Leveraging local continuity of connected measurements for dimensionality reduction and separation of backround, signal, and noise....

1. Introduction

Experiments undertaken at synchrotron and XFEL light sources can generate large volumes of data that are rich in information but plagued by noise and artifacts. There is often much continuity between successive measurements: when collecting a progression of diffraction patterns or spectra over a continuous range of physical states, the standard practice is to capture many intermediate states of the corresponding evolution of the signal. Furthermore many methods–such as powder diffraction and serial crystallography–require repeated measurements of the same system to accumulate sufficient statistics or orientational averaging.

For these two reasons large spectroscopic or diffraction datasets generated in typical experiments at third- or fourth-generation light sources are likely to contain sequences of measurements with smooth variation with respect to the underlying physical information. Equivalently, we say that such a dataset can be assembled into a high-dimensional array where the underlying signal is both *smooth* and *connected*. More explicitly, any such dataset can be represented as the tensor

$$X_{i_1 i_2 ... i_N q} \tag{1}$$

for which each distinct sequence of values for the indices $i_1, i_2, ...i_N$ corresponds to a single spectrum or diffraction pattern, q indexes momentum transfer or energy in the case of XRD or spectroscopy, respectively, and each of the indices $i_n, n \in \{1, 2, ...N\}$ corresponds to an ordering dimension along which direction the underlying signal's variation is smooth. The dimensionality of the dataset is thus simply one higher than N, the number of ordinal attributes.

This connectedness describes both the signal and any smoothly-varying background, while other contributions to the measured X-ray intensity–such as shot noise, one-off artifacts, and stochastic variations across different portions of the sample–are uncorrelated or at least substantially stochastic. (note: high-frequency variation is a more accurate description here, uncorrelated is a limiting case but it simplifies the interpretation).

In this paper, we take two examples of XRD datasets–one temperature study of xxxx material and one combinatorial study of the CoNiTi ternary system–to

