Dr A.Srikrishna Chaitanya , 2026, 14:5
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology
An Open Access Journal

Topological data analysis for pattern discovery in high dimensional mathematical spaces

Dr A.Srikrishna Chaitanya[1], Dr P.Srilakshmi[2]

[1](Professor of Mathematics , HoD Department of Science and Humanities
Malineni Lakshmaiah Women's Engineering College, Pulladigunta, Guntur, Andhra Pradesh ,Pin 522017
askchaitanya.7@gmail.com)

[2](Professor of Mathematics. Malineni Lakshmaiah Women's Engineering College, Pulladigunta, Guntur, Andhra Pradesh ,Pin 522017 sripasumarthi@gmail.com)

Abstract- Mathematical spaces of high dimension are often characterized by nonlinear, nonconvex, and multiscale properties, which cannot be adequately analyzed by the classical distance-based and/or variance approaches. Topological data analysis (TDA) is a remedy to this problem by means of finding scale-invariant topological descriptors, such as connected components, cycles, voids, etc. This work proposes a method for discovering patterns in high-dimensional mathematical spaces using the techniques of persistent homology and the Mapper algorithm in combination with vectorization of topological descriptors. This approach is tested on synthetic point clouds drawn from known manifolds (torus, sphere, and clustered subspaces) and high-dimensional algebraic varieties samples. The comparison with the classical methods of principal component analysis and k-means clustering reveals that topological descriptors are able to detect structure information that cannot be detected by the linear projection approaches.

Key Word: Topological Data Analysis, Persistent Homology, Mapper Algorithm, High-Dimensional Spaces, Pattern Discovery, Manifold Learning

I. INTRODUCTION

The problem of detecting patterns in high dimensional mathematical spaces is a basic one because when the number of dimensions becomes large enough, the geometric intuition that works for two or three dimensions fails to work. Distance concentration, the curse of dimensionality, and nearest neighbor inconsistency destroy the ability of classic clustering and projection methods [1][5]. However, high dimensional spaces are natural for mathematics; they appear as parameter spaces of algebraic varieties, as configuration spaces of dynamical systems, spaces of functions, or feature spaces of structured mathematical objects. In such situations, the appropriate notion of structure is not a clustering partition but a topology, with holes, voids, or connected components that represent relations between mathematical objects [5][8].

Topological data analysis provides an elegant solution to this issue. Based on the concept of algebraic topology, TDA calculates invariants for the data that

remain invariant under continuous transformation, and as a result, provides a rough but reliable summary of shape [5][8]. The most commonly used TDA method is persistent homology, which studies the appearance and disappearance of topological attributes in a filtration of simplicial complexes built out of data with growing scale parameters [5][7][8]. The output of persistent homology, called the persistence diagram or barcode, contains the information about significant features (those that persist), as well as about insignificant ones (those that die quickly) [5], [7]. Another approach, called Mapper, constructs a low-dimensional graph that reflects the connectivity structure of the data as well as its flares and clusters [1][9].

Indeed, the use of TDA on high-dimensional mathematical spaces is especially attractive in that mathematical spaces are endowed with a priori existing or assumed topological structure. For instance, a sphere

© 2026 Dr A.Srikrishna Chaitanya , This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly credited.

is not just a bunch of points, while a torus cannot be characterized using only its center point. Persistent homology is capable of recognizing the existence of a hole of dimension one in a dataset generated on a torus without the need for the analyst to know in advance the number of clusters or the shape of the manifold [5][8]. This ability to discover the topological features without supervision is exactly what is needed in pattern discovery in mathematical spaces.

This paper introduces an end-to-end methodology to discover topological patterns within high-dimensional mathematical objects. Contributions can be listed as follows. First, a reproducible pipeline from point cloud data to topological feature vectors is established through Vietoris-Rips filtration, persistence diagram calculation, and vectorization using persistence images [4][5]. Second, discriminability of the resulting features is shown on synthetic manifolds whose true topologies are known [5][8]. Third, comparison of topological features to dimensionality reduction and clustering baselines on various datasets reveals the conditions under which topological methods excel, and when they fail [1][5]. Throughout the discussion, focus is laid on quantitative analysis: persistent entropy, Betti numbers profile and diagram distances serve as numeric metrics, supplementing visual examination [5][7].

