Joint Geometric and Topological Analysis of Hierarchical Datasets

Lior Aloni[1] ✉, Omer Bobrowski[1], and Ronen Talmon[1]

Technion, Israel Institute of Technology, Haifa, Israel
{lioral@campus, omer@ee, ronen@ee}.technion.ac.il

Abstract. In a world abundant with diverse data arising from complex acquisition techniques, there is a growing need for new data analysis methods. In this paper we focus on high-dimensional data that are organized into several hierarchical datasets. We assume that each dataset consists of complex samples, and every sample has a distinct irregular structure modeled by a graph. The main novelty in this work lies in the combination of two complementing powerful data-analytic approaches: topological data analysis (TDA) and geometric manifold learning. Geometry primarily contains local information, while topology inherently provides global descriptors. Based on this combination, we present a method for building an informative representation of hierarchical datasets. At the finer (sample) level, we devise a new metric between samples based on manifold learning that facilitates quantitative structural analysis. At the coarser (dataset) level, we employ TDA to extract qualitative structural information from the datasets. We showcase the applicability and advantages of our method on simulated data and on a corpus of hyper-spectral images. We show that an ensemble of hyper-spectral images exhibits a hierarchical structure that fits well the considered setting. In addition, we show that our new method gives rise to superior classification results compared to state-of-the-art methods.

Keywords: manifold learning · diffusion maps · topological data analysis · persistent homology · geometric learning

1 Introduction

Modern datasets often describe complex processes and convey a mixture of a large number of natural and man-made systems. Extracting the essential information underlying such datasets poses a significant challenge, as these are often high-dimensional, multimodal, and without a definitive ground truth. Moreover, the analysis of such data is highly sensitive to measurement noise and other experimental factors, such as sensor calibration and deployment. In order to cope with such an abundance, various data-analytic approaches have been developed, aimed at capturing the structure of the data. These approaches are often unsupervised, and are designed specifically to address the "curse of dimensionality" in data.

In this paper we consider two complementing approaches for such "structural" data analysis. Both of these approaches are based on the assumption that in high-dimensional real-world data, most of the information is concentrated around an intrinsic low-dimensional structure. Recovering the simplified underlying structure may reveal the true degrees of freedom of the data, reduce measurement noise, and facilitate efficient subsequent processing and analysis. The first approach we consider focuses on the *geometry* of the data, and is called *manifold learning* [27, 24, 2, 8]. The second approach focuses on the *topology* of the data, and in known as *topological data analysis* (TDA) [5, 28]. The key difference between these approaches is the distinction between local and precise phenomena (captured by geometry), and global qualitative phenomena (captured by topology). Briefly, the goal in manifold learning is to obtain an accurate geometric representation of the manifold that best describes the data. This is commonly accomplished by approximating the Laplace-Beltrami operator of the manifold. On the other end, TDA promotes the analysis of shapes and networks using qualitative topological features that are coordinate-free and robust under various types of deformations (e.g. the existence of holes). Our goal here is to take advantage of the strengths of each of these approaches, and combine them into a powerful geometric-topological framework. Conceptually, the common thread between manifold learning and TDA is the premise that the true information underlying the data is encapsulated in the "network" of associations within the data. Here lies another key difference between these two approaches. Manifold learning methods traditionally represent such networks as graphs (i.e. nodes and edges). While graphs serve as a powerful model for various applications, this approach is limited since it can only capture *pairwise* relationships between nodes. However, it is highly conceivable that complex data and networks consist of much more intricate interactions, involving more than just two nodes at a time. The methods developed in TDA focus on hypergraphs (simplicial complexes) that allow for high-order associations to be incorporated into the model [10, 14, 19].

In this work, we propose to combine manifold learning and TDA in order to provide informative representations of high-dimensional data, under the assumption that they can be arranged into several *hierarchical datasets* as follows. We assume that we have a collection of datasets, each consists of several complex samples, where each individual sample has a distinct irregular structure that can be captured by a weighted graph. Such datasets arise in many applications from a broad range of fields such as cytometry and gene expression in bioinformatics [13, 12], social and computer network analysis [22, 18], medical imaging [16], and geophysical tomography [1]. Following this hierarchy, our proposed method operates at two separate scales. At the finer (sample) scale, we use an operator-theoretic approach to attach operators to individual samples (graphs), pairs of samples, triplets, quadruplets, etc. These operators quantitatively describe the structure of each sample separately, as well as the common structure across samples. Specifically, we use the norm of these operators as a measure of similarity between samples, facilitating a transition from operator-

theoretic analysis to affinity-based analysis. At the coarser (dataset) scale, we employ TDA to extract qualitative information from the datasets. Concretely, we use *persistent homology* [11, 29] as a topological signature for each dataset. Persistent homology is a topological-algebraic tool that captures information about connectivity and holes at various scales. It is computed over the ensemble of samples contained in each dataset, which we model as a weighted simplicial complex, where the weights are derived from the geometric operators computed at the finer scale. The signature provided by persistent homology comes with a natural metric (the Wasserstein distance [7]), allowing us at the final stage to compare the structure of different datasets.

To demonstrate the advantages of our method, we apply it to Hyper-Spectral Imaging (HSI) [6]. HSI is a sensing technique aimed to obtain the electromagnetic spectrum at each pixel within an image, with the purpose of finding objects, identifying materials, or detecting processes. We test our unsupervised method on categorical hyper-spectral images [1] and show that it accurately distinguishes between the different categories. In addition, we show that in a (supervised) classification task based on the attained (unsupervised) representation and metric, our method outperforms a competing method based on deep learning.

The main contributions of this work are: (i) We introduce a powerful combination between geometry and topology, taking advantage of both local and global information contained in data. (ii) We propose a method for analyzing hierarchical datasets, that is data-driven and "model-free" (i.e. does not require prior knowledge or a rigid model). (iii) We introduce a new notion of affinity between manifolds, quantifying their commonality.

