

Leveraging local continuity of connected measurements for dimensionality reduction and separation of background, 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 an intrinsic connectedness between successive measurements taken on a progression of different physical states: for example, a series of thermodynamic states in a phase diagram. There may also be further data redundancy, as many methods—such as powder diffraction and serial crystallography—require repeated measurements of the same system to accumulate sufficient statistics or orientational averaging. Consequently, large spectroscopic or diffraction datasets generated at third- or fourth-generation light sources are likely to contain sequences of measurements that combine smooth variation with respect to underlying physical parameters with stochastic variation due to counting statistics, detector characteristics, sample nonuniformity and other artefacts. With respect to variation in underlying physical parameters, such a spectroscopic or diffraction dataset can be assembled into a tensor of the form

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

for which each distinct sequence of values for the indices i_1, i_2, \dots, 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 $i_n, n \in \{1, 2, \dots, N\}$ 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 independent 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 xxx material and one combinatorial study of the CoNiTi ternary system—to

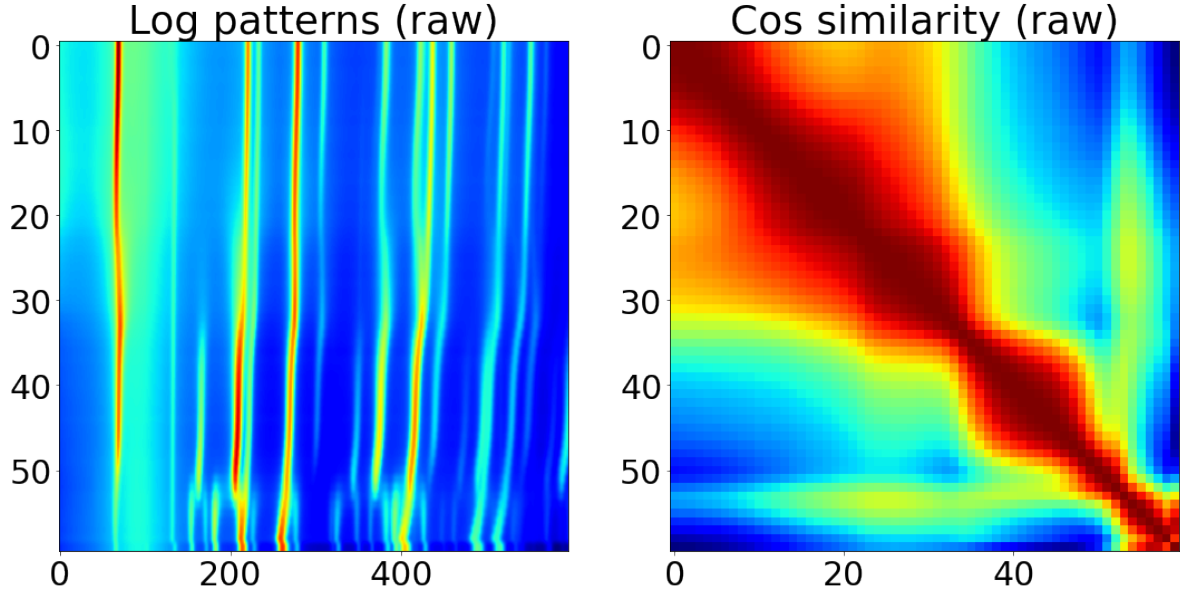


Figure 1.

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 slow-varying in the ordinal dimensions and η is a fast-varying noise contribution (again, in the ordinal dimensions). The approach depends only on the minimal assumption of connectedness and so should be applicable, with minor tuning, to any dataset.

In the particular case of an XRD dataset the signal is concentrated in peaks that are narrow in q -space, which we show allows a robust separation of any background with slow q -variation (e.g. diffuse scattering).

Finally, the most compelling exploitation of connected data emerges in its *interpretation*—especially in the propagation of noise estimates into uncertainties in any downstream model-derived quantities, and in the use of connectivity to remove domain-specific measurement artifacts. In section xxxx we show a feature extraction scheme for XRD phase mapping datasets that is based on local connectivity. The approach is robust to peak-shifting, which until now has been long-standing frustration for the interpretation of such datasets. *say why this is new and good, motivate and compare to prior efforts in the community*

2. Methods

2.1. Prior approaches

Summarize approaches using simple similarity measures, such as cosine distance in the diffraction space.

