier, J.; Hernández, S.; Garrido, B.; Magén, C.; Peiró, F.; Estradé, S.; Jetter, M.; Rühle, M.; *et al.* Retrieving the Electronic Properties of Silicon Nanocrystals Embedded in a Dielectric Matrix by Low-Loss EELS. *Nanoscale* **2014**, *6*, 14971–14983.

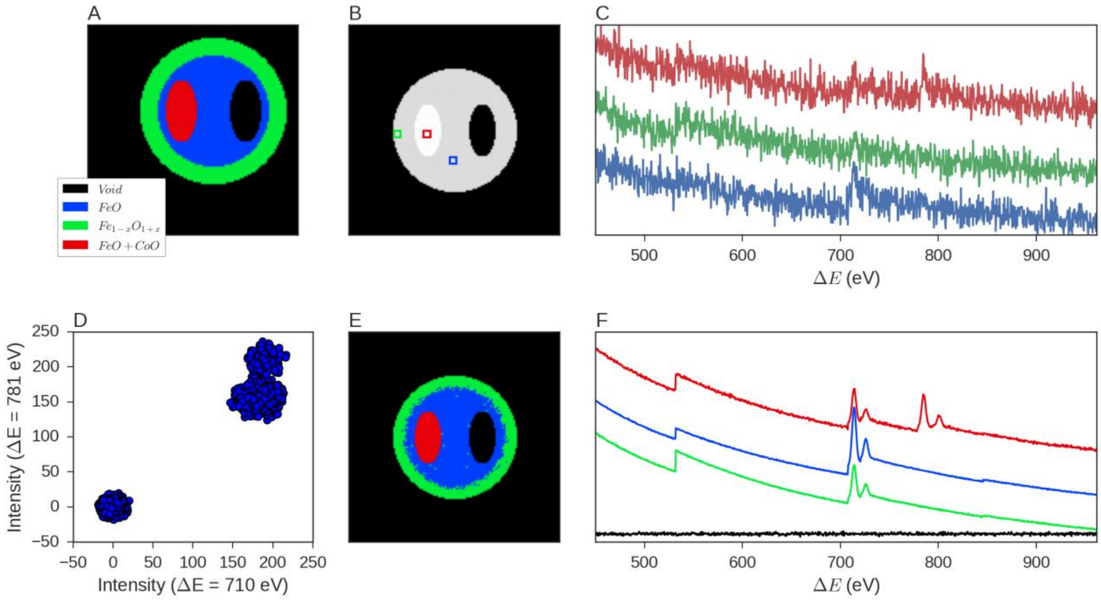


Figure 1. A) Model used for the simulated spectrum image. B) The grey level in this panel is related to the sum of all the channels of the spectrum in each pixel. C) Spectra from the markers positions in B): in red a sample spectrum of the iron-cobalt oxide, in green an iron oxide spectrum from the outer iron oxide shell and in blue a spectrum from the central iron oxide. D) Intensity of the 781 eV channel against the intensity of the 710 channel, where each point corresponds to a spectrum. E) Results of data clustering: each pixel has been given a colour based on the cluster at which its spectrum belongs. F) Averaged, normalized spectra from each cluster.

