Homology-Preserving Dimensionality Reduction
via Manifold Landmarking and Tearing

Lin Yan,[*] Yaodong Zhao,[†] Paul Rosen,[‡] Carlos Scheidegger,[§] Bei Wang[¶]

June 25, 2018

Abstract

Dimensionality reduction is an integral part of data visualization. It is a process that obtains a structure preserving low-dimensional representation of the high-dimensional data. Two common criteria can be used to achieve a dimensionality reduction: distance preservation and topology preservation. Inspired by recent work in topological data analysis, we are on the quest for a dimensionality reduction technique that achieves the criterion of homology preservation, a generalized version of topology preservation. Specifically, we are interested in using topology-inspired manifold landmarking and manifold tearing to aid such a process and evaluate their effectiveness.

Index Terms— Topological data analysis, dimensionality reduction, manifold landmarking, manifold learning, high-dimensional data visualization

1   Introduction

Dimensionality reduction (DR) is a process that obtains a *structure-preserving* low-dimensional representation of the high-dimensional data. It plays an important role in high-dimensional data visualization in both static and interactive settings. Two common criteria can be used to achieve a DR *distance preservation* and *topology preservation*. Inspired by recent work in topological data analysis, we are on the quest for a DR technique that achieves the criterion of *homology preservation*, a generalized version of topology preservation.

To motivate our work, we begin by addressing the following questions: What is homology in the context of topology? What is homology preservation in the context of structure preservation? Why does it matter (and who cares) for homology preservation in high-dimensional data analysis and visualization? See the following paragraphs on a history of homology, a discussion on homology-preservation DR, and motivations from visualization to robotics.

A brief history of homology.   Topology has been one of the most exciting research fields in modern mathematics [29]. It is concerned with the properties of space that are preserved under continuous deformations, such as stretching, crumpling, and bending, but not tearing or gluing [72].

The beginning of topology is arguably marked by Leonhard Euler, who published a paper in 1736 that solves the now famous Königsberg bridge problem. In the paper, titled *"The Solution of a Problem Relating to the Geometry of Position"*, Euler was dealing with "a different type of geometry where distance was not relevant." [45] Johann Benedict Listing was credited to be the first to use the word "topology" in print based on his 1847 work titled *"Introductory Studies in Topology"*; although many of Listing's topological ideas were due to Carl Friedrich Gauss [45]. Both Listing and Bernhard Riemann studied the *components* and *connectivity* of surfaces. Listing examined connectivity in 3-dimensional Euclidean space while Enrico Betti extended the idea to *n* dimensions. Henri Poincaré then gave a rigorous basis to the idea of connectivity in a series of papers *"Analysis situs"* in 1895. He introduced the concept of

---

[*]Lin Yan is with University of Utah. E-mail: linyan@cs.utah.edu.

[†]Yaodong Zhao is with University of Utah. E-mail: yaodong.zhao@utah.edu.

[‡]Paul Rosen is with University of South Florida. E-mail: prosen@usf.edu.

[§]Carlos Scheidegger is with University of Arizona. E-mail: cscheid@cs.arizona.edu.

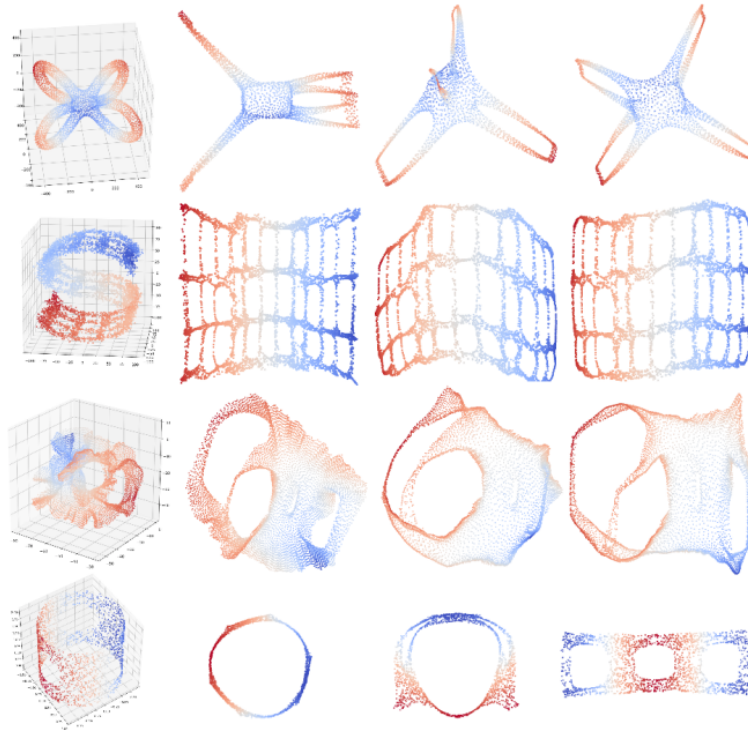[¶]Bei Wang is with University of Utah. E-mail: beiwang@sci.utah.edu.

Figure 1: In search of homology-preserving linear projections for a *bended figure eight* example. (a) Uniformly sample projection directions on a sphere. (b) The original 3-dimensional input point cloud. (c)-(e) Three instances of the 2-dimensional linear embeddings, (c) being the optimal as it produces the lowest degree-2 Wasserstein distortion.

homology and improved upon the precise definition of Betti numbers of a space [45]. In other words, it was Poincaré who "gave topology wings" [29] via the notion of homology.

The original motivation to define homology was that it can be used to tell two things (a.k.a. topological spaces) apart by examining their holes. It is a process that associates a topological space with a sequence of abelian groups called homology groups, which, roughly speaking, count and collate *holes* in a space [24]. In a nutshell, homology groups generalize a common-sense notion of connectivity. They detect and describe the connected components (0-dimensional holes), tunnels (1-dimensional holes), voids (2-dimensional holes), and holes of higher dimensions in the space.

It is not easy to give homology an intuitive and correct definition, but we will attempt to give a layman's version by quoting and modifying one given by Evelyn Lamb [33]: If it has one connected piece, it has a 0-dimensional hole (imaging a cookie); If you can put it on a necklace, it has a 1-dimensional hole; If you can fill it with toothpaste that is not exposed to the air (imaging a basketball), it has a 2-dimensional hole; For holes of higher dimensions, you're on your own.

A discussion on homology-preservation. DR techniques could be classified based on structure preservation, namely, distance preservation or topology preservation. The preservation of pairwise distances ensures that the low-dimensional embedding inherits the geometric properties of the data [25]. For instance, classical Multidimensional Scaling (MDS, a linear technique) preserves spatial distances (such as the Euclidean distances) while Isomap (a nonlinear technique) uses geodesic distances (approximated by graph distances) [36]. On the other hand, the quantitative natural of distance preservation also makes it very constraining – it is like "supporting and bolting the space with rigid steel beams"; and in nonlinear cases (such as manifolds), distances cannot be perfectly preserved [36]. The notion of topology preservation refers to the preservation of neighborhood relations between subregions of the data. Topology preservation techniques, such as locally linear embedding (LLE) [53] and Laplacian eigenmap [3], introduce some flexibility where subregions of the data could be locally stretched or shrunk in order to embed them in a

lower dimensional space [36].

In a way, many topology preservation DR techniques are concerned with a common-sense notion of connectivity; that is, the 0-dimensional connectivities among neighboring data points. In this paper, we focus on a generalized version of topology preservation, *homology preservation*, where we are interested in the preservation of both 0-dimensional and 1-dimensional homology of the data. In particular, we are in search of DR techniques that could preserve as much as possible of the 1-dimensional homology (a.k.a. *loops*) of the data.

Motivations from visualization to robotics.   Our first motivation to study homology-preserving DR is from the perspective of visualization. As technologies advance, we are collecting and generating a wide variety of large, complex, and high-dimensional datasets that demand insight-generating analysis and visualization. However, limitations on our visual systems as well as display devices have prevented us from the rapid recognition of structures beyond three dimensions. Visualization approaches therefore play an essentially role in *visually* conveying and interpreting high-dimensional structural information by utilizing low-dimensional embeddings and abstractions: from DR to visual encoding, and from quantitative analysis to interactive exploration [40]. We believe homology-preserving DR helps to expand the existing DR toolset and encodes additional structural information of high-dimensional data for visual exploration.

Our second motivation is the availability of interesting datasets with nontrivial homology, in particular, from imaging and signal processing. In studying the space of images, Lee et al. [34] have found that the majority of high-contrast 3 by 3 patches are concentrated near a circle. Follow-up work by Carlsson et al. [6] and others [73] has shown that a subspace of the space of natural image patches either exhibits circular behavior or is topologically equivalent to a Klein bottle, depending on the patch size. In signal processing, using delayed window embedding, a 1-dimensional signal can be encoded into a high-dimensional point cloud for topological data analysis. Specifically, 1-dimensional homology (i.e. loop) of such a point cloud captures the periodicity of the signal [48, 49].