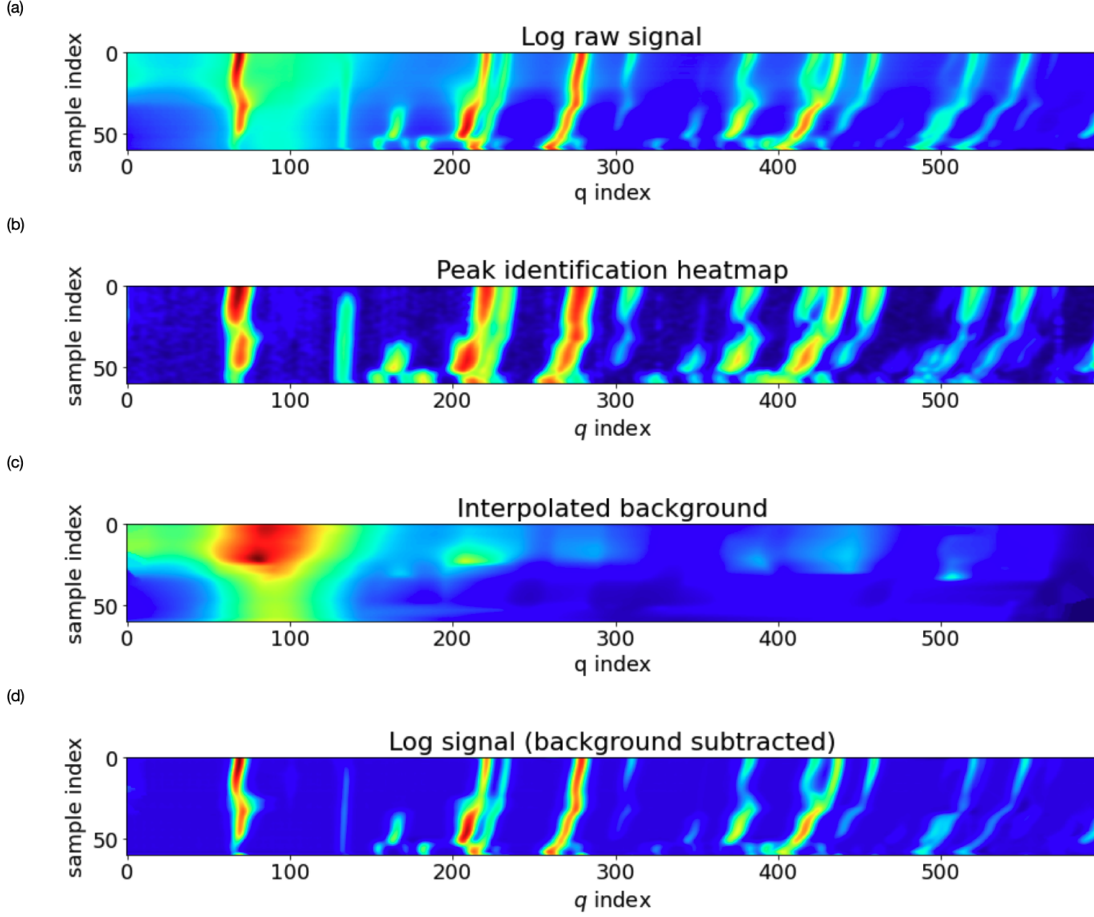


Figure 2. (a) Heatmap of XRD dataset X_{iq} corresponding to a temperature scan of XRD patterns for xxxxx system. The vertical index i and horizontal index q index temperature and momentum transfer, respectively.

2.2. 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 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

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

where N is a unit Gaussian in q with standard deviation σ is chosen to match the diffraction peak width, \otimes denotes convolution, and H is a Blackman window

this part will be fleshed out. some steps are still omitted

The background is then estimated by an interpolation using data from non-peak

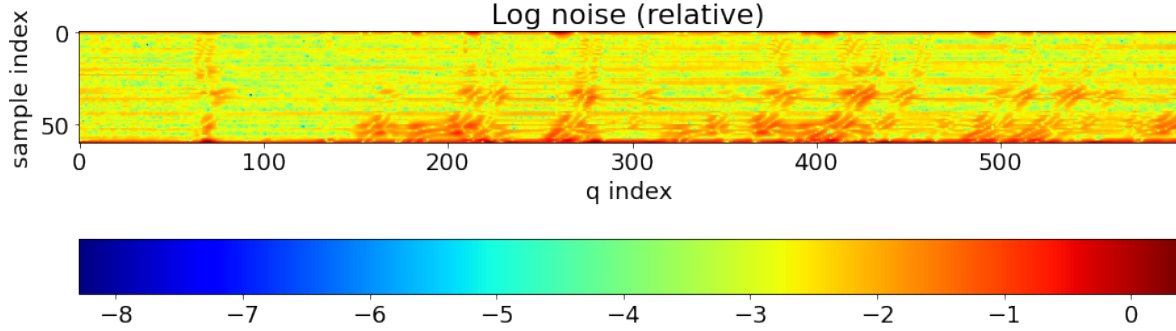


Figure 3.

