Cluster analysis of soft X-ray spectromicroscopy data

6 authors, including:

Chris Jacobsen
Argonne National Laboratory
264 PUBLICATIONS   5,949 CITATIONS

SEE PROFILE

Michael Feser
Lyncean Technologies Inc.
92 PUBLICATIONS   1,624 CITATIONS

SEE PROFILE

Mirna Lerotic
2nd Look Consulting
22 PUBLICATIONS   689 CITATIONS

SEE PROFILE

Stefan Vogt
Argonne National Laboratory
314 PUBLICATIONS   7,304 CITATIONS

SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Trace elements in groundwaters and minerals in crystalline rock fractures View project

Natural Organic Matter Spectromicroscopy View project

Cluster analysis of soft X-ray spectromicroscopy data

C. Jacobsen, M. Feser, M. Lerotic, S. Vogt[1], J. Maser[1] and T. Schäfer[2]

Department of Physics and Astronomy, SUNY Stony Brook, U.S.A.
[1] *Advanced Photon Source, Argonne National Laboratory, Argonne, IL, U.S.A.*
[2] *Forschungszentrum Karlsruhe GmbH, 76021 Karlsruhe, Germany*

Abstract

We describe the use of principle component analysis (PCA) to serve as a prefilter for cluster analysis or pattern recognition analysis of soft x-ray spectromicroscopy data. Cluster analysis provides a method to group regions with common spectral features even if no prior knowledge of their spectra is available, such as in biology or environmental science.

1 INTRODUCTION

Soft x-ray spectrum imaging involves the acquisition of a series or "stack" of images [4] across near-edge absorption features, yielding a dataset in (x, y, E) where (x, y) are position coordinates and E is photon energy. When all components with different absorption spectra in the specimen are known, curve-fitting or matrix inversion methods [2, 11] can be used to obtain maps of component thicknesses. However, it is often the case (particularly with biological or environmental science specimens) that not all of the components and their spectra are known in advance. In this case, one must look to the dataset itself for clues as to the spectroscopically-significant differences in the specimen.

2 MULTIVARIATE STATISTICAL ANALYSIS

We describe the use of multivariate statistical analysis methods to analyze soft x-ray spectro-microscopy datasets. Principal component or factor analysis [7] provides a way to characterize a dataset in terms of its most significant variations without prior knowledge of their character-istics; this has been used in x-ray absorption spectroscopy [10], electron energy-loss specrum imaging [1], and x-ray spectromicroscopy [5,8] for data analysis.

In spectromicroscopy or spectrum imaging, we acquire images with $p = 1 \ldots P$ pixels (where $p = i_{\text{col}} + (i_{\text{row}} - 1) \cdot (\# \text{ rows})$, over $n = 1 \ldots N$ energies. We assume that the heterogeneous specimen has $s = 1 \ldots S$ distinct components or phases. We can then write the optical density at a particular energy n and pixel p summed over all (as yet unknown) components as the matrix equation

$$D_{N \times P} = \mu_{N \times S} \cdot t_{S \times P} \tag{2.1}$$

If we know the exact absorption spectrum μ_{Ns} for each of the $s = 1 \ldots S$ known components in the specimen, then the matrix $\mu_{N \times S}$ is already determined. In this case, we can find the spatially-resolved thicknesses of the components by matrix inversion:

$$t_{S \times P} = (\mu_{N \times S})^{-1} \cdot D_{N \times P}. \tag{2.2}$$

The inversion of the matrix $\mu_{N \times S}$ can be accomplished in a robust fashion using singular value decomposition (SVD; see *e.g.*, [9, Sec. 2.6]), as has already been applied to x-ray spectromicroscopy composition mapping [6, 11]. Equivalent results have also been obtained using curve-fitting methods to obtain thickness maps based on known spectra (A. Hitchcock, personal communication).

What happens if we do not know all of these components and their spectra ahead of time? In this case, we will consider the specimen to be described by a set of $s = 1 \ldots S_{\text{abstract}}$ abstract components (where $S_{\text{abstract}} \leq N$), in which case we can frame the problem as one of solving

$$D_{N \times P} = C_{N \times S_{\text{abstract}}} \cdot R_{S_{\text{abstract}} \times P} \tag{2.3}$$

to find a column matrix $C_{N \times S_{\text{abstract}}}$ that gives in each column a spectrum (with N points) of one of S_{abstract} components, and a row matrix $R_{S_{\text{abstract}} \times P}$ that gives in each row an image (with P pixels) of one of S_{abstract} components. By "abstract" we mean that these components simply describe natural groupings of the data, but not necessarily in the way we might best understand them. The column matrix $C_{N \times S_{\text{abstract}}}$ can be found from eigenvectors of the covariance matrix, or from SVD decomposition of the data matrix $D_{N \times P}$. In either case, the matrix $C_{N \times S_{\text{abstract}}}$ represents a set of eigenspectra of the data set. The first of these tends to be from a sort of "average" spectrum for all the pixels, while the second component is from the most common difference from that "average," and so on. At some point, the differences between successive components simply reflect noise in the data, so that one will want to consider a reduced set of the significant components $\bar{S}_{\text{abstract}}$ rather than the full set S_{abstract} [1, 8]. Having found the eigenspectrum matrix $C_{N \times S_{\text{abstract}}}$, we can also find a corresponding eigenimage matrix $R_{\bar{S}_{\text{abstract}} \times P}$ from

