Development of Automatic Phase Change Recognition for Accelerated Materials Discovery

Sam Dawley[1], David Elbert[2], Tyrel McQueen[1,3], and Apurva Mehta[4]
[1]*Department of Chemistry, Johns Hopkins University, Baltimore, MD, 21218*
[2]*Hopkins Extreme Materials Institute, Johns Hopkins University, Baltimore, MD, 21218*
[3]*Department of Materials Science and Engineering and William H. Miller III Department of Physics and Astronomy, Johns Hopkins University, Baltimore, MD, 21218*
[4]*Stanford Synchrotron Radiation Lightsource, SLAC National Accelerator Laboratory, San Mateo County, CA, 94025*

(Dated: August, 2022)

How do we know if a spectral peak is meaningful? Answering this question is central to extraction of information such as the onset of phase changes. Recent advances in experimental methods and detectors allows collection of more data more quickly and efficiently than ever before, providing the opportunity to leverage data-intensive methods, such as artificial intelligence and machine learning, to more rigorously evaluate spectra in real, or near real-time. Such on-the-fly data analysis provides the opportunity to drive the decision-making process during experimentation with live-streamed data, as it is being collected from an instrument. On-the-fly analysis is central to creating autonomous experimental control and characterization of fundamental phenomena such as phase changes. Firstly, however, an understanding of noise and signal structure must be established so as to allow rigorous, repeatable analysis of the spectral data stream. Herein a statistical algorithm and programmatic implementation for signal structure and phase change detection is introduced as a means of advancing methods for the on-the-fly data analysis.

I. Introduction

Phase changes in diffraction patterns are commonly recognized by the onset of peaks of prominent intensity. Although this metric suffices for ideal patterns with minimal to zero background noise, in practice this criterion does not discriminate very well for signals with random noise or when the initial onset of a transition or reaction progress is sought. Instead of local prominence, local variation may be used as a metric for the presence of a peak, analogous to how peaks could be detected in an ideal pattern, using the derivative of the intensity with respect to the angle (in the case of a powder X-ray diffraction pattern).

The local variation at any point in the spectrum can be determined in a variety of ways, including the variance, signal-to-noise ratio (SNR), or coefficient of variation, which is the reciprocal of the SNR, defined as $\gamma = \sigma/\mu$ for the standard deviation, σ, and mean, μ, intensity of a signal. Here, the coefficient of variation is used as a metric for peak detection.

In particular, the coefficient of variation is measured for small partitions of the diffraction pattern, instead of the entire pattern. Then, pairwise comparisons of the sample coefficient of variation for adjacent regions is made and used as a metric for peak detection. Importantly, it is the comparison of variation in *successive* regions we are interested in, ensuring that the noise is distributed very nearly identically in either partition.

Another way of considering this test is to assume that the diffraction pattern we are testing does not contain any signal peaks. Then, when variation is compared between adjacent regions, little to no change in variability is expected. If instead there is a large change in variability, contradicting our initial assumption, there is reason to believe that the variation in signal is changing and a peak is present.

i. Theory

This method of analysis is made possible because the ratio of the estimated sample coefficients of variation in two subsets of a population has been shown to follow an F-distribution. In our case, the population is all intensity values recorded on our diffraction pattern and the subsets of the population are the partitioned regions. The null hypothesis of the test becomes equality for the coefficients of variation. So, in the case that a peak exists in region i, when the coefficient of variation in region i and $i+1$ are compared, we reject the null hypothesis and assert that a peak exists in region i. Forkman[1] derives this test statistic as Equation (1),

$$F = \frac{c_1^2/(1 + c_1^2(n_1 - 1)/n_1)}{c_2^2/(1 + c_2^2(n_2 - 1)/n_2)} \tag{1}$$

for subsets 1 and 2 with sample coefficients of variation and sample sizes c_1, c_2, n_1, and n_2.

Importantly, this test statistic allows for the assignment of probabilities to all regions within a diffraction pattern corresponding to the likelihood of the presence of a peak. In theory, such a method can be generalized to any signal with peaks and random background noise.

II. Methods

Powder X-ray diffraction patterns of lanthanum hexaboride with impurities of lanthanum tetraboride were collected and used for analysis and testing. The only variable among different experiments was exposure time, with one-, two-, and four-minute exposures being used. A plot of these patterns is shown in Figure 1.

