How to Avoid Unexpected Artifacts from Multivariate Statistical Analysis on STEM Spectrum-Imaging Datasets

Kazuo Ishizuka[1] and Masashi Watanabe[2]

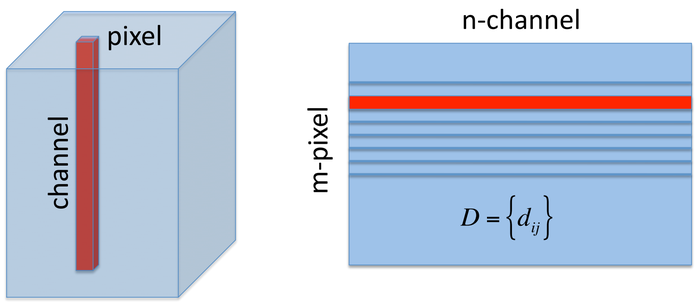
1 HREM Research Inc., Higashimatsuyama, Saitama 355-0055, Japan
2 Dept. of Materials Science and Engineering, Lehigh University, Bethlehem, PA 18015, USA

Introduction

The latest aberration-corrected scanning transmission electron microscope (STEM) makes possible to perform routinely not only atomic-scale imaging but also chemical analysis via electron energy-loss spectrometry (EELS) and/or X-ray energy dispersive spectrometry (XEDS) [e.g. 1]. The advances in the recent software developments, in combination with the latest hardware, allow us to acquire large-scale datasets such as multidimensional image series and spectrum images (SIs). Therefore, it is challenging to deal with the large-scale datasets, e.g. extraction of unknown features and estimation of dominant trends. If the datasets were relatively noisy, which is very common for atomic-resolution EELS/XEDS SIs, data analysis would be much harder tasks. Multivariate statistical analysis (MSA) is one of efficient approaches to analyse the large-scale datasets in terms of feature identification and extraction.

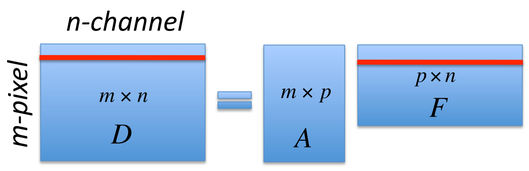
Principal component analysis (PCA)

Principal component analysis (PCA) is one of the MSA techniques [2]. The spectrum image (SI) data cube may be rearranged into 2D data matrix:



If the sample is composed of p-substances, the observed spectrum at each point will be a linear combination of the p-component spectra (Factors) f:

$$d_i(j) = \sum_{k=1}^{p} a_{ik} f_k(j)$$



Here, the component spectra will be *real* (physical) spectra, and $\{a_{ik}\}$ represent the contributions of the spectrum k to the observed spectrum i. The 2D data matrix can be decomposed to a product of the two matrices. The PCA decomposes the data matrix D to p-principal components (Loadings L) and the matrix (Score S) representing the contributions of the components as $D \approx SL$, trying to explain the data variation (variance) as much as possible using a small number of the components. The principal components (Loadings) are *abstract* (non-physical), and orthogonal to each other. Since a use of PCA is relatively straightforward, PCA has been applied to SIs as data-mining and noise-reduction tools [e.g. 3].

Problems at high noise condition

Let's consider the cases where we are interest in small amount of impurities in matrix (Fig. 1 (a)) or the composition in the interface (Fig. 1 (b)). Here, the signal itself of course contributes the data variation. However, a small amount of signal in question will be buried in the random noise of the whole data matrix.

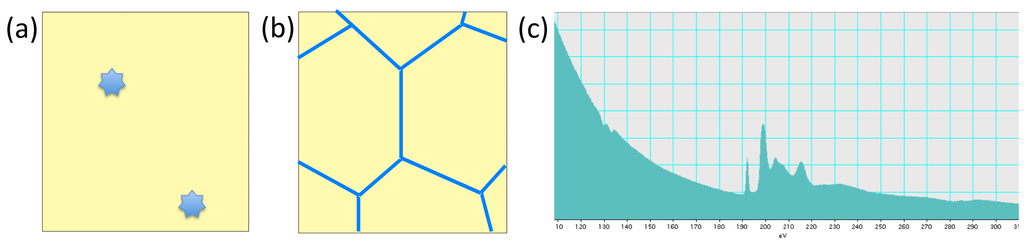


Figure 1. Some difficult cases to apply a regular PCA; (a) small amount of impurities, (b) interface composition, (c) low signal in wide dynamic range signal.