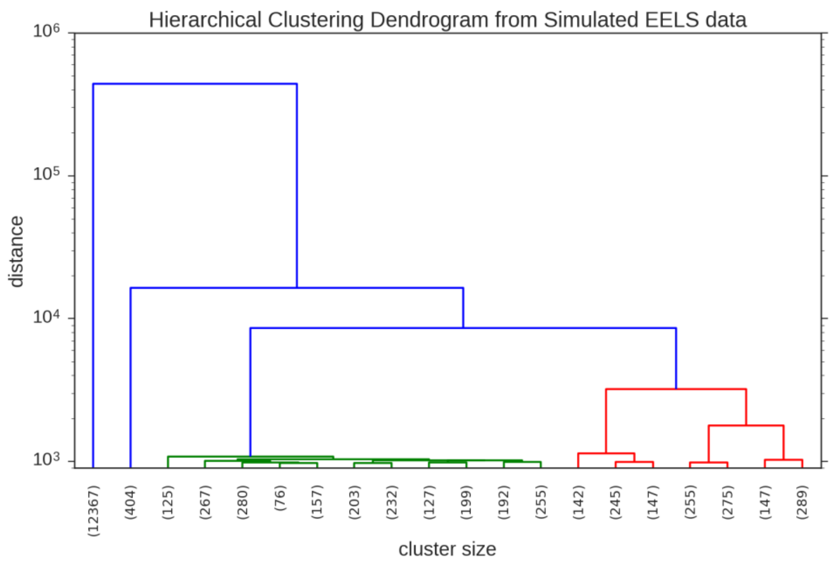


Figure 2. Dendogram plot for the last 20 hierarchical clustering links of the simulated data. Horizontal axis represents individual clusters, identified by an id number. A link is represented by a horizontal line that goes from one cluster id to another. The height at which the links form is the distance between the two linked clusters. Green and red link trees occur below the chosen distance threshold and are considered as

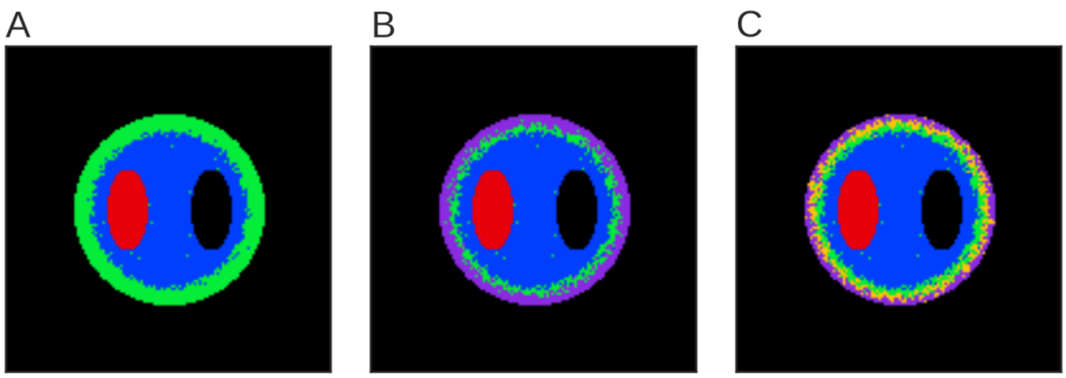


Figure 3. Clustering results of the simulated spectrum image with different distance thresholds. A) corresponds to a distance threshold of 5000 counts. B) corresponds to a distance threshold of 2500 counts. C) corresponds to a distance threshold of 1500 counts. This values correspond to the y-axis of figure 1.

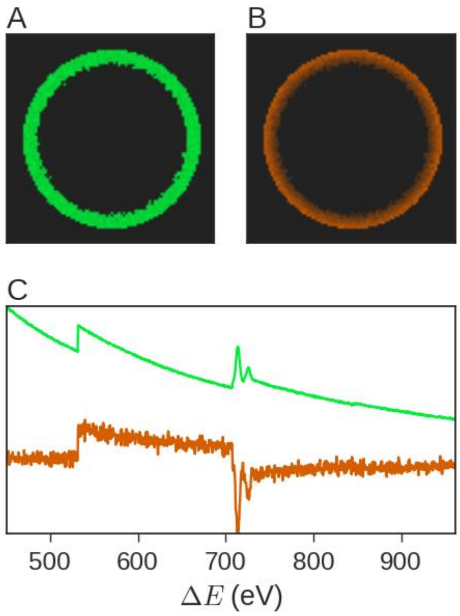


Figure 4. A) Score map from the first PCA component of the shell in the simulated spectrum image of Figure 1. B) Score map from the second PCA component of the shell in the simulated spectrum image of Figure 1. C) Factors of the corresponding score maps in panels A) and B).

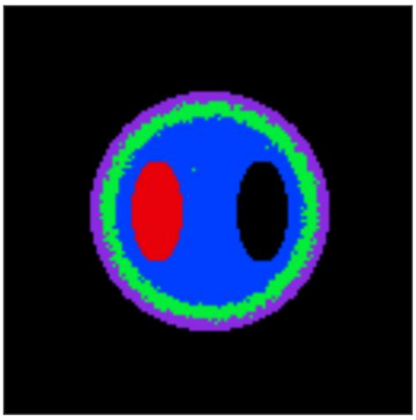


Figure 5. A) Clustering results on the first three PCA scores from figure S2. Each color corresponds to a spectra cluster.

