

Big data analysis applied to EEL spectroscopy

Author: Joan Sospedra Ramírez

Advisor: Sonia Estradé Albiol; Javier Blanco Portals

Facultat de Física, Universitat de Barcelona, Diagonal 645, 08028 Barcelona, Spain.

Abstract: In order to characterize an unknown sample, big data and machine learning methods are proposed. An electron energy loss (EEL) spectrum image obtained in the transmission electron microscope is analyzed. Applying principal component analysis (PCA) to image EEL spectra, the noise present in the raw data can be discarded, and comparing with existing datasets and alternatively through clustering analysis, the presence of vanadium and oxygen in the sample over a substrate with lanthanum and oxygen can be recognized.

I. INTRODUCTION

1.1 Introduction to EELS:

Electron energy loss spectroscopy (or EELS for short) in the transmission electron microscope (TEM), is an analytical tool that allows matter characterization at subnanometric scale. EELS studies electron beams, the energy of which is well known, once they have interacted with a sample of a given material. These fast electrons energies are of the order of 100keV. The study of the beam after the interaction consists on the classification of the received electrons by their kinetic energy in order to obtain an energy spectrum where the intensity (scattered electrons) is displayed as a function of the kinetic energy loss. ^{[1.1][3]}

A typical spectrum shows these features: ^{[1.1][4]}

- The zero loss peak, centered at zero in the abscissa axis, collects all electrons that have not suffered scattering. The width of this peak is an instrumental function and it is directly correlated with the monochromatism of the incident beam and the aberrations of the optical system.
- The following most notorious peak is the result of conduction electrons plasma resonance, the plasmon. In the solid state free electron model, plasmon excitations can be understood as harmonic oscillations of the conduction band electrons, weakly bound to a fixed ion background. This way, we can relate oscillation frequency and, consequently, its energy, to the gas electron density and effective mass of the carriers. As a result, plasmon peak position is a great indicator of the changes the sample suffers, expressed as carrier number variations.

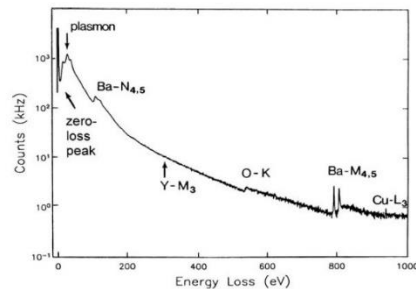


Figure 1. Electron energy loss spectrum of a high-temperature superconductor ($\text{YBa}_2\text{Cu}_3\text{O}_7$), showing zero loss and plasmon peaks and ionization edges arising from each element ^[1]

- Finally, ionization edges of the nuclei, the onset of which describe the elements present in the sample, while the area below the edge is proportional to the amount of atoms that have interacted of this specific element.

All in all, EELS goal is to map the spatial distribution of individual properties shown in individual EEL spectra.

1.2 EELS physics:

In order to introduce the physics of EEL spectroscopy, it is vital to focus on the electron-matter interaction. The kind of interactions that EELS cares about are the inelastic ones, even though elastic ones occur too: while the elastic interactions are useful in order to image the sample, the inelastic ones are those used in spectroscopy. Focusing on core loss physics, inelastic interactions are produced because of the electromagnetic repulsion between the incident electron and the electronic cloud (both internal and external layers): the emitted electron can lend a characteristic amount of energy to the atomic electron primary excitation.

A basic magnitude in scattering theory is the differential cross section, which represents the probability that an incident electron suffers scattering per solid angle unit for a given atom. The measured intensities after the collisions are compared with calculated partial ionization cross sections for limited ranges of energy and momentum transfers in the inner-shell excitation. These partial cross sections have been calculated for typical measurement conditions. [2]

