Andy Sode Anker
XVM402

Applications of principal component analysis to pair distribution function data[1]

Karena W. Chapman, Saul H. Lapidus and Peter J. Chupas

Background

While the atomic structure of solid-state crystalline materials can be characterized with routine X-ray diffraction techniques, it is much more challenging to study the structure of nanoparticles, disordered materials, solutions or amorphous materials as glasses. Nevertheless, with the increasing quality of international large-scale X-ray synchrotron facilities, it is now possible to overcome these limits by using high X-ray energy. X-ray Total Scattering with Pair Distribution Function (PDF) can be used to study molecules, ions, glasses, nanomaterials and clusters in solution at state-of-the-art synchrotron facilities. Thereby, the chemical reaction can be followed *in* situ, thus giving a deeper understanding of the chemistry in action through the reaction.

Introduction

The perspective of this article is how we deal with the enormous amount of PDF data we get with the developments in synchrotrons facilities. Especially *in situ[2] PDF* experiments were we typically measure a dataset per second is giving us loads of data and today the bottleneck is becoming the subsequently modelling.

However, the statistical method principal component analysis (PCA) can be used to give the first insight into the data without any prior knowledge of the materials structure, any modelling or fitting and the procedure is extremely fast. This article demonstrates that the method can be used to identify and quantify multiple chemical models of a series of data.

PCA is a multivariate method which converts a series of data with multiple correlated varies into a smaller number of uncorrelated principal components. The series of data is transformed into an orthogonal basis set, such that the data can be expressed as a linear combination of new orthogonal components. It is generally only a few of these new orthogonal components which contributes significantly to the data. The method has thereby reduced the variables. Rephrased does the PCA methods take advantage of correlations of the variables. It makes a correlation matrix of orthogonal components, where the eigenvectors of the matrix are the orthogonal basis set. The eigenvalues will thereby determine the importance of the orthogonal basis set, and it will typically reveal that the data can be closely reproduced by only a few of the principal components (PC's), where the one which describes most of the data will be called the first principal component (PC). The method can thereby give a strong identification of components in the data and how the ratio of them changes through the dataset.

The PCA procedure can be a helpful tool to identify chemical structures from PDF data without any prior knowledge of the structure. It can save the scientist an extremely time-consuming job of building the starting models, but also give a better estimate of where to actually start with the chemical analysis.

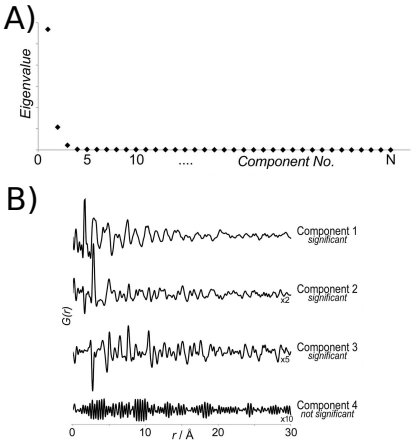

Figure 1: (a) A scree plot of the eigenvalues for all components, (b) the four components with the highest eigenvalues (shown scaled by factors of 1, 2, 5 and 10).

Principal Component Analysis of PDF data

This study presents three main advantages of using PCA to PDF data. First of all, the PCA method gives PC's, which can be related to chemical structures, phases or species in the sample. Secondly, these PC's found from PCA can be analyzed in the same way as the original PDF could, it ,therefore, gives us additional information when we start analyzing the data and also it minimizes bias in the modelling routine. Lastly, the amount of each PC can be followed through the reaction. In order to do the PCA procedure, the authors have used the program Origin Pro, which outputs PC's of the user's choice.

However, the PCA method is not always straightforward, since both the PCA procedure and the PDF data series must be robust in order to get valid results. As examples, they comment that the PDF data must be from the same sample with the same setup and the data reduction must be completely the same in all datasets, such as background subtraction, Q-range, masks etc. Furthermore, is it important that the PC's are properly scaled, which otherwise would bias the method to put more weight on some specific PC's. This can be a problem with systems with varying chemical composition, varying concentration, or if the scattering amplitude changes. Figure 1 shows the results of a PCA analyze of PDF data. The scree plot in figure 1a, gives information about the individual PC's. In this example, only the first three PC's contribute to describing the data. The

References

1. *Applications of principal component analysis to pair distribution function data.* Journal of Applied Crystallography, 2015. **48**(6): p. 1619--1626.
2. Ø., J.K.M., et al., *In Situ Studies of Solvothermal Synthesis of Energy Materials.* ChemSusChem, 2014. **7**(6): p. 1594-1611.

PC's have been plotted figure 1b, which shows that the three first curves give structural information, while, the fourth PC seem to just describe the noise.

In order to get physical values from the PCA, the authors suggest some restrictions to the PCA procedure, which all follows pure chemical logic.

