Topological Structure of Linear Manifold Clustering

Art Diky and Robert M. Haralick

Computer Science Department
The Graduate Center CUNY
New York, NY, USA
adiky@gradcenter.cuny.edu, rharalick@gc.cuny.edu

Abstract. In topological data analysis, the first step is a construction of a simplicial complex from a discrete points set $\mathcal{D}$ sampled from some manifold. In this paper, we present an algorithm for the efficient computation of such simplicial complex which utilizes a clustering structure, comprised of subspace clusters, of the point set for speeding up a complex construction procedure while keeping relevant topological invariants of the underlying sampled manifold. Experiments show that the proposed construction algorithm produces simplicial complexes with less number of simplices and noise which gives a better homological picture than other construction methods as well as an improved construction performance and a topological invariant interpretability on a geometrical level.

Keywords: linear manifold clustering · simplicial complex · topological data analysis.

1 Introduction

Given a point cloud $\mathcal{D}$ in an N-dimensional space X, we wish to construct a simplicial complex $\mathcal{S}$ such that it approximates the geometric or topological structure of the space X. In particular, we try to find efficiently and robustly a structure that contains the same topological invariants as the underlying space. In general, it is hard to construct a precise digital representation of a geometric structure in a high dimensional space. For reducing the computational complexity associated with complex description of structures in large dimensions, we will represent them as partitions with an extended description that captures the actual properties of the underlying structure. Such a partition representation can be given by bounded linear manifold clusters that carry the description of the bounded linear subspaces where the cluster is located [11]. Cluster partitioning of the geometric structure could be extended further by considering spatial relations between clusters themselves forming a piecewise linear (PL) representation of the clustering structure. In this paper we propose an efficient algorithm for the computation of the simplicial complex from the dataset geometric clustering.

1.1 Related work

There are many methods that use geometrical properties of a point set sample to discover a structure of a sampled manifold, i.e. geometric clustering or manifold learning methods.

The main challenge for clustering, especially in high-dimensional spaces, is relevance of different subsets of points are relevant to different clusters while the cluster points specified in the full space are associated with various subspaces. Moreover, various correlations between points are relevant to different clusters. There are variety of subspace and correlation clustering algorithms that try to find clusters in the axis-aligned and the arbitrarily oriented subspaces [16].

Because of the infinite number of subspaces, additional assumptions are required to overcome infinite search space. One of the assumptions is high-dimensional observations lie on or close to multiple smooth low-dimensional manifolds embedded in a full space of dataset. Thus, viewing a cluster as a collection of points on or near a compact manifold becomes a reasonable and promising extension of traditional centroid-based clustering methods, which leads to manifold clustering.

Haralick and Harpaz [11] presented a linear manifold clustering algorithm, which is a strict partitioning clustering algorithm, that performs stochastic search on the dataset in order to find the best possible location of the linear manifold clusters. Kak [15] used a linear manifold representation of a fixed number of clusters, obtained by sampling the original dataset and minimizing the reconstruction error from point assignments to cluster prototypes. Peng et al. [21] constructed linear manifold cluster prototypes by performing spectral decomposition of small random samples with subsequent assignment of the rest of the dataset points to the nearest subspace cluster prototype. Wang et al. [27] used a mixture of probabilistic PCAs to form a collection of linear manifolds on the dataset.

These linear manifolds were used to reconstruct non-linear manifolds, that reflect local geometric information of the data, and form a suitable affinity matrix that served as input for a spectral clustering technique. Moreover, many linear methods fail to provide good performance when applied to nonlinear structures. On the other hand, nonlinear methods, such as nonlinear dimensionality reduction techniques, can be naturally used on linear manifolds [10, 23, 24].

Describing the population by a probabilistic model based on the samples allows us to explore various aspects of the population and make predictions consistent with the data. Accurate modelling of the subspace clusters allows cluster points to be discriminated using the relative probability densities under the various models. Hinton et al. [14] proposed modeling methods based on mixtures of principal components and factor analyzers. Both methods are based on locally linear, low-dimensional approximations to the underlying linear manifold cluster data. Harpaz and Haralick [13] defined a non-parametric density estimation modeling technique for modeling data that lie in lower dimensional linear manifolds through a mixture of linear manifolds models. The above models are cases of latent variable modeling which allows probabilistic construction and representation of high-dimensional structures in a fewer dimensions. A latent variable model assumption states that the observed data in a high-dimensional space is generated from a low-dimensional underlying process [2]. Thus, linear latent variable models could be used under the manifold hypothesis assumption [20].

Geometric and probabilistic models of the clustering provide a good approximation for low-dimensional data. We seek for a model description that can be robust and efficient for high-dimensional data. Such model can be based on the data topological structure, and can be used to reduce high dimensional data sets into compact representation which captures topological and geometric information. A natural approach is to represent a topological structure as a simplicial complex.

There are several well-known construction techniques for a simplicial complex from the point dataset: Čech, Vietoris–Rips [30], "witness" complex [7], Delaunay triangulation, and Mapper [22]. However, many of these construction techniques are computationally intensive especially for high-dimensional spaces.

We propose a novel method for a simplicial complex construction which preserves geometric properties on a subspace level of local partition and simultaneously provides a topological description of the original space.

1.2 Overview

The rest of the paper is organised as follows. Section 2 is a technical description of the linear manifold clustering and its probabilistic model extension. Section 3 provides introductory concepts of a simplicial homology. Section 4 describes a construction of the topological structure of the linear manifold clustering. Section 5 discusses experimental results on synthetic and natural datasets and Section 6 concludes the paper.

2 Linear Manifold Clustering

In this section, we introduce a basic description of a linear manifold cluster and its probabilistic model.

2.1 Linear Manifold Cluster

While most primitive structures are associated with zero-dimensional manifolds, more complex linear structures may be described as non-zero dimensional manifolds.

Definition 1 (Linear manifold). *Λ is an unlimited M-dimensional* linear manifold *in $\mathbb{R}^N$ if and only if for some translation vector $\mathbf{t} \in \mathbb{R}^N$ and a set of orthonormal vectors $\{\mathbf{b}_i\}_{i=1,\dots,M} \in \mathbb{R}^N$,*

$$\Lambda = \{\mathbf{x} \in \mathbb{R}^N | \ \mathbf{x} = \mathbf{t} + \sum_{i=1}^{M} \alpha_i \mathbf{b}_i; \ \alpha_i \in \mathbb{R}; \mathbf{t}, \mathbf{b}_i \in \mathbb{R}^N\} \tag{1}$$

Haralick and Harpaz introduced the linear manifold cluster model that allows a cluster structure to be defined by a non-zero dimensional linear manifold [11].