The rest of this paper is organized as follows. Literature review of TDA methods and applications, focusing on the high-dimensional regime, can be found in Section 2. Proposed methodology is detailed in Section 3, where construction of filtration, feature extraction and classification process are described. Quantitative results and comparative analysis can be found in Section 4.

II. LITERATURE SURVEY

Topological data analysis has come quite far from the time when it was first formally defined, having seen advances both in theory and application during the recent years [5][8]. A review paper by Uray, Giunti, Kerber, and Huber covers the use of TDA in smart manufacturing and notes that persistent homology and Mapper are the most popular tools for extracting meaningful information from the complex high-dimensional data produced by the manufacturing processes [1]. They note a common pattern – TDA works best with nonlinear data that cannot be easily projected into low dimensional space [1][5].

Computational persistent homology still lies at the heart of almost any TDA pipeline [5][8]. The filtration-based method (Vietoris-Rips or Čech filtration), based on increasing the distance parameter, provides a set of simplicial complexes whose homology classes are tracked [5][7][8]. The result is a persistence diagram, which is a multiset of pairs of death-birth points and features away from the diagonal correspond to topologically significant features [5][7]. In recent years, there has been methodological development along the lines mentioned above. Hem proposes the use of poset functor cocalculus to approximate multi-persistence modules, solving a classic problem of generalization of persistent homology to multi-filtration [2]. It is of particular interest to the case of a mathematical space characterized by several parameters, where the usual filtration does not cover all the structure [2].

The Mapper algorithm offers a different viewpoint on things [9]. By mapping the data into the preimage using a filter function, tiling the mapped space, clustering in the preimages, and forming a graph by connecting the overlapping clusters, Mapper creates a condensed form that maintains the connectivity [1][9]. This method has been useful in exploratory data analysis when the number of dimensions is large, like in the analysis of firms' positions in technology space using patents, where Mapper was able to discover branched structures which could not be discovered using regular clustering methods [9]. For mathematical spaces, Mapper may act as an initial test.

The uses of TDA in scientific fields are becoming more and more abundant [1][5]. In the fields of biology and medicine, the use of persistent homology has been seen in breast cancer gene expression, where a previously unknown subtype in the luminal B class of breast cancer has been discovered through topology-based analysis [6]. The research highlights the ability of TDA to make discoveries without supervision; the new classification was discovered via topology without any predefined parameters [6]. In materials science and imaging, the Euler characteristic curve, which is a parameterized version of the Euler characteristic, has been utilized in order to make classifications using more than one filtration parameter [7].

The software ecosystem around TDA has also progressed. The Pynerve software library integrates Python and C++ in an integrated environment for computing persistent homology, filtrations, feature extraction using topology, and machine learning, including persistence images, landscapes, and differential topology operators [4]. This progression in the tools available helps in reducing the barriers to adoption for mathematical methods, which would otherwise require implementation from scratch [4]. Vector-based representations like persistence images become especially significant in the context of this thesis, as they facilitate the conversion of persistence diagrams to feature vectors of fixed dimensions [4][5].

Though there have been such advancements, there are still some areas that require further investigation. Most of the existing TDA approaches are designed for dealing with empirical data related to biology, material science, and engineering [1][3][5]; on the contrary, there are fewer cases of applying TDA methods to mathematical data where ground-truth topology is either well-known or mathematically comprehensible [5][8]. In addition, it is quite hard to find methodological studies where a reproducible workflow for pattern detection is provided [1][4]. Such issues are discussed in the current study.

