Manifold Alignment by Feature Correspondence

Jay S. Stanley III 1, Scott Gigante 1, Guy Wolf 4†, and Smita Krishnaswamy 2,3†
1 Computational Biology and Bioinformatics Program, 2 Department of Computer Science,
3 Department of Genetics, Yale University, New Haven, CT, USA
4 Department of Mathematics and Statistics, Université de Montréal, Montréal, QC, Canada

Abstract—We propose a novel framework for combining datasets via alignment of their associated intrinsic geometry. This alignment can be used to fuse data originating from disparate modalities, or correct batch effects while preserving intrinsic data structure. Importantly, we do not assume any pointwise correspondence between datasets, but instead rely on correspondence between a subset of features. We leverage this assumption to estimate relations between intrinsic diffusion dimensions, which are computed from graph (or manifold) Fourier coefficients of data features. These relations are then used to form an isotropic alignment between the diffusion maps of each dataset, which is subsequently used to form a unified diffusion geometry of the combined data which can be used to correct data in ambient space. We demonstrate our method on several datasets, and in particular, its effectiveness in biological applications including data fusion and batch effect removal.

I. INTRODUCTION

In many natural science settings, an often encountered problem is that while data are measured from the same system, they are collected by different equipment or on different days, where sensors are calibrated differently. This is often termed batch effect in biology and can include drastic variations between subjects, experimental settings, or even times of day when an experiment is conducted. In such settings, it is important to globally and locally align the datasets such that they can be combined for effective further analysis. Otherwise, measurement artifacts may dominate downstream analysis.

A related problem appears in multimodal data fusion, where we seek to integrate data from multiple measurements on the same underlying system. For instance, the central dogma of biology links protein expression to RNA presence; thus, protein and RNA content of similar cells should be correlated. Joint analysis allows one to relate RNA and protein dynamics in such a way that isolated experiments cannot. Joint analysis of these datasets, and in particular, its effectiveness in biological applications including data fusion and batch effect removal.

To find such a transformation, we propose to utilize the duality between diffusion coordinates and geometric harmonics that act as generalized Fourier harmonics. This duality allows us to capture a cross-sample harmonic correlation matrix \( C_{1 \times 2} = \text{corr}(\Phi_1^{(1)}, \Phi_2^{(2)}) \) between the harmonics that serve as coordinates of \( \Phi_1^{(1)} \) and \( \Phi_2^{(2)} \). Since we assume feature correspondence between samples, we can compute these correlations by expressing these eigenvectors in terms of the GFT of the data features. For each sample \( s \) we construct an \( N^{(s)} \times n \) matrix \( \tilde{X}^{(s)} \) whose \( j \)-th column is \( f_j^{(s)} \). Then, \( C \) can be computed with correlations between rows of \( \tilde{X}^{(1)} \) and \( \tilde{X}^{(2)} \).

Given the cross-sample harmonic correlation matrix \( C \), we use its SVD given by \( C = U \Sigma V^T \) to obtain its nearest orthogonal approximation \( T = U V^T \) that defines an isometric transformation between the diffusion maps. Finally, we can now compute a unified diffusion map, which can be written in (block) matrix form as

\[
\Phi_t^{(1,2)} = \begin{bmatrix} \Phi_0^{(1)} & \Phi_0^{(2)} \\ \Phi_0^{(1)} T & \Phi_0^{(2)} \end{bmatrix} \begin{bmatrix} \Lambda_0^{(1)} & 0 \\ 0 & \Lambda_0^{(2)} \end{bmatrix}^L, \tag{1}
\]

where \( \Lambda_0 \) are diagonal matrices with the diffusion eigenvalues \( \{\lambda_k^{(s)}\} \) as their main diagonal. The algorithm generalizes to multiple samples by considering \( S \times S \) blocks in \( \tilde{X} \), with computed harmonic alignment transformation \( T^{(s_1,s_2)} \) in each \( s_1 \) block with \( s_1 \neq s_2 \).