Definition 2 (Linear manifold cluster model). *Let $C \subseteq \mathcal{D}$ be a cluster of points from the dataset $\mathcal{D} \subset \mathbb{R}^N$. Then we can define an M-dimensional linear manifold cluster on the dataset $\mathcal{D}$ with a "support" linear manifold Λ as*

$$\begin{aligned} C_\Lambda = \{\mathbf{x} \in C \mid \mathbf{x} = \boldsymbol{\mu}_\Lambda + \sum_{i=1}^{M} \alpha_i \mathbf{b}_i + \sum_{j=1}^{N-M} \epsilon_j \mathbf{b}_j^\perp; \ \alpha_i, \epsilon_j \in \mathbb{R}\} = \\ \{\mathbf{x} \in C \mid \mathbf{x} = \mathbf{y} + \sum_{j=1}^{N-M} \epsilon_j \mathbf{b}_j^\perp; \ \mathbf{y} \in \Lambda; \epsilon_j \in \mathbb{R}\} = \\ \{\mathbf{x} \in C \mid \mathbf{x} = \boldsymbol{\mu}_\Lambda + \boldsymbol{\alpha} B + \boldsymbol{\epsilon} B_\perp; \ \boldsymbol{\alpha} \in \mathbb{R}^M, \boldsymbol{\epsilon} \in \mathbb{R}^{N-M}\} \end{aligned} \tag{2}$$

where $\boldsymbol{\mu}_\Lambda \in \mathbb{R}^N$ is a manifold translation vector; B is a matrix whose M columns $\mathbf{b}_i \in \mathbb{R}^N$ are orthonormal vectors that span the linear manifold Λ; $B_\perp$ is a matrix whose $N-M$ columns $\mathbf{b}_j^\perp \in \mathbb{R}^N$ are orthonormal vectors that span orthogonal complement subspace to a linear manifold Λ; $\boldsymbol{\alpha} \in \mathbb{R}^M$ and $\boldsymbol{\epsilon} \in \mathbb{R}^{N-M}$ are vectors whose components are independent random variables which characterize the position of the dataset point relative to the linear manifold, such that $Var(\boldsymbol{\alpha}) \gg Var(\boldsymbol{\epsilon})$.

In the linear manifold (LM) cluster model, see Definition 2, cluster points are positioned on or near the linear manifold. If point $\mathbf{x}$ is located near a linear manifold Λ, then the distance from the point to the manifold Λ, described by the basis B or its orthogonal compliment, is defined as

$$d_\Lambda(\mathbf{x}, B) = \|(I - BB^T)(\mathbf{x} - \boldsymbol{\mu}_\Lambda)\| \tag{3}$$

$$d_\Lambda(\mathbf{x}, B_\perp) = \|B_\perp {B_\perp}^T(\mathbf{x} - \boldsymbol{\mu}_\Lambda)\| \tag{4}$$

Usually, a similarity measure used in description of the cluster is associated with a proper distance measure, i.e. Euclidean distance. However, in geometrical clustering the Euclidean distance is usually used to create spherical clusters. In the k-means clustering, the Euclidian distance-based similarity measure, defined as $d(x, c) < \theta$ where c is the cluster center and is the θ distance threshold, creates an open ball $B_\theta(c)$ around the cluster center which geometrically corresponds to the sphere.

In case of the linear manifold cluster, a spherical cluster structure does not correspond to an actual geometrical shape of the elongated LM cluster. Thus, the Euclidean distance is not a useful measure for describing the linear manifold clusters. We use the above distance (3) to form a similarity measure which provides a better geometrical description of the LM cluster.

Definition 3 (Linear manifold cluster). *Let Λ be an M-dimensional linear manifold spanned by the basis orthonormal vectors $\mathbf{b}_i \in \mathbb{R}^N$, and θ is a distance threshold that separates points from the dataset $\mathcal{D} \subset \mathbb{R}^N$ by proximity to the linear manifold Λ. Then, the* linear manifold cluster *$C_{\Lambda,\theta} \subseteq \mathcal{D}$ is defined as follows*

$$C_{\Lambda,\theta} = \{\mathbf{x} \in \mathcal{D} \mid \mathbf{x} = \boldsymbol{\mu}_\Lambda + \boldsymbol{\alpha} B + \boldsymbol{\epsilon} B_\perp; \ \boldsymbol{\alpha} \in \mathbb{R}^M, \boldsymbol{\epsilon} \in \mathbb{R}^{N-M}, d_\Lambda(\mathbf{x}, B) \leq \theta\} \tag{5}$$

where B is a matrix composed of linear manifold basis vectors, $\{\mathbf{b}_i\}_{i=1}^{M}$, $B_\perp$ is its orthogonal compliment, and $\boldsymbol{\mu}_\Lambda \in \mathbb{R}^N$ is a manifold Λ translation vector.

Definition 4. *A collection of linear manifold clusters form a linear manifold clustering.*

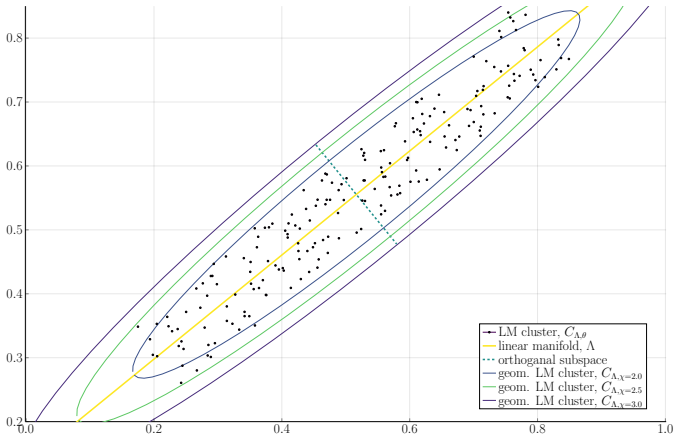


Fig. 1: Linear manifold cluster $C_{\Lambda,\theta}$ (5) with a support manifold Λ (1), and boundaries of geometric linear manifold cluster $C_{\Lambda,\chi}$ (15) created for various threshold χ.

2.2 Model-based Linear Manifold Cluster

Let $\mathcal{D} \subset \mathbb{R}^N$ be our observed N-dimensional dataset that contains multiple non-linear structures. We consider that such non-linear structure would have the form of a piecewise linear manifold, i.e. a collection of linked bounded linear manifolds. Let us consider the unknown distribution $p(\mathbf{x})$ over $\mathbb{R}^N$, which models a linear manifold cluster, from which we i.i.d. sample a set of high-dimensional points $\{\mathbf{x}_n \in \mathcal{D}\}_{n=1}^{I}$, potentially with an uncorrelated Gaussian noise. We assume that latent variables of our model come from a low-dimensional space $\mathcal{Z} \subseteq \mathbb{R}^M$.

A point $\mathbf{z}$ in the latent model space $\mathbb{R}^M$ comes from a prior distribution $p(\mathbf{z})$ and is mapped into $\mathbb{R}^N$ by a non-singular mapping $\mathbf{f} : \mathcal{Z} \to \mathbb{R}^N$. We assume that our latent and observed variables are continuous. Following the manifold hypothesis, data, that is located along a low-dimensional manifold embedded in a high-dimensional space, provides a good assumption about its intrinsic dimensionality and its embedding in full space [20]. In order for an L-dimensional manifold $\mathcal{M} = \mathbf{f}(\mathcal{Z})$ to be restored to the full space, it requires a posterior distribution $p(\mathbf{x}|\mathbf{z})$ defined on $\mathbb{R}^N$ as $p(\mathbf{x}|\mathbf{z}) = p(\mathbf{x}|\mathbf{f}(\mathbf{z}))$.