All plotting and analysis was performed using Python. Notably, partitioning the data sets was carried out using a range of angles determined by the full width at half maximum of the most intense peak in the pattern. Subsequently, the determination of the sample mean and variance for each partition was determined by bootstrapping. An example partitioning of the data set is shown in Figure 2.

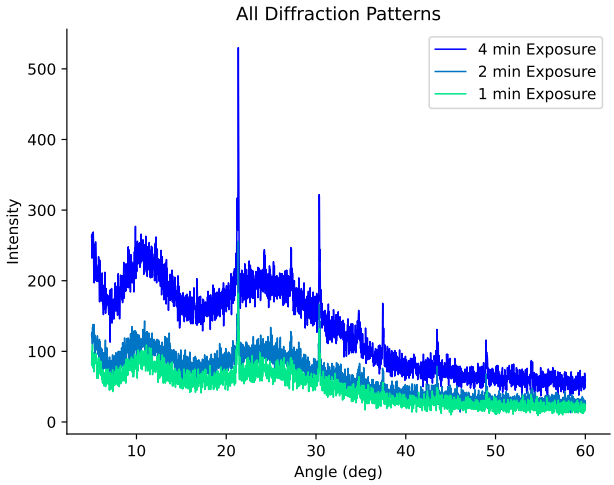


Figure 1: Diffraction patterns of lanthanum hexaboride for all exposure times considered.

After partitioning, direct comparison of the variation in adjacent partitions was done by Equation (1).

III. Results

The results of the algorithm are illustrated below. Figure 3 depicts the probabilities assigned to the partitions of a one-, two-, and four-minute expo-

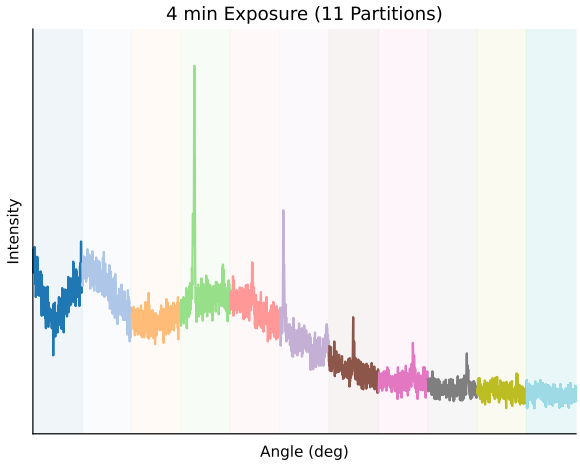
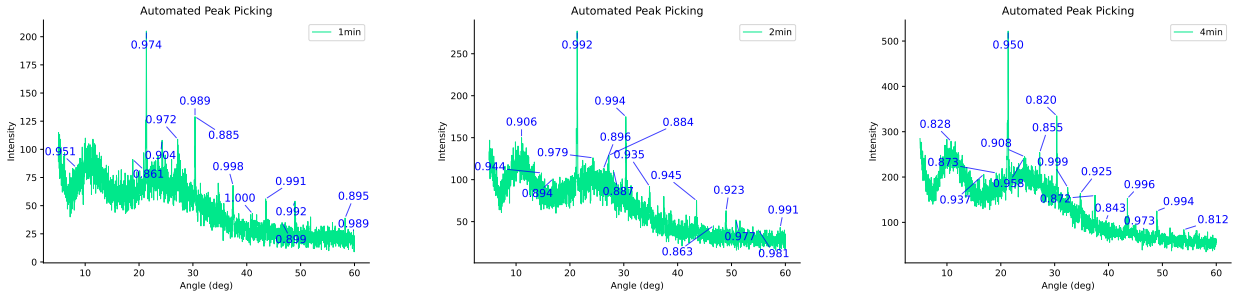


Figure 2: Example of partitioning data set into discrete regions for pairwise comparison of sample coefficients of variation. The partitions shown are artificially large. In practice, the data set is split up into more than 50 regions.



(a) Peak likelihoods based on a 1-minute exposure.

(b) Peak likelihoods based on a 2-minute exposure.

(c) Peak likelihoods based on a 4-minute exposure.

Figure 3: Performance of automated peak-picking algorithm on noisy spectra for sample of lanthanum hexaboride with minor impurities. Threshold for annotating on each plot is a peak likelihood of 0.8.

sure, i.e., the performance of the algorithm. In practice, this algorithm would return a file containing peak likelihoods for all partitions of the data set.