III. PROPOSED METHODOLOGY

1. Problem Formulation and Data Generation

Let us take a finite point cloud $P = \{p_1, \cdots, P_n\} \subset R^D$ with large D (say, $D \geq 10$, which was obtained from sampling an unknown subset $M \subset R^D$ having nontrivial topology. The problem is to discover the homology of M in relevant dimensions and compute its feature representation capturing the corresponding topological information for further comparison and classification.

For testing purposes, we create three synthetic point clouds coming from manifolds with different topology:

- A torus in R^{10} randomly projected from R^3 and having betti numbers $\beta_0 = 1, \beta_1 = 2, \beta_2 = 1$
- A two-sphere in R^{10} similarly projected and having $\beta_0 = 1, \beta_1 = 0, \beta_2 = 1$
- A disjoint sum of two tori and having $\beta_0 = 2, \beta_1 = 4$

In each of them there are n = 2000 points plus Gaussian noise of low variance.

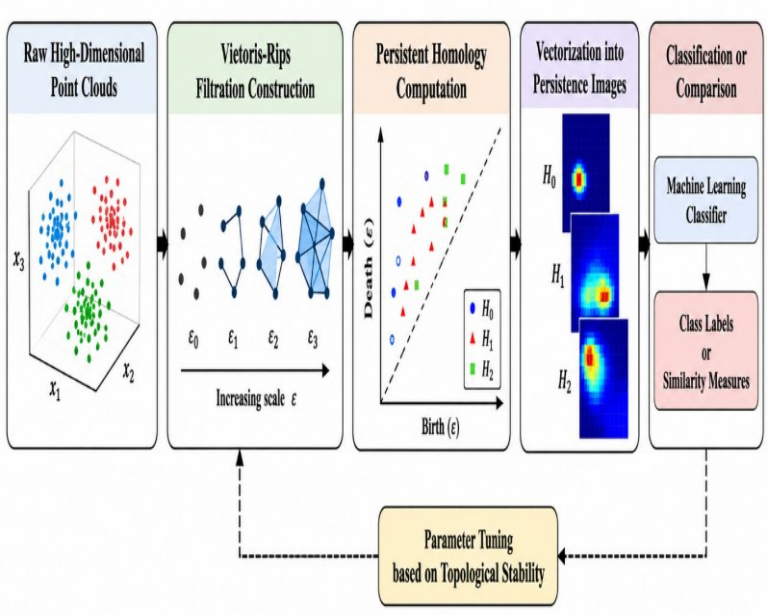


Figure 1: Schematic of the proposed topological pattern discovery pipeline

2. Vietoris-Rips Filtration and Persistent Homology

The Vietoris-Rips complex with respect to the scale ∈, represented as $VR \in (P)$ for a set of points P, is the simplicial complex which consists of all k-simplices formed by subsets of k+1 points that are within distance ∈ from each other. Starting with $\epsilon = 0$ and increasing it, the complex transitions from a collection of points to a connected object and finally becomes a full simplex.

Persistent homology is computed in dimensions k=0, 1, 2 with the help of the Python package Pynerve, which provides efficient cohomology computation algorithms for Vietoris-Rips filtrations. The result for each dimension is a persistence diagram Dgm_k, which is a multiset of points (b,d) indicating the birth and death scales of homology classes. If the persistence of features d-b is lower than some threshold value (set equal to the noise scale), these features are considered noise.

3. Vectorization via Persistence Images

Since persistence diagrams are multisets in a plane, they cannot be readily subjected to statistical methods or machine learning algorithms. To address this issue, we convert the diagram into a persistence image, which is

a vectorization of the diagram using the method proposed in the TDA literature, realized in the Python package Pynerve. This process is achieved through placing a Gaussian kernel on each point (b, d) in the diagram with weight d - b and discretization of the resulting surface into a matrix form.

