SLAC-TN-12-024

Using Dynamic Quantum Clustering to Analyze Hierarchically Heterogeneous Samples on the Nanoscale

Allison Hume

Office of Science, Science Undergraduate Laboratory Internship (SULI)
Princeton University
SLAC National Accelerator Laboratory
Menlo Park, California

August 19, 2011

Prepared in partial fulfillment of the requirements of the Office of Science, Department of Energy's Science Undergraduate Laboratory Internship under the direction of Marvin Weinstein at the SLAC National Accelerator Laboratory.

Participant: ____________________________________
Signature

Research Advisor: ____________________________________
Signature

Work supported by US Department of Energy contract DE-AC02-76SF00515.

Table of Contents

Abstract

Using Dynamic Quantum Clustering to Analyze Hierarchically Heterogeneous Samples on the Nanoscale. ALLISON HUME (Princeton University, Princeton, NJ 08544) MARVIN WEINSTEIN (SLAC National Accelerator Laboratory, Menlo Park, CA 90425)

Dynamic Quantum Clustering (DQC) is an unsupervised, high visual data mining technique. DQC was tested as an analysis method for X-ray Absorption Near Edge Structure (XANES) data from the Transmission X-ray Microscopy (TXM) group. The TXM group images hierarchically heterogeneous materials with nanoscale resolution and large field of view. XANES data consists of energy spectra for each pixel of an image. It was determined that DQC successfully identifies structure in data of this type without prior knowledge of the components in the sample. Clusters and sub-clusters clearly reflected features of the spectra that identified chemical component, chemical environment, and density in the image. DQC can also be used in conjunction with the established data analysis technique, which does require knowledge of components present.

Introduction

The ability to study hierarchically heterogeneous materials on the nanoscale can lend insight to problems ranging from the chemistry of batteries to techniques of ancient artists. A hierarchically heterogeneous sample is a sample that contains multiple chemical components that appear homogenous on the nanoscale, but heterogeneous on a micron or millimeter scale. The Transmission X-ray Microscopy (TXM) group at beam line 6-2c in the Stanford Synchrotron Radiation Light-source (SSRL) has the ability to image such samples with a 30 nanometer resolution and field of view large enough to see the hierarchical structure.

The data collected for a sample is a set of images taken over a range of energies from which the TXM group can determine a full X-ray Absorption Near Edge Structure (XANES) spectrum for each pixel. XANES goes beyond identifying the elements present in a sample, determined by the edge-jump in the spectrum, and provides information about the surrounding chemistry, such as the oxidation state of the element. This information is present in the fine structure in the energy range higher than the edge-jump. [1] The TXM group, however, is not able to analyze this data without knowing what chemical components are present in the sample. The current data analysis procedure compares the absorption spectrum for each pixel to spectra for compounds known to be in the sample using R^2 fitting. A colored map of the sample is created by assigning the known spectra to colors such as red and blue and coloring each pixel according to how similar it is to the "red" or "blue" spectra. Any pixels with extremely poor R^2 values are examined to find unknown components. This method is effective for samples in which the composition is known and the only information needed is the distribution, but a new method is necessary for samples of unknown composition.

Dynamic Quantum Clustering (DQC) is a method of data mining that does not require any previous information about the composition of the sample. It performs unsupervised clustering to identify energy spectra that have similar characteristics. To test the applicability of DQC to TXM-XANES data, a sample was used that had been analyzed by the original method as well. The sample came from a piece of ancient Roman pottery made from hematite (iron(III) oxide) and hercynite (iron(II) aluminate). The images are taken as a cross-section of the interface between the two oxides with hematite on the surface and hercynite underneath. This data set consists of a series of 146 images, each with close to one million pixels, which determines the spectra for each pixel. The spectra are noisy and the combination of this noise and the large data size creates a significant data mining challenge to find groups of pixels with similar spectra.

Methods

DQC is an algorithm that maps a problem of unsupervised clustering to a problem in quantum mechanics and provides a visual animation of the cluster formation. An outline of the algorithm appears below and a full description of the theory and methods of computation can be found in *Dynamic quantum clustering: a method for visual exploration of structures in data* by Marvin Weinstein and David Horn. [2]

In DQC each data point is described as a Gaussian wave function in an n-dimensional parameter space. The full data set is then associated with the sum of these individual Gaussians:

$$\Psi(\vec{x}) = \sum_i e^{\frac{-(\vec{x}-\vec{x_i})^2}{2\sigma^2}} = \sum_i \psi_i(\vec{x_i}) \tag{1}$$

Choosing a larger value for σ causes the individual wave functions to have more overlap while a smaller value decreases overlap. This in turn affects the level of definition of the extrema of the full wave function.