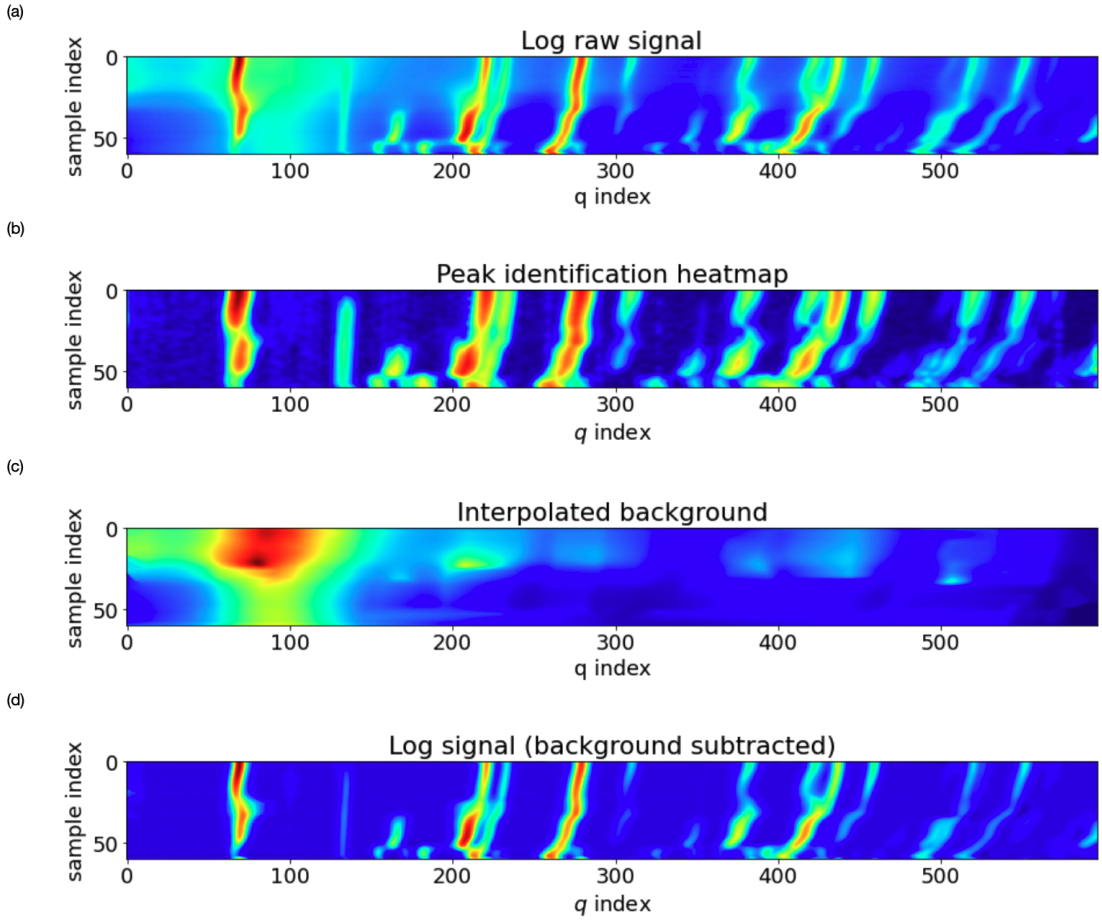

Figure 1. (a) Heatmap of XRD dataset X_{iq} corresponding to a temperature scan of XRD patterns for xxxxx system. The vertical index i and horizontal index q index temperature and momentum transfer, respectively.
(b) ...
(c)
(d)

demonstrate the separation of connected data into a physically 'meaningful' component and a component that comprises noise and artifacts. At the most basic level, we see that the signal is readily separated into two components: $X = \hat{X} + \eta$, where $\hat{X}$ is a slow-varying signal and η is a fast-varying noise contribution. The approach depends only on the minimal assumption of connectedness and so should be applicable, with minor tuning, to any dataset.

2. Methods

2.1. Background subtraction

The first step to estimating the background is to identify the peak regions, which must be excluded before calculating the background. In the case of typical diffraction data a simple high-pass DFT filter does not cleanly extract the peaks when phase is retained

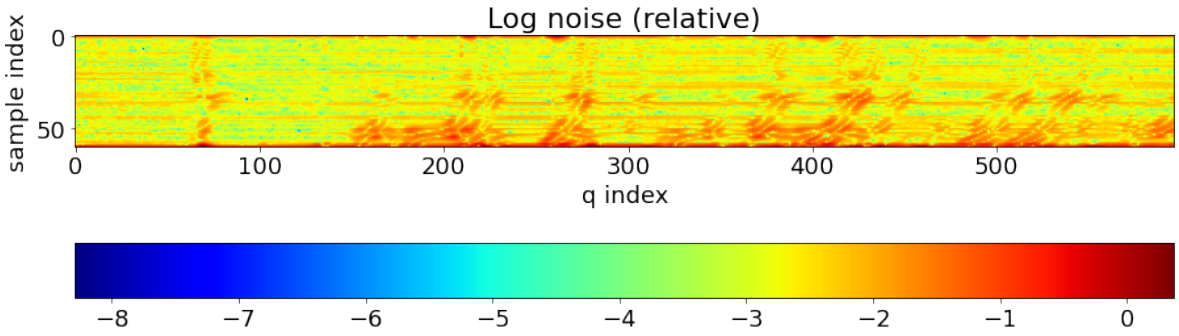

Figure 2.

in the inverse Fourier transform. This is because of the presence of ringing artifacts as well as the high concentration of signal in the diffraction peaks, which pile on top of the background in the low-frequency region of the power spectral density. On the other hand we can identify peak regions by applying a threshold on the transformed signal illustrated in Fig. 1(b), which we calculate as

$$N(0,\sigma) \circledast |F^{-1}(HF(q))|, \tag{2}$$

where N is a unit Gaussian in q with standard deviation σ is chosen to match the diffraction peak width, $\circledast$ denotes convolution, and H is a high-pass Gaussian window. The background can then be simply estimated by an interpolation using data from non-peak regions to fill in background intensities within the peak regions (Fig. 1(c)) and finally subtracted from the denoised data (Fig. 1(d)).