2   Problem Formulation

In this work, we consider the following hierarchical structure. At the top level, we have a collection of N_D *datasets*

$$\mathcal{D} = \{D_1, \ldots, D_{N_D}\}.$$

These datasets may vary in size, shape and origin. However, we assume that they all have the same prototypical structure, as follows. Each dataset D consists of a collection of N *samples*

$$D = \{S_1, \ldots, S_N\},$$

and each sample is a collection of L *observations*

$$S_i = \{x_{i,1} \ldots, x_{i,L}\}, \ 1 \leq i \leq N.$$

Note that N and L may vary between datasets (for simplicity we omit the dataset index), but all the samples within a single dataset are of the same size L.

Next, we describe the structure of a single dataset D. Let $\{\mathcal{M}_\ell\}_{\ell=1}^M$ be a set of latent manifolds, and let Π be their product

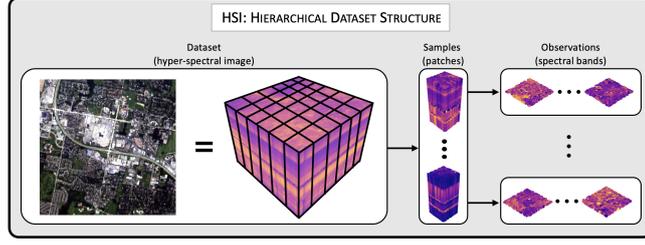$$\Pi = \mathcal{M}_1 \times \cdots \times \mathcal{M}_M. \tag{1}$$

Fig. 1: The specification of the considered hierarchical structure for HSI.

We use Π as a model for the common hidden space underlying the dataset D. Let $\mathcal{X} = \{x_1, \ldots, x_L\}$ be a set of points sampled from Π, where each point can be written as a tuple $x_j = (x_j^{(1)}, \ldots, x_j^{(M)})$ and $x_j^{(\ell)} \in \mathcal{M}_\ell$ for $\ell = 1, \ldots, M$.

Our main assumption here is that all samples S_i are generated by the same set $\mathcal{X}$, while each sample contains information only about a subset of the manifolds in the product (1). The subset of manifolds corresponding to a sample S_i is represented by a tuple of indices $I_i = (\ell_1^{(i)}, \ldots, \ell_m^{(i)})$ ($m \leq M$), and a product manifold

$$\Pi_{I_i} = \mathcal{M}_{\ell_1^{(i)}} \times \cdots \times \mathcal{M}_{\ell_m^{(i)}}. \tag{2}$$

For convenience, for each $x \in \Pi$, we define the projection

$$x_{I_i} = \left(x^{(\ell_1^{(i)})}, \ldots, x^{(\ell_m^{(i)})}\right) \in \Pi_{I_i}.$$

Next, for each sample S_i (and a corresponding subset I_i) we assume there is a function $g_i : \Pi_{I_i} \to \mathcal{O}_i$, for some target metric space $\mathcal{O}_i$. We define the observation function $f_i : \Pi \to \mathcal{O}_i$ as

$$f_i(x) = g_i(x_{I_i}) + \xi_i,$$

where $\xi_i \in \mathcal{O}_i$ denotes a random independent observation noise. Finally, the sample S_i is defined as

$$S_i = \{x_{i,1}, \ldots, x_{i,L}\} = \{f_i(x_1), \ldots, f_i(x_L)\}.$$

In other words, each sample S_i in the dataset D reveals partial and noisy information about Π. As stated earlier, we assume that each of the datasets $D_k \in \mathcal{D}$ is generated by the model described above. However, the manifold Π, the functions f_i, g_i, and the parameters L, M, N, I_i and ξ_i may differ between datasets. Note that within a single dataset there is a correspondence between all samples $S_1, \ldots, S_N$, as they are generated by the same set of realizations $\mathcal{X}$.

In the context of HSI, the hierarchical structure described above is as follows. Each hyper-spectral image is viewed as a single dataset, the full spectrum of a single patch as a sample, and different spectral bands within a patch as observations (see Figure 1 and Additional Materials for more details).

3 Background

In this section, we review some preliminaries required to describe our proposed method. Section 3.1 presents the diffusion and the alternating diffusion operators. Section 3.2 provides a brief introduction to persistent homology.

3.1 Multiple manifold learning and diffusion operators

Manifold learning is a class of unsupervised nonlinear data-driven methods for discovering the geometric structure underlying high dimensional data [27, 24, 2, 8]. The main assumption in manifold learning is that high-dimensional data lie on a hidden lower-dimensional manifold.

One of the notable approaches in manifold learning is *diffusion maps* [8], in which diffusion operators built from data are shown to approximate the Laplace-Beltrami operator. This differential operator contains all the geometric information on the manifold [4, 15], and thus its approximation provides means to incorporate geometric concepts such as metrics and embedding into data analysis tasks. In [17, 26], an extension of diffusion maps for multiple datasets termed 'alternating diffusion' was introduced. This extension, which is based on the product of diffusion operators, was shown to recover the manifold structure common to multiple datasets. In this work, we utilize a variant of alternating diffusion, proposed in [25], which is briefly described in the remainder of this subsection.

Consider two diffeomorphic compact Riemannian manifolds without a boundary, denoted by $(\mathcal{M}_1, g_1)$ and $(\mathcal{M}_2, g_2)$, and a diffeomorphism $\phi : \mathcal{M}_1 \to \mathcal{M}_2$. For each manifold $\mathcal{M}_\ell$ ($\ell = \{1, 2\}$) and a pair of samples $x, x' \in \mathcal{M}_\ell$, let $k_\ell(x, x')$ be a Gaussian kernel based on the distance induced by the metric g_ℓ with a kernel scale $\epsilon_\ell > 0$. Define $d_\ell(x) = \int_{\mathcal{M}_\ell} k_\ell(x, x')\mu_\ell(x')dv_\ell(x')$, where $v_\ell(x)$ is the volume measure, and $\mu_\ell(x')$ is the density function of the samples on $\mathcal{M}_\ell$. Consider two kernel normalizations: $a_\ell(x, x') = \frac{k_\ell(x, x')}{d_\ell(x)}$ and $b_\ell(x, x') = \frac{k_\ell(x, x')}{d_\ell(x')}$. Based on the normalized kernel $a_\ell(x, x')$, define the *forward diffusion operator* by $A_\ell f(x) = \int_{\mathcal{M}_\ell} a_\ell(x, x')f(x')\mu_\ell(x')dv_\ell(x')$ for any function $f \in C^\infty(\mathcal{M}_\ell)$. Similarly, based on $b_\ell(x, x')$, define the *backward diffusion operator* by $B_\ell f(x) = \int_{\mathcal{M}_\ell} b_\ell(x, x')f(x')\mu_\ell(x')dv_\ell(x')$. In a single manifold setting, it is shown that when $\epsilon_\ell \to 0$, A_ℓ converges to a differential operator of an isotropic diffusion process on a space with non-uniform density $\mu_\ell(x)$, and B_ℓ converges to the backward Fokker-Planck operator, which coincides with the Laplace-Beltrami operator when the density $\mu_\ell(x)$ is uniform [21].