Finally, we are motivated by the large collection of datasets that arise from robotics. Homological concepts naturally arise in robotics from the perspective of *motion planning*, that is, a process that aims to "design a trajectory of robot states from a given initial state to a specified goal state through a complex configuration space" [26]. Two trajectories are considered being topologically equivalent, if the boundaries formed by both trajectories do not contain any obstacle. The notion of *punctured Euclidean space*, that is, $R^D - O$, occurs frequently in the configuration spaces of robots, where O represent either the physical obstacles (such as humans, chairs, or other robots) that the robots need to avoid in the 2- or 3-dimensional configuration spaces, or illegal states (such as all legs off the ground) for the higher-dimensional configuration spaces [4]. Homological information can be used to cluster and classify different classes of trajectories in complex configuration spaces and to find representative (optimal) trajectory for each class [4]. We envision homology-preserving DR techniques could help capture the nontrivial homologies in the environments for path planning and state planning. As part of future work, we are interested in knowing how topology-inspired landmarks and skeletons (Section 5) can be used as representative trajectories in the configuration space.

Contributions.   The goal of this paper is to generalize topology preservation to homology preservation, much as generalizing connectivity to the notion of homology. Given that distance preservation maintains the geometric and therefore topological properties of the data with some sense of rigidity, this could be the end of the story.

However, we present examples in the paper illustrating that we can achieve homology preservation while at the same time maintaining (and sometimes even improving) the preservation of distances. Our contributions are:

- We introduce a new class of homology-preserving DR techniques that combine the strengths of landmark Isomap (L-Isomap) with the power of homology-preserving landmarks.

- For complex data such as circular manifolds, we provide a simple and fast procedure that can tear those manifolds, while at the same time preserving as much homology as possible.

- We conduct experiments for homology-preserving manifold landmarking and manifold tearing to evaluate their effectiveness.

Outline.   We motivate our problem of interest in Section 2 with a naive interpretation of homology preservation in the setting of linear projection. We discuss related work in Section 3 that covers topics such as manifold landmarking, manifold tearing, and quality assessment of DR techniques. We introduce
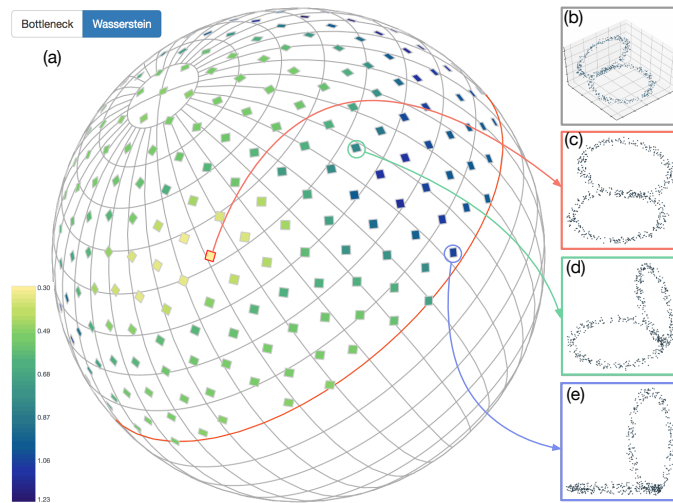
Figure 2: In search of homology-preserving linear projections for a *bended figure eight* example. (a) Uniformly sample projection directions on a sphere. (b) The original 3-dimensional input point cloud. (c)-(e) Three instances of the 2-dimensional linear embeddings, (c) being the optimal as it produces the lowest degree-2 Wasserstein distortion.

homology-based measure for the assessment and comparison of DR techniques in Section 4. We detail a novel landmark selection method for L-Isomap based upon data skeletonization, and describe a new class of homology-preserving DR techniques in Section 5. We describe homology-preserving manifold tearing in Section 6. We conclude with real-world examples (Section 7) and a discussion on future directions (Section 8).

2 A Motivational Example Using Linear Projections

To motivate our research objective, we start with a naive interpretation of homology-preservation DR using linear projects. Given a high-dimensional point cloud, we have created an interactive visualization prototype, where we observe and quantify the results of DR using simple linear projections. We illustrate the idea of homology preservation using a 3-dimensional point cloud sampled from two perpendicular rings joined at a point, referred to as the *bended figure eight* in Figure 2(b).

We begin by sampling a number of projection directions uniformly from a sphere in Figure 2(a). Given a particular projection direction, we use a couple of homology-centric criteria (detailed in Section 4) to assess the quality of each linear embedding. Here, we focus on the preservation of 1-dimensional homology. The optimal projection direction is marked by a red diamond in Figure 2(a), whereas Figure 2(c) corresponds to the optimal 2-dimensional embedding. Figure 2(d) and (e) give two examples of not-so-optimal embeddings that correspond to the blue and green projection directions in (a), respectively. Each embedding shows a certain amount of deformation to at least one of the two loops.

It is clear from the visualization that the optimal linear projection direction tries to preserve the shape of the two loops as much as possible, therefore producing the most optimal homology preserving embedding.

3 Related Work

Dimensionality reduction. DR techniques can be studied following various taxonomies. For instance, they are considered as linear (resp. nonlinear) methods if they produce low-dimensional linear (resp. nonlinear) mapping of the input high-dimensional data that preserve certain features of interest. They can be thought of as conducting convex or nonconvex optimizations, full or sparse spectral eigendecompositions, global or local structure preservation; see [12, 66] for thorough reviews. We largely follow the classification from [36] in terms of distance or topology preservation (see Section 1).

Quality assessment and visualization. To assess the performance of DR techniques, different quality measures have been proposed that can be roughly classified as global- or local-based approaches. The former quantifies the preservation of local neighborhoods/subregions, and the latter studies the preservation of global shape of data. Global measures include Shepard diagram [55, 56], stress [31, 32], and residual variance [64] (as described in Section 4), and local measures consist of rank-based criteria such as co-ranking matrix [37], mean relative rank errors [37] and Spearman's rho [60], normalization independent embedding quality assessment [76], and many more [25].

On the other hand, the characteristics of various quality measures can be linked with fine-grained visual analytics. For instance, point-wise quality measures can be augmented in the visualization to highlight erroneous local regions [42].

Manifold landmarking. Our proposed strategy takes advantage of *manifold landmarking*, that is, finding a subset of points along the manifold that captures its structural characteristics [39]. Manifold landmarking is useful for dimensionality reduction, for example, in the case of landmark MDS and landmark Isomap [13, 14]. It can also be employed to generate sparse manifolds for machine learning tasks [43] or sparse matrices for semidefinite programming [71], as well as supervised learning [65].

Previous landmarking methods can be classified as geometric or statistical approaches. Geometrically, landmarks can be selected randomly [13] or using *maxmin* methods [14]. Selection can focus on the boundaries using minimum spanning tree, rather than randomly selected points or cluster centers [9]. Landmarks can also be chosen based on the minimum set cover problem [38, 57]. In addition, modification to existing landmark-based DR method such as L-Isomap can be done by introducing modifications to the distance matrices involving landmarks and its related spectral problems [59, 68]. Statistically, manifold landmarking utilizes regression [61], mixed-integer optimization [46], active learning [74], and Gaussian processes [39].

Topology-Inspired data skeletonization. Compared to exiting landmarking approaches, our strategy is one that is topological in nature. Our work utilizes advances in topology-inspired data skeletonization, that is, the process of extracting the topological structure of data using a low-dimensional even 1-dimensional representation, in order to better interpret complex, noisy, nonlinear, and high-dimensional data.

Data skeletonization can be broadly considered as a graph-extraction problem. The work in [28, 30] develops the notions of *principal curves* and *principal graphs*, which are roughly smooth curves that pass through the middle of a cloud of points. Metric graph reconstruction [1] also helps to skeletonize data based on inspecting and classifying local neighborhoods.

Topology-inspired data skeletonization from [22, 44] are most relevant to our framework. Ge et al. [22] give a framework to extract and simplify a 1-dimensional skeleton using the Reeb graph. Natali et al. [44] introduce a *Point Cloud Graph* as a data abstraction that is a generalization of the Reeb graph to arbitrary high-dimensional point clouds. Reeb graphs play a fundamental role in computational topological, topological data analysis and shape analysis; see [5] for a survey.

In this paper, we extract a 1-dimensional skeleton (referred to as *skeleton* for the remaining of the paper) from the input space based on an approximation of the Reeb graph. Compared to previous work, our work is novel in the sense that it utilizes such a skeleton for the purpose of landmark section and DR.