To make for a simple extension to datasets of arbitrary dimensions we estimate the background by linear interpolation in the q dimension alone, together with an N-dimensional nearest-neighbor assignment to fill in points out of range for interpolation.

2.2. Noise estimation

While in the case of intensity variation in the q dimension we assume that high-frequency features belong to the informative (i.e. physical) component of the signal, in the case of the ordering dimensions we make the opposite assumption that the physically relevant part of the signal varies smoothly from one XRD pattern to any adjacent one, while any high-frequency, uncorrelated deviations from this progression are due to either noise (from e.g. detector characteristics or Poisson statistics) or artifacts (such as insufficient orientational sampling of the diffracting crystallites). Under this assumption the signal and noise lack substantial overlap in the N-1-dimensional Fourier transform along ordering dimensions; thus we can use a simple DFT filter to separate them (Fig. 2)

The above estimations of background and noise components separate the signal into estimates of the physical components (separated into background + diffuse scattering and diffraction) and uncertainty values that correspond to single samples

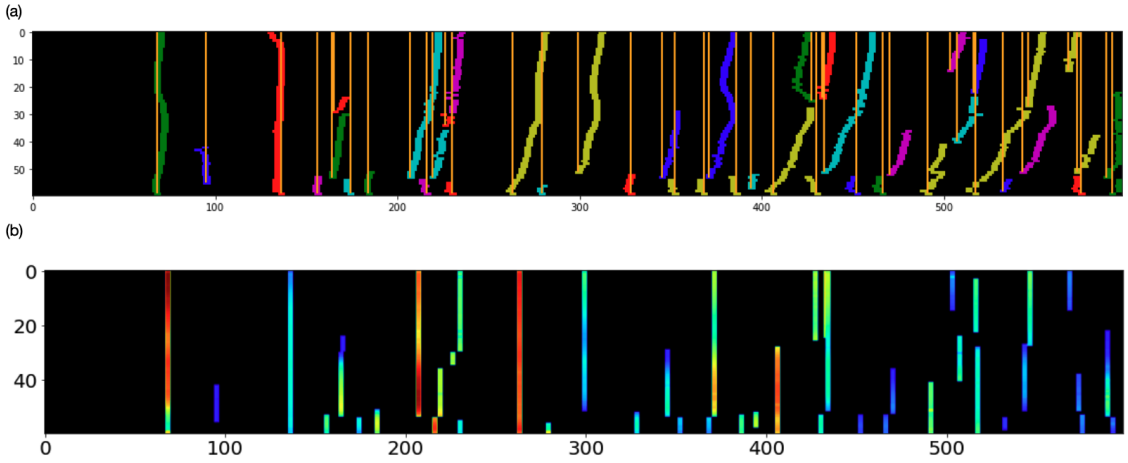

Figure 3. (a) Color-coded features identified from peaks connected in i, q space for XRD dataset X_{iq} corresponding to a temperature scan of XRD patterns for xxxxxx system. Index i denotes temperature.
(b) Intensity profile along each of the features in (a).

of an underlying noise distributions. Despite its incompleteness, the latter can be used as an input for sensitivity analysis on downstream models.

3. Discussion

3.1. Application: feature extraction

The above analysis addresses the data itself, with no scientific interpretation aside from the separation between crystalline and diffuse contributions to the scattering signal. However, we find that the same ordering and smoothness properties are useful for reducing the data into salient *information*, where the goal is to find physically-meaningful boundaries in the ordering dimensions corresponding to, e.g., the surface separating a single-phase region from surrounding multi-phase regions in a combinatorial diffraction dataset. Specifically, we can identify diffraction peaks in every XRD pattern, independently, and then link each peak in a pattern to any peaks in adjacent patterns centered at a nearby q value. We next define each set of linked peaks as a *feature* . ..... utility of this is that it removes peak-shifting, etc. . . .

4. Conclusions and future work

5. References