Next, for two manifolds, consider the following $C^\infty(\mathcal{M}_1) \to C^\infty(\mathcal{M}_1)$ composite operators,

$$Gf(x) = \phi^* A_2 (\phi^*)^{-1} B_1 f(x),$$

and

$$Hf(x) = A_1 \phi^* B_2 (\phi^*)^{-1} f(x)$$

for any function $f \in C^\infty(\mathcal{M}_1)$, where $\phi^* : C^\infty(\mathcal{M}_2) \to C^\infty(\mathcal{M}_1)$ denotes the pullback operator from $\mathcal{M}_2$ to $\mathcal{M}_1$ and $(\phi^*)^{-1}$ denotes the push-forward from $\mathcal{M}_2$ to $\mathcal{M}_1$, both corresponding to the diffeomorphism ϕ.

In [17, 26], it was shown that the two composite operators G and H recover the common structure between $\mathcal{M}_1$ and $\mathcal{M}_2$ and attenuate non-common structures that are often associated with noise and interference. In [25], the following symmetric alternating diffusion operator was introduced

$$S_{1,2}f(x) = \frac{1}{2}\Big(Gf(x) + Hf(x)\Big),$$

which, in addition to revealing the common structure as G and H, has a real spectrum – a convenient property that allows to define spectral embeddings and spectral distances.

In practice, the operators defined above are approximated by matrices constructed from finite sets of data samples. Let $\{(x_i^{(1)}, x_i^{(2)})\}_{i=1}^L$ be a set of L pairs of samples from $\mathcal{M}_1 \times \mathcal{M}_2$, such that $x_i^{(2)} = \phi(x_i^{(1)})$. For $\ell = \{1, 2\}$, let $\boldsymbol{W}_\ell$ be an $L \times L$ matrix, whose (i, i')-th element is given by $W_\ell(i, i') = k_\ell(x_i^{(\ell)}, x_{i'}^{(\ell)})$. Let $\boldsymbol{Q}_\ell = \mathrm{diag}(\boldsymbol{W}_\ell \mathbf{1})$ be a diagonal matrix, where $\mathbf{1}$ is a column vector of all ones. The discrete counterparts of the operators A_ℓ and B_ℓ are then given by the matrices $\boldsymbol{K}_\ell = (\boldsymbol{Q}_\ell)^{-1}\boldsymbol{W}_\ell$ and $\boldsymbol{K}_\ell^T$, respectively, where $(\cdot)^T$ is the transpose operator. Consequently, the discrete counterparts of the operators G and H are $\boldsymbol{G} = \boldsymbol{K}_2 \boldsymbol{K}_1^T$ and $\boldsymbol{H} = \boldsymbol{K}_1 \boldsymbol{K}_2^T$, respectively. The symmetric matrix corresponding to the operator $S_{1,2}$ is then

$$\boldsymbol{S}_{1,2} = \boldsymbol{G} + \boldsymbol{H}.$$

3.2   Simplicial complexes and persistent homology

At the heart of the topological layer in our proposed method, we will use (abstract) simplicial complexes to represent the structure of a dataset. Briefly, a simplicial complex is a discrete structure that contains vertices, edges, triangles and higher dimensional simplexes (i.e. it is a type of hypergraph). This collection has to be closed under inclusion – for every simplex we must also include all the faces on its boundary. See Figure 2 in the Additional Materials for an example.

One of the many uses of simplicial complexes is in network modeling. While graphs take into account pairwise interactions between the nodes, simplicial complexes allow us to include information about the joint interaction of triplets, quadruplets, etc. We will use this property later, when studying the structure of samples within a dataset.

Homology is an algebraic structure that describes the shape of a topological space. Loosely speaking, for every topological space (e.g. a simplicial complex) we can define a sequence of vector spaces $H_0, H_1, H_2, \ldots$ where H_0 provides information about connected components, H_1 about closed loops surrounding holes, H_2 about closed surfaces enclosing cavities. Generally, we say that H_k provides information about *k-dimensional cycles*, which can be thought of as k-dimensional surfaces that are "empty" from within. We describe homology in more detail in Section 2.2 of the Additional Materials. In this work we will mainly use H_0 and H_1, i.e. information about connectivity and holes. However, the framework we develop can be used with any dimension of homology.
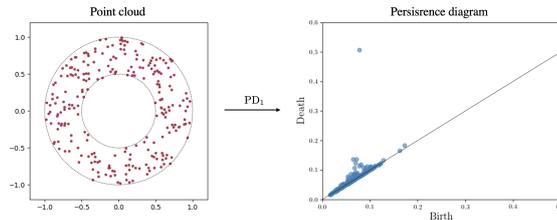
Fig. 2: Persistence diagram. Left: a point cloud generated in an annulus with a single hole (1-cycle). The filtration used is the union of balls around the points, for an increasing radius. Right: persistence diagram for 1-cycles. The birth/death axes represent radius values. The single feature away from the diagonal represents the hole of the annulus, while other cycles are considered "noise".

