Treating Outliers in the Course of Principal Component Analysis of EELS Spectrum-Images

Pavel Potapov[1,2]

[1] Leibniz Institute for Solid State and Materials Research (IFW), Dresden, Germany.
[2] Technical University of Dresden, Department of Physics, Dresden, Germany.

EELS STEM spectrum-images can be drastically denoised by application of the Principal Component Analysis (PCA) that retains a meaningful fraction of data while removes irrelevant uncorrelated noise. Unfortunately, EELS spectra often suffer of outliers in data due to X-rays accidentally hitting the EELS spectrometer. Even a singular outlier can distort dramatically the PCA results and cause artefacts in the reconstructed data. Several techniques, like robust PCA [1,2], have been suggested to deal with outliers, however, a pitfall of these methods is a significant complication of the optimization problem.

The present work describes a simple practical method to account for outliers *within* the course of the *standard* PCA treatment. The common NIPALS algorithm finds a loading and a score of a desired component through the series of consequent matrix multiplications. Inserting an additional step among iterations, namely the removal of the most outlying data points, drastically improves the efficiency and stability of the algorithm. The outlying pixel or energy channels can be "sealed" by averaging the neighbouring data points that were proved to be located within the reasonable range of the data variation. An important parameter of this algorithm is a threshold for classifying a given data point as an outlier. Analysis of the real-life EELS spectrum-images suggests that a reasonable threshold value is $\mathbf{4\sigma}$, where $\boldsymbol{\sigma}$ is the standard deviation of a score within a component.

An example of EELS elemental maps denoised with PCA is shown in Fig.1. The outliers distort noticeably the retrieved PCA components with the index greater than 2. This causes the overestimation of the variances in the range of 3rd - 9th PCA components and might lead to the incorrect interpretation of the scree plot (Fig.2). The dataset reconstructed from 4 major components eventually shows the artefactual peaks (Fig.3) due to the contamination of the 3rd and 4th components with outliers. "Sealing" the outliers in the course of NIPALS iterations allows for accurate determining the number of meaningful components in the scree plot and removes the artefacts from the reconstructed data.

References:

[1] E. Candes, X. Li, Y. Ma and J. Wright, J. ACM **58** (2011) p.1.
[2] T. Zhou and D. Tao, ICML-11 (2011) p.33
[3] The author acknowledges funding from DFG "Zukunftskonzept" (F-003661-553-Ü6a-1020605).

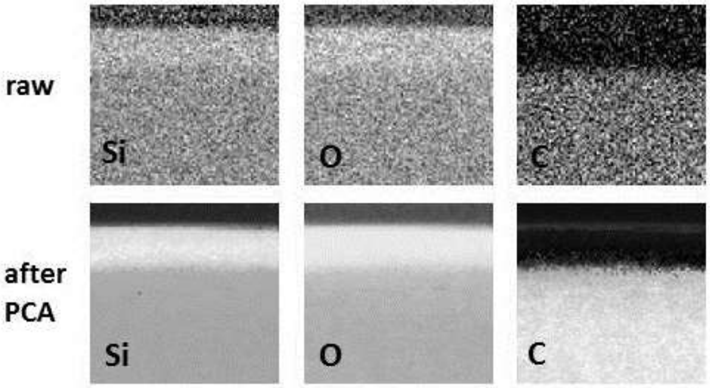

Figure 1. Elemental maps extracted from the raw (upper row) and PCA-denoised (lower row) EELS spectrum-images of the Si-C-O sample.

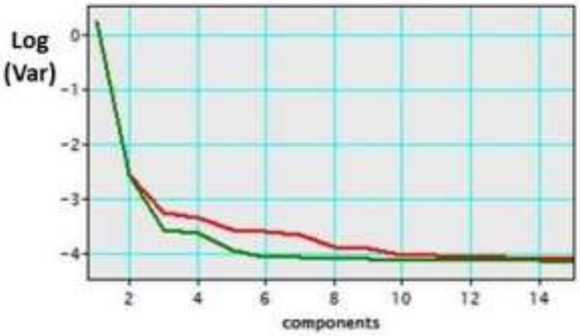

Figure 2. Scree plots of plain PCA (red) and PCA with accounting for outliers (green).

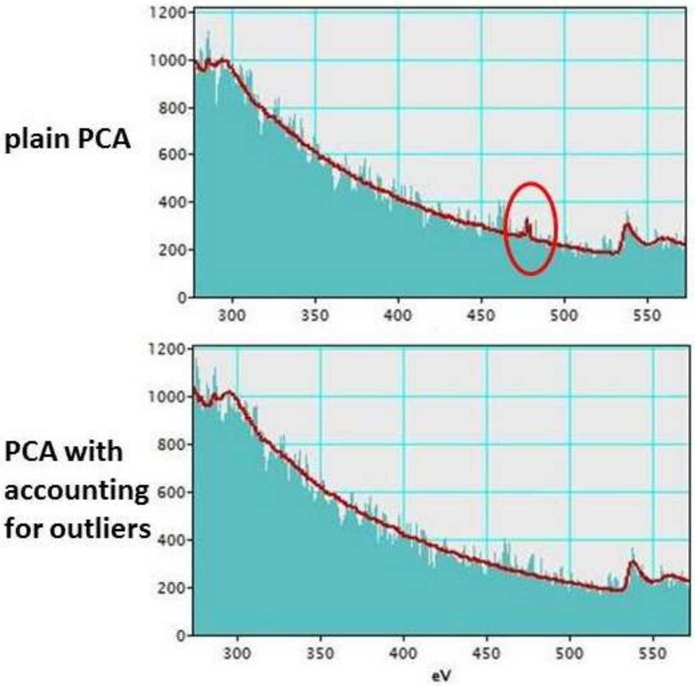

Figure 3. Example of overlaid single pixel spectra: initial (blue-filled) and those denoised by PCA (red). The upper figure represents the plain PCA where the treatment artefact is outlined by the red oval. The lower figure is the result of the PCA accounting for outliers. Both datasets were reconstructed using 4 major PCA components.