doi:10.1017/S1431927610057260
Microsc. Microanal. 16 (Suppl 2), 2010

Multivariate Statistics in Microanalysis: From Abstract to Useful

John Henry J. Scott and Jeffrey M. Davis

National Institute of Standards and Technology, Gaithersburg, MD 20899

Multispectral X-ray maps with several X-ray spectral channels at each pixel are ubiquitous, and many instruments now routinely produce hyperspectral maps with entire 2048-channel X-ray spectra at every pixel. Figure 1 shows an example dataset produced by stage scanning a semiconductor device wafer and acquiring an X-ray excited energy-dispersive X-ray spectrum (XEDS) at every pixel. This sample is richly heterogeneous with several distinct material phases juxtaposed. Non-linear excitation and absorption effects lead to an even wider variety of spectral signatures in the hyperspectral data cube, presenting significant challenges to the analyst during interpretation.

To meet such challenges the microanalysis community has begun to utilize multivariate statistical analysis techniques, and the number of publications applying such methods to X-ray compositional imaging is growing rapidly. However, these methods have received a mixed reception among traditional analysts, in part because they are mathematically complex and difficult to understand, and in part because they are often too abstract to yield a satisfying interpretation when applied to "real world" analysis problems. Principal component analysis (PCA) is perhaps the most glaring example of this trend. As a mathematical transformation of the raw data, PCA is beyond criticism and like many dataset decomposition methods it can be a powerful tool for providing insight into the multivariate structure of the data. It is also an absolutely invaluable component in more complex data workflows. However, the principal component vectors that result from the PCA rotation are abstract. They rarely have a clean physical interpretation, and when they do it is often by accident.

In other fields of research, such as chemometrics and remote sensing, mature and effective multivariate analysis procedures have been developed over the course of several decades. In most cases the techniques have been sifted and refined based on rigorous measures of accuracy and effectiveness when applied to domain-specific data types, atmosphere-corrected visible wavelength reflectance data for example. It may be possible to adapt some of these workflows to microanalysis data, but there are pitfalls because the existing tools are not adapted to the unique non-linearities inherent in X-ray compositional imaging. To move our field forward, good measurement science demands that we push past abstract data manipulations to reconnect with physically interpretable results and that we assess the quality of novel methods on known/standard datasets.

Figure 2, for example, depicts the results of three supervised classification experiments on the data from Figure 1. In supervised classification the analyst defines exemplar spectral signatures by marking regions of interest in the raw data, then asks the computer to label each pixel as a member of the class with the closest spectral match. In this example, although multivariate tests estimate the intrinsic dimensionality of the data to be ~30 or so, for simplicity only three classes were defined (red, green, blue). The quality differences between the three algorithms is visually obvious and can be quantified by confusion matrices and other post-classification yardsticks. Critical evaluation of multivariate technique performance using practical and meaningful metrics is essential if we hope to convert more of these tools from the abstract to the useful.

References

[1] SID: H. Du, C.-I. Chang, H. Ren, F.M. D'Amico, J. O. Jensen, J., "New Hyperspectral Discrimination Measure for Spectral Characterization." *Optical Engineering* **43** (2004) 1777-1786.
[2] SVM: A.J. Izenman, *Modern Multivariate Statistical Techniques*, Springer, 2008, 369.