Persistent homology is one of the most heavily used tools in TDA [11, 29]. It can be viewed as a *multi-scale* version of homology, where instead of considering the structure of a single space, we track the evolution of cycles for a nested sequence of spaces, known as a *filtration*. As the spaces in a filtration grow, cycles of various dimensions may form (born) and later get filled in (die). The k-th persistent homology, denoted PH_k, keeps a record of the birth-death process of k-cycles. Commonly, the information contained in PH_k is summarized using a *persistence diagram* – a set of points in $\mathbb{R}^2$ representing all the (birth,death) pairs for cycles in dimension k, and denoted PD_k (see Figure 2). The motivation for using persistent homology is that it allows us to consider cycles at various scales, and identify those that seem to be prominent features of the data. In order to compare between persistence diagrams, we will employ the Wasserstein distance [7] defined as follows. Suppose that PD and PD$'$ are two persistence diagrams, then the p-Wasserstein distance is defined as

$$d_{W_p}(\mathrm{PD}, \mathrm{PD}') = \inf_{\phi:\widetilde{\mathrm{PD}}\to\widetilde{\mathrm{PD}}'} \Big(\sum_{\alpha\in\widetilde{\mathrm{PD}}} \|\alpha - \phi(\alpha)\|^p \Big)^{\frac{1}{p}}, \qquad (3)$$

where $\widetilde{\mathrm{PD}}$ is an augmented version of PD that also includes the diagonal line $x = y$ (and the same goes for $\widetilde{\mathrm{PD}}'$). This augmentation is taken in order to allow cases where the $|\mathrm{PD}| \neq |\mathrm{PD}'|$. In other words, the Wasserstein distance is based on an optimal matching between features in PD and PD$'$.

4   Proposed Method

Recall the hierarchical dataset structure presented in Section 2. The processing method we propose for such datasets is hierarchical as well. At the fine level, each sample S_i is treated as a weighted graph, which we analyze geometrically using a diffusion operator. At the coarse level, each dataset D is considered as a weighted simplicial complex, from which we extract its persistent homology,
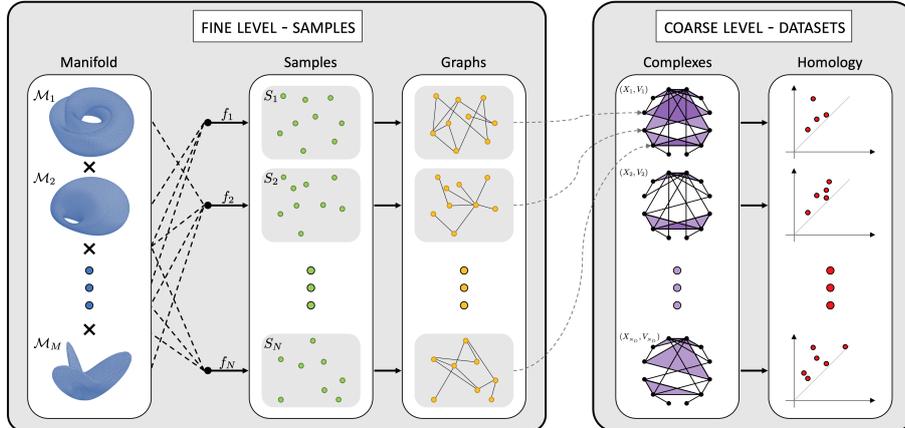
Fig. 3: Method outline. Each dataset (at the coarse level) is represented as a weighted simplicial complex, where the weights are calculated using the alternating diffusion operator between the sample graphs (see (8), (10)). The output is a set of persistence diagrams in the rightmost column, where each diagram summarizes a single dataset. We can compare the datasets using the Wasserstein distance (12).

enabling us to compare between different datasets using the Wasserstein distance. Figure 3 summarizes this pipline. The motivation for this analysis is the following. At the fine level, we use geometry in order to capture the detailed structure of a sample. Since all the samples within a dataset D are assumed to be generated by the same set of realizations $\mathcal{X} \subset \Pi$, their geometry provides a solid measure of inter-sample similarity. Conversely, at the coarse level, the geometry of different datasets can be vastly different. Thus, in order to compare datasets, we propose to use topology as an informative representation of the global qualitative structure, rather than geometry.

4.1 The sample diffusion operator

We treat each sample S_i as a weighted graph, with weights calculated using a Gaussian kernel, forming an affinity matrix $\boldsymbol{W}_i \in \mathbb{R}^{L \times L}$, whose (j_1, j_2)-th element is given by

$$W_i(j_1, j_2) = \exp\left(-\frac{d_i^2(x_{i,j_1}, x_{i,j_2})}{\epsilon}\right), \tag{4}$$

where d_i is a distance suitable for the observation space $\mathcal{O}_i$.

Next, following [8], we apply a two-step normalization. The first step is designed to handle a possibly non-uniform density of data points on the manifold. Let $\boldsymbol{Q}_i = \mathrm{diag}(\boldsymbol{W}_i \mathbf{1})$ be a diagonal matrix that approximates the local densities

of the nodes, where $\mathbf{1}$ is a column vector of all ones, and define

$$\widetilde{\boldsymbol{W}}_i = \boldsymbol{Q}_i^{-1} \boldsymbol{W}_i \boldsymbol{Q}_i^{-1}. \tag{5}$$

In the second step, we build another diagonal matrix $\widetilde{\boldsymbol{Q}}_i = \mathrm{diag}(\widetilde{\boldsymbol{W}}_i \mathbf{1})$, and form the following stochastic matrix

$$\boldsymbol{K}_i = \widetilde{\boldsymbol{Q}}_i^{-1} \widetilde{\boldsymbol{W}}_i, \tag{6}$$

which is called the *diffusion operator* of sample S_i. We note that the construction of $\boldsymbol{K}_i$ is similar to the construction described in Section 3.1 that follows [25] with only one difference – the first normalization that copes with non-uniform sampling.

4.2   The dataset simplicial complex

We construct a weighted simplicial complex for every dataset D, whose vertex set consists of the samples $\{S_i\}_{i=1}^N$. We assume that the simplicial complex is given (and depends on the problem at hand but not on the data), and we only need to determine the weights on the simplexes.