Manifold tearing and loop detection. Most classic DR techniques do not perform well when the data manifolds contain essential (i.e., non-contractable) loops, such as cylinders, tori or spheres. What sometimes are referred to as loopy manifolds [41] are in fact manifolds with nontrivial homology. Such manifolds typically cannot be embedded into the target space without introducing significant distortions. Some recent efforts have been made to detect and cut essential loops in such manifolds.

Lee and Verleysen [35] introduce a two-stage tearing procedure: first, a k-nearest neighbor (kNN) graph among the point cloud sample is used to represent the underlying space; second, a minimum spanning tree (MST) or a shortest path tree (SPT) that contains no cycles is computed on the kNN graph; Finally, edges that do not generate non-contractible cycles with more than 4 edges are reintroduced to form the torn graph for downstream DR.

Our work differs from [35] significantly in the following sense. We use a topology-inspired data skeleton that consists of landmarks and landmark connections to describe all candidate essential loops, and employ a homological criterion to choose the proper loop to tear while preserving as much as possible the homological characteristics of the data. Whereas other techniques cut all or a large number of loops, we try to cut, roughly, as few loops as possible while preserving the remaining homology.

4 Homology-Based Quality Assessment

4.1 Background

Before we present various homology-based criteria in assessing the quality of DR, we need to introduce a few relevant topological notions such as homology and persistent homology [17], and a distance-based evaluation criterion.

Homology and Betti numbers.   Given a topological space X, the 0-, 1- and 2-dimensional homology groups are denoted as $H_0(X)$, $H_1(X)$ and $H_2(X)$, respectively. Betti numbers b_i count the number of i-dimensional holes, and are used to distinguish spaces based on the connectivity across all dimensions. Formally, it is defined as the rank of the i-dimensional homology groups, $b_i = rank(H_i(X))$. For a torus, $b_0 = 1$, $b_1 = 2$ and $b_2 = 1$; this means that a torus has 1 connected components, 2 holes and 1 void.

Persistent homology.   Simply put, persistent homology studies homology at multiple scales. As illustrated in Fig. 3, we begin with a point cloud X equipped with a distance metric D_X (i.e. Euclidean distance). We study the homology of a sequence of spaces formed by a union of balls of increasing radius t centered at the points. Using persistent homology, we investigate the homological changes within this growing sequence of spaces indexed by time (this is referred to as a *filtration*).
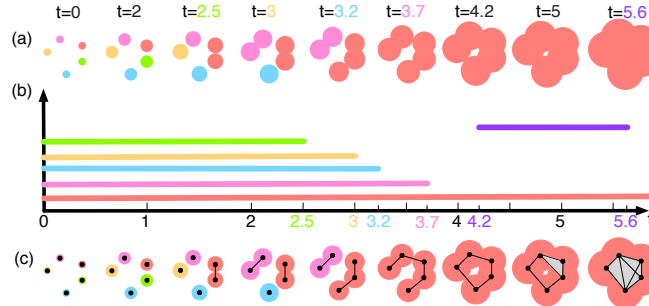

Figure 3: Computing persistent homology of a point cloud and the barcode.

In Fig. 3(a), at time $t = 0$, each colored point is *born* (appears) as its own (connected) component. As t increases, we focus on the important events when components merge with one another to form larger components or tunnels. We begin by tracking the birth and death times of each component or tunnel as well as its lifetime in the filtration. At $t = 2.5$, the green component merges into the red component and *dies* (disappears); therefore the green component has a lifetime (i.e., *persistence*) of 2.5. At $t = 3$, the orange component merges into the pink component and dies; therefore it has a persistence of 3. Similarly, the blue component dies at $t = 3.2$ while the pink component dies at $t = 3.7$. At time $t = 4.2$, the collection of components forms a tunnel; and the tunnel disappears at $t = 5.6$. The red component born at time 0 never dies, therefore it has a persistence of ∞. We record and visualize the appearance (birth), the disappearance (death), and the persistence of homological features in the filtration via persistence diagrams [10] (Fig. 4), or equivalently, persistence *barcodes* [23]. A point $p = (a, b)$ in the persistent diagram of X records a homological feature that is born at time a and dies at time b. 0- and 1-dimensional persistence diagrams, denoted as $PD_0(D_X)$ and $PD_1(D_X)$, captures the births and deaths of components and tunnels, respectively. Equivalently in the barcode of Fig. 3(b), such a feature is summarized by a horizontal bar that begins at a and ends at b.

Computationally, the above nested sequence of spaces can be combinatorially represented by a nested sequence of simplicial complexes (i.e. collections of vertices, edges and triangles) with a much smaller footprint, as illustrated in Fig. 3(c), see [18] for computational details.

Distance-based quality measure.   To evaluate the fits of various DR techniques on comparable grounds, Tenenbaum et al. [64] introduce the *residual variance* (*RV*):

$$RV(X,Y) = 1 - R^2(D_X, D_Y).$$

D_Y is the matrix of Euclidean distances in the low-dimensional embedding produced by a DR technique, and D_X is a best estimate of the intrinsic manifold distance for a given technique. In the case of Isomap,
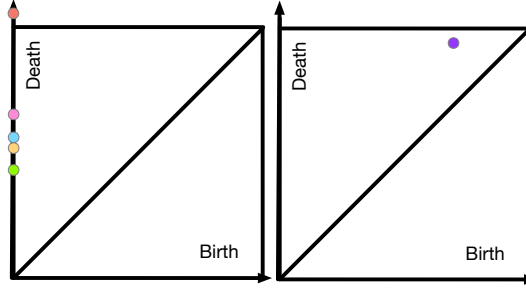
Figure 4: From left to right, 0- and 1-dimensional persistence diagrams.

D_X corresponds to the geodesic distance matrix approximated by the graph distance matrix D_G. For MDS, it is the Euclidean distance from the input data. R is the standard Pearson correlation coefficient that measures the linear correlation between all entries of D_X and D_Y (that are reshaped into vectors). Such a measure can also be used as a practical approach to select an appropriate neighborhood size for noiseless and noisy data, where lower residuals indicate betterfitting solutions with less metric distortion [2].

4.2 Homology-Based Quality Measures

There have been a few recent works in the direction of a homology-based evaluation criterion for DR. Paul and Chalup [47] compare the Betti numbers (b_0 and b_1) of spaces before and after DR and study their convergence as the number of sample points increases. Rieck and Leitte [52] present an evaluation scheme based on persistent homology. Specifically, density functions of the input space and the embedding space are estimated based on neighborhood graphs in both spaces; then persistence diagrams are computed for these density functions. The global quality of each embedding is assessed by computing the degree-2 Wasserstein distance between the two persistence diagrams. The local quality at each embedded point is estimated based on a mean matching cost between the diagrams.

For the purpose of this paper, we employ a few homology-based quality measures described below.

Wasserstein distance. Our first criterion is the Wasserstein distance for persistent homology [11]. Recall from the previous section, we could obtain a i-dimensional persistence diagram $PD_i(D_X)$ from a point cloud X described by a pair-wise distance matrix D_X. Let D_X and D_Y denote the distance matrices describing the structure relations among points within the high-dimensional input space and those within the low-dimensional embedding, respectively.

To measure similarities between the persistent homology of space X and Y, for a fixed dimension i, we define the *degree-p Wasserstein distance* as the Wasserstein distance between the i-th persistence diagrams of D_X and D_Y:

$$W_p(X,Y) = \left[\inf_{\eta} \Sigma_x \|x - \eta(x)\|_{\infty}^p \right]^{1/p},$$

where the infimum is over all bijections $\eta : PD_i(D_X) \to PD_i(D_Y)$, and the sum is over all points $x \in PD_i(D_X)$ [70]. Note that taking the limit $p \to \infty$ yields the bottleneck distance [50].

In the context of this paper, we always use $p = 2$, leading to the degree-2 *Wasserstein distance* measure, denoted as WD. Depending on the chosen dimension i, WD_0 and WD_1 roughly capture the homological distortions of 0- and 1-dimensional features before and after DR. For the assessment of Isomap and L-Isomap, D_X is the Euclidean distance matrix in the input space and D_Y is the Euclidean distance matrix in the embedding.

Similarly, we could consider the *bottleneck distance* by taking the limit $p \to \infty$, $W_{\infty}(X,Y) = \inf_{\eta} \Sigma_x \|x - \eta(x)\|_{\infty}$. As illustrated in Figure 2, we find the (near) optimal linear embedding by minimizing the Wasserstein distance measure (resp. bottleneck distance measure). For the remaining of this paper, we use the Wasserstein distance.

Persistent Betti numbers. Betti numbers count the number of homological features and can be used as summary statistics to differentiate topological spaces. However, Betti numbers alone do not differentiate between significant and noisy homological features. We use the *persistent Betti numbers* (*PB*) as a way to quantify how much homological information is preserved during the DR. Let PB_i denote the number

of significant i-dimensional homological features, that is, the number of points in the i-dimensional persistence diagram that is above a certain threshold that separates features from noise.