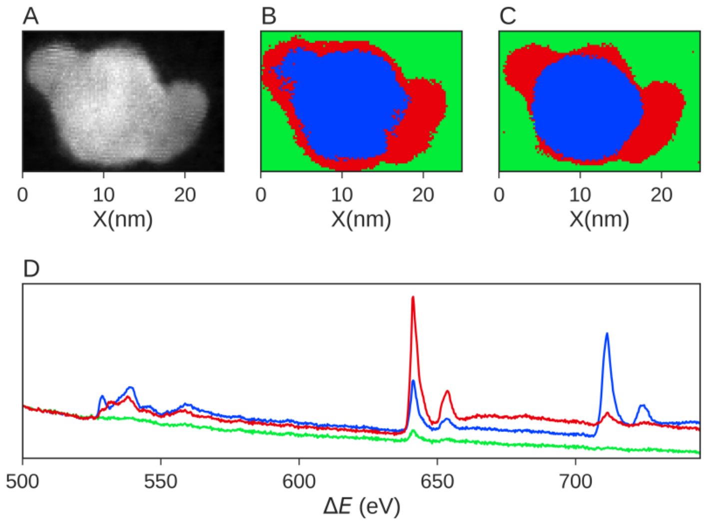


Figure 6. A) High Angle Annular Dark Field image simultaneously acquired with the spectrum image for the Fe_3O_4/Mn_3O_4 core/shell nanoparticle. B) Clustering results of the spectrum image without normalization. Each color corresponds to a cluster. C) Clustering results of the spectrum image with intensity normalization. D) Averaged spectra from each cluster.

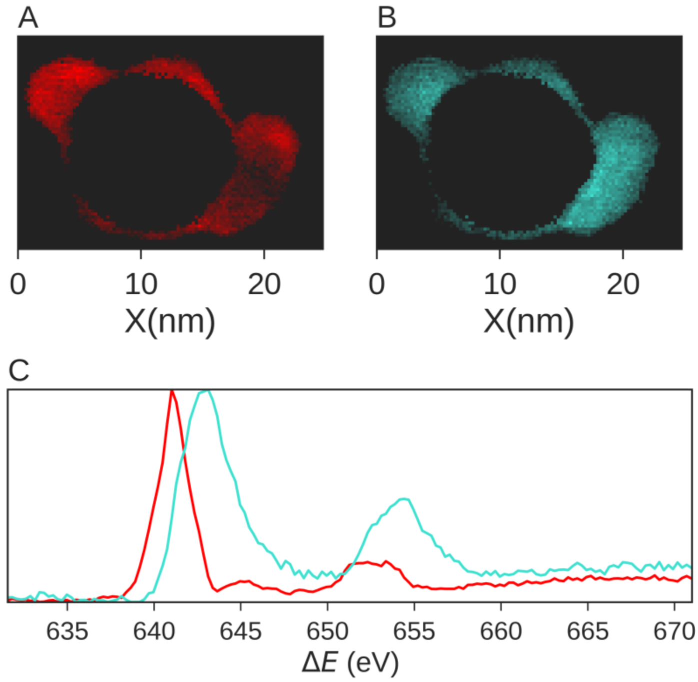


Figure 7. A) Score map from the first PCA component of the manganese oxide shell of Figure 6C. B) Score map from the second PCA component of the manganese oxide shell of Figure 6C. C) Factors of the corresponding score maps in panels A) and B).

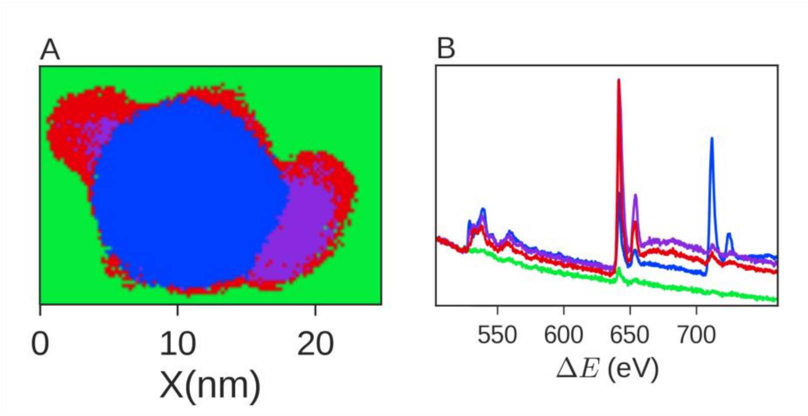


Figure 8. A) Clustering results on the first six PCA scores from figure S4. Each color corresponds to a cluster. C) Clustering results of the spectrum image with intensity normalization. D) Averaged spectra from each cluster.