Considering our model in (1) and (2), we propose weights that are inversely correlated with the number of common hidden variables between the samples. Denote by V the weight function for the simplexes representing dataset D. Ideally, for any d-dimensional simplex $\sigma = [i_1, \ldots, i_{d+1}]$ we want to have

$$V(\sigma) = U(|I_{i_1} \cap \cdots \cap I_{i_{d+1}}|), \tag{7}$$

where I_i is the tuple of indexes corresponding to sample S_i (see Section 2) and U is a decreasing function. While this condition cannot hold in a strict sense (mainly due to observation noise), the method we devise below provides a close approximation.

We start with the edges. Let $\boldsymbol{K}_{i_1}, \boldsymbol{K}_{i_2}$ be a pair of diffusion operators for samples $S_{i_1}, S_{i_2} \in D$ (see Section 4.1). As described in Section 3.1, the work in [25], based on the notion of alternating diffusion [17], showed that one can reveal the common manifold structure between S_{i_1} and S_{i_2} by considering the symmetric alternating diffusion operator

$$\boldsymbol{S}_{i_1,i_2} = \boldsymbol{K}_{i_1} \boldsymbol{K}_{i_2}^T + \boldsymbol{K}_{i_2} \boldsymbol{K}_{i_1}^T. \tag{8}$$

As a heuristic, we propose to set the weight function V to be the inverse of the Frobenius norm, i.e.

$$V([i_1, i_2]) = \|\boldsymbol{S}_{i_1,i_2}\|_F^{-1}. \tag{9}$$

The rationale behind this heuristic stems from the common practice in kernel methods. Typically, the eigenvalues of the kernel are used for evaluating the dominance of the component represented by the corresponding eigenvectors. Indeed, using the spectral distance was proposed in [23] in a setting where the samples

form individual graphs, as in the current work. Here, we follow the same practice but with a kernel that captures only the common components. We will show empirically in Section 5 that indeed $V([i_1, i_2])$ inversely correlates with $|I_{i_1} \cap I_{i_2}|$, as desired.

Next, we consider triangles in our complex. In a similar spirit to (8), we define the three-way symmetric alternating diffusion operator by

$$\boldsymbol{S}_{i_1,i_2,i_3} = \boldsymbol{S}_{i_1,i_2}\boldsymbol{K}_{i_3}^T + \boldsymbol{K}_{i_3}\boldsymbol{S}_{i_1,i_2} + \boldsymbol{S}_{i_2,i_3}\boldsymbol{K}_{i_1}^T + \boldsymbol{K}_{i_1}\boldsymbol{S}_{i_2,i_3} + \boldsymbol{S}_{i_1,i_3}\boldsymbol{K}_{i_2}^T + \boldsymbol{K}_{i_2}\boldsymbol{S}_{i_1,i_3}. \tag{10}$$

The weight function of the corresponding triangle is then set as

$$V([i_1, i_2, i_3]) = \|\boldsymbol{S}_{i_1,i_2,i_3}\|_F^{-1}. \tag{11}$$

In Section 5 we also show empirically that $V([i_1, i_2, i_3])$ inversely correlates with $|I_{i_1} \cap I_{i_2} \cap I_{i_3}|$. In particular, we have $V([i_1, i_2]) \leq V([i_1, i_2, i_3])$ for all i_1, i_2, i_3, which is required in order to have a filtered complex.

In a similar spirit, one can define V for simplexes of any dimension. However, for the simulation and application we consider here, edges and triangles suffice.

4.3   Topological distance between datasets

The proposed pipeline concludes with a numerical measure of structural similarity between two datasets D and D'. Recall that the output of the previous section are weighted simplicial complexes, denoted by the pairs (X, V) and (X', V'), where X and X' are complexes and V and V' are the weight functions. We use each weight function to generate a filtration that in turn serves as the input to the persistent homology computation (see Section 3.2). The filtration we take is the *sublevel* set filtration $\{X_v\}_{v \in \mathbb{R}}$, where $X_v = \{\sigma : V(\sigma) \leq v\}$. Considering the weights constructed in (9) and (11) , this implies that simplexes that represent groups of samples that share more structure in common will appear earlier in the filtration.

Let PD_k and PD'_k be the k-th persistence diagrams of (X, V) and (X', V'), respectively. We can then compare the topology of two datasets by calculating the Wasserstein distance

$$d_{\text{dataset}}(D, D') = d_{W_p}(\mathrm{PD}_k, \mathrm{PD}'_k). \tag{12}$$

The choice of p and k depends on the application at hand. The entire pipeline is summarized in Algorithm 1. Note that it is currently described for $k = \{0, 1\}$, but once the weight function V in Subsection 4.2 is extended beyond edges and triangles to higher orders, the algorithm can be extended for $k \geq 2$ as well.

5   Simulation Study

In this section, we test the proposed framework on a toy problem, where we can manipulate and examine all the ingredients of our model and method.

Algorithm 1: A geometric-topological distance between two datasets

---

Input   : Two hierarchical datasets: D and D'
Output   : Distance between the datasets: $d_{dataset}(D, D')$
Parameters: $k = \{0, 1\}$ (homology degree), p (Wasserstein distance order), ϵ (kernel scale)

1. Construct a simplicial complex X for each dataset as follows:
   (a) For each sample S_i, $i = 1, \ldots, N$ compute the diffusion operator $\boldsymbol{K}_i$, with scale parameter ϵ according to (4)–(6)
   (b) For all edges (i_1, i_2):
       i. Compute the symmetric alternating diffusion operator $\boldsymbol{S}_{i_1,i_2}$ according to (8)
       ii. Set the weight function: $V([i_1, i_2]) = \|\boldsymbol{S}_{i_1,i_2}\|_F^{-1}$
   (c) For all triangles (i_1, i_2, i_3):
       i. Compute the three-way symmetric alternating diffusion operator $\boldsymbol{S}_{i_1,i_2,i_3}$ according to (10)
       ii. Set the weights of $V([i_1, i_2, i_3]) = \|\boldsymbol{S}_{i_1,i_2,i_3}\|_F^{-1}$
2. Compute the k-th persistence diagram of the weighted complexes (X, V) and (X', V'), corresponding to D and D', respectively.
3. Compute the distance between the persistence diagrams: $d_{W_p}(\mathrm{PD}_k, \mathrm{PD}'_k)$

