doi:10.1017/S1431927618002945

A Natural Basis for Unsupervised Machine Learning on Scanning Diffraction Data

Paul Cueva[1], Elliot Padget[1] and David A. Muller[1,2]

[1.] School of Applied and Engineering Physics, Cornell University, Ithaca, NY, USA
[2.] Kavli Institute at Cornell for Nanoscale Science, Ithaca, NY, USA

The advent of high-speed, high dynamic range, and low-noise pixelated detectors ushers in a new regime for scanning transmission electron microscopy (STEM) [1]. The efficient acquisition of a full diffraction pattern at each point in a scan enables analyses across a vast range of length scales. The information-rich 4D datasets are well suited to factor analysis, and machine learning techniques have proven invaluable for feature extraction in multidimensional datasets [2]. Efficient and physically transparent analyses remain challenging. Here we discuss a physically-motivated basis for efficiently representing scanning diffraction data and present interpretable results of component analysis.

Factor analysis is a popular approach for extracting features in large datasets. However, the inherently nonlinear relation between diffraction data and the underlying properties of a material makes linear methods less than ideal. While the nonlinearities can be accounted for by the careful training of supervised machine learning algorithms such as convolution neural networks, the need for extensive simulation can make this approach prohibitively time consuming [3]. Here, we transform the data into a basis where linear operations map to the correct nonlinear interaction, allowing standard component analysis to provide more physically meaningful results.

Linear decomposition aims to extract characteristic components and their concentration maps. An example ideal result for scanning diffraction data would give unique local structure factors as components and the corresponding thicknesses as concentrations. The relation between the measured diffraction patterns and the structure factors/thicknesses is highly nonlinear, but can be approximately linearized by applying a physically-motivated homomorphic filter to the data. Once the data is in this natural basis, linear techniques such as principal component analysis (PCA) and non-negative matrix factorization (NNMF) can yield interpretable components and concentrations. Figure 1 demonstrates the results of non-negative matrix factorization on raw data and on data in the proposed natural basis. The concentration map from the raw data is plagued by the usual artifacts of diffractive imaging whereas the natural basis data shows clean distinctions between grains without significant dynamical diffraction artifacts.

One of the greatest challenges in the application of unsupervised machine learning and factor analysis is the choice of metaparameters, such as the number of components. A common approach to determining the number of components is to first apply PCA and to examine the variance captured by each component. Once a suitable cutoff has been selected, then constrained analysis such as NNMF can be performed. Figure 1g compares the variance captured by components from raw data and those from the filtered data. The natural-basis data requires up to 68x fewer components to capture the same amount of variance.

Finally, analysis in this basis allows for the easy identification of structural changes such as strain relaxation, as demonstrated in Figure 2, even though the nanoparticle is not on zone axis. In conclusion, careful consideration of bases must be taken to maximize the utility of linear signal processing. Exploiting *a priori* connections between experiment and data permits more effective analyses [4].

References:

[1] M Tate, *et al.*, Microscopy and Microanalysis **22** (2016), pp. 237-249.
[2] A Belianinov, *et al.*, Advanced Structural and Chemical Imaging **1(1)** (2015), p. 6
[3] W Xu and J LeBeau, Microscopy and Microanalysis **23(S1)** (2017) pp: 120-121
[4] PC was supported by the Center for Bright Beams, an NSF STC (PHY-1549132). This work made use of the Cornell Center for Materials Research Shared Facilities, an NSF MRSEC (DMR-1719875 and NSF-MRI-1429155). EP supported by an NSF Graduate Research Fellowship (DGE-1650441).