Finding a suitable threshold requires checking the separation of points in the persistence diagram [51]. Significant features can be extracted from the persistence diagram if it has an empty band (of a certain width) parallel to the diagonal that does not contain any points [8]. More sophisticated methods from statistics based on bootstrapping can be used to improve the threshold estimation that separate signals from noise, based on the notion of a *confidence band* [20], this is left to the future work.

In this paper, we use PB_1 as a rule of thumb to assess the quality of DR in terms of its preservation of significant 1-dimensional features.

Comparisons to prior work. We compare our DR framework with the prior work that use homology-based quality measures [47,52]. Although it is generally true that data with "more holes require higher sample sizes" [47], we demonstrate (in Section 7) that, by using carefully selected landmarks, we can preserve holes even with a small number of landmark points.

Comparatively speaking, there are a few critical differences between our work and the work in [52]. The global and local quality measures from [52] deal only with 0-dimensional while our proposed measure focuses on 1-dimensional homology preservation. Most importantly, [52] uses homology-based quality measures for the evaluation of existing DR techniques, while our work is actively searching and developing new DR techniques that maximize homology-based quality measures via manifold landmarking and tearing.

5 Homology-Preserving Manifold Landmarking

We begin this section with a review of the nonlinear DR techniques known as the Isomap [64] and landmark Isomap (L-Isomap) [13]. We then discuss homology-preserving landmark selection based on the Reeb graph and its discrete approximation. We summarize the pipeline for a new class of techniques by combining the utilities of homology-preserving landmarks with the efficiency of landmark-based DR.

5.1 Isomap and L-Isomap

Isomap. Suppose the original input data contains N samples in D dimensions, $X \in R^{D \times N}$. Isomap embeds the points onto a lower dimensional space $Y \in R^{d \times N}$ ($d < D$) while preserving geodesic distances between all input points [64]:

1. *Construct neighborhood graph*. A weighted, undirected k-nearest neighbor (kNN) graph G is constructed over all data points, where an edge between a point $x_i \in X$ and its neighbor $x_j \in X$ is assigned a weight that represents the Euclidean distance between them. An appropriate k can be chosen based on the residual variance [2,64].

2. *Compute shortest paths*. All pairwise shortest paths between points in the KNN graph G are computed to approximate the geodesic distances between them, which leads to an $N \times N$ graph distance matrix D_G.

3. *Construct a d-dimensional embedding*. Classical MDS is applied to the above graph distance matrix D_G to obtain a low-dimensional embedding.

Isomap suffers from two computational inefficiencies: calculating the shortest-paths distance matrix and eigenvalues within MDS. The former has a complexity of $O(kN^2 \log N)$ using Dijkstra's algorithm with Fibonacci heaps, while the latter takes $O(N^3)$ [13].

L-Isomap. L-Isomap [13] addresses the two inefficiencies of Isomap at once. It is based on the landmark MDS (L-MDS) [14]:

1. *Construct neighborhood graph* (same as in Isomap).

2. *Select landmarks*. n-points ($n \ll N$) from X are randomly selected to be landmark points.

3. *Compute shortest paths.* Compute the shortest paths from each data point to the landmarks, resulting in a $n \times N$ geodesic distance matrix. Also compute the $n \times n$ shortest paths distance matrix between pairs of landmarks.

4. *Apply L-MDS to obtain d-dimensional embedding*. First, apply classical MDS to the landmarks only, embedding them in R^d using as input the $n \times n$ distance matrix between pairs of landmarks. Second, the embedding coordinates for the remaining data points are computed based on a fixed linear transformation of their geodesic distances to the landmarks [58].

5. *PCA normalization (optional)*. This normalization is to re-orient the axes of the embedding to reflect the overall distribution, rather than the distribution of the set of landmarks; see [13, 14] for details.

L-Isomap leads to enormous savings when $n \ll N$: Computing the shortest paths in step 3 takes $O(knN \log N)$ using Dijkstra's algorithm and L-MDS in step 4 runs in $O(n^2 N)$ [13].

Here, to differentiate different versions of L-Isomap based on various landmark selection schemes, L-Isomap using randomly selected landmarks is referred to as the *random L-Isomap*, while the one using homology-preserving landmarks in the next section is called the *homology L-Isomap*.

5.2 Homology-Preserving Landmark Selection

Our work uses the idea of a data skeleton based on the Reeb graph for the purpose of landmark selection in DR. Although Reeb graphs have been used in the context of shape abstraction and comparison [22, 44], to the best of our knowledge this is the first time they have been used in the context of landmark-based DR.

In this section, we first review relevant topological notions and computations for Reeb graphs. We then describe our landmark section algorithm using a skeleton based on the Reeb graph.

Reeb graph. Let $f : X \to R$ be a continuous function defined on a manifold X. The level set of f at a value $a \in R$ is defined as $f^{-1}(a) = \{x \in X \mid f(x) = a\}$. The *Reeb graph* of f is constructed by identifying every connected component in a level set to a single point [22].

Extracting homological skeleton Given point cloud data, the domain can be approximated by a neighborhood graph (such as the kNN graph or the ε-neighborhood graph) among the data points, and efficient algorithms exist [22, 27, 44] to approximate the Reeb graph in such a discrete setting.

In this paper, we employ a mapper-based implementation to approximate the Reeb graph [54] as our homology-preserving data skeleton, refered to as the *homologicla skeleton* (or simply skeleton). The mapper algorithm [62] approximates the Reeb graph by considering the connected components of interval regions (i.e. $f^{-1}(a, b)$) instead of the connected components of level sets (i.e., $f^{-1}(a)$).

We start with a function $f : X \to R$ defined on a point cloud X, and a cover $\mathcal{U}$ of $f(X)$ consisting of finitely many open intervals $\mathcal{U} = \{(a_i, b_i)\}_{i=1}^n$. To specify such as cover, we pick two resolution parameters n and p, where n is the number of intervals and p is the percentage of overlap between a pair of adjacent intervals. *Pulling back* the cover $\mathcal{U}$ through f gives an open cover of the point cloud X, which is then refined into a connected cover by splitting each cover element into various clusters using a user-defined clustering algorithm [7]. Such a cover of X is denoted as $\mathcal{V} = f^*(\mathcal{U})$. In this paper, we use DBSCAN [19] for clustering. It is a widely used density-based clustering algorithm that groups together points that are closely packed together; the choice of clustering algorithm is not essential to our experiments.

The 1-dimensional skeleton of the nerve of $\mathcal{V}$ is considered a discrete approximation of the Reeb graph of f on X; it is referred to as the *homological skeleton* for the remainder of the paper. Such a skeleton is a graph with nodes representing the elements of $\mathcal{V}$, and edges representing the pairs of cover elements in $\mathcal{V}$ with nonempty intersections.

In the original mapper algorithm, the node of a skeleton represents abstractly a cover element of the point cloud, that is, a cluster of points in X. However, in our setting, the nodes of a homological skeleton are considered as the landmarks for DR; therefore, we choose the *centroid* of each cluster as its representative. The centroid of each cluster is part of the original point cloud, and serves as a landmark for L-Isomap.

Extracting homological skeleton does not increase the asymptotic complexity of L-Isomap. In our experiments, the running time of H-L-Isomap is comparable with that of R-L-Isomap.

Filter function. The key idea behind the Reeb graph is that it explores the topology of a space by analyzing the behavior of a possibly varying real-valued function (referred to as a *filter function*) defined on it [5]. Reeb graphs encode topological information on data in a 1-dimensional structure, disregarding the dimension of the data in the ambient space [5]. The data can be regarded as being parameterized with

respect to the filter function being used, in other words, the filter function "plays the role of the *lens*" through which we look at the properties of the data [5].

Different filter functions lead to different insights into the point cloud. It remains an open question as how to choose an appropriate filter function beyond a best practice or a guesstimation. Commonly used options include height functions, distances from the barycentre of a space, surface curvature, integral or average geodesic distances and geodesic distances from a source point in the space [5].

In this paper, we use mainly the geodesic distance from a source point as the filter function, referred to as the distance-to-the-base-point or simply the DTB function. Such a filter function has shown desirable properties in capturing the 1-dimensional homological information of the space [5, 22]. We demonstrate via experiments in Section 7 that a skeleton induced by DTB is homology-preserving for landmark-based DR; it also serves as a compact and informative summary for guiding the manifold tearing process.

Landmark selection pipeline.   Given a point cloud X in R^D, we now summarize our homology-preserving landmark selection pipeline and its combination with L-Isomap (referred to as the homology L-Isomap):

1. *Construct neighborhood graph* (same as Isomap). Let G denote the resulting kNN graph.