In the case of EELS (Electron Energy Loss Spectrum) the signal range is relatively wide as shown in Fig. 1 (c). Here, the high intensity signal will contributes more variance in the data matrix than the low signal in question. Then, the low signal may be hidden by the the variance of the strong signal. Although the PCA approach is very efficient and useful, therefore, it may create unexpected artifacts especially in higher noise conditions [4] (see Fig. 3). Since these artifacts might mislead results, it is essential to find a way to avoid such artifacts.

How to improve the sensitivity

There may be two approaches to improve the PCA sensitivity: (1) reduction of random noise and (2) enhancement of true variations. The former requires modifications in experimental conditions (higher currents and/or a longer acquisition time). Conversely, the latter can be achieved by performing PCA analysis to divided small segments of a SI data, which we call the local PCA approach (Fig. 2). Here, the division can be made spatially (a and b) and spectrally (c). The spatially local PCA will be especially useful to detect segregated element in the matrix. The spectrally local PCA is useful to detect a weak signal, if the weak signal is spectrally separated from the strong signal.
We have implemented the local PCA into the MSA plug-in [3], where the whole spectrum is spatially or spectrally divided into local regions suggested by the user. Then, all the local regions are processed sequentially, and the reconstructed result is obtained by pasting the local regions together.

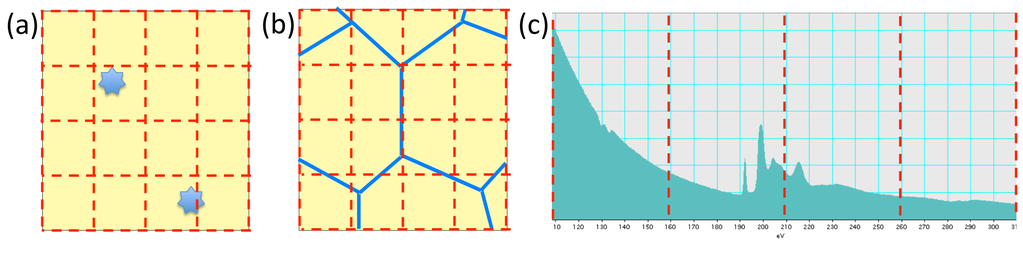


Figure 2. (a and b) Spatially local PCA and (c) spectrally local PCA.

Unexpected artifacts obtained by PCA

Recently, an interested unexpected results was reported by Lichtert and Verbeeck [4]. Here, an EELS SI data was created from the artificial sample of hexagonal boron nitride (h-BN) sheet as shown in Fig. 3 (a). All the atomic columns contain two B and N atoms, expect for the columns labelled by 1 and 2, where an extra N and B atoms are added respectively.

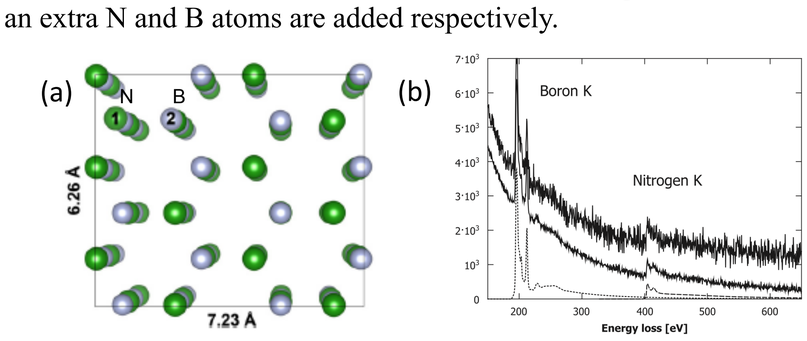


Figure 3. (a) BN model, where there are one excess N or B atom at the positions 1 and 2, respectively. (b) Typical spectra containing high noise (2 x 10^2) and low noise (5 x 10^1). The mean pixel value of the average spectrum is 1.2 x 10^3. (Reproduced from Ref [4])

Using the simulated data set they performed PCA, and obtained a reasonable result for the low noise case (Fig. 4, left column), where an extra N appears correctly on position 1. However, in the high noise case, the weighted PCA gives an extra N signal on position 2 as shown in Fig. 4 (right column), even when they used 10 components.

