Cluster analysis in soft X-ray spectromicroscopy: Finding the patterns in complex specimens

M. Lerotic a,⁎, C. Jacobsen a, J.B. Gillow b, A.J. Francis b, S. Wirick a, S. Vogt c, J. Maser c

a Department of Physics and Astronomy, State University of New York at Stony Brook, Stony Brook, NY 11794–3800, USA
b Environmental Sciences Department, Brookhaven National Laboratory, Upton, NY 11973, USA
c Advanced Photon Source, Argonne National Laboratory, Argonne, IL 60439, USA

Available online 26 February 2005

Abstract

Soft X-ray spectromicroscopy provides spectral data on the chemical speciation of light elements at sub-100 nanometer spatial resolution. If all chemical species in a specimen are known and separately characterized, existing approaches can be used to measure the concentration of each component at each pixel. In other situations such as in biology or environmental science, this approach may not be possible. We have previously described [M. Lerotic, C. Jacobsen, T. Schäfer, S. Vogt, Ultramicroscopy 100 (1–2) (2004) 35] the use of principle component analysis (PCA) to orthogonalize and noise-filter spectromicroscopy data, and cluster analysis (CA) to classify the analyzed data and obtain thickness maps of representative spectra. We describe here an extension of that work employing an angle distance measure; this measure provides better classification based on spectral signatures alone in specimens with significant thickness variations. The method is illustrated using simulated data, and also to examine sporulation in the bacterium Clostridium sp.

© 2005 Elsevier B.V. All rights reserved.
PACS: 07.05.Kf; 07.85.Tt; 61.10.Ht; 78.70.Dm

Keywords: X-ray microscopy; X-ray spectromicroscopy; Principal component analysis; Cluster analysis

1. Introduction

Spectromicroscopy provides information on chemical speciation at the microscopic scale; it encompasses all techniques which deliver spatially resolved spectra, or spectrally resolved images. Soft X-ray spectromicroscopy is especially well suited to studying the distribution of organic K and metal L-edge speciation at a scale of tens of nanometers [1].

When the spectra of all major constituents are known in advance, compositional maps can be obtained by existing methods (see e.g., [2]). However, it is often the case (particularly in biology and environmental science) that the complexity of the specimen precludes this method of analysis. For these cases, we have described the use of principal component analysis (PCA) to reduce and noise-filter a data set, followed by cluster analysis (CA) which finds natural groupings of data and calculates averaged spectra and thickness maps associated with these [3].

While cluster analysis has proven to be very successful at revealing previously-unnoticed but chemically distinct regions in many specimens, we have also found that variations in thickness are often classified as distinct regions as well. This is undesirable, since our goal is to identify representative spectra and then produce thickness maps corresponding to the spectra. Therefore, we have sought an improvement in the criterion used for clustering. By using an angle distance measure [4] rather than a conventional Euclidean distance measure in clustering, we are able to greatly suppress thickness variations and cluster data based more completely on chemical speciation alone.

This method will be illustrated using two data sets. We first use simulated data to illustrate the problems of thickness versus compositional variations, and demonstrate the improve-
ment that the angle distance measure provides. The simulated data are similar to what we have used in previous work [3].

The data set makes use of experimentally measured absorption spectra of pure thin films of collagen and the amino acids leucine and tyrosine. Letters A, B, and C contain leucine and D, E, F tyrosine with composition of 90%, 50%, and 10% respectively, while the rest of the composition of each letter and the rest of the sample was assumed to be collagen.

The specimen consists of a 128 x 128 pixel array for each of the 133 photon energies evenly spaced between 282 and 302 eV; a 10 x 10 pixel region in the lower left corner was assumed to have no absorption for normalization purposes. To create a sample with large thickness variations, each pixel was assigned a random thickness in the range 0–200 nm. We then applied the method to a study of sporulation in the bacterium Clostridium sp. which being studied for its properties of uranium reduction at nuclear waste disposal sites [5]. Clostridium sp. was grown to late log-phase (~22 h) and whole cells were recovered by centrifugation at 6,000 x g and washed two times in 0.05 M NaCl. The cells were dried onto a Si3 N 4 window prior to analysis. These bacteria were imaged across the carbon K absorption edge using the Stoney Brook Scanning Transmission X-ray Microscope (STXM) IV at the beamline X1A1 at the National Synchrotron Light Source [6].

2. Spectromicroscopy data

Spectromicroscopy data consist of a series of energy spectra at adjacent positions that form an image, or a “stack” of images [7,1]. Each data point is indexed spatially by its pixel p = 1, . . . , P and spectrally by its energy n = 1, . . . , N. Experimentally, we measure the optical density D = log I 0/I which for a single pixel can be written as an energy-dependent linear absorption coefficient µ multiplied by the specimen thickness t, or D = µ t. For a specimen with s = 1, . . . , S compositional components, the entire data can then be represented in matrix notation as Ds,p = µN,s tS,p. If we know the spectra of all components (so that µN,s is determined), thickness maps of these components can be obtained by matrix inversion as

\[ tS,p = (µN,s)^{-1} \cdot Ds,p \] (1)

where singular value decomposition provides a good means of matrix inversion.