In a product space $\mathbb{R}^N \times \mathcal{Z}$, we consider a joint distribution $p(\mathbf{x}, \mathbf{z})$ which after marginalization of the latent space variable allows us to find a model of $p(\mathbf{x})$:

$$p(\mathbf{x}) = \int_{\mathcal{Z}} p(\mathbf{x}, \mathbf{z}) d\mathbf{z} = \int_{\mathcal{Z}} p(\mathbf{x}|\mathbf{z}) p(\mathbf{z}) d\mathbf{z} \tag{6}$$

A linear manifold cluster model (2), under the assumption of $\boldsymbol{\alpha}$ and $\boldsymbol{\epsilon}$ being modeled by the Gaussian distribution, can be defined as a generative latent model, in particular, as a factor analysis model [29]:

$$p(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\mathbf{0}, \mathbf{I}) \tag{7}$$

$$p(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{x}|\mathbf{B}\mathbf{z} + \boldsymbol{\mu}, \boldsymbol{\Psi}) \tag{8}$$

where $N \times M$ matrix $\mathbf{B}$ is composed of M orthonormal basis vectors $\mathbf{b}_j \in \mathbb{R}^N$ which span the linear manifold, and $\mathbf{\Psi}$ is a diagonal covariance matrix that captures variance of $\boldsymbol{\alpha}$ and $\boldsymbol{\epsilon}$ random variables in the linear manifold cluster model (2).

A marginal distribution of $p(\mathbf{x})$, that is also Gaussian, can be computed analytically from (6):

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \mathbf{B}\mathbf{B}^T + \mathbf{\Psi}) \tag{9}$$

The posterior over the latent space is also Gaussian:

$$p(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mathbf{z}|\mathbf{A}(\mathbf{x} - \boldsymbol{\mu}), \mathbf{C}^{-1}) \tag{10}$$

where

$$\mathbf{C} = \mathbf{I} + \mathbf{B^T}\mathbf{\Psi}^{-1}\mathbf{B}$$
$$\mathbf{A} = \mathbf{B^T}(\mathbf{B}\mathbf{B^T} + \mathbf{\Psi})^{-1}$$

The model parameters may be determined by a maximum-likelihood estimation iterative procedure because there is no closed-form analytic solution for $\mathbf{B}$ and $\mathbf{\Psi}$. However, the closed-form solution exists for an isotropic variance model, $\mathbf{\Psi} = \sigma^2\mathbf{I}$, which transforms the above model into the probabilistic principal component analysis problem [26].

Harpaz and Haralick [13] proposed a non-parametric density model for linear manifold cluster where the total density estimate for a point $\mathbf{x}$, given that it came from cluster C, is given by

$$p(\mathbf{x}|C) = \left(\prod_{j=1}^{M} h(\mathbf{b}'_j(\mathbf{x} - \boldsymbol{\mu})) \right) h(||(\mathbf{I} - \mathbf{B}\mathbf{B}')(\mathbf{x} - \boldsymbol{\mu})||^2) \tag{11}$$

where $\{\mathbf{b}_i\}_{i=1,\ldots,M} = \mathbf{B}$ are the basis vectors that span the linear manifold cluster C, μ is a translation vector of the cluster C, $h(\mathbf{b}'_j(\mathbf{x} - \boldsymbol{\mu}))$ is a histogram pdf estimate of the projection of points onto the j-th spanning vector $\mathbf{b}_j$ of the "support" linear manifold of the cluster C, $h(||(\mathbf{I} - \mathbf{B}\mathbf{B}')(\mathbf{x} - \boldsymbol{\mu})||^2)$ is the histogram pdf estimate of the distances from the cluster points to the "support" linear manifold of the cluster C, and M is a dimension of the cluster C.

A marginal distribution of $p(\mathbf{x})$ can be computed as a total mixture density estimate of all clusters:

$$p(\mathbf{x}) = \sum_{i=1}^{K} \frac{|C_i|}{\sum_{j=1}^{K} |C_j|} p(\mathbf{x}|C_i) \tag{12}$$

A cluster probabilistic model does not require any geometric similarity measure, but a distance measure is required if we want provide a geometrical interpretation. Such distance measure that preserves probabilistic properties of a cluster point distribution and provides a proper geometric distance measure is the Mahalanobis distance.

Definition 5 (Mahalanobis distance). *Given a multivariate normal distribution D with a mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, the Mahalanobis distance between this distribution mean and a point $\mathbf{x}$ defined as follows:*

$$d_{\boldsymbol{\Sigma}}(\mathbf{x}, \boldsymbol{\mu}) = \sqrt{(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})} \tag{13}$$

We adopt this distance to use with a probabilistic model of the linear manifold cluster C as the geometric similarity measure to define a distance from the center of the cluster to a point $\mathbf{x}$ as follows:

$$d_C(\mathbf{x}) = \sqrt{(\mathbf{x} - \boldsymbol{\mu}_C)^T \boldsymbol{\Sigma}_C^{-1} (\mathbf{x} - \boldsymbol{\mu}_C)} \tag{14}$$

where $\boldsymbol{\Sigma}_C$ is a covariance matrix calculated from the cluster C, and $\boldsymbol{\mu}_C$ is its center.

Given a linear manifold cluster covariance matrix $\boldsymbol{\Sigma}_C$, we define a geometric linear manifold cluster $C_{\Lambda,\chi}$, similarly to (5), using the distance measure (14) as follows:

$$C_{\Lambda,\chi} = \{\mathbf{x} \in \mathcal{D} \mid d_C(\mathbf{x}) \leq \chi\} \tag{15}$$

Choice of the Mahalanobis distance allows us correctly represent an elongated elliptical structure of the linear manifold cluster which is encoded by model's covariance. When the covariance matrix is isotropic, the above geometric cluster model produces spherical clusters similar to the k-means algorithm [3]. Moreover, the squared distance (14) follows chi-squared distribution, thus providing the threshold parameter χ a probabilistic interpretation, a probability of the point being in the cluster.

3 Simplicial Homology

We begin by introducing some concepts of a simplicial homology which are required to specify properties of topological invariants and a topology of linear manifold structures.

Similarly to clustering, topological data analysis provides generalization over point dataset by connecting the neighbouring points into the shapes, simplices. These shapes eventually make up an abstract topological description of an underlying dataset geometry – an abstract simplicial complex.