Supplementary information

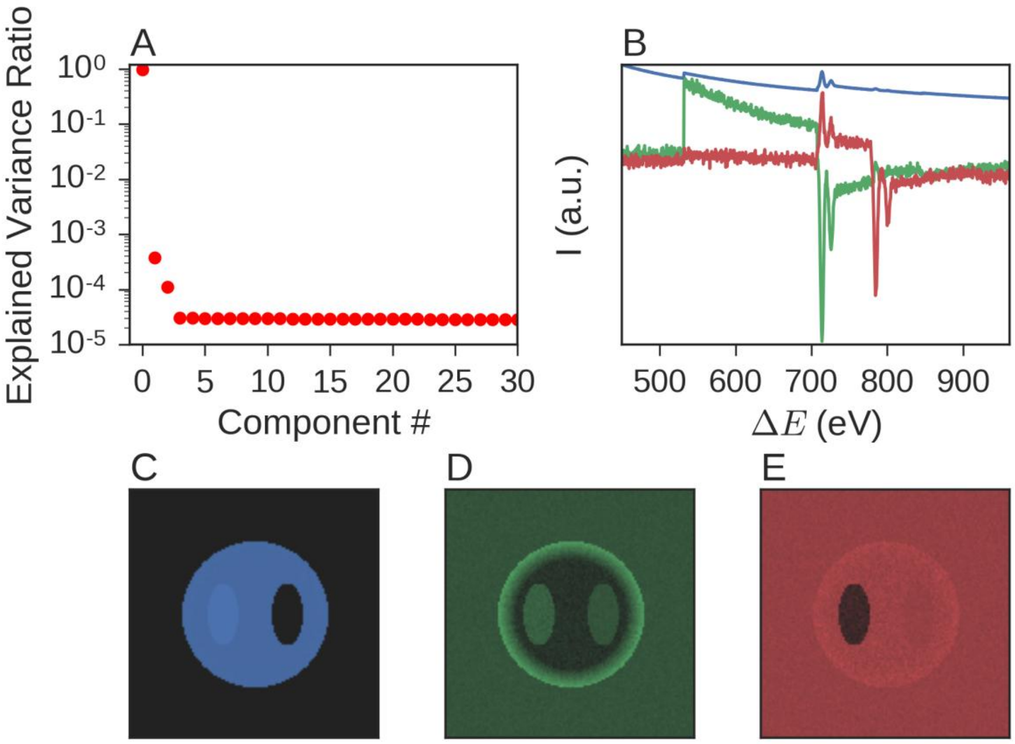


Figure S1. PCA decomposition results of thesimulatedspectrum image. A) Normalized variance ratio of each PCA component from the SI of figure 1A. B) The three first factors of the PCA decomposition. C-E) Score maps for each factor. Colour coded.