We start with the description of a single dataset D. Revisiting the notation in Section 2, we assume that the latent manifold for each dataset is of the form $\Pi = \mathcal{M}_1 \times \cdots \times \mathcal{M}_M$, where each $\mathcal{M}_\ell$ is a circle of the form

$$\mathcal{M}_\ell = \{(\cos(\theta_\ell), \sin(\theta_\ell)) \mid 0 \leq \theta_\ell < 2\pi\}. \tag{13}$$

In other words, Π is an M-dimensional torus. In this case, the latent realization set $\mathcal{X} = \{x_1, \ldots, x_L\}$ is a subset of $\mathbb{R}^{2M}$. We generate $\mathcal{X}$ by taking a sample of iid variables $\theta_\ell \sim U[0, 2\pi]$ for $\ell = 1, \ldots, M$. For all the samples S_i ($1 \leq i \leq N$) we take $|I_i| = 3$, and the observation space is then $\mathcal{O}_i = \mathbb{R}^6$. For every $x \in \Pi$, we define

$$g_i(x_{I_i}) = \big(R_1 \cos(\theta_{\ell_1}), R_1 \sin(\theta_{\ell_1}), R_2 \cos(\theta_{\ell_2}), R_2 \sin(\theta_{\ell_2}), R_3 \cos(\theta_{\ell_3}), R_3 \sin(\theta_{\ell_3})\big)$$

where $I_i = (\ell_1, \ell_2, \ell_3)$ are the indexes of the subset of manifolds viewed by sample S_i. The radii R_1, R_2, R_3 are generated uniformly at random in the interval $[1, R_{\max}]$, for each i independently. The indexes ℓ_1, ℓ_2, ℓ_3 are also chosen at random. Finally, the sample observation function is given by $f_i(x) = g_i(x_{I_i}) + \xi_i$, where $\xi_i \sim \mathcal{N}(0, \sigma_i^2 I)$ (independent between observations).

For the generation of all datasets, we use $N = 40$, and $L = 200$. The torus dimension M varies between 3 and 30 across the datasets. As M increases, the chances that the pair of samples (S_i, S_j) has underlying circles in common decreases. Subsequently, the connectivity of the simplicial complex of the respective dataset decreases. Thus, M strongly affects the affinity between the datasets.

In Section 4 we argued that the weight function V defined in (9) and (11) is roughly decreasing in the number of common variables (7). Here, we provide an
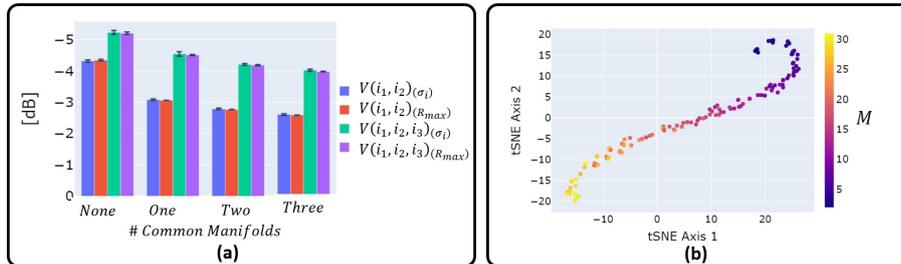
Fig. 4: (a) The mean and standard deviation of the weight function V (for edges and triangles). We take 20 random realizations, as well as varying $R_{\max}$ between $1 - 15$ and σ_i between $0.001 - 1000$. Note that we plot the value $\log(1 - V)$, and the y axis is flipped, so this is indeed a monotone-decreasing behavior. (b) t-SNE embedding based on the Wasserstein distance for H_1 (holes), where the color indicates the size of the pool of underlying manifolds M.

experimental evidence for that heuristic. In Figure 4(a), we plot the mean and standard deviation of the weight function (in dB) as a function of the number of common indices (manifolds/circles). We calculate V across 20 realizations, and across various choices of $R_{\max}$ $(1 - 15)$ and σ_i $(0.001 - 1000)$. The results clearly indicate a monotone decreasing relationship between V and the number of common manifolds. In addition, robustness to noise (σ_i) and to the particular observation space $(R_{\max})$ is demonstrated.

For each $3 \leq M \leq 30$ we generate 5 datasets with $R_{\max} = 15$ and $\sigma_i = 0.1$, so that overall there are $N_D = 140$ datasets. For each dataset, we follow Algorithm 1 and calculate the topological distance d_{dataset} (12) using $k = 1$ (holes) and $p = 2$. Figure 4(b) presents the t-SNE [20] embedding based on the obtained distance matrix between all datasets. The color of each dataset indicates the value of M. Indeed, we observe that the datasets are organized according this value. In other words, our hierarchical geometric-topological analysis provides a metric between datasets that well-captures the similarity in terms of the global structure of the datasets.

6   Application to HSI

In this section we demonstrate the performance of our new geometric-topological framework on Hyper-Spectral Imaging (HSI). The structure of hyper-spectral images fits well with the considered hierarchical dataset model – a dataset here is a single image, and a sample S_i within the dataset is a single square patch. The observations in each patch correspond to the content of the patch at separate spectral bands (see Figure 1).

The HSI database contains images of various terrain patterns, taken from the NASA Jet Propulsion Laboratory's Airborne Visible InfraRed Imaging Spectrometer (AVIRIS) [1]. In [3] these images were classified into nine categories:

`agriculture`, `cloud`, `desert`, `dense-urban`, `forest`, `mountain`, `ocean`, `snow` and `wetland`. The database consists of 486 hyper-spectral images of size 300×300 pixels and the spectral radiance is sampled at 224 contiguous spectral bands (365 nm to 2497 nm). See Figure 5(a). In terms of our setting, we have $N_D = 486$ datasets, $N = 3600$ samples in each dataset (taking a patch size of $n = 5$), and each sample consists of $L = 224$ observations. In this case, each of the observations is a vector in $\mathcal{O}_i = \mathbb{R}^{25}$ (corresponding to the patch-size).