2. *Compute a filter function f on X. f* captures certain desirable structural information of X suitable for DR. In our experiments, we use DTB as the filter function and a base point is chosen from extreme points or barycenters (see Section 7 for details). DTB can be computed based on G from a given base point.

3. *Compute skeleton and landmarks.* Compute a discrete approximation of the Reeb graph of f as a homological skeleton, using the mapper algorithm. The cover $\mathscr{U}$ of $f(X)$ is given by user-specified resolution parameters n and p. The nodes of the skeleton correspond to clusters of points in X; the cluster centroids are chosen as the landmarks for L-Isomap, denoted as $X_L \subseteq X$.

4. *Apply L-Isomap.* Replace randomly generated landmarks (step 2 of L-Isomap) with the homology-preserving landmarks (a.k.a. homological landmarks) X_L and apply the rest of L-Isomap algorithm.

The above pipeline is not restricted to L-Isomap. We believe it can be easily extended to other graph-based DR techniques [75].

A simple example.   Our pipeline is illustrated with a simple example in Fig. 5. We begin with a noisy point cloud with 1983 points sampled from a swiss roll with an irregular, hard-to-spot hole in the middle. First, a DTB filter function is computed with respect to an extremal base point. Second, a homological skeleton connecting a set of 22 landmarks is obtained using the mapper approximation. Third, we apply L-Isomap using these homological landmarks in the skeleton (black stars). In Fig. 5, the black homological skeleton is highlighted in the input space, as well as the Isomap and L-Isomap embeddings, which clearly captures the location of the significant hole in the data.

This simple example demonstrates that the homological L-Isomap has the potential to preserve as much as possible the 1-dimensional homological feature even with a smaller number of landmarks than the Random L-Isomap algorithm (roughly $n = O(\sqrt{N})$). However, given the dataset contains a single loop, homological L-Isomap does introduce distance distortion away from the loop. Surprisingly, homological L-Isomap is shown to outperform both Isomap and random L-Isomap in certain datasets using the widely accepted residual variance (see Section 7).

6   Homology-Preserving Manifold Tearing

Complex data that contain essential loops may or may not be embedded into low-dimensional space without introducing significant distortions. An alternative and complementary approach for homology-preserving DR is through manifold tearing. In this section, we give a simple and fast manifold tearing procedure guided by the homology-preserving skeleton of the point cloud data. Different from prior work, our procedure tries to cut as few loops as possible, using homology-based quality assessment (described in Section 4) while at the same time preserving as much as possible the remaining homological features.

Suppose we have a point cloud equipped with a pre-computed homology-preserving skeleton (see Section 5). Our homology preserving tearing process is as follows:

- *Construct neighborhood graph* (same as Isomap). The neighborhood graph is denoted as G. A slightly larger k can be chosen to account for the tearing process (optional).
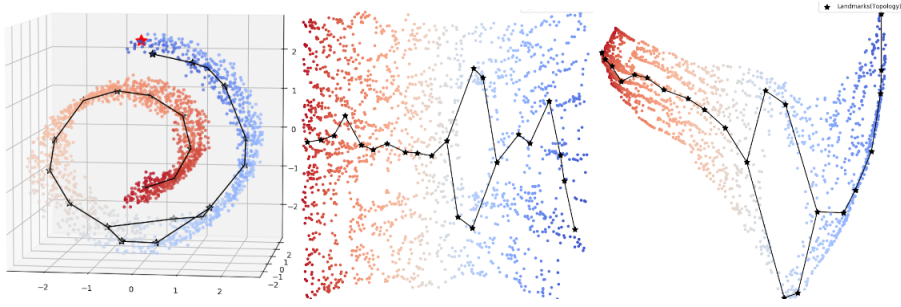
Figure 5: A *swiss roll with a hole* example. (a) The original point cloud is colored by the *distance to a base point* function. The base point is marked by a red star. (b) Isomap embedding. (c) Homological L-Isomap embedding. The homological skeleton is highlighted in black.
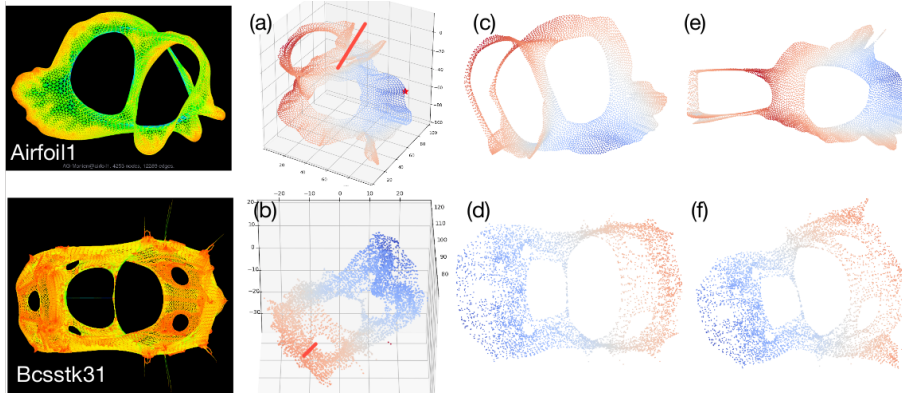

Figure 6: Results without and with manifold tearing for *Airfoil1* (top) and *Bcsstk31* (bottom). (a)-(b) original point clouds with marked cutting location. (c)-(d) Isomap embeddings without tearing. (e)-(f) Isomap embedding with tearing.

- *Tear the neighborhood graph*. A cut plane can be specified based on the homology-preserving skeleton. Specifically, a cut location can be specified on an edge of the skeleton, which then defines a cut plane that is orthogonal to the edge to be cut. An edge that spans a pair of nodes on the opposite side of the cut plane is removed from G, resulting a new graph G'.

- *Compute shortest paths*. Compute shortest paths between all nodes in G' and obtain a geodesic distance matrix $D_{G'}$.

- *Apply Isomap* to $D_{G'}$.

The above process is exploratory in nature, that is, we can use different evaluation criteria of the resulting embeddings to rank the potential cut locations. In this paper, we use the number of significant homology classes as the criterion. In addition, we envision such a process could be embedded into any interactive visualization framework for DR that involves human-in-the-loop.

7   Results

We present examples in this section illustrating that we can achieve homology preservation while at the same time maintaining (and sometimes improving) distance preservation. We explore complementary procedures with a common goal: homology-preserving manifold landmarking and manifold tearing, both aided by the use of a small homological skeleton. Such a skeleton is compact, fast to extract and does not introduce significant distance distortions.
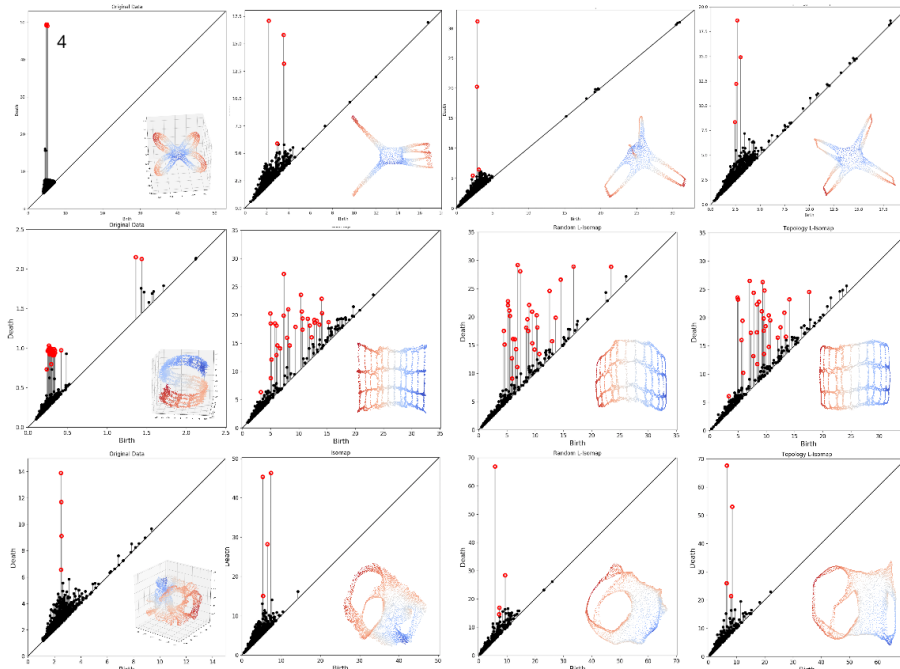
11

Figure 7: Homology-based quality assessment of DR for *Octa* (row 1), *Fishing Net* (row 2) and *4elt* (row 3). For each each, from left to right: persistence diagrams for the original data, Isomap embedding, random L-Isomap embedding and homological L-Isomap embedding.

	octa	fish.	4elt	cyl.-3	cyl.-5	air.	bcsstk	mice
$|X|$	2994	6188	7807	2K	2K	8034	8030	674
$|X_L|$	76	63	18	54	82	60	53	18

Table 1: The number of landmarks $|X_L|$ for each dataset of size $|X|$.