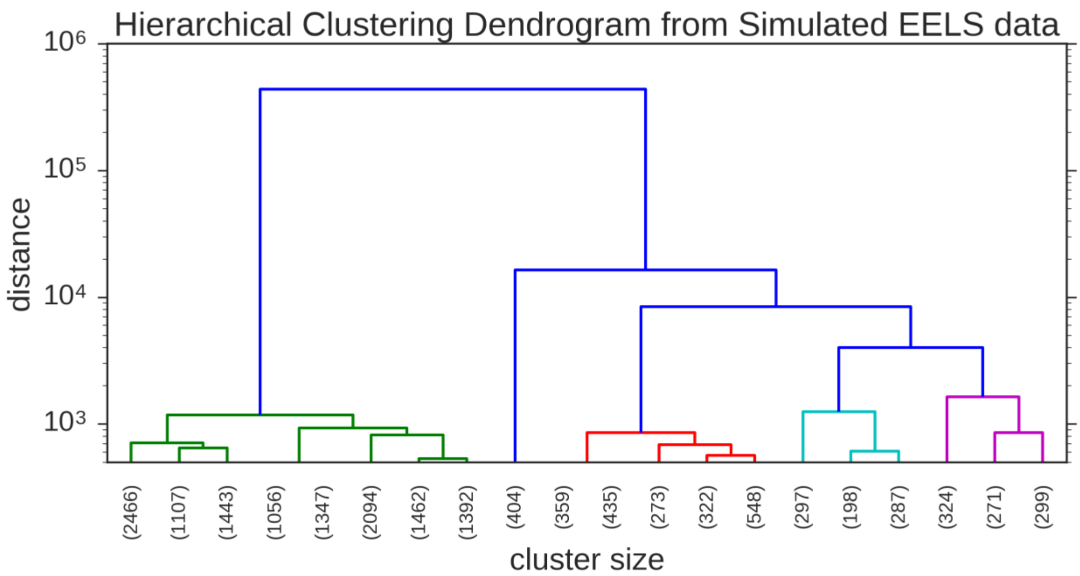


Figure S2. Dendogram plot for the last 20 hierarchical clustering links of the PCA scores from the simulated data decomposition. Horizontal axis represents individual clusters, identified by an id number. A link is represented by ahorizontal line that goes from one cluster id to another. The height at which the links form is the distance between them.