The wave function is then taken to be the ground state of the Hamiltonian:

$$(-\frac{\sigma^2}{2}\boldsymbol{\nabla}^2 \ + V(\vec{x}))\Psi(\vec{x}) = E\Psi(\vec{x}), E = 0 \tag{2}$$

The full wave function therefore determines some potential $V(\vec{x})$ that will have minima determined by the maxima of the wave function, which correspond to the most most dense regions of the data in parameter space:

$$V(\vec{x}) = \frac{\sigma^2}{2\Psi(\vec{x})}\boldsymbol{\nabla}^2\Psi(\vec{x}) \tag{3}$$

The potential, is less sensitive to the value of σ than the full wave function so the choice of value for that variable is less important than it is for other data mining methods that use a sum of Gaussians.

The individual data points are them evolved in time according to the time-dependent Schrödringer equation:

$$i\frac{\partial\psi_i(\vec{x_i},t)}{\partial t} = (-\frac{\mathbf{\nabla}^2}{2m} + V(\vec{x}))\psi_i(\vec{x_i},t) \tag{4}$$

The value chosen for the mass of the "particle" has an effect on the amount of tunneling that occurs as the data evolve, as well as the speed at which the evolution occurs.

The data is a matrix: each row corresponds to the 146 point energy spectrum of a single pixel. After removal of the pixels for which there was no data, slightly over half a million pixels were left. Singular value decomposition (SVD) was then performed. SVD is a method of factorizing a matrix that allows the original matrix to be approximated as a sum of a series of terms that show increasing detail. In this case, since each row of the matrix is a spectrum, recreating the matrix from a subset of the 146 singular components reduced the noise in the spectra. The more singular components used to recreate the original matrix, the closer the spectra will be to their original shape. The first five components were used for the rest of the initial analysis. Five features were enough to retain the important features in the shape of the spectra, while reducing the noise of the spectra greatly. This can be seen in Figure 1 which shows an example spectrum taken from the original matrix in green and the same spectrum reconstructed from the first give components in black.

DQC is able to operate in high dimension; however, the speed of computation scales linearly with the number of dimensions. So reduction from 146 to five dimensions in the data matrix increased the speed of computation. The most important step taken to reduce the time of computation, however, was the selection of template states. Template states are essentially an approximate basis: a subset of data points, linear combinations of which can reproduce all of the other states up to a given accuracy. By choosing about 1500 states that spanned the full data and only evolving these states, the computation became possible on a desktop workstation. A description of the DQC Maple library, including further detail on the process of selecting template states can be found on Weinstein's website. [3] In this analysis, all of the states were used to construct the potential, the template states were evolved through 75 time-steps using the potential, and the remaining states were expressed as approximate linear combinations of the template states to fully evolve the data.

After the 75 frame animation of the evolution was completed, the result was examined, and clusters and sub-clusters were visually identified. In this analysis, a cluster is considered as a group of data points which is separate from all others in any dimension and a sub-cluster is any visible structure within a cluster such as connected strands. The data were then colored based on the visible structure, the original image was reconstructed from those colors, and the clusters were used to understand the structure and distribution of the data. A selection of reconstructed spectra, as well as the average spectra, was examined from each cluster and sub-cluster to determine the aspects of the spectra that affected the clustering of the data. Four clusters and the sub-structure of two of the clusters were examined in depth. The colors used are shades of red, shades of blue, and shades of green. The three shades of red (dark red, red, and pink) and the shades of blue (blue and cyan) each correspond to the sub-structure of a cluster while the shades of green (green and dark green) correspond to two clusters.

Results

As DQC had never been used for this type of data, the first important result was that the data clustered. Figure 2 shows the final frame of the evolution and illustrates how both the clustering and sub-clustering appear. It is important to note that the size of the cluster in the final frame does not necessarily correspond to the number of points it contains. In theory, each isolated point in the final frame must be considered as an individual cluster. Watching the animation is useful to estimate the size of clusters. Figure 3 shows the final frame again but illustrates how the clusters and sub-clusters were colored. In practice, most of the isolated points were grouped with other clusters.