The simplicial complex X we use here is a standard triangulation of the 2-dimensional grid of patches. This way the spatial organization of patches in the image is taken into account in the computation of the persistent homology.

We apply Algorithm 1 to the images (datasets) and obtain their pairwise distances. Figure 5(b) demonstrates how our new topological distance arranges the images in space. Specifically, we plot the t-SNE embedding [20] of the images (datasets) based on d_{dataset} for $k = 1, p = 2$. Each point in the figure represents a single hyper-spectral image, colored by category. Importantly, the category information was not accessible to the (unsupervised) algorithm, and was added to the figure in order to evaluate the results. We observe that most images are grouped by category. In addition, the embedding also conveys the similarity between different categories, implying that this information is captured by d_{dataset}. For example, `agriculture` images (blue points) and `dense-urban` images (purple crosses) are embedded in adjacent locations, and indeed they share common patterns (e.g., grass areas). Conversely, `snow` images form their own separate cluster, as most of the snow instances do not have any common structure with the other categories.

For an objective evaluation of the results, we train an SVM classifier [9]. Prior to computing the SVM, we embed the images (datasets) into a Euclidean space using diffusion maps [8], and apply the classifier to the embedded images. For diffusion maps, the distance obtained by Algorithm 1 d_{dataset} is used as input, and we generate embedding with 20 dimensions (see Section 2.1 in Additional Materials). For the classification, we divide the datasets into a train set and a test set with 10-fold cross validation; the reported results are the average over all folds. We use mean Average Precision (mAP) as the evaluation score.

We compare our results to the results reported in [3], where a deep learning approach was used for the classification of the images. To the best of our knowledge, the results in [3] are considered the state of the art for the ICONES dataset. In Table 1, we present the obtained classification results. In order to make a fair comparison with the reported results in [3], we show the mAP obtained on the train sets. We observe that our method achieves superior results. In addition, we report that our method obtains 0.81 mAP on the test sets.

In order to test the sensitivity of the proposed algorithm to the choice of hyper-parameters, in Figure 5(c) we present the train scores (top) and test scores (bottom) as a function of the two key hyper-parameters – the patch size n and and kernel scale ϵ (normalized by the median of the distances in the affinity matrix (4)). The correspondence of the colors between the two figures, as well as the apparent smoothness of the color gradient within each image imply robust-
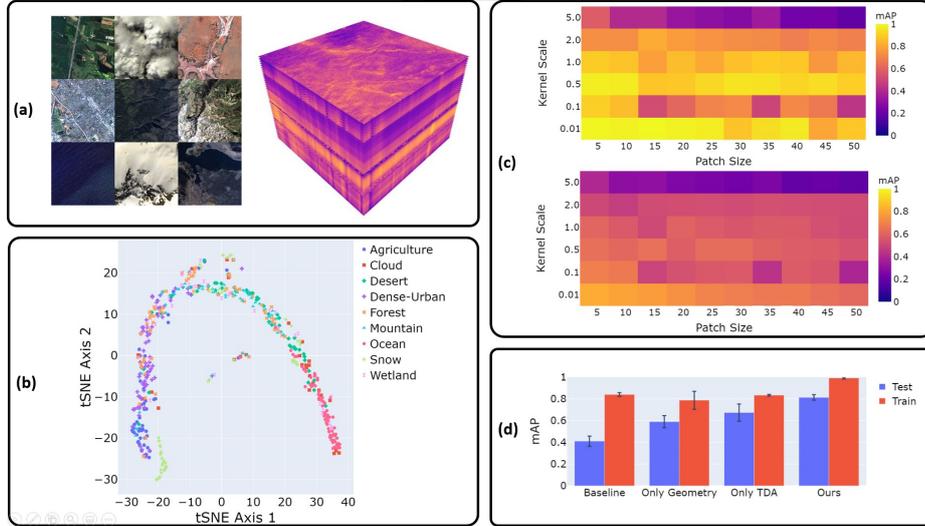
Fig. 5: (a) Left: examples of the terrains of hyper-spectral images from different categories (RGB). Right: hyper-spectral image example, stack of 224 spectral bands. (b) t-SNE embedding based on the Wasserstein distance for H_1 (holes) between the hyper-spectral images. Each category is denoted by a different color and marker. (c) Classification results (mAP score) as a function of two hyper-parameters: kernel scale and patch size. Top: train score. Bottom: test score. (d) Numerical ablation study – classification results (mAP score) evaluating the contribution of each component in our method.

ness to hyper-parameter tuning. Further, we can optimize the hyper-parameters using the train set without leading to an overfit.

Next, we perform an ablation study to evaluate the contribution of the geometric and topological analyses separately. In order to do so, we consider three variants of the algorithm. (i) A *'baseline'* solution: here we replace both the geometric and the topological components with the following implementation which was inspired by [3]. We split the $L = 224$ spectral bands into 5 contiguous ranges. For each range, we apply Principal Component Analysis (PCA) and keep only the principal component in order to reduce the clutter and to get the essence of the spectral information. Next, the pairwise Euclidean distance between the principal components is considered as the counterpart of d_{dataset}

Table 1: Classification Results (mAP).

	agric.	cloud	desert	dense-urban	forest	mountain	ocean	snow	wetland	All
[3]	0.48	0.66	0.5	0.86	0.57	0.64	0.83	0.57	0.23	0.59
Ours	1.0	0.95	1.0	1.0	0.99	1.0	1.0	1.0	0.82	0.98

(the output of Algorithm 1). (ii) *Geometry-based* solution: in Step 1 of Algorithm 1, the weighted simplicial complex is replaced by a weighted graph, taking into account only the weights on the edges $V([i_1, i_2])$. Step 2 is removed, and in Step 3, we use a spectral distance (the L_2 distance between the eigenvalues of graphs as in [23]) between the graphs as the output of the algorithm d_{dataset}. (iii) *Topology-based* solution: in Algorithm 1, the weight function is set to be the cross-correlation between the samples (rather than using alternating diffusion).