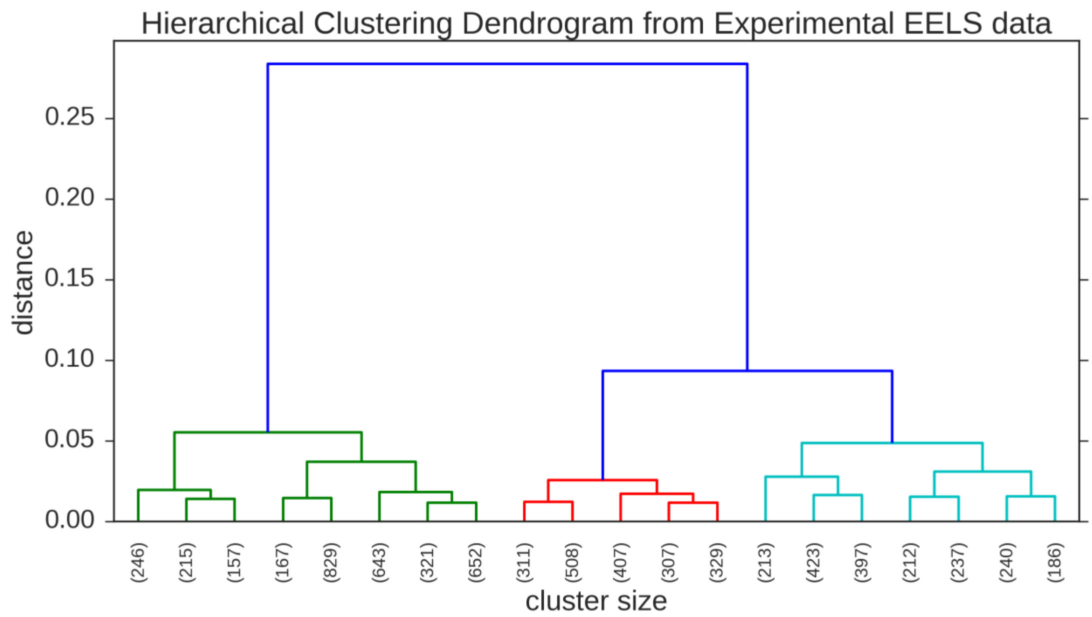
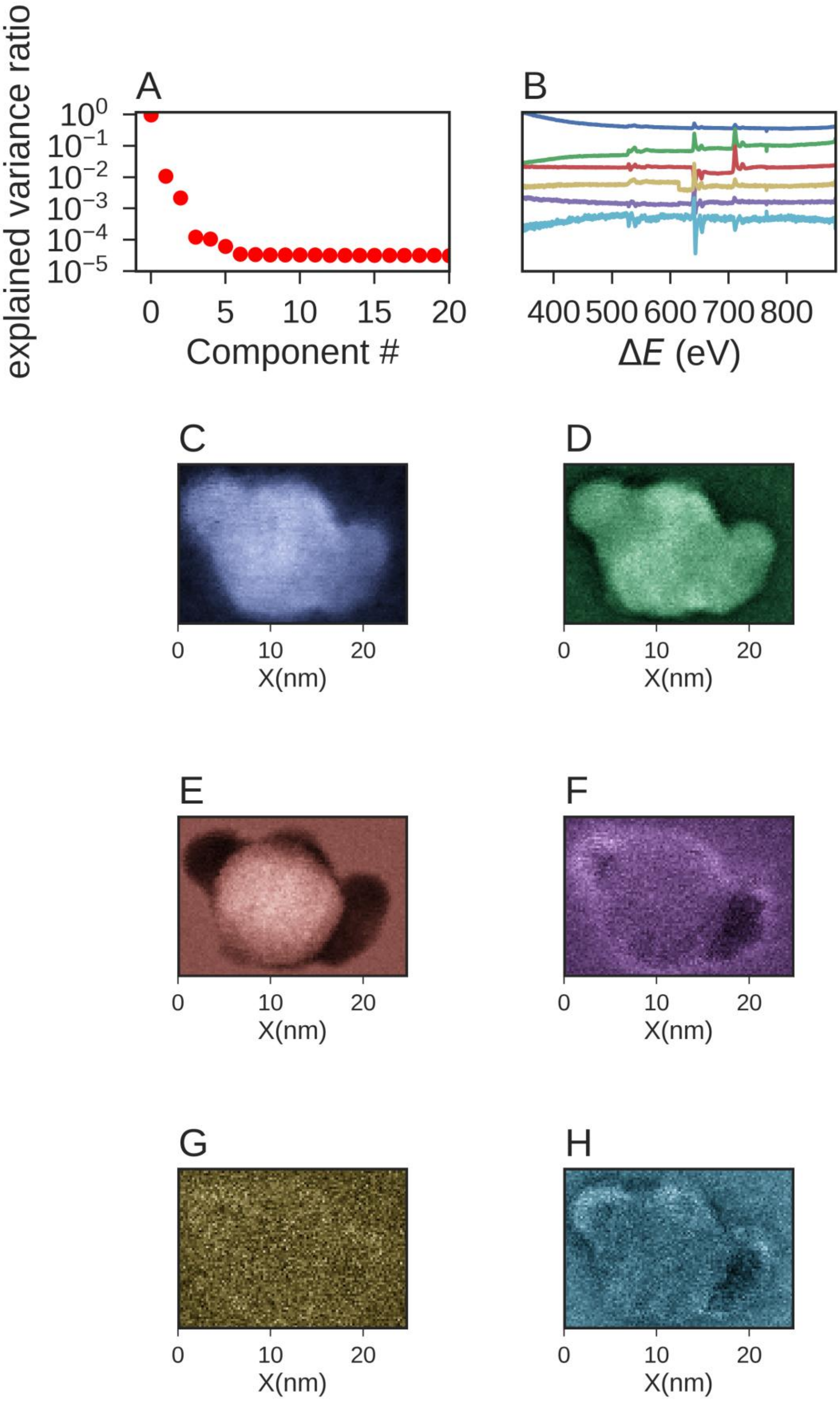


Figure S3. Dendogram plot for the last 20 hierarchical clustering links of the experimental spectrum image. Horizontal axis represents individual clusters, identified by an id number. A link is represented by ahorizontal line that goes from one cluster id to another. The height at which the links form is the distance between the two linked clusters.



A
B
C
D
E
F
G
H
explained variance ratio
10^{0}
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
0 5 10 15 20
Component #
400 500 600 700 800
ΔE (eV)
0 10 20
X(nm)

Figure S4. PCA decomposition results of the experimental spectrum image. A) Normalized variance ratio of each PCA component from the SI of figure 3A. B) The six first factors of the PCA decomposition. C-H) Score maps for each factor. Colour coded.