2.1. Principal component analysis

In the raw data, the signal at one energy n may be strongly related to the signal at another energy, while at other energies the signals may show significant variations throughout all the pixels p of the data. A classification method that treats the signal at each energy n on an equal footing is not well suited to optimum classification of the image. We, therefore, use principal component analysis (PCA) [3,8] to represent a sample with a small number SPCmax of abstract spectra instead of large number P of real spectra. The principal components are orthogonal abstract spectra, where the first principal component is equal to the average of the spectra in the sample, the second principal component represents the most common deviations from this average, and all subsequent components give further deviations from the set of previous components until all the variations in the data are represented. In practice, only a small number (typically 2–8) of these components are required to fully represent all the meaningful variations in the data, while subsequent components simply describe variations due to noise. As a result, PCA enables us to fully represent the sample with not N > 100 energies but 3 < 10 spectral signatures [9,3]. However, since all but the first component represent differences in spectral signatures, they are not physically interpretable spectra and indeed they will contain negative (and thus physically meaningless) optical densities. Therefore, PCA provides a nicely orthogonal and noise-filtered representation of the data, although it is difficult to interpret directly.

2.2. Cluster analysis

Our goal is to find pixels with common spectral signatures by some classification scheme. By using the PCA representation of the data, we have a nice orthogonal and noise-filtered search space for cluster analysis (a form of unsupervised pattern recognition) in which to carry this out. Each pixel is characterized by its weights of the significant components; we then use a learning vector quantization (LVQ) algorithm [10] to carry out the clustering as described in detail in [3]. This algorithm requires a measure of the distance from one pixel to another in the search space; we have used a scaled Euclidian distance for this purpose in previous work [3]. The scaling involves subtracting the mean <Rs> of each component, and a power-law scaling using eigenvalues according to \[ \bar{D}_{rs} = (R_{rs} - <R>) / (λs)^(1/2) \]. This scaling provides enhanced sensitivity to the more subtle variations of the data, and provides excellent clustering in many applications. However, while we have tried various approaches (including the exclusion of the first component from the clustering algorithm [3]) to reduce the sensitivity of Euclidian distance measure clustering to thickness variations in the specimen, there has been opportunity for further improvement of the method.

2.3. Angle distance measure

Let us consider the case of a simple specimen with two different, unknown compositions and varying thickness. PCA will deliver a representation of the data with just two significant components, and the data can be represented by a two-dimensional scatterplot where each pixel is represented by a dot located at its weight in each component (Fig. 1). Pixels with the same composition but different thickness should
have the same ratio of principal components, so one would expect pixels of this composition to be distributed on a line from the origin. This simple example illustrates that the usual Euclidean distance measure may not be appropriate for classification of the data; pixels of roughly similar composition may end up being clustered according to their distance from the origin, rather than their differing ratio of PCA components. If one were instead to exclude the radial position of the pixel, and only look at its angular position, one would be more likely to classify the data according to its chemical composition as opposed to its local thickness. We have, therefore, made use of an angle distance measure $\theta$, or normalized scalar product, defined as

$$\theta = \arccos \left( \frac{\sum_i x_i y_i}{\sqrt{\sum_i x_i^2} \sqrt{\sum_i y_i^2}} \right).$$

(2)

Examination of Fig. 1 suggests that this measure will better distinguish between different compositions independent of their thickness. However, caution must be used: at low optical densities (thin regions of the sample), the presence of noise can lead to an unphysically large range of angles. To avoid classification errors from these thin sample regions, angular cluster centers are first calculated only using those pixels above a threshold optical density value $D_{\text{min}}$ and those angular cluster centers are subsequently applied to all data. With the data clustered in this fashion, representative spectra can be obtained by averaging all the pixels in the cluster, and this set of spectra $\mu$ can then be used to obtain thickness maps according to Eq. (1).

3. Illustration: simulated data

To test and illustrate the use of the angle distance measure, we created a simulated data set with known compositions and random thickness as described in Section 1. Principal component analysis gave the expected result of three components to this data (since the sample was “built” out of collagen, tyrosine, and leucine). Cluster analysis was then carried out, with map results shown in Fig. 2 with the corresponding scatterplots presented in Fig. 3. The first two cases were calculated...
Fig. 3. Scatterplots of clustered pixel distributions corresponding to the clustering maps of the simulated data of Fig. 2. For this figure, case C has been split into two separate displays: case C1 is the usual scatterplot, while case C2 shows all pixels projected out to a unit sphere radius so that angular separations can be more easily visualized. As can be seen, pixels with similar composition fall along straight lines when plotted according to PCA components, and the angle distance measure is better matched to clustering such compositional similarities.

using the Euclidean distance. In the first case all significant component were included, while in the second case the first principal component was excluded. In the third case, the angle measure was used, with first principal component included.