$$R_{\bar{S}_{\text{abstract}} \times P} = C^T_{\bar{S}_{\text{abstract}} \times N} \cdot D_{N \times P}. \tag{2.4}$$

where we have used the fact that C is orthogonal so that its inverse is the transpose, or $C^{-1} = C^T$. If in fact we know all physical components that make up the specimen and their spectra as a matrix $\mu_{N \times S_{\text{physical}}}$, we can also find a transformation matrix T between physical and abstract spectra as

$$T_{\bar{S}_{\text{abstract}} \times S_{\text{physical}}} = C^T_{\bar{S}_{\text{abstract}} \times N} \cdot \mu_{N \times S_{\text{physical}}} \tag{2.5}$$

and thereby use our PCA representation of the data to obtain thickness maps using SVD, where the benefit provided by PCA is one of filtering out much of the noise in the dataset.

3 CLUSTER ANALYSIS

What if the spectra of all physical components are in fact not known? This problem is common in the case of "natural" specimens such as those found in biology or environmental science. The eigenspectra can be useful in finding energies where unique variations appear in the data, but the eigenspectra cannot be interpreted in a straightforward manner since they represent in some sense successive differences in spectra. The approach we are developing is to use cluster analysis [3] to seek the natural groupings of the data. If we consider the eigenimage matrix $R_{\bar{S}_{\text{abstract}} \times P}$, we realize that for each pixel $p = 1 \ldots P$, we have weighting coefficients in

each of $\bar{S}_{\text{abstract}}$ dimensions. We then seek natural groupings of these pixels using a Euclidian distance-based learning algorithm:

1. We start by placing G cluster centers at random locations in the unit-normalized, $\bar{S}_{\text{abstract}}$-dimensional space of eigenspectra weights per pixel $R_{\bar{S}_{\text{abstract}} \times P}$.

2. For one pixel p in the data, we calculate the Euclidian distance between this pixel's position vector $R_{\bar{S}_{\text{abstract}} \times p}$ and each of the G cluster centers.

3. We then choose the cluster center g_{match} which is closest to this pixel's position vector, and move this cluster center some fraction of the way towards it. The fraction used (the "learning rate") is a value between 1 and 0. Other cluster centers are left untouched.

4. We now repeat steps 2 and 3 for the subsequent pixels in the dataset.

5. We now repeat steps 2–4 for a number of iterations, adjusting the "learning rate" each time. In the work shown here, 20 iterations were used, and the "learning rate" was adjusted from 0.5 to 0.1 over these iterations.

6. We now classify pixels in terms of their closest cluster center. We can now calculate the average spectrum for these pixels, and display clusters as separate colors in a composite image.

While more advanced clustering algorithms exist [3], the algorithm described above has thus far appeared to be robust in our limited application to soft x-ray spectromicroscopy data. We note that the application of principal component analysis to orthogonalize and reduce the data is all-important; without it, we have had no success in obtaining reliable clustering of either simulated or experimental data.

4 ILLUSTRATION: LUTETIUM IN HÆMATITE

As a brief illustration, we consider an example drawn from environmental science experiments by Schäfer *et al.* Lutetium can be used as a homologue for Americium and Curium, which are of interest in studies of colloidal transport of radionuclides at nuclear waste disposal sites. Lutetium was placed in solution with hæmatite, and the solution was then dried on a silicon nitride window. An oxygen edge XANES spectromicroscopy data set was then acquired using the Stony Brook STXM IV microscope at the National Synchrotron Light Source. Principal component analysis of the data showed that there were $\bar{S}_{\text{abstract}} = 4$ significant components, and trials of different numbers of clusters showed that $G = 3$ clusters nicely sorted the dataset into three regions: a mostly open region with very little optical density (cluster 1 in Fig. 1), and regions with less or more absorption at 530.0 eV relative to 531.7 eV, indicating less (cluster 2 in Fig. 1) or more (cluster 3 in Fig. 1) pure hæmatite. These data suggest that Lutetium has been incorporated into hæmatite, which could imply increased mobility of radionuclides from nuclear waste repositories due to colloidal transport in groundwater.