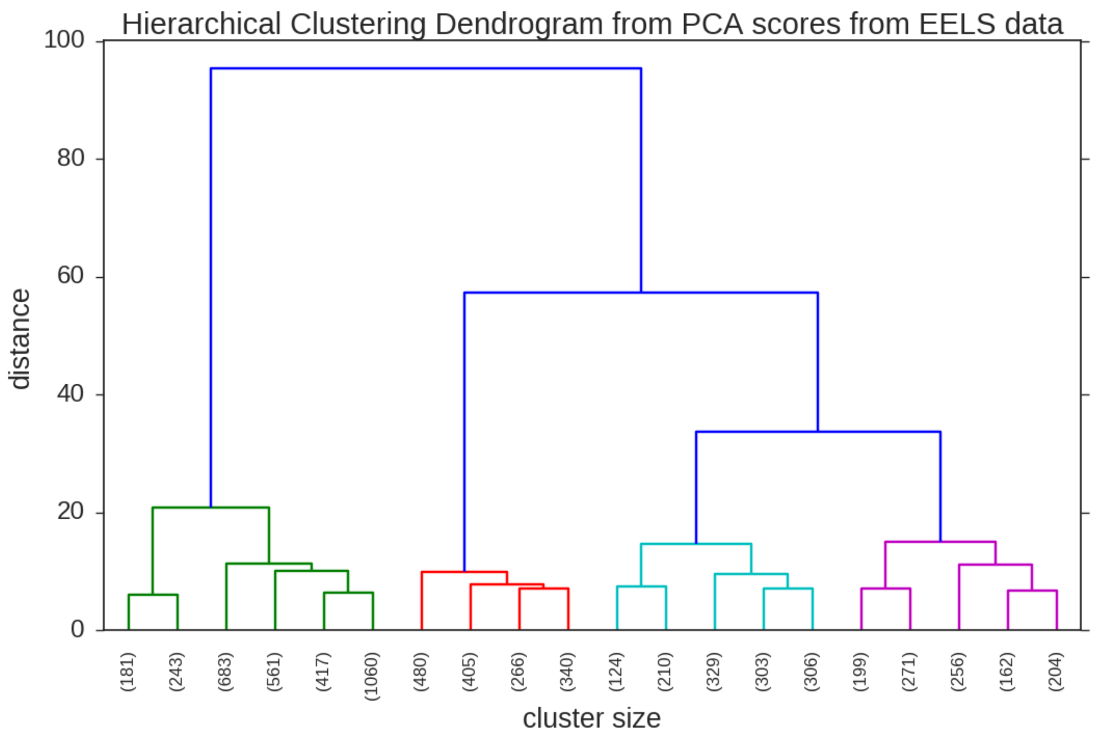


Figure S5. Dendogram plot for the last 20 hierarchical clustering links of the six first PCA scores from the experimental spectrum image. Horizontal axis represents individual clusters, identified by an id number. A link is represented by ahorizontal line that goes from one cluster id to another. The height at which the links form is the distance between the two linked clusters.