From Morse theory for elastic scattering, Lenz (1954) arrived at an expression that can be read as follows [1.2]

$$\frac{d\sigma_i}{d\Omega} = \frac{4\gamma^2 Z}{a_0^2 q^4} \left(1 - \frac{1}{[1 + (qr_0)^2]^2} \right) \quad (1)$$

where γ corresponds to the relativistic Lorentz factor, a_0 is Bohr's radius, r_0 is the screening radius, equivalent to $a_0 Z^{-1/3}$ according to the Thomas-Fermi model, and q is given by [1.3]

$$q^2 \approx k_0^2 (\theta^2 + \bar{\theta}_E^2) \quad (2)$$

where $k_0 = \frac{2\pi}{\lambda} = \frac{\gamma m_0 v}{\hbar}$ is the incident electrons wavevector magnitude, θ is the dispersion angle and $\bar{\theta}_E = \frac{E}{\gamma m_0 v^2}$ is the characteristic angle associated to an average energy loss \bar{E} . Thus, equations (1) and (2) can be combined into [1.4]

$$\frac{d\sigma_i}{d\Omega} = \frac{4\gamma^2 Z}{a_0^2 k_0^4} \frac{1}{(\theta^2 + \bar{\theta}_E^2)^2} \left(1 - \left[\frac{\theta_0^4}{\theta^2 + \bar{\theta}_E^2 + \theta_0^2} \right] \right) \quad (3)$$

with $\theta_0 = \frac{1}{k_0 r_0}$.

To further proceed, it will be convenient to introduce the Bethe theory (Bethe, 1930), as it provides a convenient and consistent quantum-mechanical basis for electronic excitation in atoms. [5]

Several cross section models have been developed, all of which are modifications to the Bethe approach (e.g., Newbury 1986, Goldstein et al. 1986), which allow us to properly describe the interaction phenomenon through various approaches, but whatever method we use, obtaining a relationship between the signal intensity received in terms of the number of atoms of a given element is now a matter of applying a probabilistic approach to the system. This goal could be reached with the Hartree-Fock method, for instance.

1.3 ELNES physics:

Until now, we have worked in the framework of atomic physics. The next step in order to characterize more thoroughly the intensity variations of the core loss region involves understanding the structure of the material at solid state level. This structure, produced by the second order corrections of the solid state system, is called energy loss near edge structure (ELNES). It gives us data on density of states, coordination or even the kind of bonds of ionized atoms of every species described by the edge. ELNES compiles the stronger oscillations, ranging from 30-50eV from the onset, whereas further oscillations from the edge onset that can extend out several hundred eVs away are under EXELFS domain (extended energy loss fine structure).

This fine structure appears when inner core electrons are promoted to higher states, although the ionization process can impart more than the critical ionization energy E_c needed by the core to be ejected from its inner shell. For instance, phenomena like white lines are characteristic edges that appear in some spectra when the promotion is subject to selection rules from quantum mechanics (L3, L2 of transition metals).

It is because of different densities of unoccupied states from atom to atom that the ELNES allows us to distinguish their coordination.

Hence, in terms of solid state, we can then correlate the fine structure of the edge, and, in particular, the chemical shift (small variations in edge onset) with such variables as oxidation states, atomic charge and coordination with the data from reference materials in order to characterize our sample [6].

1.4 Data processing:

The obtained data analysis is not trivial, though. Because of every packet of data is typically given in the form of a two-dimensional grid (the spectrum image), usually with a magnitude order of hundreds of pixels long, every one of them containing the energy spectra expressed through 2048 channels, this analysis becomes a problem that has to be tackled with big data approaches. As hyperspectral data can even hold more dimensions, such as momenta of the electrons, a way of restraining and easily visualizing said dimensions is very much needed.

The chosen tool to analyze the collected data from the sample is the compilation library for Python Hyperspy^[7]. This programming package, based on the libraries NumPy, SciPy, matplotlib and scikit-learn, will allow us to study, plot and correct spectra and, in last instance, manipulate the multidimensional data collected in the laboratory through mathematical algorithms or analysis methods. Namely:

1.4.1 PCA:

Once our experimental data has been merged together in Hyperspy, our first step in order to apply the statistical treatment is to manipulate it through PCA, or principal component analysis. PCA must arguably be the most popular multivariable analytic method. It consists on finding a new parametric model for the dataset, where every spectrum can be described as a weighted sum of a finite number of components and noise.^[8] PCA looks for the minimum number of variables that describe the original data in order to reduce the problem's dimensionality without the loss of physical information. The model assumes that the problem is linear, and the signal variance is higher than the noise. Therefore, we use PCA as a way to reduce our signal's noise with a clever choice of the components we keep. The result is a clean signal with most of its experimental and background noise removed, but this comes with a downside that has to be noted: the ultimate decision of which components are relevant and which are not resides on the hands of the person who executes the PCA, a decision that is not always easy. On the one hand, leaving too many components still leaves undesired noise, while on the other hand, cutting too many components away means a loss of relevant physical information.^[9]

1.4.2 Cluster analysis:

Cluster analysis (or clustering for short) is a well-known procedure in data science and it aims to classify individual pixel spectra in groups according to similarity in attributes among them. The first step is turning our 3D data into a 2D dataset. This is accomplished through merging the position axes into the new position axis n as follows: $n = X \cdot Y$, which we will treat as our objects, versus our remaining

dimension, the intensity value in each channel $p = E$, which is our attribute. In our new formed matrix $n \cdot p$ each individual spectrum is now a row, and it is now a suitable input for most data clustering algorithms. By considering the spectra in a spectrum image as a collection of p -dimensional points the algorithm can be easily applied to EELS data.^[10] In addition, if we apply a noise reduction method such as the aforementioned PCA, and we only take a subset of components, computation time is greatly reduced. Finally, by assembling all spectra by similarities in shape, we can characterize the material setting the number of clusters, which will describe different zones of the sample. Studying the average spectrum of every cluster, we are now able to map composition vs position. One last thing to note is that clustering does not alter in any way the input data, making it a very powerful resource, because results are components with physical meaning.

II. DATA ANALYSIS

2.1 Early preparation and PCA:

A problem dataset is studied. It corresponds to an interface containing a layer and a substrate of unknown composition.

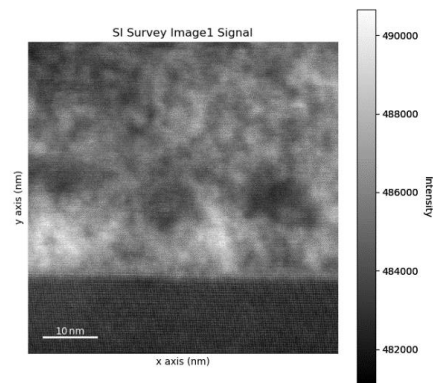


Figure 2. Survey image of the sample.

The first steps in order to characterize the composition and structure of the sample is to discard the noise of the datasets. Our first concern are the spikes. Spikes in the data are very high and slim peaks that are collected and read but don't have an electronic origin. They are mostly produced by cosmic gamma rays that may happen to impact on the detector. By reading the results of the derivatives of the datasets we can easily find high spikes that most certainly do not belong to the EEL spectra and eliminate them.

Once our spectra are only made up by our EEL data, we can now apply PCA.

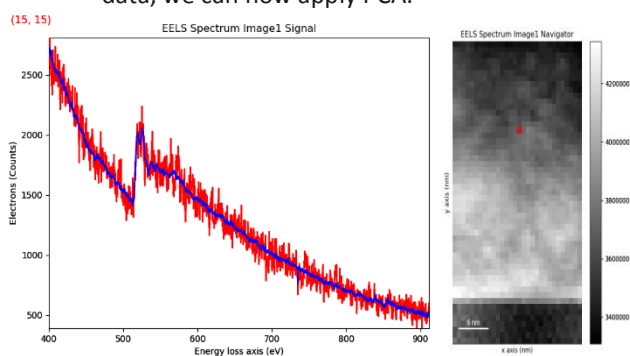


Figure 3. Comparison between raw data (red) and PCA-treated data (blue) in the sample region.

Even though it may be tempting to correct the visible defects of the detector in the form of burns or dead pixels, it would result in a fatal error. If we manipulate in any way the original dataset in order to embellish it, the results will strongly vary. Hence, we will first treat the data looking for clever ways to avoid the defects instead of correcting them. The correction will come once we have the final plots.