Definition 6 (Abstract Simplicial Complex [19]). *An abstract simplicial complex is a collection $\mathcal{S}$ of finite nonempty sets, such that if σ is an element of $\mathcal{S}$, so is every nonempty subset of σ, i.e. for any $\sigma \in \mathcal{S}, \sigma' \subseteq \sigma, \sigma' \in \mathcal{S}$.*

The element of σ of $\mathcal{S}$ is called a *simplex* of $\mathcal{S}$; its *dimension* is one less than the number of its elements. The *vertex set* V of $\mathcal{S}$ is the union of the one-point elements of $\mathcal{S}$, such that $v \in V$ is a 0-simplex $v \in \mathcal{S}$. We specify a p-simplex $\sigma = [v_0 v_1 \dots v_p]$ such that vertices listed in some order which is permanently fixed for all vertices. The dimension of $\mathcal{S}$ is the largest dimension of any simplex, or infinite, $\dim \mathcal{S} = \max\{\dim \sigma \mid \sigma \in \mathcal{S}\}$. Each nonempty subset of σ is called a *face* of σ.

Given a linear manifold clustering $\mathcal{C}$, we wish to construct a topological model description of a clustered data which is derived from a cover formed by bounded linear manifold clusters. In general, such a cover is required for the construction of an abstract simplicial complex which provides a topological description of the underlying data [19].

Definition 7 (Cover). *Let X be a topological space, then a* cover *$\mathcal{U}$ of X is a collection of sets whose union contains X as a subset,*

$$\mathcal{U} = \{U_\alpha\}_{\alpha \in A} \text{ such that } X \subseteq \bigcup_{\alpha \in A} U_\alpha \tag{16}$$

where A is an index set.

For a given a dataset cover, its nerve provides a compact combinatorial description of the connectivity relationship between cover sets based on an existence of their non-empty intersections.

Definition 8 (Nerve [5]). *The* nerve *of a cover $\mathcal{U}$, denoted by $N(\mathcal{U})$, is the abstract simplicial complex $\mathcal{S}$ with vertex set V, and where a family $\{v_0, \dots, v_k\}$ spans a k-simplex σ if and only if $U_{v_0} \cap \dots \cap U_{v_k} \neq \emptyset$.*

The Nerve Theorem (1) is a fundamental theorem in algebraic topology relates the topology of the nerve of a cover to the topological space. In order to formally state it, we need to introduce a notion of weak equivalence between topological spaces to provide a spaces' equivalence on a level of topological invariants.

Definition 9 (Homotopy [6]). *Given two maps $f, g : X \to Y$ of topological spaces, f and g are* homotopic*, $f \simeq g$, if there is a continuous map $H : X \times [0,1] \to Y$ so that $H(x,0) = f(x)$ and $H(x,1) = g(x)$ for all $x \in X$.*

The relationship of being homotopic is an equivalence relation.

Definition 10 (Homotopy Equvalance [9]). *Two topological spaces X and Y are* homotopy equivalent *if there are continuous maps $f : X \to Y$ and $g : Y \to X$ such that $g \circ f$ is homotopic to the identity map id_X, $g \circ f \simeq id_X$, and $f \circ g \simeq id_Y$. This gives an equivalence relation to topological spaces, $X \simeq Y$, and we say that they have the same* homotopy type *if they are homotopy equivalent.*

For any good cover, we can define a nerve $N(\mathcal{U})$, which is an abstract simplicial complex, that is homotopy equivalent to the underlying topological space X by the means of following theorem:

Theorem 1 (Nerve Theorem [5]). *Suppose that X and $\mathcal{U}$ are a topological space and its cover, and suppose that the cover consists of open sets and is numerable. Suppose further that for all $\emptyset = S \subseteq A$, we have $\cap_{s \in S} U_s$ being either contractible or empty. Then $N(\mathcal{U})$ is homotopy equivalent to X.*

The construction of the cover defines the structure of the corresponding simplicial complex and provides equivalency to an underlying topological space. From the multitude of cover construction methods, the Čech complex construction provides a proper procedure for a nerve construction of a topological manifold in a metric space X from the family of open balls $\mathcal{B}_\epsilon(X) = \{B_\epsilon(x)\}_{x \in X}$ where a N-dimensional open ball for any value of $\epsilon > 0$ is defined as

$$B_\epsilon^N(x) = \{y \in \mathbb{R}^N \mid d(x,y) < \epsilon\} \tag{17}$$

Definition 11 (Čech Complex). *Let (X,d) be a metric space, S is a finite set of points in X, $S \subset X$. The Čech complex of S at scale $\epsilon > 0$ is the abstract simplicial complex whose p-simplices correspond to a non-empty intersection of $(p+1)$ balls of radius ϵ centered at the $(p+1)$ distinct points of X.*

$$\check{C}_\epsilon(S) = \left\{ \sigma = [x_0 x_1 \dots x_p] \subseteq S \mid \bigcap_{x \in \sigma} B_\epsilon(x) \neq \emptyset \right\} \tag{18}$$

We can view the Čech complex $\check{C}_\epsilon(S)$ as the nerve, see Definition 8, of the collection of balls $\{B_{\epsilon/2}(x)\}_{x \in S}$, thus it has the same homotopy type as the union of these balls, and often has the same homotopy type as X. If ball cells have the same size, the Čech complex is called standard. If they are different, then this complex is defined as a generalized Čech complex [17].

We can simplify construction of the simplicial complex if we check only pairs of ball intersections for n-simplices when $n > 1$. The *Vietoris-Rips complex* $VR_\epsilon(X)$ is the set of simplices $[x_0 x_1 \dots x_p]$ such that $d_X(x_i, x_j) \leq \epsilon$ for all (i,j).

The Vietoris-Rips complex $\mathrm{VR}_\epsilon(S)$ can be viewed as the largest simplicial complex having the same 1-skeleton as Čech complex $\check{C}_\epsilon(S)$.

Lemma 1. *Letting S be a finite set of points in some Euclidean space and $\epsilon \geq 0$, then $VR_\epsilon(S) \subseteq \check{C}_{\sqrt{2}\epsilon}(S)$ [9].*

From an abstract simplicial complex $\mathcal{S}$, we can calculate a topological invariant, a number of k-dimensional holes in this simplicial complex, which is designated as *Betti number*, β_k. In particular, for every topological space X and every non-negative integer k, there is a vector space $H_k(X)$, a homology group, whose dimension is intuitively interpreted as the number of independent k-dimensional cycles in X, which is designated as the k-dimensional Betti number of X.

Definition 12 (Betti number [8]). *The kth Betti number β_k of the simplicial complex $\mathcal{S}$ is the rank of the kth homology group $H_k(\mathcal{S})$.*

The details of the Betti number calculation can be found in any standard text of classical algebraic topology, such as [19].

4 Topology of Linear Manifold Clustering

We propose a construction of a simplicial complex which is based on a cover created from the linear manifold clusters – a linear manifold cluster complex. Many of the construction methods use raw dataset points as base elements for the construction procedure [7, 17, 30]. We argue that a reduced presentation of the dataset and a simplicial complex constructed from it have the same homotopy type as the underlying space from where the dataset is sampled.

There exists only one simplicial complex construction method that uses a reduced data representation – Mapper [22]. This method combines dimensionality reduction and clustering techniques to transform a point dataset in a low-dimensional aggregated representation, i.e. clustering, which are used in a simplicial complex construction. Such low-dimensional data representation is required for simplifying cover construction, which can grow exponentially with the dimension of the reduced space.