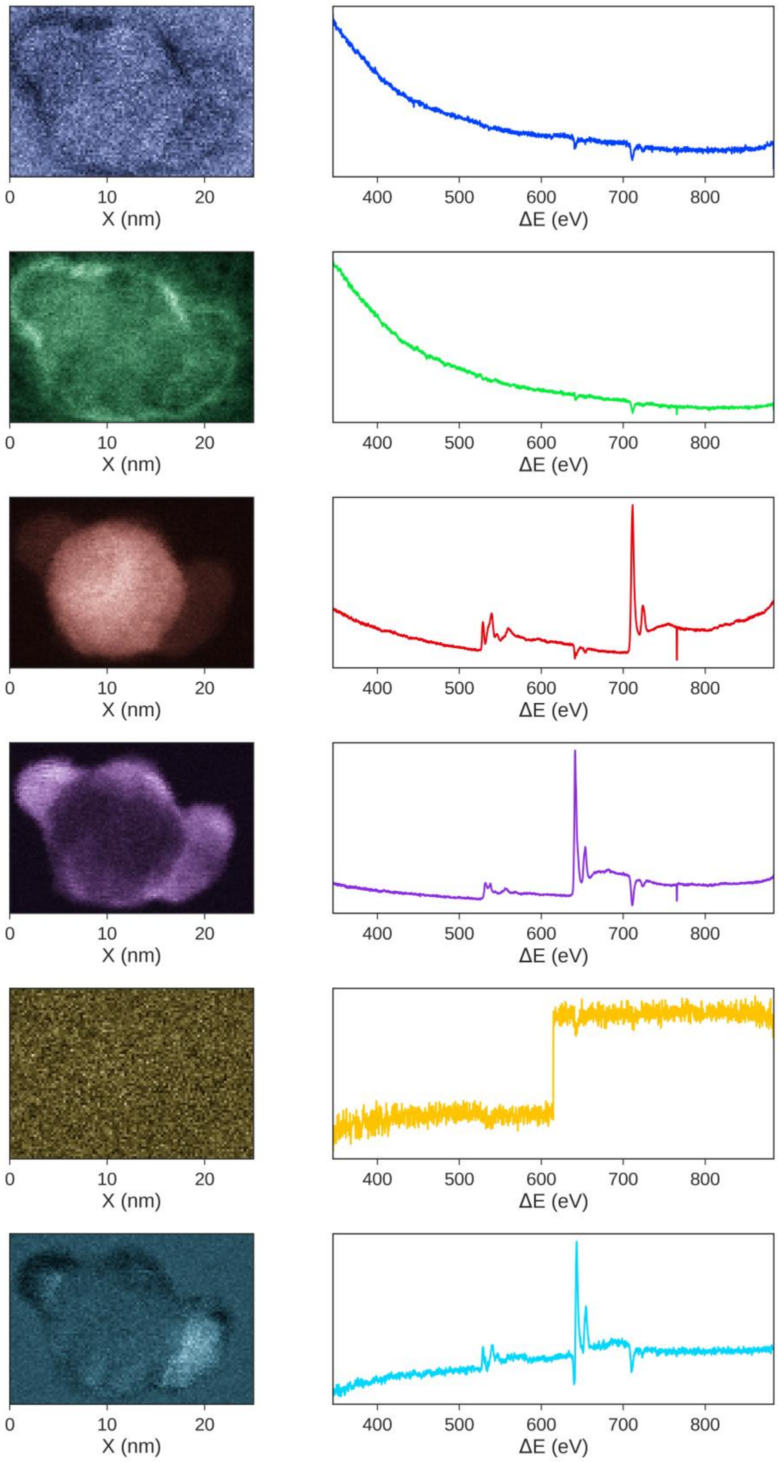


Figure S6. ICA results using 6 components an the default settings of the Hyperspy toolbox.

Code Snippets

Agglomerative hierarchical clustering of a spectrum image.

```
import hyperspy.api as hs
from scipy.cluster.hierarchy import linkage, fcluster

s = hs.load('SI.hdf5')
```

```
sn=s.deepcopy()
sn.data/= np.dstack([s.data.sum(-1)]*s.data.shape[-1]) #normalized image

sn.unfold()
X = sn.data #n by p matrix
sn.fold()

Z = linkage(X) # Z is the clustering tree object.

labels = fcluster(Z,6000,criterion='distance') # labels is a vector that assigns a number to each
#spectre according to which cluster they belong

plt.imshow(labels.reshape(sn.data.shape[:-1])) #show as an image
```

Clustering on PCA score maps.

```
import hyperspy.api as hs
from scipy.cluster.hierarchy import linkage, fcluster

s = hs.load('SI.hdf5')

s.decomposition() # Perform PCA.

scores= s.get_decomposition_loadings()[:n] # Take  score maps of first n components.

X=np.rollaxis(scores.unfold().data , 1)

for i in range(n):

   X[:,i]/=X[:,i].max()  # normalization

Z = linkage(X,) # Z is the clustering tree object.

labels = fcluster(Z,distance,criterion='distance') # labels is a vector that assigns a number to
each  #spectre according to which cluster they belong

plt.imshow(labels.reshape(s.data.shape[:-1])) #show as an image
```

PCA on clusters.

```
import hyperspy.api as hs
from scipy.cluster.hierarchy import linkage, fcluster

results=[]

for i in set(labels): # labels obtained from a previous clustering method

  s.decomposition (navigation_mask=(labels!=i))) #do PCA for each cluster.
```

```
results.append(  s.deepcopy())

s.plot_decomposition_results()
```