The picture that was recreated based on the clustering of the data appears in Figure 4(a) colored according to how each pixel was colored in the final frame of the animation. It is important to note that no information about the location of the pixels in the image was used during the clustering process. Figure 4(b) shows the picture that was created from the original data analysis method for comparison. Information about pixel location *was* used by the TXM group to create this picture. Again, the different shades of color in the DQC picture come from the structure of the data. The different shades of red and green in the original, however, were added by examining the amplitude of the curves after the fitting was performed and adding white to the lower amplitude curves to represent lower intensity pixels. The blue piece of the picture was identified in the original analysis through the poor fit those pixels had with both the hematite and hercynite spectra. In the DQC analysis those pixels clustered to a point within four frames and remained separated for the rest of the animation.

Examining the spectra in each cluster provided information about what aspects of the curves were important in the clustering process. Several groups of clusters and sub-clusters were seen to have spectra of similar shape, but different intensities. Figure 5 shows the averages of these groups, as well as a selection of curves from each group to illustrate how the intensity of spectra comes out in the clustering.

Most of the large clusters corresponded to groups of spectra with different shapes. Figure 6 shows 100 random spectra from each of the three clusters whose spectra had the most distinct shapes. Upon examining the spectra in the green and red clusters, it was confirmed that the green spectra correspond to the spectra of hematite, and the spectra in the main red band correspond to the spectra of hercynite. The locations of hematite and hercynite in the image (locations of green and red pixels in Figure 4(a)) matched what was expected based on the original analysis. The information about the location of hematite and hercynite, although confirmed by the other analysis method, were therefore determined purely from similarities in spectra in the DQC analysis.

It is clear (both in Figure 6 and Figure 5) that the green spectra are tightly grouped while the red and blue spectra contain a much larger spread of curve shapes. This information is reflected in the sub-clustering of the three clusters. The green cluster shows no sub-structure while the red cluster shows three large strands, among other features, and the blue shows a tight node and two diffuse strands. The sub-structure of the red and blue clusters were examined in depth. One strand of the red cluster was determined to correspond to a different chemical environment, as seen in

Figure 7. Figure 7(a) shows only the red cluster in a two dimensional projection of the last frame of the animation with one strand colored yellow. Figure 7(b) shows 20 random spectra from both the red strand and the yellow strand. It shows that the yellow strand corresponds to an apparently different chemical environment from the hercynite from the rest of the red. This is visible in the image in Figure 4(a) as the yellow pixels mostly appear on the interface between the hematite and the hercynite. The blue cluster has a band that resembles the spectrum of mostly pure iron but also contains curves that resemble hercynite. Although the blue cluster appeared as a point on the animation by the fourth or fifth frame, the sub-structure of this cluster is visible with the other clusters removed. Figure 8(a) contains a view of the sub-structure of this cluster from the fifth frame of the animation. Figure 8(b) contains sample spectra from the different components of the structure and Figure 8(c) contains the average spectra from the two components. Both of these plots show that the curves more closely resembling hercynite or pure iron separate in the sub-clustering. It is interesting to note in the image in Figure 4(a) that the light blue or pure iron appears as a coherent band.

Discussion and Conclusion

The DQC algorithm successfully identified structure in the XANES data. The animation of the evolution shows detailed clustering and sub-clustering which correspond to physically important information in the curves. Upon examination of the spectra present in the various clusters it appears that the most important information in forming the clusters comes from the shape of the pre-edge and edge jump. Thus, different components are the first groups to cluster out. Clusters containing spectra for hematite, hercynite, and relatively pure iron contained pixels from areas where those components were expected to appear based on the original analysis, which confirms that DQC is able to correctly identify different chemical components in the data. Some sub-structure of clusters identified the chemical environment of a compound that is present. This was seen in the mapping of apparent differences in the hercynite spectra to the visible strands of the red cluster.

The other information that is important in the cluster formation is the intensity of the spectra.

This was seen in how well some separate clusters and large sub-features corresponded to the intensity information that had been added in the original picture. Although learning that clustering occurs partially based on intensity was informative for understanding how DQC responds to XANES data, in the future clustering will most likely be performed on normalized data. This is because intensity is not physically interesting, and it can be added to the image afterwards, in the same way it was added to the TXM image. Clustering on normalized data will most likely simplify the structure that needs to be examined which will make it easier to find physically interesting information.