From scatterplots in Fig. 3 it can be seen that each letter was represented by pixels scattered along a line. Therefore, for the same composition, pixels with different optical densities differ only in their distance from the origin. Since we would like to exclude thickness from the analysis, the angle distance measure is the natural choice. This is proved by the results of cluster analysis in the Figs. 2 and 3. In two cases where the Euclidean distance measure was used, pixels belonging to each letter were not all grouped together, as can be seen from the scatterplots: only some parts of the pixels scattered along the lines belong to clusters which represent letters. In other words, clustering failed to recognize all pixels that belong to each letter. When the angle distance measure was used, the situation was greatly improved: almost all pixels which belong to each letter were clustered in corresponding groups.

4. Clustering of Clostridium sp.

Having verified that the angle distance measure provides improved clustering results on simulated data, we now examine carbon edge data of an anaerobic spore-forming bacterium, *Clostridium* sp., for which PCA found four significant components. Cluster analysis was performed using both the Euclidean distance measure (Figs. 4 and 5), and the angle distance measure (Figs. 6 and 7). The results of cluster
Fig. 4. Optical density image (A) of a Clostridium sp. bacterium imaged at 288.20 eV. This image shows a mix of thickness and chemical variations in the complex specimen. Following cluster analysis using a Euclidean distance measure, a cluster index map and corresponding spectra (B) were obtained. In the cluster analysis, the lipid membrane is nicely separated, but no compositional variations are shown inside the bacterium except for the spore, and the spore itself is represented with two clusters with similar spectra but different thickness. The corresponding scatterplots are shown in Fig. 5.

Analysis with the Euclidean distance measure is largely influenced by the thicknesses of the sample regions. While the Euclidean distance measure was able to identify the spore within the bacterium, it clustered the spore into two separate regions (a thicker center region and a thinner outer region, corresponding to thickness projected through the roughly spherical shape of the spore). The use of the angle distance measure produced a single cluster for the immature spore which is presumed to be of approximately uniform composition (prior to outer spore membrane formation) [11], and the angle distance measure was better able to identify compositional variation within the cytoplasm of the bacterium that is consistent with a non-uniform ultrastructure (such as an asymmetric distribution of DNA in the region of the immature spore (forespore)) [12]. These results on real data reinforce the idea that an angle distance measure is more appropriate for spectromicroscopy analysis of samples with large thickness variations.

Fig. 5. Scatterplots of the Euclidean distance measure clustering of Clostridium sp. The first plot, of PCA component 1 versus component 2, clearly shows different thicknesses of similar composition (compare with Fig. 4), but the Euclidean distance measure places these into two separate clusters (4 and 5). The corresponding cluster index maps and spectra are shown in Fig. 4.
Fig. 6. Cluster indices and corresponding spectra of cluster analysis of *Clostridium* sp. using an angle distance measure. Two different compositional types are found within the bacterium, and the spore is identified with a single cluster. The corresponding scatterplots are shown in Fig. 7.

Fig. 7. Scatterplots of the angle distance measure clustering of *Clostridium* sp. The compositional variations within the bacterium body are most clearly revealed by looking at how clusters 2 and 3 are separated in the scatterplot of component 3 versus component 4. The corresponding cluster index maps and spectra are shown in Fig. 6.

5. Conclusion

Soft X-ray spectromicroscopy can provide rich information on compositional variations at the nanoscale. For those samples for which a complete set of reference spectra are not available (such as is usually the case in biology or environmental science), principal component analysis can be used to orthogonalize and noise-filter the data, and cluster analysis can be used to classify the data and provide representative spectra and corresponding thickness maps. By using an angle distance measure rather than the usual Euclidean distance measure, we show here an improvement in clustering based on compositional variations in specimens with significant thickness variations. This is illustrated in clustering of a simulated data set of known, constructed composition. It is also demonstrated in a study of the bacterium *Clostridium* sp., where a spore is correctly identified and its spectral signature obtained in unsupervised cluster analysis without employing...
any prior information. Angle distance measure clustering provides a powerful tool for the analysis of spectromicroscopy data of biological and environmental science specimens.

Acknowledgements

We wish to thank Michael Feser for many helpful discussions. We gratefully acknowledge funding from the National Institutes for Health under contract R01 EB00479-01A1, the National Science Foundation under contracts OCE-021029 and CHE-0221934, and in part by Brookhaven National Laboratory, Laboratory Directed Research and Development (LDRD) and Environmental Management Science Program (EMSP), Environmental Remediation Sciences Division, Office of Biological and Environmental Research, Office of Science, U.S. Department of Energy, under contract No. DE-AC02-98CH10886.

References