Contrary to Mapper, our methods keeps compact and efficient description of geometric properties of the original dataset. The linear manifold cluster model [11] provides a simple analytical as well as probabilistic interpretation which we use to construct an abstract simplicial complex. Moreover, a linear manifold cluster support subspace can be viewed as a tangent space and serve to construct a tangent complex [4].

4.1 Construction of Piecewise Linear Manifold Complex

The main step in any simplicial complex construction procedure is a construction of a cover. Generally, the construction is performed in a metric space with a particular metric d, i.e. standard Čech construction. Often the metric for a cover is selected to be a Euclidean, which results in defining an equiradial hyperspherical cover. However, as it was discussed in section 2.1, usage of the Euclidean distance does not allow to represent correctly the elongated structures, i.e. a linear manifold cluster.

In a piecewise linear manifold (PLM) complex construction method, we use the Mahalanobis distance (13) to define element of the cover, a hyperelliptical open neighborhood, which is derived directly from the cluster description, $E_\epsilon(C_\Lambda) = C_{\Lambda,\epsilon}$, see (15). Thus, we are able to use a clustering $\mathcal{C}$ of the original dataset $\mathcal{D} \subset \mathbb{R}^n$ as a basis of a cover where each element is guided by the parameter ϵ that controls its size. So, for any $\epsilon > 0$, a cover $\mathcal{E}_\epsilon(\mathcal{D})$ is created from a model-based linear manifold clusters of a clustering $\mathcal{C}$, $\mathcal{E}_\epsilon(\mathcal{D}) = \{E_\epsilon(C_\Lambda)\}_{C_\Lambda \in \mathcal{C}(\mathcal{D})}$.

Using the above cluster cover and following the Čech construction (11), we can create a PLM complex which has the same homotopy type as an original dataset $\mathcal{D}$. Moreover, if covariance matrices of clusters are isotropic then the resulted clusters are hyperspherical, as if produced by the k-means algorithm, and the PLM complex becomes a generalized Čech complex [17] with simplices constructed from the intersections of the linear manifold clusters. However, a straightforward application of the Čech construction increasingly complicates a process of simplex discovery because an intersection of the hyperellipses is an increasingly hard problem especially in high dimensional spaces.

To overcome such shortcomings, the piecewise linear manifold complex can be viewed as a version of the witness complex [7] computed in the intrinsic geometry of the dataset $\mathcal{D}$ with a set of landmark points corresponding to the centers of linear manifold clusters.

Definition 13 (Piecewise Linear Manifold Complex). *Let $\mathcal{D}$ be a dataset, and $\mathcal{C}$ be its clustering of size L. Given a distance matrix D of the dimension $L \times N$ calculated using* (14) *between the cluster set $\{C_i \in \mathcal{C}\}_{i \in I}$ and the dataset $\mathcal{D} = \{\mathbf{x}_j\}_{j \in J}$ where $I = \{1, 2, \ldots, L\}$ and $J = \{1, 2, \ldots, N\}$ are index sets. We define a* piecewise linear manifold (PLM) complex *$PL_\epsilon(\mathcal{C}) = \{V, S\}$ with a vertex set $V = I$, and a finite collection of simplices S constructed for some $\epsilon > 0$ as follows:*

- *the edge $\sigma = [ab]$ belongs to S if and only if for $a, b \in V$ there exists a $j \in J$ such that:*

$$\max(D(a, j), D(b, j)) \leq \epsilon$$

– *A p-simplex* $\sigma = [a_1 a_2 \dots a_{p+1}]$ *belongs to* S *if and only if for* $a_1, \dots, a_{p+1} \in V$ *all its edges belong to* S*; or there exists a* $j \in J$ *such that:*

$$\max(D(a_1, j), \dots, D(a_{p+1}, j)) \leq \epsilon$$

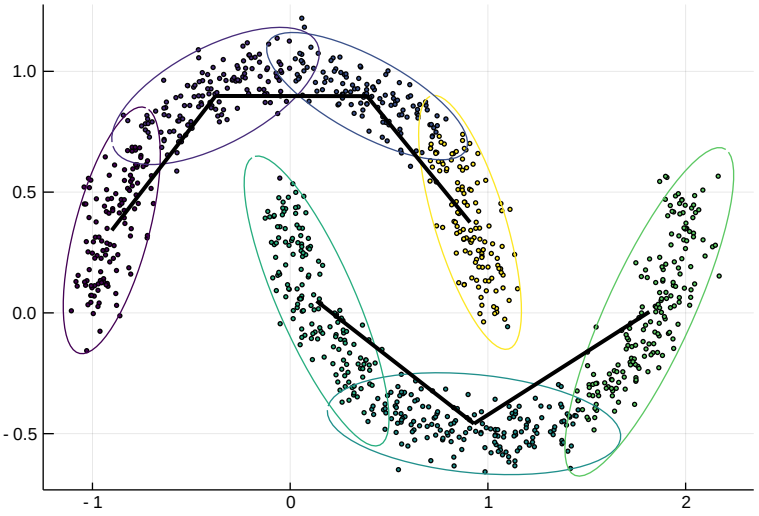


Fig. 2: A piecewise linear manifold complex $PL_\epsilon(\mathcal{C})$ (black) constructed from the clustering $\mathcal{C}$ (colors) of the dataset $\mathcal{D}$ (points).

In [7], it was suggested that the landmark points for a witness complex, initialized based on clustering, poorly reflect the underlying space properties due to the variation of sample density. Even though it might be true for k-means clustering because it uses a similarity measure that only accounts for geometrical position of the points, we believe that usage of the linear manifold clusters, which have explicit probabilistic interpretation, for the landmarks is appropriate. A linear manifold cluster provides a truthful representation of the part of the dataset in the vicinity of the cluster center and can be viewed as an approximation of the tangent space in this point. In its turn, the local tangent space provides a low-dimensional linear approximation of the local geometric structure of the nonlinear smooth manifold from which the dataset was sampled [28].

5 Results

5.1 Experimental Protocol

We performed the series of experiments to show that the constructed piecewise linear manifold (PLM) complexes exhibit the same topological properties as the complexes created by other construction methods: Vietoris-Rips [30] and witness [7]. We used the synthetic and real datasets with known topological properties to compare the construction results.