regions to fill in background intensities within the peak regions (Fig. 1(c)). Finally, it is 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.

2.3. 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 of an underlying noise distributions. The latter can be used to propagate uncertainties to any downstream analyses.

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

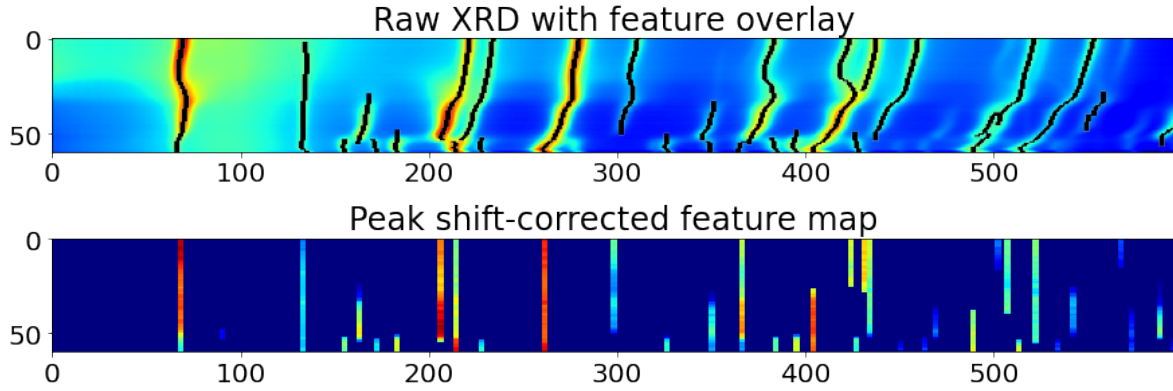
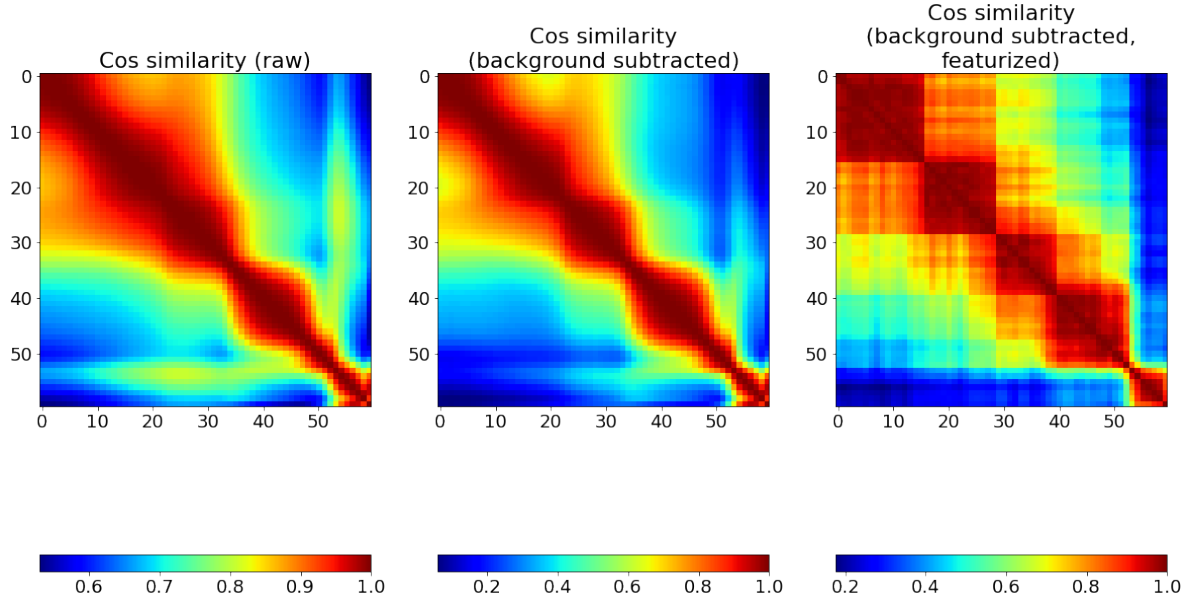


Figure 4. (a) Color-coded features identified from peaks connected in i, q space for XRD dataset X_{iq} corresponding to a temperature scan of XRD patterns for xxxxxx system. Index i denotes temperature.