Examples

In the article, the authors show 3 examples of situations where PCA has contributed to the data analysis. The first one is an example of a complex system with a large peak overlaps, which makes it very difficult for the human eye to distinguish phases. The PCA procedure could do this without any further problems though. The second example is on a large *in situ* dataset of a chemical system with changing structural motifs. Here the authors could relate the PC's to structures derived from an earlier conventional analysis. The last example is on a battery material, which is a mix of Li^+ and RuO_2, that changes through 3 different phases. In the article, the authors had success with identifying the PC's and determining what the ratio of them are through the reaction, figure 2. The reaction mechanism could thereby be established of this reaction by PCA, without any further conventional analysis. All examples were done on earlier experiments, where the solution was found, such that it was possible to benchmark the method.

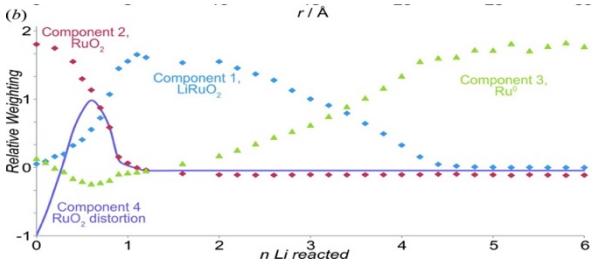

Figure 2: PCA analysis of PDF data obtained during electrochemical lithiation of RuO_2The weighting of the components during the reaction, where the distortion component varies during the transition from RuO_2 to

The last example is of my own data, which I did PCA of in order to get a feeling of how difficult it is to interpret. The results were quite amazing, for me at least, since the first PC showed the same model as I have found after months of work. Figure 3a shows the scree plot and 3b the first three PC with the $\{Bi_{38}O_{45}\}$ cluster. However, the second PC shows something, that I cannot identify. The authors of the article make the point that a chemical structure can be a linear combination of PC's, however, I suspect that the reason for my odd looking second PC is caused by a shift in the chemical composition, which means that I should have scaled the PC's relative to this chemical change. However, I had no success to do that. The third component is clearly noise, which is confirmed by the scree plot.

Instead, I used the method on newly published[3] results of data from a colleague' of mine, where the chemical composition does not change significantly. Here, I could identify multiple chemical structures with the PCA procedure, which is plotted in figure 3c and figure 3d.

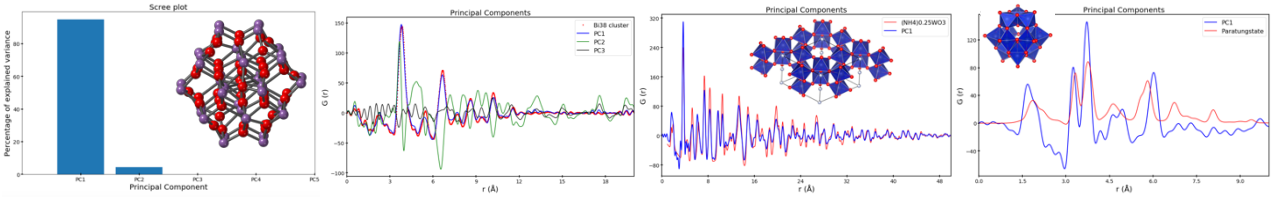

Figure 3: A) Scree plot of PCA of my own PDF data on bismuth cluster indicates that there are 2 significant principal components. B) The first principal component can be modelled to the $\{Bi_{38}O_{45}\}$ cluster, which I have also found with conventional fitting techniques. C, D) Two different structures identified with PCA, which match the structure of my colleagues newly published article.[3]

Conclusion

This article demonstrates, that multivariate procedures as PCA can benefit research areas with large amounts of data as materials chemistry with *in situ* PDF data. In the "Combinatorial appraisal of transition states for *in situ* pair distribution function analysis"[4], they used another multivariate method, CATS, to extract information about atomic to nanoscale details of crystallization from the PDF data. However, it seems like the method of choice is normally PCA, which seems to be a quit upcoming method to simplify large amounts of data in various research fields.[5] To my knowledge, it is not tremendous important which multivariate method we use, as long as they are only used to simplify the data in order to give better starting models for analysis and the results of the statistical method is not used as an ultimate solution. While the methods are extremely fast, can save scientist a lot of time of building models and can give a precise starting model, it does lack of chemical knowledge of how the atoms interfere. Therefore, these methods, which has been proven efficient and powerful, must be used with caution.

References

3. Juelsholt, M., T. Lindahl Christiansen, and K.M.Ø. Jensen, *Mechanisms for Tungsten Oxide Nanoparticle Formation in Solvothermal Synthesis: From Polyoxometalates to Crystalline Materials.* The Journal of Physical Chemistry C, 2019.
4. *Combinatorial appraisal of transition states for \it in situ pair distribution function analysis.* Journal of Applied Crystallography, 2017. **50**(6): p. 1744--1753.
5. Fernandez-Garcia, M., C. Marquez Alvarez, and G.L. Haller, *XANES-TPR Study of Cu-Pd Bimetallic Catalysts: Application of Factor Analysis.* The Journal of Physical Chemistry, 1995. **99**(33): p. 12565-12569.