Finally, we construct a new kernel over the combined diffusion coordinates and build a robust unified diffusion geometry over \( X = \bigcup_{s=1}^S X^{(s)} \) that is invariant to batch effects and also naturally denoises various sample-specific artifacts. This diffusion geometry can naturally be incorporated in diffusion-based methods for several data processing tasks. Results show that our method successfully aligns artificially misaligned samples, as well as biological data containing batch effects and multiple data modalities, and outperforms state-of-the-art manifold alignment methods on these tasks.

Algorithm 1 Harmonic Alignment

**Require:** Dataset \( \tilde{X} = X^{(1)} \cup X^{(2)} \) with \( n \) features and two batches (i.e., sub-datasets), where each batch \( X^{(s)} \) has \( N^{(s)} \) observations

**Ensure:** Aligned diffusion map \( x \mapsto \Phi_t^{(1,2)}(x), x \in \tilde{X} \)

1: for \( s \in \{1,2\} \) do
2: Compute the \( N^{(s)} \times N^{(s)} \) anisotropic kernel \( k^{(s)} \) over \( X^{(s)} \).
3: Compute the diffusion map \( \Phi_t^{(s)} \).
4: end for
5: Compute the \( N^{(1)} \times N^{(2)} \) bandlimited correlation matrix \( C \).
6: Orthogonalize via SVD \( C = U \Sigma V^T \) to get \( T = U V^T \)
7: Compute the unified diffusion map \( \Phi_t^{(1,2)} \) (Eq. (1))
Input data with the neighborhood of were classified using nearest neighbors in \( X \) the cell neighborhoods. MNN and Harmonic Alignment. Harmonic Alignment most accurately recovers shows the average percentage overlap of the neighborhood of \( f_i \) between data modalities. Let them, we align the two manifolds in order to recover the known bijection randomly permuting the datasets to scramble the correspondence between and chromatin profiling of 11,296 cells from adult mouse kidney [13]. After Fig. 2. Percentage overlap of cell neighborhoods from joint gene expression Fig. 1. Recovery of \( k \)-neighborhoods under feature corruption. (a) Two sets of 1000 points were sampled from MNIST, one of which was distorted by a corruption matrix \( O_p \) for various identity percentages \( p \) such that \( p \% \) of the columns \( O_p \) are drawn from the identity matrix, and the remainder of the columns are drawn from a random orthogonal matrix. Points in \( X^{(2)} O_p \) were classified using nearest neighbors in \( X^{(1)} \) after harmonic alignment with different bank sizes, mutual nearest neighbors (MNN) [12], and no alignment. From left to right: Input; 75% corrupted pixels; Reconstruction by averaging \( k \)-neighborhoods without alignment fails to recover digit identity; Reconstruction with harmonic alignment correctly recovers the corrupted digits.

Reconstructions

Input data

Raw Corrupted

Naive

Aligned

No alignment

MAGAN

MNN

Uncorrupted Features (%)

5-nn classification accuracy (%)

100%

75%

50%

25%

0%

0% 25% 50% 100%

75%

Noisy Inflammatory

Response

Denoised Response

Aligned Response

TFNα Abundance

TFNα Abundance

IFNγ Abundance

Denoised Response

Sample 1

Sample 2

P(x)

P(x)

P(x)

Fig. 3. Batch effect removal in 4000 from two single-cell immune profiles obtained via mass cytometry on blood samples of two patients infected with Dengue fever [14]. (a) Data before denoising shows a mild batch effect. (b) Denoised data enhances a batch effect in IFNγ. (c) Harmonic alignment corrects the shift.

Fig. 1. Recovery of \( k \)-neighborhoods under feature corruption. (a) Two sets of 1000 points were sampled from MNIST, one of which was distorted by a corruption matrix \( O_p \) for various identity percentages \( p \) such that \( p \% \) of the columns \( O_p \) are drawn from the identity matrix, and the remainder of the columns are drawn from a random orthogonal matrix. Points in \( X^{(2)} O_p \) were classified using nearest neighbors in \( X^{(1)} \) after harmonic alignment with different bank sizes, mutual nearest neighbors (MNN) [12], and no alignment. From left to right: Input; 75% corrupted pixels; Reconstruction by averaging \( k \)-neighborhoods without alignment fails to recover digit identity; Reconstruction with harmonic alignment correctly recovers the corrupted digits.

REFERENCES