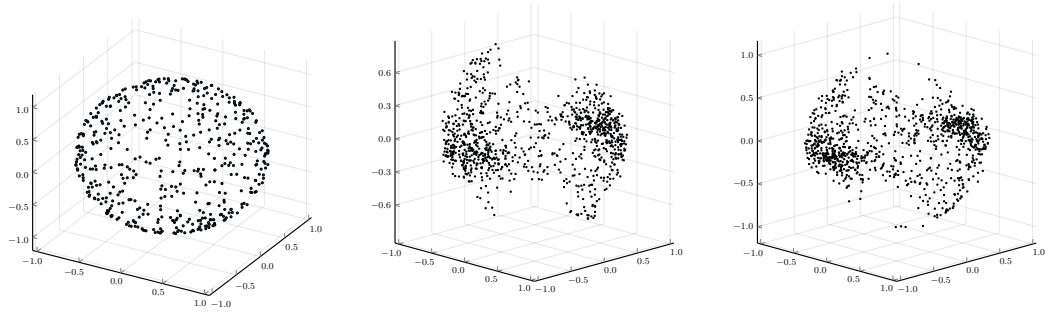
Following the experimental procedure in [7], we evaluated how well a PLM construction algorithm captures the topological properties of the underlying datasets. The correctness of the constructed simplicial complex $\mathcal{S}$, produced by the PLM construction, is evaluated by computing the Betti numbers β_i of the complex which is a standard procedure in algebraic topology. The resulting set of Betti numbers was compared with the known Betti numbers profile, $(\beta_0, \beta_1, \beta_2)$, of the topological space from where the dataset is sampled. In addition, we measured a relative dominance of the topological

profile which compares the length of the persistence of the profile interval with the interval when the complex becomes one connected component, $\beta_0 = 1$.

In addition to the linear manifold clustering (LMCLUS) method for dataset partitioning, we used the k-means clustering algorithm to generate a distance matrix for a PLM complex, see Definition 13, applying (14) to spherical k-means clusters. We did this to evaluate stability of the PLM complex construction, as LMCLUS algorithm has larger variability of produced clusterings.

5.2 Results

"Two Moons" In the first expersment, we used "Two Moons" dataset, see Figure 3d, which is synthetic dataset composed of 1000 2D points that are divided into two non-linear shapes, the two interleaving half circles, of equal size. This dataset can be used to provide a test for binary classification as well as to structure detection or clustering. This dataset has following Betti profile $(2, 0, 0)$, which corresponds to two connected components represented by half circled shapes.



(a) Sphere, $(\beta_0, \beta_1, \beta_2) = (1, 0, 1)$

(b) OIP, $K = 300$, $(\beta_0, \beta_1, \beta_2) = (1, 1, 0)$

(c) OIP, $K = 15$, $(\beta_0, \beta_1, \beta_2) = (1, 5, 0)$

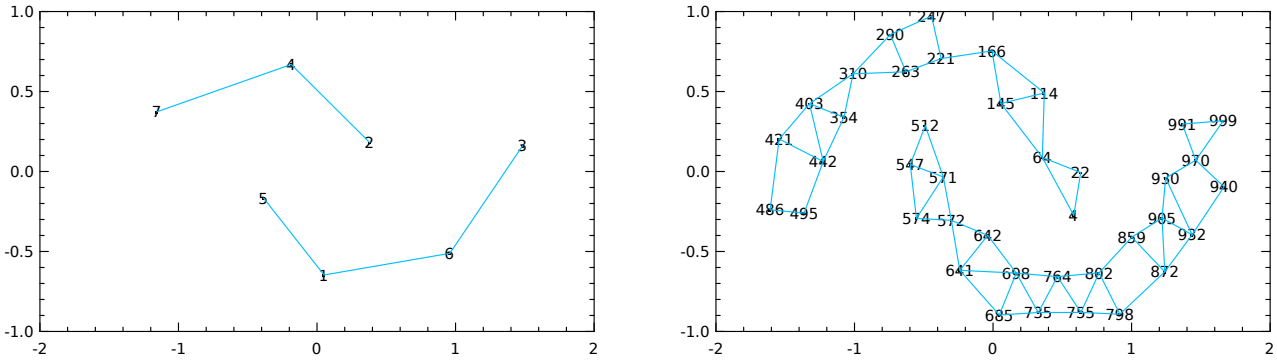
(d) Two Moons, $(\beta_0, \beta_1, \beta_2) = (2, 0, 0)$

(e) Circles, $(\beta_0, \beta_1, \beta_2) = (2, 2, 0)$

Fig. 3: Experimental datasets with corresponding Betti profiles

We generated a 1000 2D point "Two Moons" dataset with an added Gaussian noise with standard deviation 0.07 see Figure 3d. This dataset was clustered by the original LMCLUS algorithm [11] and the resulting clustering refined to produce bounded linear manifold clusters with the best clustering selected by a linear manifold minimum description length score [12]. For clustering with bounded LMCLUS algorithm, we used the following settings: *best_bound = 0.25*, *sampling_factor = 0.3*, *min_cluster_size = 20*. The rest of the setting were set to default values. We also performed the partitioning by the k-means algorithm with $k = 12$ which created a spherical clustering [1].

Using the piecewise linear manifold construction procedure described in Section 4, we generated a PLM complex, see Definition 13, from the intersections of the boundaries of the linear manifold



(a) PLM+LMCLUS, $\chi = 0.11$

(b) Witness, $l = 40, r = 0.2$

Fig. 4: Simplicial complex for "Two Moons" dataset, see figure 3d, created using the piecewise linear manifold construction, a, with Mahalanobis distance $\chi = 0.11$ in distance space $\mathbb{D}$, and using the witness complex construction [7], b, with 40 maxmin landmarks and landmark ball radius $r = 0.2$.

clusters. We used a sequence of values to threshold the Mahalanobis distance for each cluster to receive a hyperelliptical neighborhood around the center of each cluster for construction of a filtration complex F from the PLM complex. The filtration complex F was used to determine a filtration value, a boundary radius, when a particular Betti profile appears to evaluate the relative dominance of such profile. Similar procedure was done with k-mean clusters to generate the corresponding PLC complex. We also repeated same filtration generation procedure for other constructions: Vietoris-Rips [30] and witness [7]. We used different a filtration parameters for different construction procedures: for the Vietoris-Rips construction – a distance between dataset points, for the witness construction – a landmark radius between specified number of the landmarks ($l = 40$). We pervormed construction procedures 100 times, and evaluated the median values of a relative dominance and a complex cell number. Table 1 shows the results of these trials, as well as a percentage of successfully recovered Betti profiles for a particular dataset, reported in "% success".

Figure 4 shows an example of the complex produced by the PLM construction (4a) and the witness construction (4b). Figure 5a shows a persistence barcode of the PLM complex filtration. This barcode shows of two connected components, homology group H_0, for boundary radius $\chi \in [0.11, 0.24)$. The homology group H_1, that corresponds to 1D topological hole or circle, also appears on the barcode which can be explained by overlapping cluster boundaries from the dataset top and bottom half when boundary radii are large enough, $\chi \geq 2.2$. For the witness complex construction with 40 landmarks, see Figure 5b, we observe the similar homological groups.

Circles This synthetic dataset, see Figure 3e, composed of 1000 2D points that are divided into two concentric circles of equal size. This dataset has Betti profile $(2, 2, 0)$, which corresponds to two connected components and two 1-dimensional holes. We used the similar experimental protocol and parameters as for "Two Moons" dataset. The experimental results are presented in Table 1. We were not able to acquire results for the witness construction due to suboptimal designation of landmark points in the dataset.