The blue cluster (pixels containing mostly iron) represented an example of a "strong" cluster, or a cluster that forms almost immediately and does not join other clusters later. It was important to see that DQC was sensitive to this cluster because iron and hercynite have similar spectra and the pixels in the iron cluster represented less than 1% of the data set. This was a confirmation of how well DQC works with this type of data. The characteristic differences in the spectra is barely visible to the eye when examining the spectra reconstructed from five SVD components but the algorithm is sensitive enough to small changes in the spectra that this tiny subset of the data clustered immediately and remained separate from the data.

It is also important to note that the two data analysis methods can be used well in tandem - DQC provides information about what parts of the image might be interesting to study in depth but, by using the traditional method to create a second picture, the user can also identify areas of the image that should be examined further in the DQC animation. Any area of the image identified by the user can be examined in the animation and so the user can request specific information such as how a smaller group of pixels clusters. The differences between the two pictures are also important and will generally reflect areas that should be examined in depth. This provides another motive for using the two data analysis techniques together. The traditional method is currently faster but DQC is capable of providing more detail and work is continuing to speed up the algorithm. DQC holds great potential for use in the analysis of XANES data in the TXM group.

Acknowledgements

I would like to thank my mentor Marvin Weinstein for all of his help and support as well as Apurva Mehta, Florian Meirer, and Yijin Liu for their time and assistance. I would also like to thank the Department of Energy for funding this program and for everyone at the SLAC National Laboratory for being welcoming and helpful. Thank you to Stephen Rock, Maria Mastrokyriakos and Anita Piercy for all of the work they put into making the summer go smoothly.

REFERENCES

[1] F. Meirer, Y. Liu, A. Mehta. Mineralogy and morphology at nanoscale in hierarchically heterogeneous materials. June 24, 2011.

[2] M. Weinstein, D. Horn. Dynamic quantum clustering: a method for visual exploration of structures in data. SLAC-PUB-13759.

[3] M. Weinstein. DQCOverview1. `http://www.slac.stanford.edu/~niv/index_files/DQCOverview1.html`

FIGURES

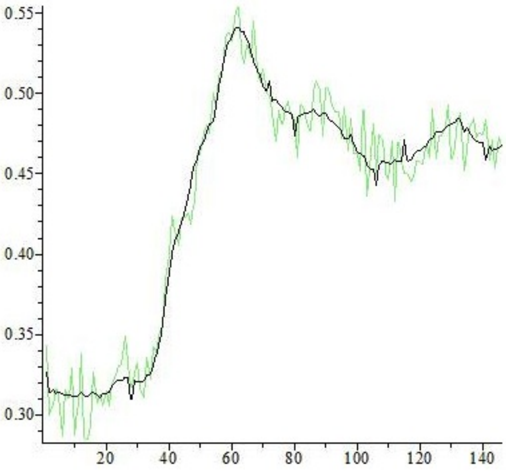
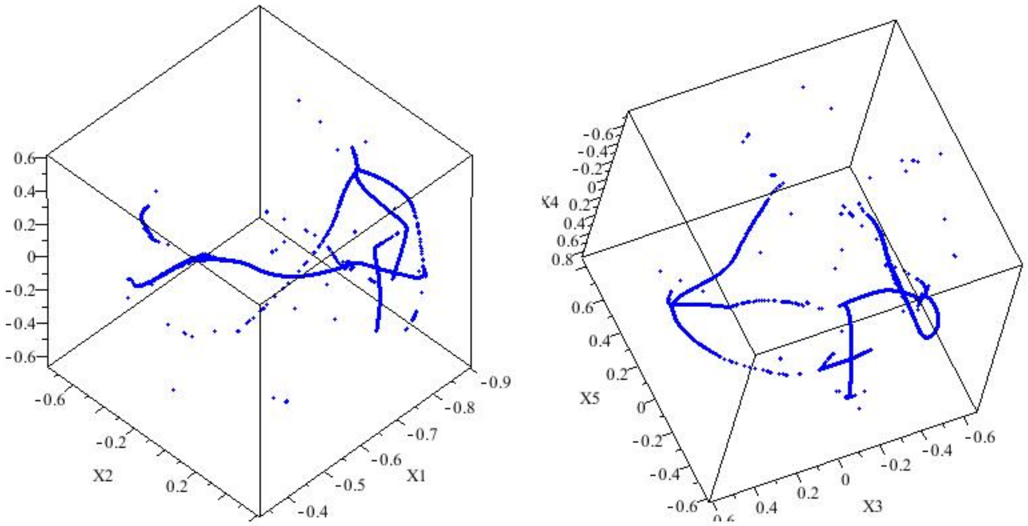