7.1 Data

For manifold landmarking, we demonstrate our technique with datasets that contain nontrivial homology. Simplify put, datasets with loops are more likely to benefit significantly from our technique.

Octa is a point cloud sampled from a mesh of octahedron handles. The original mesh contains up to 41*K* vertices. *Fishing Net* is a synthetic, noisy point cloud sampled from a "S"-shaped surface that contains 3×11 irregular holes. *4elt* is derived from a 3-dimensional embedding of the 4elt graph used in [21]. The original graph from [69] contains 15606 nodes and 45878 edges, and is a mesh created to study fluid flow around a 4-element airfoil. As stated in [21], the original graph "exhibits extreme variation in the spatial density of nodes".

Finally, *Mice* dataset contains 300-dimensional point clouds derived from time-varying temperature measurements of pregnant mice [63]. The point cloud is generated by standard delayed window embedding with a window size of 300 in signal processing. We run our experiment on a particular pregnant mouse that is not jet-lagged, and hope to detect and preserve 1-dimensional homological features in the input space that capture periodicity in the signal.

For manifold tearing, we use datasets that contain essential loops for demonstration. *Cylinder-3* is a point cloud sampled from a cylinder with 3 holes carved out, and *Cylinder-5* is created similarly. *Airfoil1* comes from a 2-dimensional finite element problem under the AG-Monien Matrix group from the SuiteSparse Matrix Collection [15]. *Bcsstk31* is derived from a 3-dimensional embedding of a stiffness matrix for automobile component [16].

7.2 Dimensionality Reduction with Manifold Landmarking

Results and evaluation with persistence diagrams. For each dataset, 2-dimensional embeddings obtained using homological L-Isomap are compared with Isomap and random L-Isomap in Fig. 1.
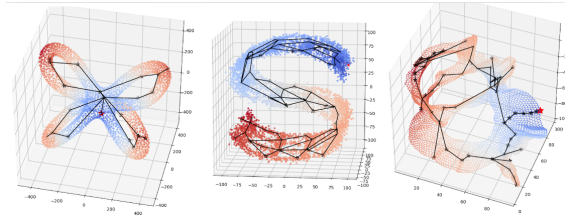
Figure 8: Homological skeletons for *Octa* (left) using 21 landmarks, *Fishing Net* (middle) and *4elt*(right).

Evaluation using persistent Betti numbers are illustrated by the 1-dimensional persistence diagrams in Fig. 7. We determine the number of persistent (significant) features by looking at the separation between points in the diagram. Suppose each dataset contains m persistent features in the original point cloud, then top m features with the highest persistence are marked in red within the persistence diagram associated with each embedding.

For *Octa*, as shown in Fig. 7 (row 1), the original data contains 8 significant features, 4 of which (colored red) correspond to the visible loops via embeddings (4 other features are the interior tunnels within each handle). All 4 of the red features are preserved (i.e. they remain significant, that is, well-separated from the diagonal of the persistence diagram) using homological L-Isomap. Isomap preserves 3 loops while random L-Isomap preserves only 2. For more detailed analysis, it is remarkable to see that using only 21 landmarks, the homological skeleton is able to summarize the homological features reasonably well (Fig. 8, left).

For *Fishing Net*, as shown in Fig. 7 (row 2), the original data has 33 significant features in the persistence diagram; and both Isomap and homological L-Isomap perform comparatively in terms of preserving the shape of each feature in the embeddings. However, homological L-Isomap uses only 66 points as landmarks (roughly 1% of the size of the point cloud), and is therefore more computationally efficient. Furthermore, its homological skeleton in Fig. 8 (middle) captures the underlying homological feature pretty well.

For *4elt*, as shown in Fig. 7 (row 3), the original data contains 4 significant features; 3 of which are readily visible in the 3-dimensional embedding of its homological skeleton in Fig. 8 (right). Both Isomap and homological L-Isomap preserve these features reasonably well, while random L-Isomap preserves only 2. In addition, homological L-Isomap does slightly better in preserving the shape of a couple of features.

For *Mice*, we combine the results of DR with 1-dimensional persistence diagrams in Fig. 9. There are 2 significant features in the original 300-dimensional input space. Such features likely correspond to periodicity in the temperature profile of the mice that corresponds to circadian or ultradian rhythms. Both Isomap and homological L-Isomap perform comparatively in terms of preserving the most dominant feature, while homological L-Isomap only uses a small fraction of the points as landmarks. On the other hand, random L-Isomap is able to detect the significant feature but does not preserve its shape as well.

Quality assessment with residual variance. We also assess the quality of DR using quality measures introduced in Section 4, in particular, the RV measure. WD_0 and WD_1 measures are also computed (but not reported here), as they are partially encoded by persistence diagrams in the previous section.

For *Octa*, we evaluate the quality of each embedding using the RV measure by varying the number of landmarks, see Fig. 10. As the number of chosen landmarks increases, we are interested in how well homological L-Isomap preserves distances, when compared with Isomap and random L-Isomap. For a fixed landmark size, the blue box plot corresponds to the RV measures for 20 instances of random L-Isomap, each drawing landmarks randomly from a fixed point cloud: solid blue line in the box plot is the median, dotted blue line is the mean, the boundary of the box is the standard deviation, and black hollow circles are outliers. Solid red circles are RV measures for homological L-Isomap (denoted as Topology L-Isomap in the figure), while solid green circles are for Isomap.

A surprising observation is that homological L-Isomap outperforms Isomap and random L-isomap in terms of distance preservation, when the number of landmarks is small (below 100). In fact, homological L-Isomap beats random L-Isomap with just 21 landmarks and it outperforms Isomap with 42 landmarks. The optimal landmark size that achieves both computational efficiency and quality is at around 76 landmarks. When the number of landmarks goes beyond 150, homological L-Isomap does not seem to
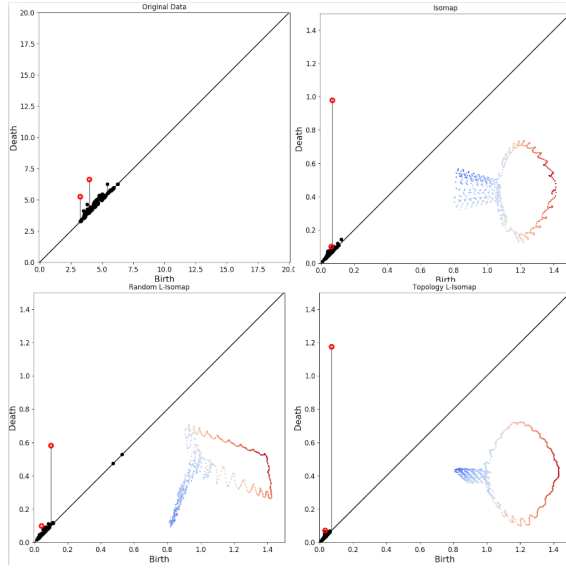
Figure 9: For *Mice*, persistence diagrams for original data (a), Isomap embedding (b), random L-Isomap embedding (c) and homological L-Isomap embedding (d) are combined with DR results.
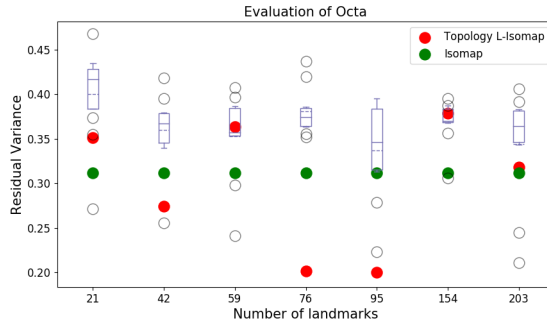

Figure 10: Quality assessment of embeddings using residual variance for *Octa*.

have an obvious advantage over other methods. In fact, at 203 landmarks, homological L-Isomap performs comparably with Isomap. This is not surprising, with a large number of landmarks, both L-Isomap and Isomap preserve the geometry of the data equally well.

We also compare several datasets, the *swiss roll with a hole*, *Fishing Net* and *Octa*, using their respective optimal landmark size, in FIg. 11. Notice that homological L-Isomap outperforms the others for both *Fishing Net* and *Octa*, while it does not do well with *Swiss roll with a hole*. Intuitively, homological L-Isomap performs best when the data is complex, and has nontrivial homological features. In this case, both *Fishing Net* and *Octa* are a lot more complex and homologically interesting than the *Swiss roll with a hole*.

7.3 Dimensionality Reduction with Manifold Tearing

Manifold tearing results are shown in Fig. 14 for *Cylinder-5* and Fig. 6 for *Airfoil1* and *Bcsstk31* respectively. See Fig. 12 and Fig. 13 for quality assessment using persistence diagrams.