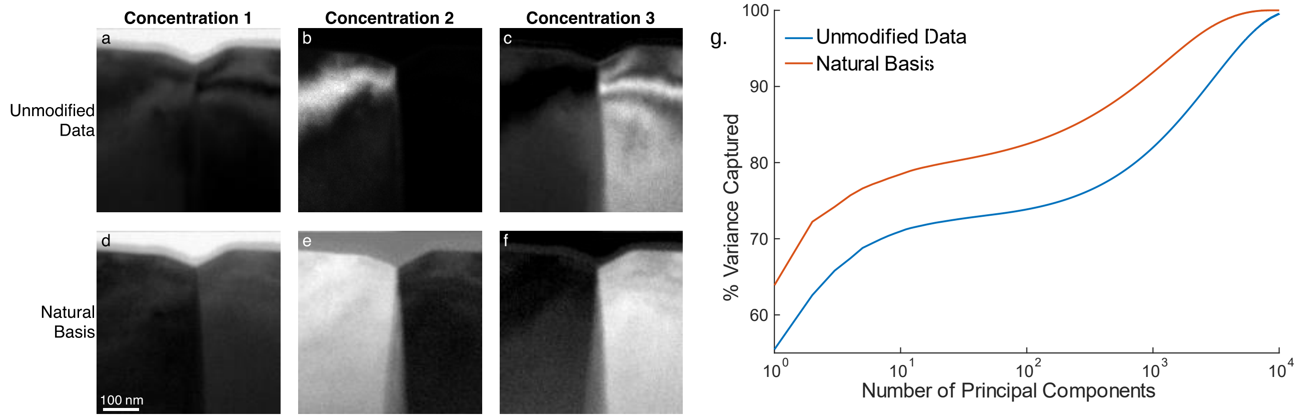

Figure 1. Comparison of the concentrations in the three-component Non-Negative Matrix Factorization of unmodified data (a,b,c) & of the same data in the natural basis (d,e,f). The 124x124x128x128 300kv nanodiffraction data is of a grain boundary in Nb_3Sn, sample courtesy of Daniel Hall & Mattias Liepe. NNMF on the unmodified data produces concentrations that are dominated by changes in thickness and tilt. Using the natural basis instead yields clear distinction between vacuum and the two grains, without diffraction contrast artifacts. (g) Comparison of the variance captured with each principal component. The unmodified data requires up to 68x more components to capture the same percentage of variance.

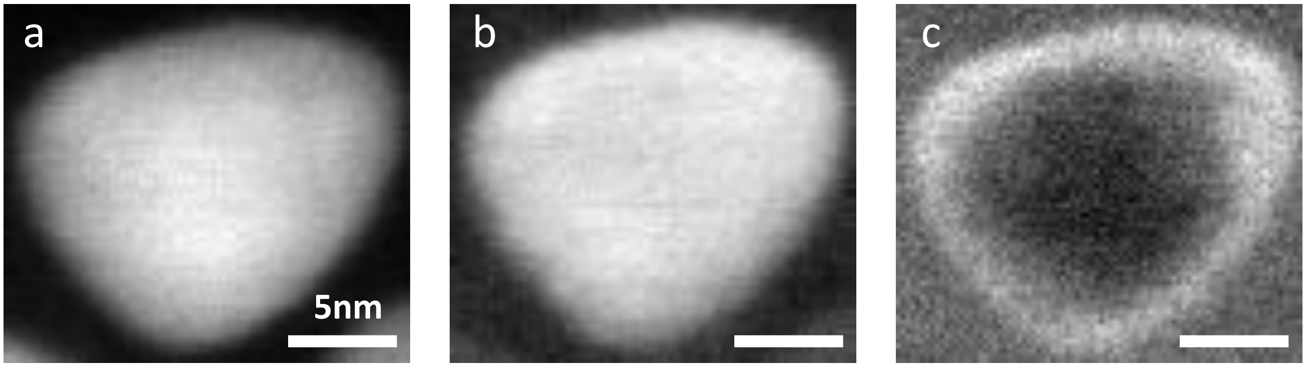

Figure 2. EMPAD data recorded from a randomly-oriented PtCo@Pt core-shell nanoparticle. (a) The extracted annular dark field image. (b,c) After natural basis transformations, the principal component analysis concentrations physically identifiable as Pt+Co (b) and the Pt shell showing strain relaxation (c).

https://doi.org/10.1017/S1431927618002945 Published online by Cambridge University Press