5 CONCLUSION

We have described here the beginning of an ongoing investigation of the use of clustering methods for the analysis of soft x-ray spectromicroscopy data. Cluster analysis appears promising for problems where one does not have mixtures of pure materials with known spectra. We gratefully acknowledge funding from the National Institutes for Health under contract R01 EB00479-01A1, and from the National Science Foundation under contracts OCE-0221029 and CHE-0221934.
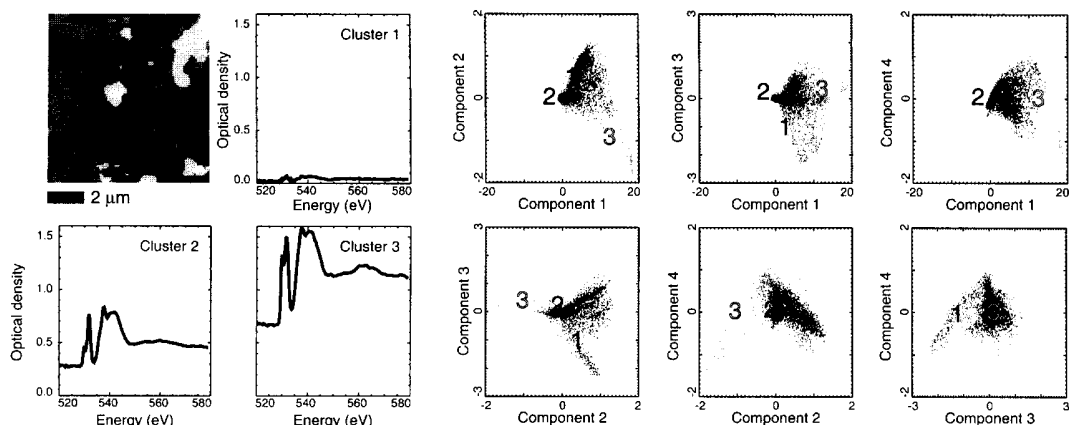
Fig. 1. Cluster analysis results on Lu incorporated into hæmatite. At upper left we show the 3 clustered regions, with cluster 1 being the darkest and cluster 3 being the lightest; the optical density spectra for each of these regions are also shown. Cluster 2 shows a mix of Lu and hæmatite, while cluster 3 shows mostly pure hæmatite. At right we show the distribution of weights of various components for all pixels $R_{\bar{S}_{\text{abstract}} \times P}$ as plotted over pairs of components $\bar{S}_{\text{abstract}}$ for the Lutecium/Haematite specimen. That is, the plot of component 2 versus component 1 shows values of $R_{s_{\text{abstract}}=1 \times P}$ plotted as a function of $R_{s_{\text{abstract}}=2 \times P}$. The position of each cluster center is also plotted for the two components shown. This figure shows that it is not always possible to recognize cluster center positions based on just two variables; instead, the data are clustered in a multidimensional space of *all* the components $\bar{S}_{\text{abstract}}$.

References

[1] N. Bonnet, N. Brun, and C. Colliex. Extracting information from sequences of spatially resolved EELS spectra using multivariate statistical analysis. *Ultramicroscopy*, 77:97–112, 1999.

[2] C. J. Buckley, N. Khaleque, S. J. Bellamy, M. Robins, and X. Zhang. Mapping the organic and inorganic components of tissue using NEXAFS. *Journal de Physique*, IV 7 (C2 Part 1):83–90, 1997.

[3] B.S. Everitt, S. Landau, and M. Leese. *Cluster Analysis*. Arnold Publishers, London, fourth edition, 2001.

[4] C. Jacobsen, G. Flynn, S. Wirick, and C. Zimba. Soft x-ray spectroscopy from image sequences with sub-100 nm spatial resolution. *Journal of Microscopy*, 197(2):173–184, 2000.

[5] P. L. King, R. Browning, P. Pianetta, I. Lindau, M. Keenlyside, and G. Knapp. Image-processing of multispectral x-ray photoelectron-spectroscopy images. *Journal of Vacuum Science and Technology A*, 7(6):3301–3304, 1989.

[6] I. N. Koprinarov, A. P. Hitchcock, C. T. McCrory, and R. F. Childs. Quantitative mapping of structured polymeric systems using singular value decomposition analysis of soft x-ray images. *Journal of Physical Chemistry B*, 106:5358–5364, 2002.

[7] E. R. Malinowski. *Factor analysis in chemistry*. John H. Wiley & Sons, New York, 2nd edition, 1991.

[8] A. Osanna and C. Jacobsen. Principle component analysis for soft x-ray spectromicroscopy. In W. Meyer-Ilse, T. Warwick, and D. Attwood, editors, *X-ray Microscopy: Proceedings of the Sixth International Conference*, pages 350–357, Melville, NY, 2000. American Institute of Physics.

[9] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling. *Numerical Recipes in C*. Cambridge University Press, Cambridge, UK, 1988.

[10] S. R. Wasserman. The analysis of mixtures: Application of principle components analysis to XAS spectra. *Journal de Physique IV France*, 7 (C2):203–205, 1997.

[11] X. Zhang, R. Balhorn, J. Mazrimas, and J. Kirz. Mapping and measuring DNA to protein ratios in mammalian sperm head by XANES imaging. *Journal of Structural Biology*, 116:335–344, 1996.