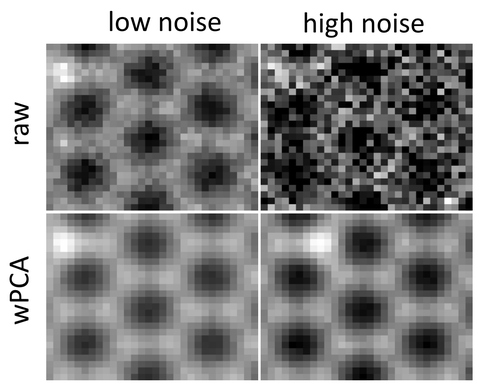


Figure 4. N elemental maps of untreated and weight PCA cases (top and bottom rows) for low noise and high noise cases (left and right columns).

Application of local PCA

We applied our local PCA to the same data set. Here, we used spectrally local PCA, where the whole spectral was divided into four regions (Fig. 5 (a)). The reconstructed N map even for the high noise case shows an extra N signal on the correct position 1 using only three components as shown in (b).

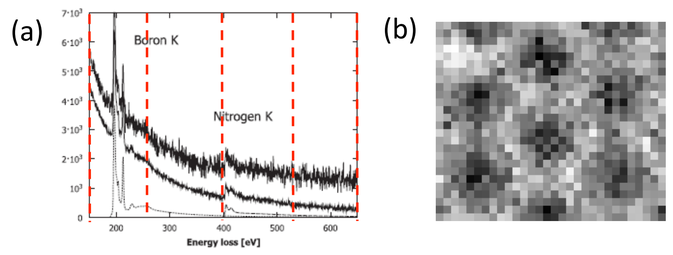
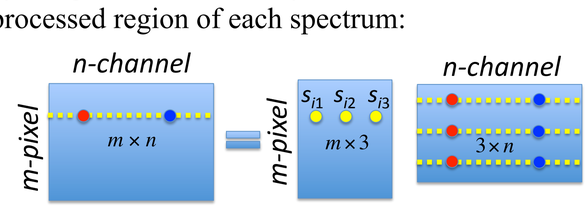


Figure 5. (a) Spectrally local PCA applied for the h-BN spectrum by dividing the whole spectrum into four regions. (b) N map for the high noise case reconstructed by the spectrally local PCA.

It is interesting to note that the fluctuation of the reconstructed N map (Fig. 5 (b)) is higher (worse) than that obtained by a regular PCA (Fig. 4). However, this does not mean the result obtained by the local PCA is inferior to the result obtained by a regular PCA. We may note that the score of each component (loading) is used for the whole processed region of each spectrum:

$$d_i(j) = s_{i1} l_1(j) + s_{i2} l_2(j) + s_{i3} l_3(j)$$



In the case of a regular PCA, the score that will be mainly determined by the strong B signal is used for the whole spectrum including a nitrogen region. Thus, the reconstructed N map shows the same intensity variation with the reconstructed B map for the high noise case. In the same way, the fluctuation of the N map is controlled by the strong B signal that shows a lower noise level (fluctuation) than the N signal itself. Contrary to this, in the case of the local PCA, the score is determined by each local signal, and thus the reconstructed N map as well as its fluctuation is controlled by the N signal itself.

Conclusions

It has been demonstrated that the spectrally local PCA is useful especially for EELS SIs, since an acquired raw signal intensity varies significantly. The advantage of the spatially local PCA to detect minor signal has been discussed for XEDS [5].

References

[1] S.J. Pennycook & P.D. Nellist ed. *Scanning Transmission Electron Microscopy: Imaging and Analysis*, Springer, NY (2011).
[2] E.R. Malinowski, *Factor Analysis in Chemistry*, 3rd ed., Wiley, New York (2002).
[3] M Watanabe et al., *Microscopy and Analysis*, **23**, Issue 7 (2009), 5-7.
[4] S. Lichtert & J. Verbeeck, *Ultramicrosc*., **125** (2013), 35-42
[5] M. Watanabe & K. Ishizuka, M&M 2014 (2014), 112-113.

Acknowledgements

The authors acknowledge J. Verbeeck for providing the simulated BN test data. M.W. wishes to acknowledge financial support from the NSF through grants DMR-0804528 and DMR-1040229.