An interesting aspect of the assigned likelihoods is how large they are. Many of the falsely assigned points on the pattern maintain probabilities well over 0.8, and while the true peaks typically result in more significant test statistics and higher probabilities, they are only marginally more significant. This result is potentially due to the very short exposure times chosen for testing, those being no greater than four minutes.

Another interesting aspect is in Figure 3a, with the double assignment of a peak to the same region at around 30 degrees. This could lead to issues if we are trying to convert the file of peak likelihoods into a crystal structure.

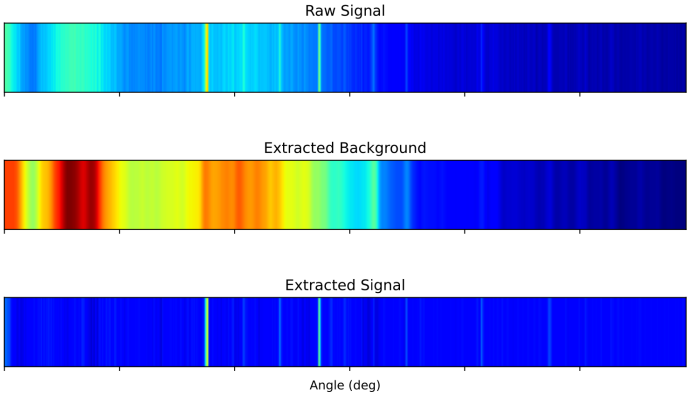


Figure 4: Noise extracted and subtracted from XRD diffraction of lanthanum hexaboride after four minute exposure. Method for noise extraction and visualization taken from collaborators at SLAC, with permission.

IV. Discussion

i. Future Directions

As mentioned previously, there are clear outstanding issues with the current algorithm, namely, the double assignment of peaks as well as parameter optimization. The former is not a large issue, as a simple check to determine if a peak is assigned to the same angle can be performed to ensure that this does not happen.

The latter may require more work, either by theoretical or brute force computational means. An obvious parameter that is pattern-dependent and needs to be optimized to the data set is the partition size. A good standard that was used here is the full-width at half-maximum of the most intense peak in the signal. Though, tests to see how the half-width at half-max, or the third-width at half-max perform may be beneficial, for example.

Another parameter which could be optimized is exposure time of the experiment. More specifically, determining the exposure time for a given sample which results in peak likelihoods which exceed some specified threshold. For example, finding the time which leads to the most intense signal having a peak likelihood of over 95%. This direction parallels work that may be performed with collaborators at the SLAC National Laboratory.

ii. Broader Context

In the broader context of on-the-fly data analysis for data-driven experimentation, work presented here is focused on analyzing and identifying peak structure in noisy signals. In combination with an understanding of *noise* structure in these signals, picking out significant deviations from the baseline of a signal, i.e., picking out peaks, can be performed instantaneously, as data is collected and streamed from an instrument.

In fact, our collaboration with the SLAC National Laboratory has afforded a method for extracting noise from a signal[2], shown in Figure 4. The corresponding processed pattern, as well as the raw data, is shown in Figure 5.

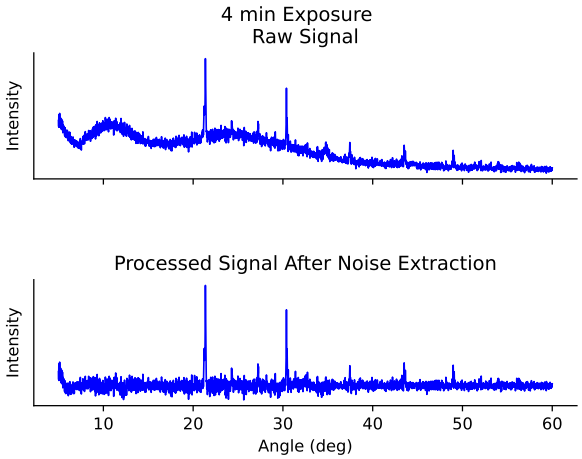


Figure 5: Alignment of raw data with processed data of lanthanum hexaboride diffraction pattern after noise extraction.

References

[1] Johannes Forkman (2009) Estimator and Tests for Common Coefficients of Variation in Normal Distributions, *Communications in Statistics - Theory and Methods*, 38:2, 233-251, DOI: 10.1080/03610920802187448.

[2] Hoidn, Oliver. (2022) *xrd_clustering*. Retrieved from https://github.com/hoidn/xrd_clustering.