Sphere This synthetic dataset, see Figure 3a, composed of 500 3D points sampled from $\mathbb{S}^2$. This dataset has Betti profile $(1, 0, 1)$, which corresponds to a connected component and one 2-dimensional hole. We used the similar experimental protocol and parameters as for "Two Moons" dataset. The experimental results are presented in Table 1.

Dataset	Construction			
	Vietoris-Rips	Witness	PLM + LMCLUS	PLM + k-means
Two Moons				
% success	100.0	98.0	100.0	100.0
median relative dominance	0.63	1.0	0.06	0.44
median number of cells	10948	187	43	22
Sphere				
% success	100.0	100.0	100.0	100.0
median relative dominance	0.03	0.98	0.65	0.73
median number of cells	32710	62	52	62
Circles				
% success	100.0	0.0	99.0	42.0
median relative dominance	0.17	0.0	0.04	0.21
median number of cells	11934	0	48	24
OIP300				
% success	-	100.0	84.0	100.0
median relative dominance	-	1.0	0.5	0.38
median number of cells	-	612	43	54
OIP15				
% success	-	100.0	47.0	28.0
median relative dominance	-	1.0	0.57	0.52
median number of cells	-	420	85	60

Table 1: Recovering the homology profile of various datasets using 4 different constructions

Optical image patches The optical image patches (OIP) dataset is a large collection of high-contrast 3×3 optical image patches which after normalization are represented by points on the unit sphere in $\mathbb{R}^8$ [18]. For our experiments, we used a sample of 30% densest vectors from OIP dataset based the density estimator $\rho_K(x) = |x - x_K|$ where x_K is the K-th nearest neighbor of x for some K. We used two samples of 15×10^3 points for $K = 15$ and $K = 300$ provided by JavaPlex library tutorial [25].

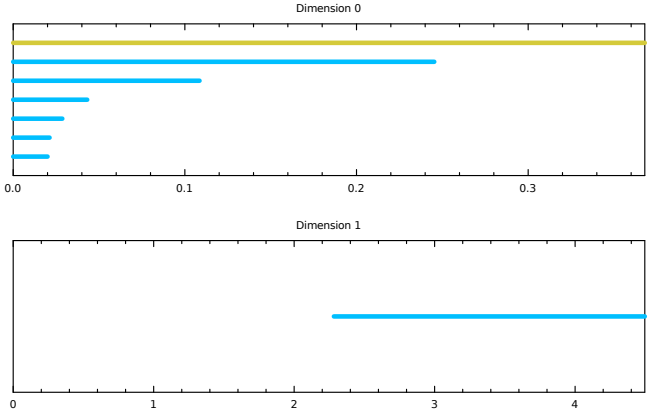
For the sample from the estimator ρ_{300}, see Figure 3b, the Betti profile is (1,1,0). We clustered this dataset by the LMCLUS algorithm and generated bounded linear manifold clusters using the following parameters: *best_bound = 0.2*, *sampling_factor = 0.3*, *min_cluster_size = 50*. The above experimental protocol was used for the sample from the estimator ρ_{15}, which has the Betti profile (1,5,0). Following parameters for LMCLUS algorithm were used: *best_bound = 0.3*, *sampling_factor = 0.1*, *min_cluster_size = 30*. The maximal filtration value was set to 1.0. We also used the witness construction algorithm with 50 landmarks to compare correctness of the results reported in [18] and complexity of simplicial complexes constructed by our algorithm.

The experimental results presented in Table 1 show that the PLM construction algorithm was able to create simplicial complexes with smaller number of cells that exhibit the same topological properties as the witness construction method, see Figure 5.

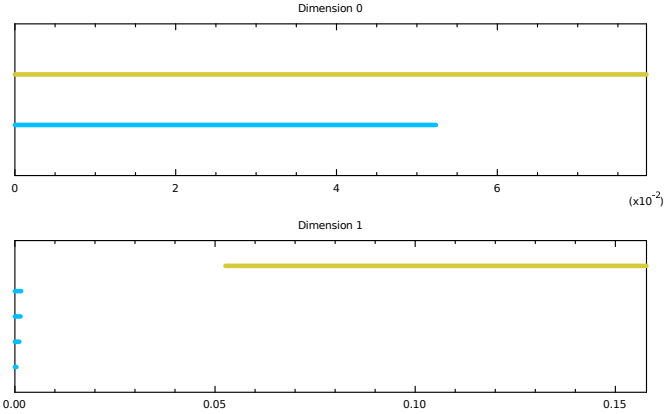
6 Conclusion

We described a novel simplicial complex construction technique based on the linear manifold clustering which provides a comprehensive geometric and topological structural descriptions.

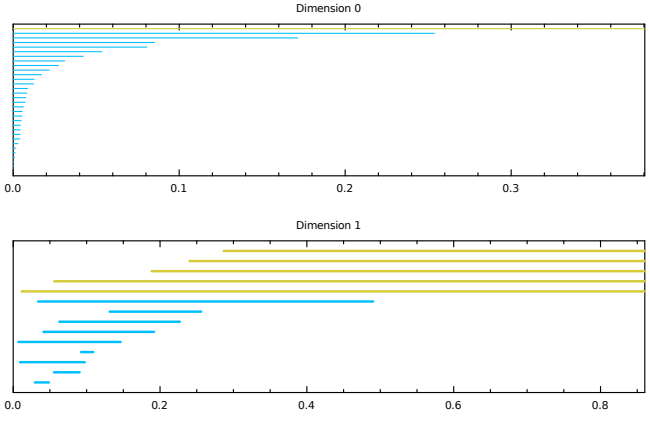
We confirmed that the PLM complex construction method produces reasonable results for various datasets, and showed that such construction method generates more compact, informative and efficient simplicial complexes in comparison to other methods while retaining all topological invariant data.



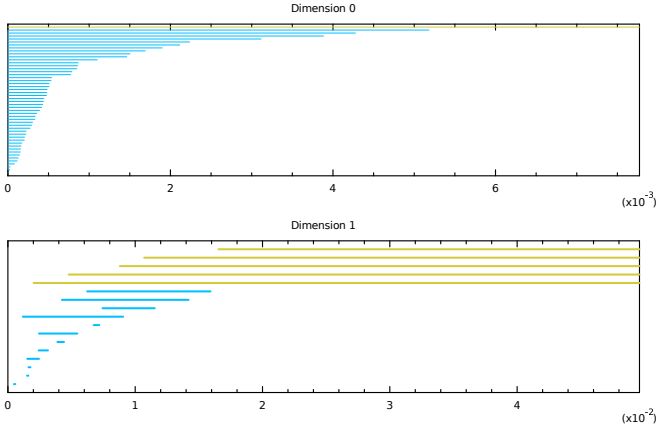
(a) TM, PLMC



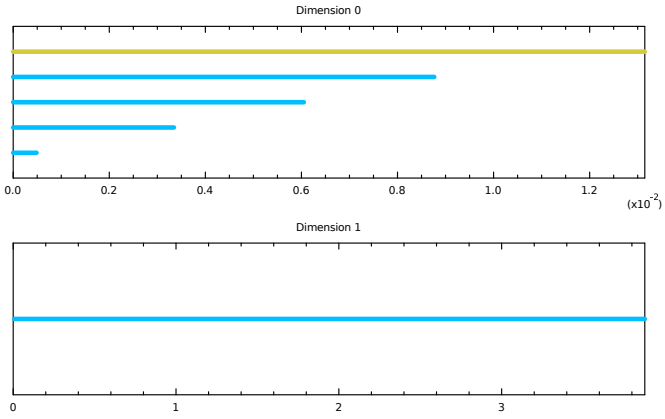
(b) TM, WTC($l = 40$)