Given a homological skeleton for *Cylinder-5*, we apply multiple cutting options to the skeleton and rank the resulting embeddings by homology preservation. As shown in Fig. 14, without manifold tearing, the Isomap embedding destroys 5 out of 6 homological features, while optimal tearing preserves 5 out of 6 persistent homological features.

While Isomap preserves reasonably well the 3 persistent features for *Airfoil1*, manifold tearing further preserves 2 of the 3 homological features if we are willing to destroy one of them (Fig. 6 top, and Fig. 13).
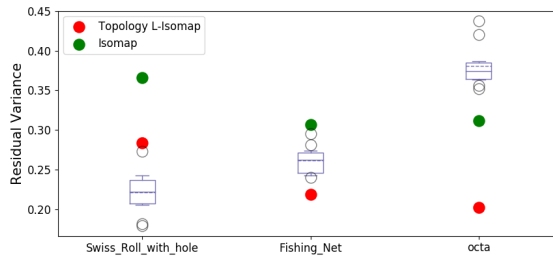
Figure 11: Quality assessment of embeddings using residual variance for different datasets, including *Swiss roll with a hole*, *Fishing Net* and *Octa*.
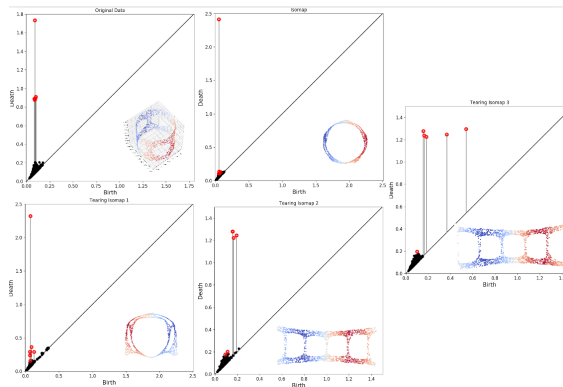

Figure 12: Homology based quality assessment of DR for *Cylinder-5*.

For the case of *Bcsstk31*, we can focus on manifold tearing by cutting a short edge in the homological skeleton of *Bcsstk31* as shown in Fig.15, therefore "open up" the space further to reveal more geometric structures of the data.

8 Discussion

We demonstrate in this paper that we can achieve homology preservation while maintaining and possibly improving distance preservation using homological L-Isomap and (almost) homology-preserving manifold tearing. Many research questions remain. In particular, we are interested in exploring higher-dimensional homological skeletons [67] for DR to preserve homological features beyond 1-dimensions.

References

[1] M. Aanjaneya, F. Chazal, D. Chen, M. Glisse, L. Guibas, and D. Morozon. Metric graph reconstruction from noisy data. *International Journal of Computational Geometry and Applications*, 22(04):305–325, 2012.

[2] M. Balasubramanian, E. L. Schwartz, J. B. Tenenbaum, V. de Silva, and J. C. Langford. The isomap algorithm and topological stability. *Science*, 295(5552):7, 2002.

[3] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, 2003.

[4] S. Bhattacharya, D. Lipsky, R. Ghrist, and V. Kumar. Invariants for homology classes with application to optimal search and planning problem in robotics. *Annals of Mathematics and Artificial Intelligence*, 67(3-4):251–281, 2013.

[5] S. Biasotti, D. Giorgi, M. Spagnuolo, and B. Falcidieno. Reeb graphs for shape analysis and applications. *Theoretical Computer Science*, 392:5–22, 2008.

[6] G. Carlsson, T. Ishkhanov, V. De Silva, and A. Zomorodian. On the local behavior of spaces of natural images. *International journal of computer vision*, 76(1):1–12, 2008.
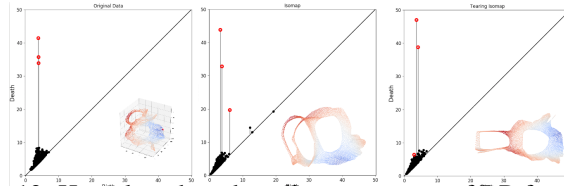
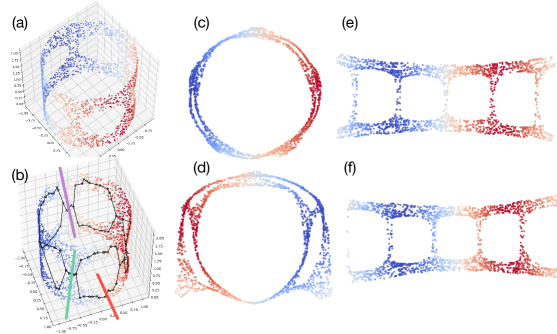Figure 13: Homology based quality assessment of DR for *Airfoil1*.

Figure 14: Results without and with manifold tearing for *Cylinder-5*. (a) Original point cloud. (b) Homological skeleton with 3 cutting options colored red, purple and green. (c) Isomap embedding without tearing. (d) Partial tearing with the red option. (e) Non-optimal tearing with purple option. (f) Optimal tearing with the green option.

[7] M. Carriére, B. Michel, and S. Oudot. Statistical analysis and parameter selection for mapper. *ArXiv: 1706.00204*, 2017.

[8] F. Chazal, L. J. Guibas, S. Y. Oudot, and P. Skraba. Persistence-based clustering in Riemannian manifolds. *Journal of the ACM*, 60(6), 2013.

[9] Y. Chen, M. M. Crawford, and J. Ghosh. Improved nonlinear manifold learning for land cover classification via intelligent landmark selection. *IEEE International Conference on Geoscience and Remote Sensing Symposium*, 2006.

[10] D. Cohen-Steiner, H. Edelsbrunner, and J. Harer. Stability of persistence diagrams. *Discrete and Computational Geometry*, 37(1):103–120, 2007.

[11] D. Cohen-Steiner, H. Edelsbrunner, J. Harer, and Y. Mileyko. Lipschitz functions have l_p-stable persistence. *Foundations of Computational Mathematics*, 10(2):127–139, 2010.

[12] J. P. Cunningham and Z. Ghahramani. Linear dimensionality reduction: Survey, insights, and generalizations. *Journal of Machine Learning Research*, pp. 2859–2900, 2015.

[13] V. de Silva and J. B. Tenenbaum. Global versus local methods in nonlinear dimensionality reduction. *Advances in Neural Information Processing Systems*, pp. 705–712, 2003.

[14] V. de Silva and J. B. Tenenbaum. Sparse multidimensional scaling using landmark points. Technical report, Stanford University, 2004.

[15] R. Diekmann and R. Preis. Ag-monien graph collection. `https://www.cise.ufl.edu/research/sparse/matrices/AG-Monien/airfoil1_dual.html`.
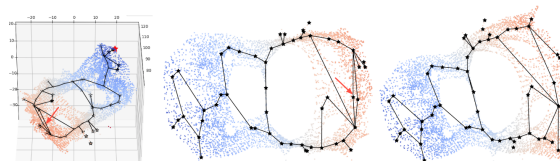
Figure 15: Using homological skeleton to aid manifold tearing for *Bcsstk31*. The location marked by the red arrow is where skeleton cutting takes place.

16

[16] I. Duff, R. Grimes, and J. Lewis. Sparse matrix problems. *ACM Trans. on Mathematical Software*, 14(1):1–14, 1989.

[17] H. Edelsbrunner and J. Harer. Persistent homology – a survey. *Contemporary mathematics*, 453:257–282, 2008.

[18] H. Edelsbrunner and J. Harer. *Computational Topology: An Introduction*. American Mathematical Society, Providence, RI, USA, 2010.

[19] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. *International Conference on Knowledge Discovery and Data Mining*, pp. 226–231, 1996.

[20] B. T. Fasy, F. Lecci, A. Rinaldo, L. Wasserman, S. Balakrishnan, and A. Singh. Confidence sets for persistence diagrams. *The Annals of Statistics*, 42(6):2301–2339, 2014.

[21] E. R. Gansner, Y. Koren, and S. C. North. Topological fisheye views for visualizing large graphs. *IEEE Transactions on Visualization and Computer Graphics*, 11(4):457–468, 2005.

[22] X. Ge, I. Safa, M. Belkin, and Y. Wang. Data skeletonization via Reeb graphs. *Advances in Neural Information Processing Systems*, 2011.

[23] R. Ghrist. Barcodes: The persistent topology of data. *Bullentin of the American Mathematical Society*, 45:61–75, 2008.

[24] R. Ghrist. Three examples of applied & computational homology. *Nieuw Archief voor Wiskunde (The Amsterdam Archive, Special issue on the occasion of the fifth European Congress of Mathematics)*, pp. 122–125, 2008.