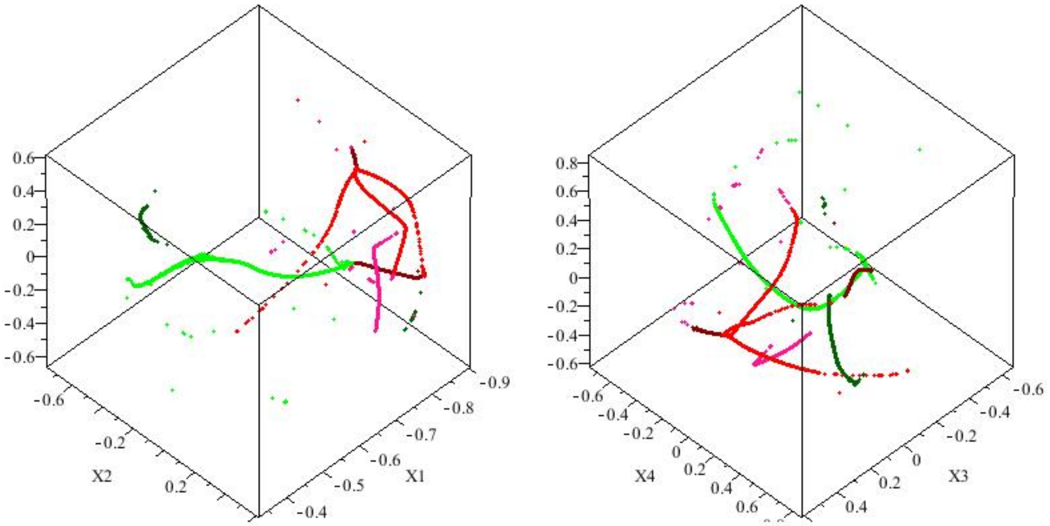
Figure 1: In this illustration of SVD reconstruction the black curve is a sample spectrum taken from the data and the green curve is the same spectrum reconstructed from the first five singular components.



(a) Dimensions 1, 2, 3

(b) Dimensions 3, 4, 5

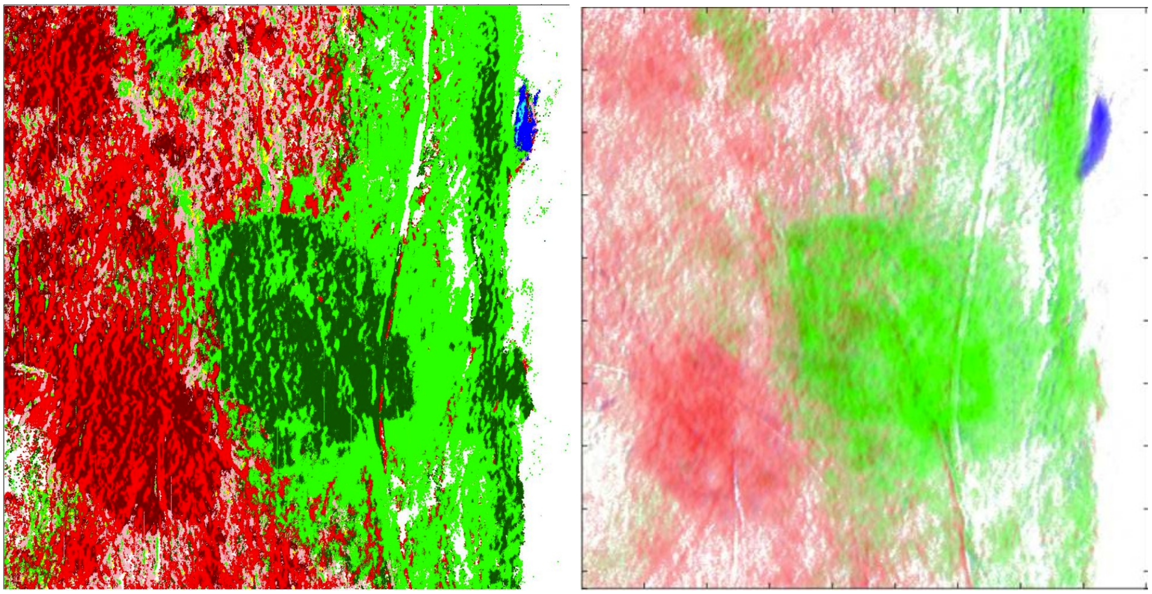
Figure 2: Two views of the last frame of the DQC animation that show all five dimensions of the data and illustrate how the data separated into clusters, as well as how many of the clusters contain visible sub-structure.