We repeat the use of an SVM classifier as described above using the output of each of three variants. Figure 5(d) shows the results. First, we observe that the simple baseline based on PCA attains a test score of only 0.41 mAP. Second, the addition of the geometric analysis and the topological analysis significantly improves the results. Third, the combination of the analyses in Algorithm 1 gives rise to the best results.

To conclude, rather than relying on the measured values alone, the application of our method to HSI emphasizes associations within the data. This concept is embodied in the proposed hierarchical manner. At the fine scale, graphs based on local spectral associations are constructed. At the coarse scale, simplicial complexes based on global spatial associations are formed. The combination of the structures at the two scales, involving both spectral an spatial information, is shown to be beneficial and gives rise to an informative and useful representation of the images.

Acknowledgement

References

1. NASA jet propulsion laboratory's airborne visible infrared imaging spectrometer (AVIRIS). https://aviris.jpl.nasa.gov/
2. Belkin, M., Niyogi, P.: Laplacian eigenmaps for dimensionality reduction and data representation. Neural Comput. **15**(6), 1373–1396 (2003)
3. Ben-Ahmed, O., Urruty, T., Richard, N., Fernandez-Maloigne, C.: Toward content-based hyperspectral remote sensing image retrieval (cb-hrsir): A preliminary study based on spectral sensitivity functions. Remote Sens. **11**(5), 600 (2019)
4. Bérard, P., Besson, G., Gallot, S.: Embedding Riemannian manifolds by their heat kernel. Geom. Funct. Anal. **4**(4), 373–398 (1994)
5. Carlsson, G.: Topology and data. Bull. Am. Math. Soc. **46**(2), 255–308 (2009)
6. Chang, C.I.: Hyperspectral imaging: techniques for spectral detection and classification, vol. 1. Springer Sci. & Business Media (2003)
7. Cohen-Steiner, D., Edelsbrunner, H., Harer, J., Mileyko, Y.: Lipschitz functions have L p-stable persistence. Found. Comput. Math. **10**(2), 127–139 (jan 2010)

8. Coifman, R.R., Lafon, S.: Diffusion maps. Appl. Comput. Harmon. Anal. **21**(1), 5–30 (2006)
9. Cortes, C., Vapnik, V.: Support-vector networks. Mach. Learn. **20**(3), 273–297 (1995)
10. Dabaghian, Y., Mémoli, F., Frank, L., Carlsson, G.: A topological paradigm for hippocampal spatial map formation using persistent homology. PLoS Comput. Biol. **8**(8), e1002581 (2012)
11. Edelsbrunner, H., Harer, J.: Persistent homology-a survey. Contemp. Math. **453**, 257–282 (2008)
12. Edgar, R., Domrachev, M.e.a.: Gene expression omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. **30**(1), 207–210 (2002)
13. Giesen, C., Wang, Hao AO, e.a.: Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. Nat. Methods **11**(4), 417–422 (2014)
14. Giusti, C., Pastalkova, E., Curto, C.: Clique topology reveals intrinsic geometric structure in neural correlations. Proc. Natl. Acad. Sci. **112**(44), 13455–13460 (2015)
15. Jones, P.W., Maggioni, M., Schul, R.: Manifold parametrizations by eigenfunctions of the laplacian and heat kernels. Proc. Natl. Acad. Sci. **105**(6), 1803–1808 (2008)
16. LaMontagne, P.J., Benzinger, L.: Oasis-3: longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and alzheimer disease. MedRxiv (2019)
17. Lederman, R.R., Talmon, R.: Learning the geometry of common latent variables using alternating-diffusion. Appl. Comput. Harmon. Anal. **44**(3), 509–536 (2018)
18. Leskovec, J., Krevl, A.: SNAP datasets: Stanford large network dataset collection. http://snap.stanford.edu/data (2014)
19. Lum, P.Y., Singh, G., Lehman, A., Ishkanov, T., Vejdemo-Johansson, M., Alagappan, M., Carlsson, J., Carlsson, G.: Extracting insights from the shape of complex data using topology. Sci. Rep. **3**(1), 1–8 (2013)
20. Van der Maaten, L., Hinton, G.: Visualizing data using t-SNE. J. Mach. Learn. Res. **9**(11) (2008)
21. Nadler, B., Lafon, S., Coifman, R.R., Kevrekidis, I.G.: Diffusion maps, spectral clustering and reaction coordinates of dynamical systems. Appl. Comput. Harmon. Anal. **21**(1), 113–127 (2006)
22. Naitzat, G., Zhitnikov, A., Lim, L.H.: Topology of deep neural networks. J. Mach. Learn. Res. **21**(184), 1–40 (2020)
23. Rajendran, K., Kattis, A., Holiday, A., Kondor, R., Kevrekidis, I.G.: Data mining when each data point is a network. In: Int. Conf. Patterns of Dynamics. pp. 289–317. Springer (2016)
24. Roweis, S.T., Saul, L.K.: Nonlinear dimensionality reduction by locally linear embedding. Sci. **290**(5500), 2323–2326 (2000)
25. Shnitzer, T., Ben-Chen, M., Guibas, L., Talmon, R., Wu, H.T.: Recovering hidden components in multimodal data with composite diffusion operators. SIAM J. Math. Data Sci. **1**(3), 588–616 (2019)
26. Talmon, R., Wu, H.T.: Latent common manifold learning with alternating diffusion: analysis and applications. Appl. Comput. Harmon. Anal. **47**(3), 848–892 (2019)
27. Tenenbaum, J.B., De Silva, V., Langford, J.C.: A global geometric framework for nonlinear dimensionality reduction. Sci. **290**(5500), 2319–2323 (2000)
28. Wasserman, L.: Topological data analysis. Annu. Rev. Stat. Appl. **5**, 501–532 (2018)
29. Zomorodian, A., Carlsson, G.: Computing persistent homology. Discrete & Computational Geometry **33**(2), 249–274 (2005)