A Novel Method for Automatic Determination of the Number of Meaningful Components in the PCA Analysis of Spectrum-Images

Pavel Potapov[1,2], Paolo Longo[3] and Axel Lubk[2]

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

The Principal Component Analysis (PCA) allows to denoise drastically STEM spectrum-images by extracting the meaningful fraction of data while cutting off the irrelevant noise. The number of meaningful PCA components is usually estimated through the evaluation of a scree plot – a dependence of the log eigenvalues (variances) on the component index [1]. Noise components are assumed to exhibit the lower variance and therefore could be visually isolated from meaningful ones on the scree plot. This strategy however introduces some subjectivity in the treatment and is hardly implemented into automatic flow algorithms. Recently, several approaches for automatic determination of the optimal "cut-off" based on the Gaussian nature of noise have emerged [2]. These methods perform nicely in the situation of the homoscedastic Gaussian noise, in particular for simulated spectrum-images. However, they tend to overestimate dramatically the "cut-off" for real STEM EELS and XEDS data because the real noise deviates from the Gaussian one. In this paper, we propose a novel practical method that utilizes other property of uncorrelated noise – directional isotropy.

For any couple of PCA components, a joint bivariate distribution of scores can be plotted. In case when both components represent pure noise, such a scatterplot must be isotropic in all directions (see the examples in Fig.1a, 2a). Therefore the domains of meaningful and noise-like components can be accurately isolated by tracking the anisotropy of scatterplots for the sequential couples of components: 1-2, 2-3, 3-4 et cet. How to quantify such anisotropy? Consider a series of histogram resulted from projections of a scatterplot onto differently oriented axes. The orientation of an axis is determined by a polar angle φ sketched in Fig.1a. We define a suitable measure for anisotropy of a scatterplot as

$$A = \frac{1}{mk} \sum_{\varphi=0}^{\varphi=2\pi} \sum_{i=1}^{s} (H_\varphi(i) - H(i))^2 - 1$$

where $H_\varphi(i)$ represent an s-channels histogram of projection on the axis oriented at angle φ, $H(i)$ is a histogram averaged over all φ, k is a number of projections and m is a number of pixels in a spectrum-image. It can be shown that criterion A approaches **0** for a perfectly isotropic scatter plot. This gives a great advantage over the traditional analysis of the variance that decays continuously from the unknown level in the noise region.

Fig.1 shows an example of atomically resolved XEDS spectrum-image where the anisotropy analysis detects clearly 3 meaningful PCA components. For the noise region, A oscillates in the close vicinity of **0** while the variance decreases gradually making the scree plot analysis more ambiguous. Fig. 2 represents a similar analysis for the EELS spectrum-image with 4 meaningful PCA components. The oscillation of A in the noise (isotropic) region looks quantitatively similar in XEDS and EELS data. Our experience suggests that noise can be successfully isolated by setting a threshold $A < $**0.5-1.0** for diverse EELS and XEDS spectrum-images. This can be easily implemented into an algorithm for automatic detection of the number of meaningful PCA components.

References:

[1] exemplary: M. Watanabe, E. Okunishi and K. Ishizuka, Microsc. Anal. **23/7** (2009), p. 5.
[2] exemplary: M. Gavish and D.L. Donoho, IEEE Trans. Information Theory **60** (2014) p.5040.
[3] PP and AL designed the treatment algorithms and PL acquired the experimental spectrum-images.
[4] PP acknowledges funding from DFG "Zukunftskonzept" (F-003661-553-Ü6a-1020605).
[5] AL received funding ERC under the Horizon 2020 program (grant 715620).
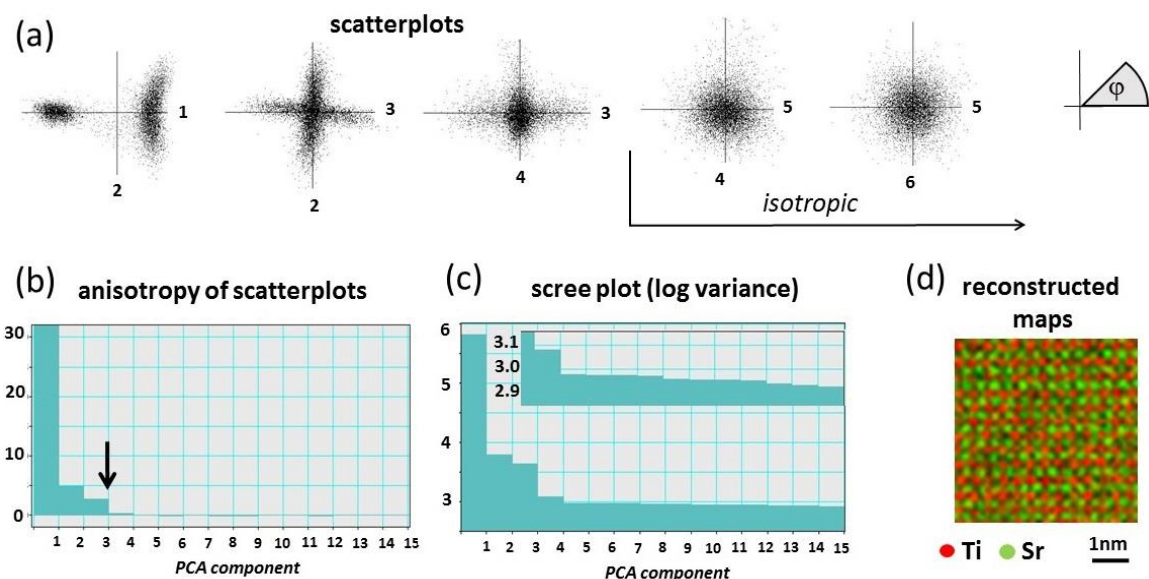
Figure 1. (a) Scatterplots of the PCA-decomposed XEDS spectrum-image of $SrTiO_3$ showing isotropic noise beyond the 3rd component. (b) Analysis of the scatterplots anisotropy allows to determine the "cut-off" component more reliably than (c) the classical scree plot approach. (d) Elemental maps reconstructed with the 3 PCA components.
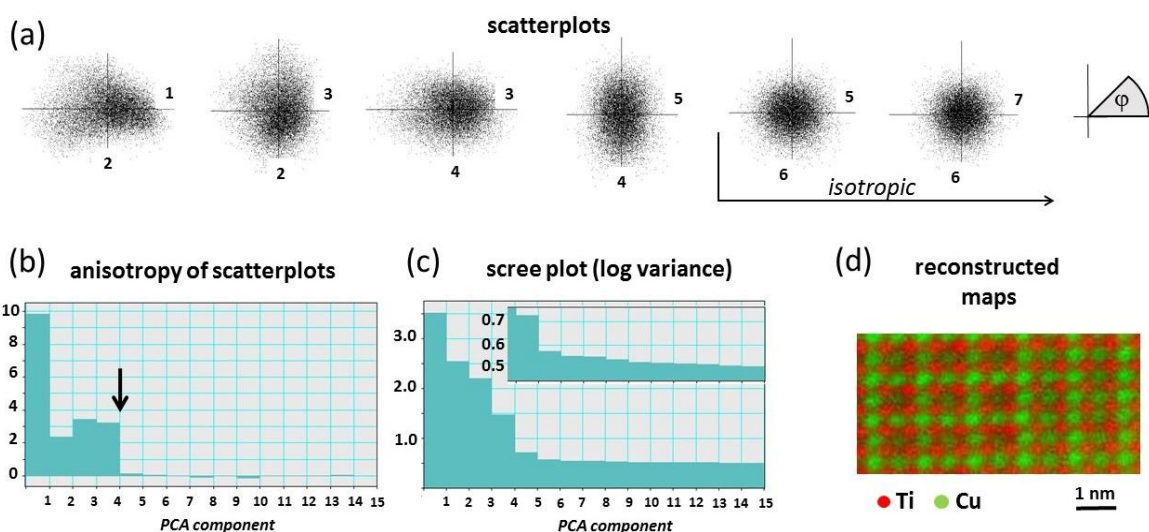

Figure 2. (a) Scatterplots of the PCA-decomposed EELS spectrum-image of $CaCu_3Ti_4O_3$ showing isotropic noise beyond the 4th component. (b) Analysis of the scatterplots anisotropy allows to determine the "cut-off" component more reliably than (c) the classical scree plot approach. (d) Elemental maps reconstructed with the 4 PCA components.