(a) Dimensions 1, 2, 3

(b) Dimensions 3, 4, 5

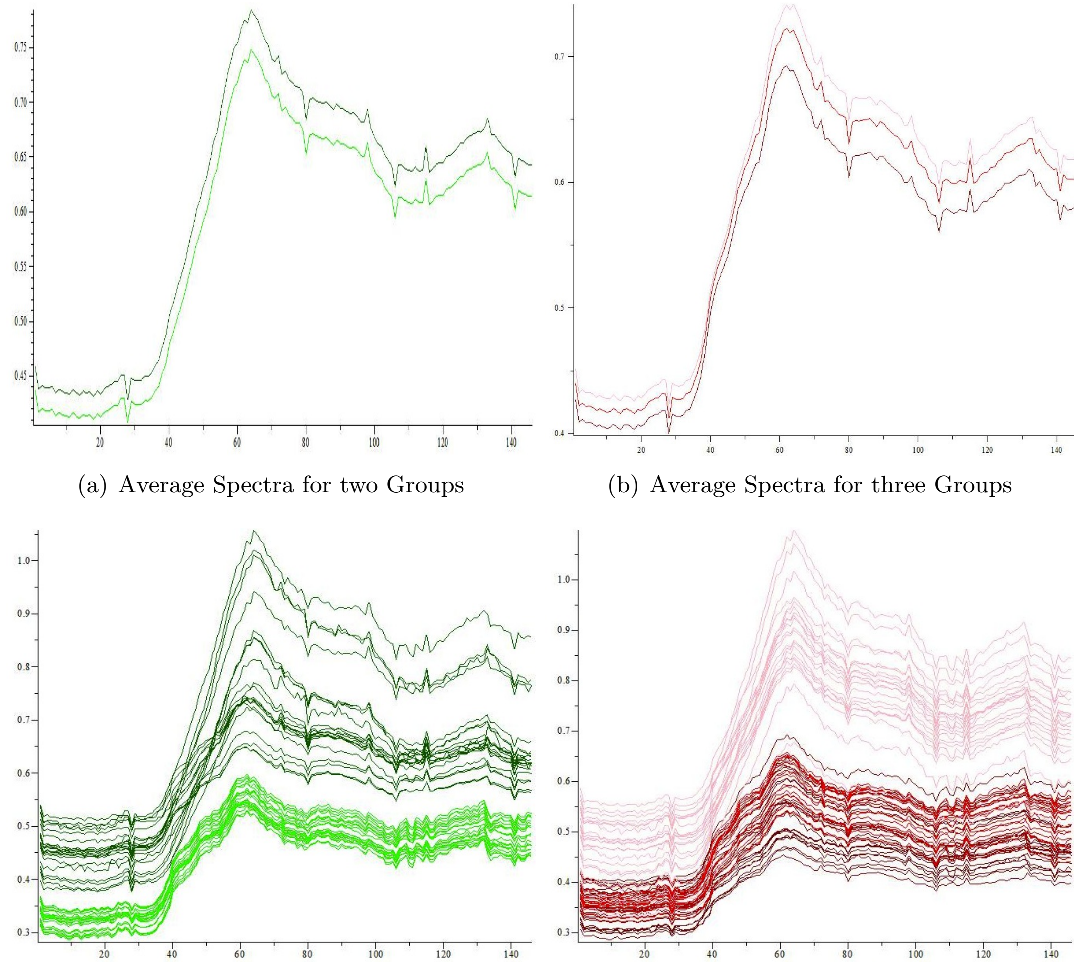
Figure 3: Two views of the last frame of the DQC animation that show all five dimensions of the data and illustrate how clusters and sub-clusters were colored.