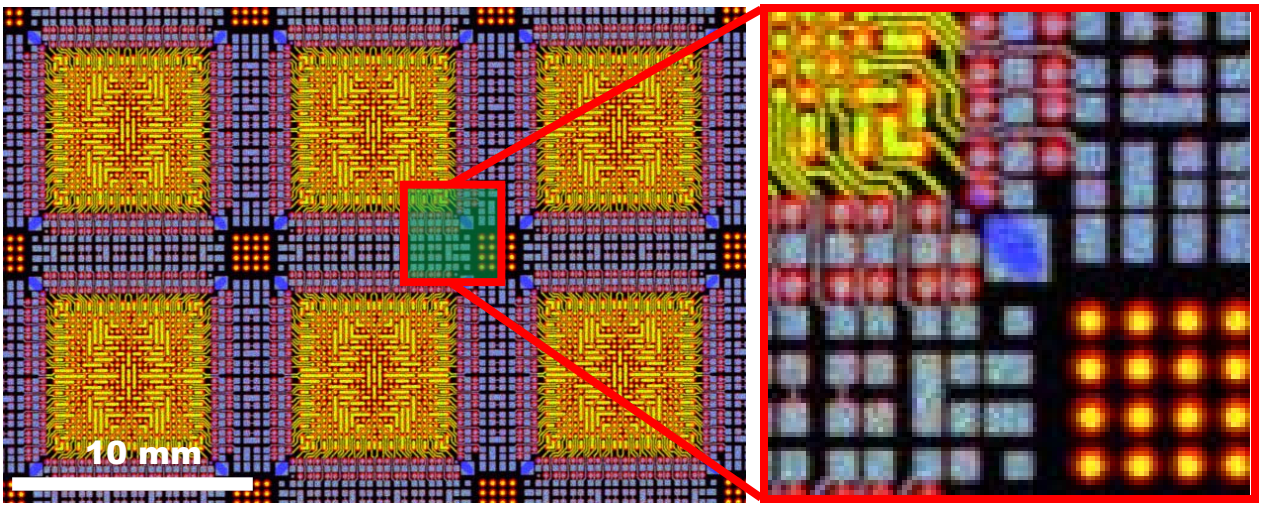

Figure 1. Hyperspectral dataset of semiconductor circuitry acquired on a stage-scanned milli-XRF instrument (left), and a detail view of the same data (right). Red, green, and blue represent fluoresced X-ray intensity at 8.05 keV, 7.47 keV, and 9.72 keV respectively. Horizontal full width of left image is 30 mm.

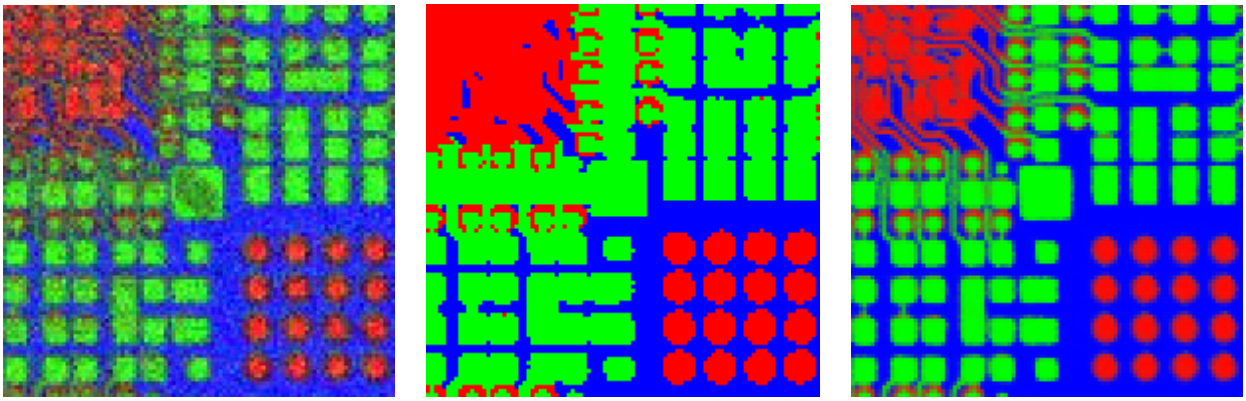

Figure 2. Results from supervised classification of pixels into three classes based on the x-ray spectral properties at each pixel. Three different classification decision rules were used, differing widely in both performance and complexity. Fast and efficient parallelepiped class boundaries (left) are noisier than classification based on spectral information divergence [1] (center), while a well-trained support vector machine (SVM) classifier [2] (right) displays some of the best properties of each. Rigorous and quantitative evaluation of these competing methods, combined with traditional microanalysis expertise, is necessary to translate abstract multivariate statistical methods into practical tools for solving microanalysis problems in the real world.

https://doi.org/10.1017/S1431927610057260 Published online by Cambridge University Press