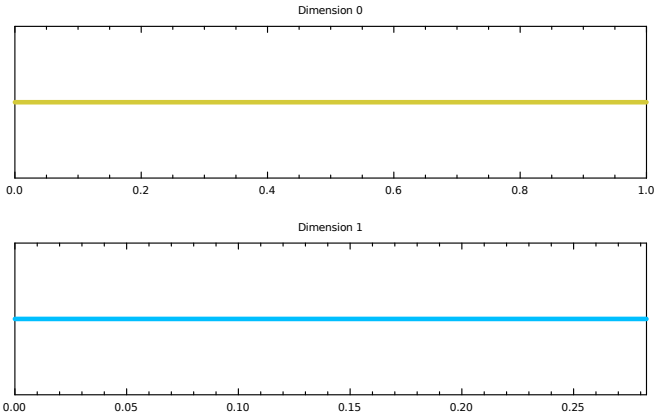
(c) OIP, $K = 15$, PLMC



(d) OIP, $K = 15$, WTC($l = 50$)



(e) OIP, $K = 300$, PLMC



(f) OIP, $K = 300$, WTC($l = 50$)

Fig. 5: Persistence barcodes of datasets "Two Moons" (a, b) and "Optical Image Patches" (c, d, e, f) from the filtered simplicial complexes created by the piecewise linear manifold construction (a, c, e) and the witness constructions (b, d, f). *Note: Yellow color identifies right semi-infinite interval.*

References

[1] Aggarwal, C.C., Reddy, C.K.: Data clustering: algorithms and applications. CRC press (2013)

[2] Bartholomew, D.J., Knott, M., Moustaki, I.: Latent variable models and factor analysis: A unified approach, vol. 904. John Wiley & Sons (2011)

[3] Bishop, C.M.: Pattern Recognition and Machine Learning. Springer (2006)

[4] Carlsson, E., Carlsson, G., De Silva, V.: An algebraic topological method for feature identification. International Journal of Computational Geometry & Applications **16**(04), 291–314 (2006)

[5] Carlsson, G.: Topology and data. Bulletin of the American Mathematical Society **46**(2), 255–308 (2009)

[6] Carlsson, G.: Topological pattern recognition for point cloud data. Acta Numerica **23**, 289–368 (2014)

[7] De Silva, V., Carlsson, G.E.: Topological estimation using witness complexes. SPBG **4**, 157–166 (2004)

[8] Edelsbrunner, H.: A Short Course in Computational Geometry and Topology. Springer (2014)

[9] Edelsbrunner, H., Harer, J.: Computational topology: an introduction. American Mathematical Soc. (2010)

[10] Goh, A., Vidal, R.: Clustering and dimensionality reduction on riemannian manifolds. In: Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on. pp. 1–7. IEEE (2008)

[11] Haralick, R., Harpaz, R.: Linear manifold clustering in high dimensional spaces by stochastic search. Pattern recognition **40**(10), 2672–2684 (2007)

[12] Haralick, R.M., Diky, A., Su, X., Kiang, N.Y.: Inexact MDL for linear manifold clusters. In: 23rd International Conference on Pattern Recognition, ICPR 2016, Cancún, Mexico, December 4-8, 2016. pp. 1345–1351 (2016). https://doi.org/10.1109/ICPR.2016.7899824, `http://dx.doi.org/10.1109/ICPR.2016.7899824`

[13] Harpaz, R., Haralick, R.: Modeling high-dimensional probability distributions via linear manifold clusters (2007)

[14] Hinton, G.E., Dayan, P., Revow, M.: Modeling the manifolds of images of handwritten digits. IEEE transactions on Neural Networks **8**(1), 65–74 (1997)

[15] Kak, A.: Clustering Data That Resides on a Low-Dimensional Manifold in a High-Dimensional Measurement Space. Purdue University (February 2016)

[16] Kriegel, H., Kröger, P., Zimek, A.: Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. ACM Transactions on Knowledge Discovery from Data **3**(1), 1 (2009)

[17] Le, N.K., Martins, P., Decreusefond, L., Vergne, A.: Construction of the generalized cech complex (2014)

[18] Lee, A.B., Pedersen, K.S., Mumford, D.: The nonlinear statistics of high-contrast patches in natural images. International Journal of Computer Vision **54**(1), 83–103 (Aug 2003). https://doi.org/10.1023/A:1023705401078, `https://doi.org/10.1023/A:1023705401078`

[19] Munkres, J.R.: Elements of algebraic topology, vol. 2. Addison-Wesley (1984)

[20] Niyogi, P., Smale, S., Weinberger, S.: A topological view of unsupervised learning from noisy data. SIAM J. Comput. **40**, 646–663 (2011)

[21] Peng, X., Zhang, L., Yi, Z.: Scalable sparse subspace clustering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 430–437 (2013)

[22] Singh, G., Mémoli, F., Carlsson, G.E.: Topological methods for the analysis of high dimensional data sets and 3d object recognition. In: SPBG. pp. 91–100 (2007)

[23] Souvenir, R., Pless, R.: Manifold clustering. In: Tenth IEEE International Conference on Computer Vision. vol. 1, pp. 648–653 (2005)

[24] Subbarao, R., Meer, P.: Nonlinear mean shift for clustering over analytic manifolds. In: Computer Vision and Pattern Recognition. vol. 1, pp. 1168–1175. IEEE (2006)
[25] Tausz, A., Vejdemo-Johansson, M., Adams, H.: JavaPlex: A research software package for persistent (co)homology. In: Hong, H., Yap, C. (eds.) Proceedings of ICMS 2014. pp. 129–136. Lecture Notes in Computer Science 8592 (2014), software available at `http://appliedtopology.github.io/javaplex/`
[26] Tipping, M.E., Bishop, C.M.: Probabilistic principal component analysis. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **61**(3), 611–622 (1999)
[27] Wang, Y., Jiang, Y., Wu, Y., Zhou, Z.: Spectral clustering on multiple manifolds. Neural Networks, IEEE Transactions on **22**(7), 1149–1161 (2011)
[28] Zhang, Z., Zha, H.: Principal manifolds and nonlinear dimensionality reduction via tangent space alignment. SIAM Journal on Scientific Computing pp. 313–338 (2004)
[29] Zhao, J.H., Yu, P.L., Jiang, Q.: Ml estimation for factor analysis: Em or non-em? Statistics and computing **18**(2), 109–123 (2008)
[30] Zomorodian, A.: Fast construction of the vietoris-rips complex. Computers & Graphics **34**(3), 263–271 (2010)