2.2 Core-loss edges identification:

To facilitate the study of the edges in the core-loss region, it is useful to subtract the background, thus leaving above the energy loss axis only the electron counts coming from the interaction with the atom nuclei that we want to study. Looking at Figure 3, we can appreciate a detector burn at values prior to 469 eV approximately, adding more counts than the real amount. Knowing this, we can interpolate an exponential curve to fit the background, but instead of taking the whole available background to plot a more accurate function, we will only take the region between the burn and the first edge. With this decision we are sacrificing precision for the fitting function to

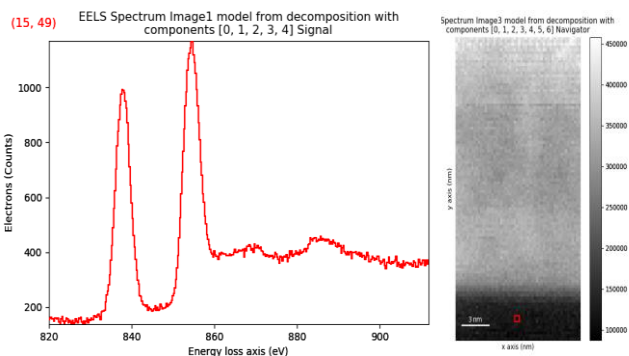
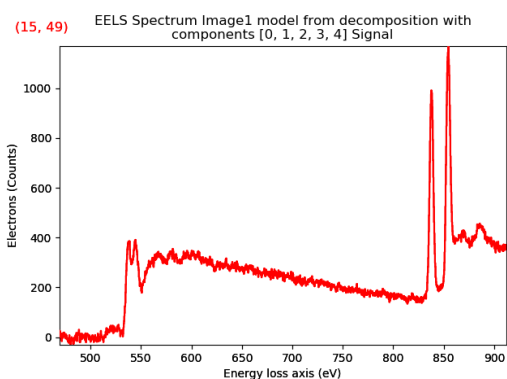


Figure 4. (1) PCA-treated data of the substrate region of the sample. (2) Detail of the 850 eV region. Background removal treatment has been applied again to the Figure 4.1 in order to obtain this plot. Exponential fitted with the data from 700 and 800 eV

gain physical meaning, not using the burnt area.

The next step is to identify the nuclei corresponding to these edges.^[13] Edge identification can be difficult when presented with complex spectrum edges, plural scattering or close lying edges.^[11] The best identification tool is comparing with existing data sources, such as the online EELS atlas ^[12]. Therefore, examining the species from Figure 4, we can easily identify the lanthanum white lines: lanthanum shows its two characteristic

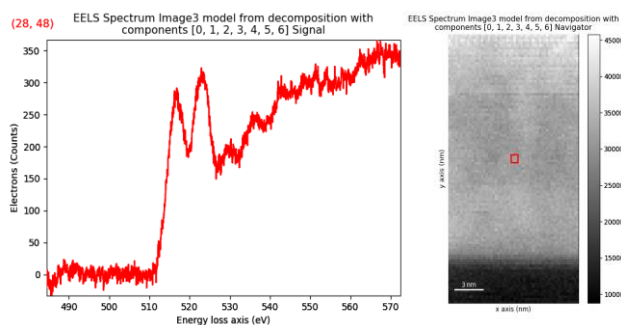


Figure 5. PCA-treated data for the indicated region in the layer. $M_{4,5}$ edges at 849 and 832 eV respectively.

In a similar way, in the same spectrum image and focusing on the 520 eV area we can study the other family of edges. Using a database, we can identify vanadium ($L_{2,3}$ edges at 521 and 513 eV respectively) as well as oxygen. For reference, a similar oxygen edge shape can be found in the spectra for the thin layer of titanium oxide (TiO_2) in ^[12.1].

To sum things up, this sample is made of a layer of a vanadium oxide over a lanthanum oxide.