(a) Image Created by DQC Analysis

(b) Image Created by Fitting Analysis

Figure 4: Two images of the data with the left image colored based on the unsupervised clustering of the data using DQC analysis and the right image colored based on R^2 fitting to previously known components.



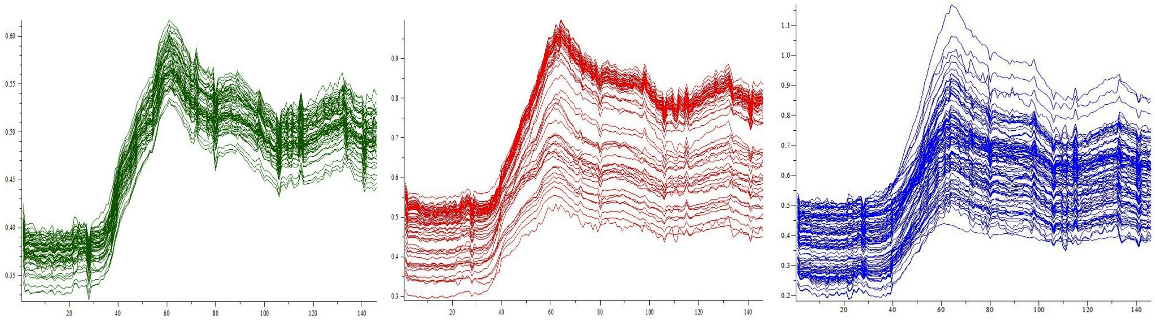
(a) Average Spectra for two Groups

(b) Average Spectra for three Groups

(c) 100 Random Spectra for two Groups

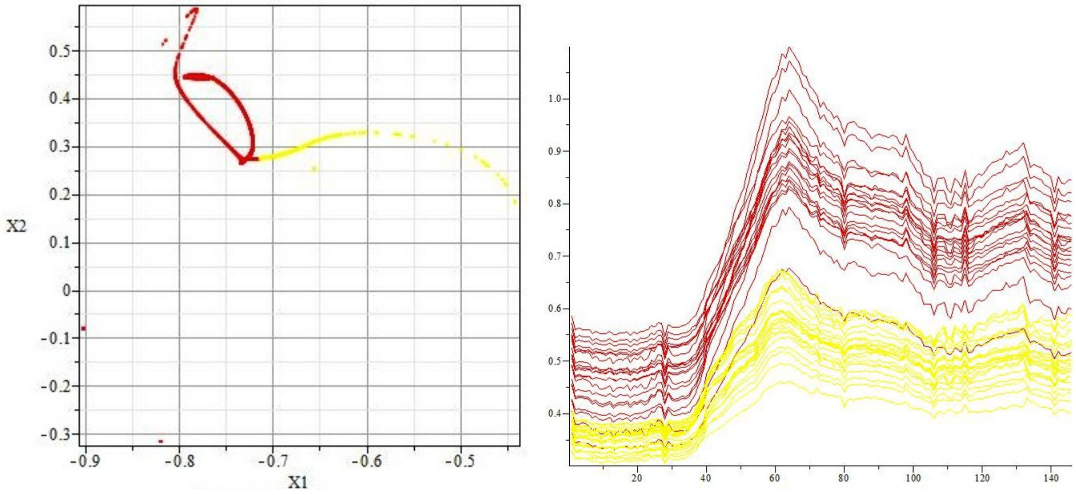
(d) 100 Random Spectra for three Groups

Figure 5: The amplitude differences seen between spectra of several clusters and sub-clusters that have spectra of similar shape. There were two major intensities seen in the shape in green and at least three seen in the shape in red. The different intensities are illustrated using the average reconstructed spectra for a group and 100 random reconstructed spectra for each group.