[25] A. Gracia, S. González, V. Robles, and E. Menasalvas. A methodology to compare dimensionality reduction algorithms in terms of loss of quality. *Information Sciences*, 270(20):1–27, 2014.

[26] J.-S. Ha, S.-S. Park, and H.-L. Choi. Asymptotically optimal sampling-based algorithms for topological motion planning. *ArXiv: 1603.05099*, 2016.

[27] W. Harvey, Y. Wang, and R. Wenger. A randomized $O(m \log m)$ algorithm for computing Reeb graphs of arbitrary simplicial complexes. *ACM Symposium on Computational Geometry*, pp. 267–276, 2010.

[28] T. Hastie and W. Stuetzle. Principal curves. *Journal of the American Statistical Association*, 84(406):502–516, 1989.

[29] I. M. James, ed. *History of Topology*. Elsevier B.V., 1999.

[30] B. Kégl and A. Krzyżak. Piecewise linear skeletonization using principal curves. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(1):59–74, 2002.

[31] J. Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1):1–27, 1964.

[32] J. Kruskal. Nonmetric multidimensional scaling: a numerical method. *Psychometrika*, 29(2):115–129, 1964.

[33] E. Lamb. What we talk about when we talk about holes. Scientific American Blog Network, `https://blogs.scientificamerican.com/roots-of-unity/what-we-talk-about-when-we-talk-about-holes/`, December 2014.

[34] A. B. Lee, K. S. Pedersen, and D. Mumford. The non-linear statistics of high-contrast patches in natural images. *International Journal of Computer Vision*, 54:83–103, 2003.

[35] J. A. Lee and M. Verleysen. Nonlinear dimensionality reduction of data manifolds with essential loops. *Neurocomputing*, 67:29–53, 2005.

[36] J. A. Lee and M. Verleysen. *Nonlinear Dimensionality Reduction*. Springer, 2007.

[37] J. A. Lee and M. Verleysen. Quality assessment of dimensionality reduction: Rank-based criteria. *Neurocomputing*, 72(7):1431–1443, 2009.

[38] Y.-K. Lei, Z.-H. You, T. Dong, Y.-X. Jiang, and J.-A. Yang. Increasing reliability of protein interactome by fast manifold embedding. *Pattern Recognition Letters*, 34(4):372–379, 2013.

[39] D. Liang and J. Paisley. Landmarking manifolds with gaussian processes. *Proceedings of Machine Learning Research*, 37:466–474, 2015.

[40] S. Liu, D. Maljovec, B. Wang, P.-T. Bremer, and V. Pascucci. Visualizing high-dimensional data: Advances in the past decade. *IEEE Transactions on Visualization and Computer Graphics*, 23(3):1249–1268, 2017.

[41] D. Meng, Y. Leung, and Z. Xu. Detecting intrinsic loops underlying data manifold. *IEEE Transactions on Knowledge and Data Engineering*, 25(2):337–347, 2013.

[42] B. Mokbel, W. Lueks, A. Gisbrecht, and B. Hammer. Visualizing the quality of dimensionality reduction. *Neurocomputing*, 112:109–123, 2013.

[43] J. C. Nascimento and G. Carneiro. Deep learning on sparse manifolds for faster object segmentation. *IEEE Transactions on Image Processing*, 26(10):4978–4990, 2017.

[44] M. Natali, S. Biasotti, G. Patané, and B. Falcidieno. Graph-based representations of point clouds. *Graphical Models*, 73(5):151–164, 2011.

[45] J. J. O'Connor and E. F. Robertson. A history of topology. MacTutor History of Mathematics, `http://www-groups.dcs.st-and.ac.uk/history/PrintHT/Topology_in_mathematics.html`, 1996.

[46] C. Orsenigo and C. Vercellis. Landmark selection for isometric feature mapping based on mixed-integer optimization. *Lecture Notes in Computer Science*, 8234, 2013.

[47] R. Paul and S. K. Chalup. A study on validating non-linear dimensionality reduction using persistent homology. *Pattern Recognition Letters*, 100:160–166, 2017.

[48] J. A. Perea, A. Deckard, S. B. Haase, and J. Harer. SW1PerS: Sliding windows and 1-persistence scoring; discovering periodicity in gene expression time series data. *BMC Bioinformatics*, 16(1):257, 2015.

[49] J. A. Perea and J. Harer. Sliding windows and persistence: An application of topological methods to signal analysis. *Foundations of Computational Mathematics*, 15(3):799–838, 2015.

[50] J. Reininghaus, S. Huber, U. Bauer, and R. Kwitt. A stable multi-scale kernel for topological machine learning. In *IEEE conference on computer vision and pattern recognition*, pp. 4741–4748, 2015.

[51] B. Rieck and H. Leitte. Agreement analysis of quality measures for dimensionality reduction. *Topological Methods in Data Analysis and Visualization IV*, pp. 103–117, 2015.

[52] B. Rieck and H. Leitte. Persistent homology for the evaluation of dimensionality reduction schemes. *Computer Graphics Forum*, 34(3):431–440, 2015.

[53] S. Roweis and L. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

[54] N. Saul and H. J. van Veen. Mlwave/kepler-mapper: 186f (version 1.0.1). zenodo. `http://doi.org/10.5281/zenodo.1054444`, November 2017.

[55] R. Shepard. The analysis of proximities: multidimensional scaling with an unknown distance function. i. *Psychometrika*, 27(2):125–140, 1962.

[56] R. Shepard. The analysis of proximities: multidimensional scaling with an unknown distance function. ii. *Psychometrika*, 27(3):219–246, 1962.

[57] H. Shi, B. Yin, Y. Bao, and Y. Lei. A novel landmark point selection method for L-ISOMAP. *IEEE International Conference on Control & Automation*, pp. 621–625, 2016.

[58] H. Shi, B. Yin, Y. Kang, C. Shao, and J. Gui. Robust L-Isomap with a novel landmark selection method. *Mathematical Problems in Engineering*, 2017.

[59] L. Shi, P. He, and E. Liu. An incremental nonlinear dimensionality reduction algorithm based on ISOMAP. *Lecture Notes in Computer Science*, 3809, 2015.

[60] S. Siegel and N. Castellan. *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill Inc., 1988.

[61] J. Silva, J. Marques, and J. ao Lemos. Selecting landmark points for sparse manifold learning. In Y. Weiss, B. Schölkopf, and J. C. Platt, eds., *Advances in Neural Information Processing Systems*, pp. 1241–1248. MIT Press, 2006.

[62] G. Singh, F. Mémoli, and G. Carlsson. Topological methods for the analysis of high dimensional data sets and 3d object recognition. *Eurographics Symposium on Point-Based Graphics*, 22, 2007.

[63] B. L. Smarr, I. Zucker, and L. J. Kriegsfeld. Detection of successful and unsuccessful pregnancies in mice within hours of pairing through frequency analysis of high temporal resolution core body temperature data. *PloS one*, 2016.

[64] J. B. Tenenbaum, V. d. Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

[65] M. E. Tipping. Sparse bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, 1:211–244, 2001.

[66] L. van der Maaten, E. Postma, and J. van den Herik. Dimensionality reduction: A comparative review. Technical report, Tilburg University, 2007.

[67] S. K. Verovšsek, V. Kurlin, and D. Lešnik. A higher-dimensional homologically persistent skeleton. *ArXiv: 1701.08395*, 2017.

[68] M. Vladymyrov and M. A. Carreira-Perpinán. Locally linear landmarks for large-scale manifold learning. *European Conference on Machine Learning*, 2013.

[69] C. Walshaw. A multilevel algorithm for force-directed graph drawing. *International Symposium on Graph Drawing*, pp. 171–182, 2000.

[70] L. N. Wasserstein. Markov processes over denumerable products of spaces describing large systems of automata. *Problems of Information Transmission*, 5:47–52, 1969.

[71] K. Q. Weinberger, B. D. Packer, and L. K. Saul. Nonlinear dimensionality reduction by semidefinite programming and kernel matrix factorization. *International Workshop on Artificial Intelligence and Statistics*, 2005.

[72] Wikipedia contributors. Topology. Wikipedia, The Free Encyclopedia, March 2018.

[73] S. Xia. A topological analysis of high-contrast patches in natural images. *Journal of Nonlinear Sciences and Applications*, 9:126–138, 2016.

[74] H. Xu, L. Yu, M. Davenport, and H. Zha. Active manifold learning via a unified framework for manifold landmarking. ArXiv: 1710.09334, 2017.

[75] S. Yan, D. Xu, B. Zhang, H. jiang Zhang, Q. Yang, and S. Lin. Graph embedding and extensions: A general framework for dimensionality reduction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(1):40–51, 2007.

[76] P. Zhang, Y. Ren, and BoZhang. A new embedding quality assessment method for manifold learning. *Neurocomputing*, 97(15):251–266, 2012.