2.3 Clustering application:

Once the clustering algorithms have been run for two clusters, we obtain the following figures

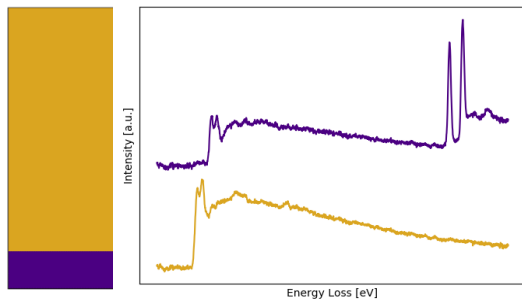


Figure 6. (Left) Intensity labels: average of the spectra (arbitrary units) in terms of the energy loss in eV. The height of the labels is arbitrary, just for comparison reasons. (Right) Representation of the cluster distribution for the sample for two clusters.

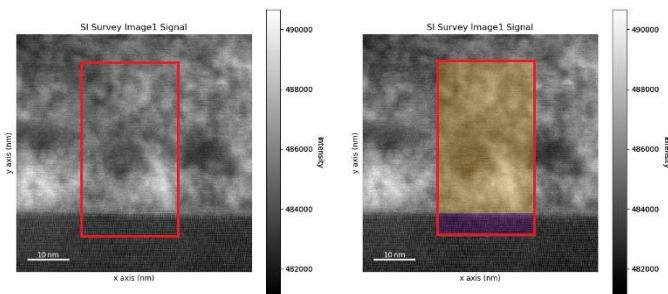


Figure 7. Comparison between the survey image and the survey image with the cluster distribution superposed.

Clustering analysis for two clusters produced Figure 6. Each clusters show the two principal species conforming the sample: a typical spectrum for vanadium oxide represented in yellow and lanthanum oxide in purple.

III. CONCLUSIONS AND DISCUSSION:

I had a sample of material with an unknown composition. After removing the spikes present in the spectra and then applying PCA to remove its inherent noise, I could identify its edges, compare them with online databases and thus determine its composition. Alternatively, applying clustering, two intensity labels have been compiled to further characterize the average composition of the layer and the substrate. The problem has in fact been resolved, having now the core loss edges for both the sample and the substrate identified.

The sample clearly shows vanadium and oxygen whereas the substrate contains lanthanum and oxygen. [12.2] [12.3] [12.4]

IV. BIBLIOGRAPHY:

4.1 Books:

[1] Egerton, R. F. *Electron Energy Loss Spectroscopy in the Electron Microscope*. Springer. 3rd ed. New York: 2011.

[1.1] Chapter 1; [1.2] pg 124, eq (3.13); [1.3] pg 124, eq (3.14); [1.4] pg 125, eq (3.15)

[2] *Cross Sections for Inner-Shell Ionization by Electron Impact*. Xavier Llovet, Cedric J. Powell, Francesc Salvat, and Aleksander Jablonski, pg 40.

[5] *Electron Impact Ionization*, Edited by T.D. Märk and G.H. Dunn, Springer-Verlag Wien GmbH, pg 199.

[6] *Transmission Electron Microscopy, A Textbook for Materials Science*. David B. Williams & C. Barry Carter. Springer. (Chapters 39 and 40)

4.2 Scientific papers and thesis:

[8] *EEL spectroscopic tomography towards a new dimension in nanomaterials analysis*, Luis Yedra, 2012 (pg. 2)

[9] *Towards a new dimension in analytical TEM: EELS, tomography and the spectrum volume*. Luis Yedra, PhD thesis. Chapter 4.3.2

[10] *Clustering analysis strategies for electron energy loss spectroscopy (EELS)*. Pau Torruella, Marta Estrader, Alberto López-Ortega, Maria Dolors Baró, Maria Varela, Francesca Peiró, Sònia Estradé.

Websites:

[3] <http://www.gatan.com/techniques/eels>

[4] <http://www.eels.info/about/techniques/eels-0>

[7] <http://hyperspy.org>

[11] <http://www.eels.info/uses/identify-elements-within-sample>

[12] <http://www.eels.info/atlas>

[12.1] <http://www.eels.info/atlas/titanium>

[12.2] <http://www.eels.info/atlas/vanadium>

[12.3] <http://www.eels.info/atlas/lanthanum>

[12.4] <http://www.eels.info/atlas/oxygen>

[13] <http://www.eels.info/how/quantification/workflow/choose-suitable-edges-quantification>