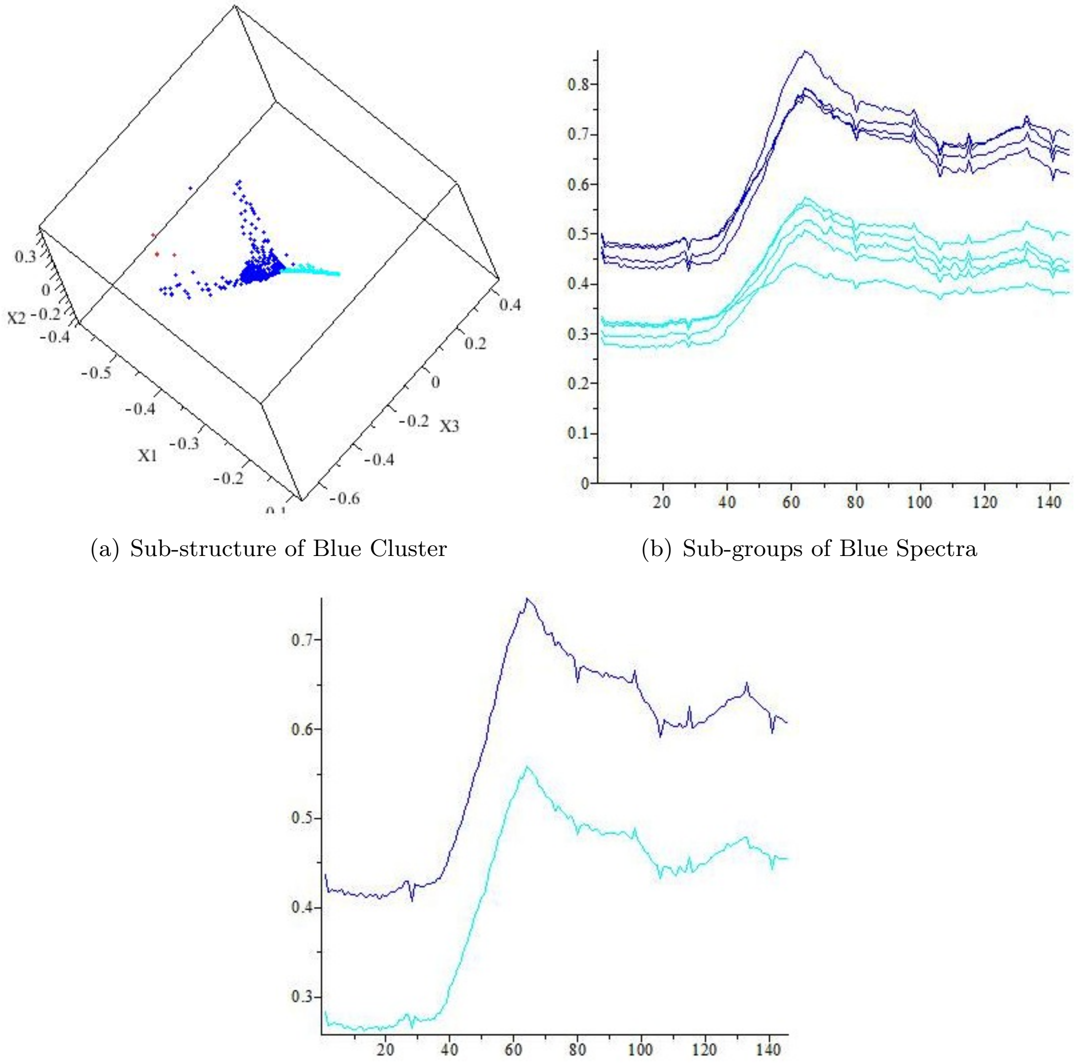
(a) Hematite (b) Hercynite (c) Mostly Iron

Figure 6: 100 random reconstructed spectra from three clusters that illustrate the three main spectra shapes, and therefore the three main chemical components seen in the data.



(a) Sub-structure of Red Cluster (b) Sub-groups of Red Spectra

Figure 7: The sub-structure of the red cluster is visible in the last frame of the animation. This sub-structure reflects differences in the shape of the spectra, corresponding to different phases of hercynite.



(a) Sub-structure of Blue Cluster

(b) Sub-groups of Blue Spectra

(c) Averages of Sub-groups of Blue Spectra

Figure 8: The sub-structure of the blue cluster is visible in the fifth frame of the animation when the rest of the data is removed. This sub-structure reflects differences in the shape of the spectra, corresponding to how much iron is present. Curves more closely resembling hercynite or pure iron separate in the sub-clustering. The lighter blue curves show the flattening of the spectra after the edge-jump that is characteristic of iron.