(b) Intensity profile along each of the features in (a).



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. . . .

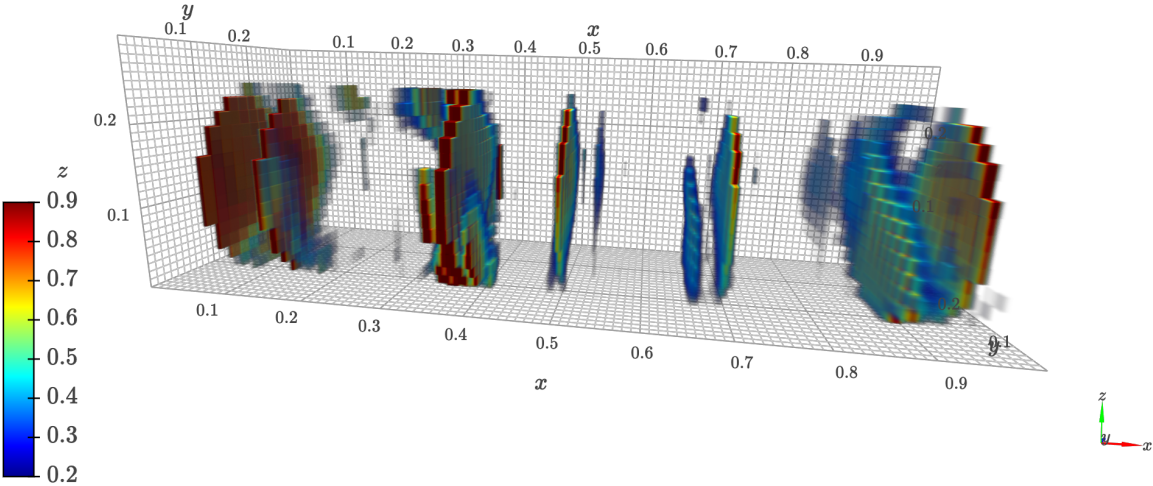


Figure 5. caption

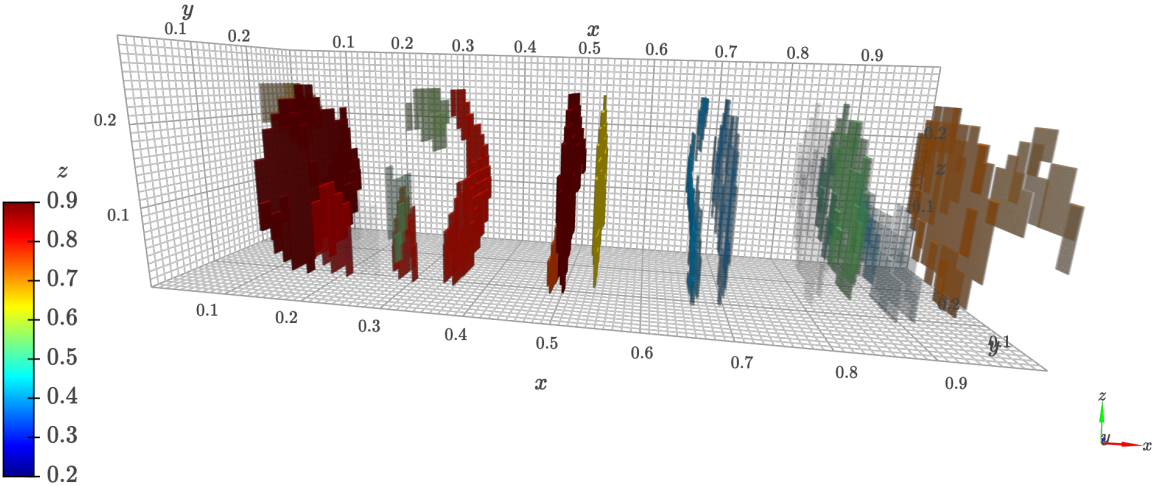


Figure 6. caption

- 4. Conclusions and future work
- 5. References