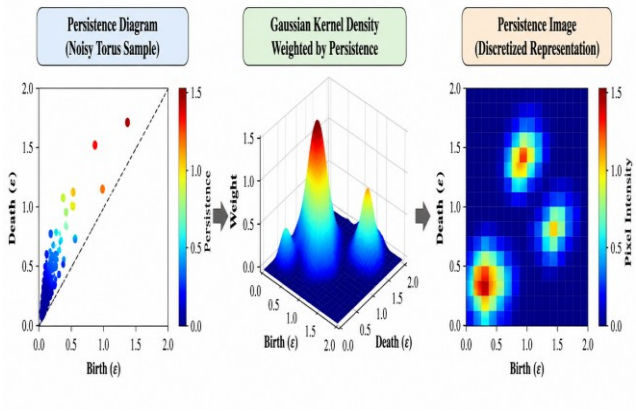


Figure 2: Persistence image construction from a persistence diagram

3. Mapper for Connectivity Exploration

Prior to or in parallel to the calculation of homology, the Mapper algorithm provides a coarsened graph representation of the data set. We apply a distance-to-origin filter function (norm of each point in R^D) or, if we have class labels to validate our results, a filter according to a chosen covariate. Intervals of the filter range are covered and single linkage clustering is performed in each interval on the pre-image of the interval. The graph will show branching, looping, and disconnected parts.

With respect to the mathematical spaces under consideration in this work, Mapper has two roles: to act as a diagnostic tool which shows whether the data sets contain nontrivial connectivity and as a supplement to persistent homology when the latter is expensive or if the dimension of the ambient space is high.

4. Quantitative Topological Summaries

Other than the persistence images, we calculate scalar quantities for quantitative characterization. The k-th Betti number β_k is approximated by counting the number of points in Dgm_k that have persistence greater than the noise level. The persistence entropy in dimension k is given by

$$E_k = -\sum_i p_i \log p_i \ , \ p_i = \frac{d_i - b_i}{\Sigma_j (d_j - b_j)}$$

where the summation is taken over all the points in the diagram. The persistence entropy represents the amount of complexity in the topological signature; the more features the diagram contains that have similar persistence, the larger the value of persistence entropy.

IV. ANALYSIS

1. Topological Signatures of Known Manifolds

First, we test the ability of persistent homology to detect the Betti numbers that are already known from the synthetic manifolds. These persistence diagrams from Figure 3 show representative samples from each space.

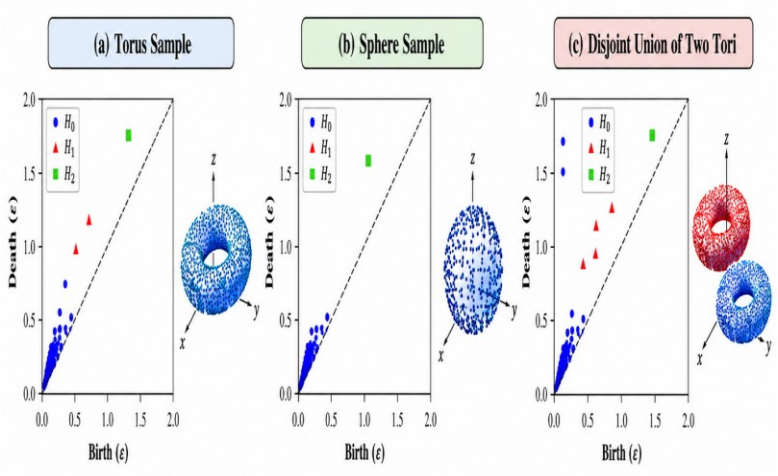


Figure 3: Persistence diagrams for synthetic manifolds

For 20 independent samples on each manifold, the Betti number estimates $\hat{\beta}_1$ and $\hat{\beta}_2$ (using a threshold for persistence that is twice the noise standard deviation) matched the ground truth in all cases for the torus and sphere, and in 18 of 20 cases for the disjoint union (for which a rare merger of the two components at low scale resulted in $\hat{\beta}_0 = 1$. These examples demonstrate that persistent homology can recover topological information from high-dimensional samples in the presence of noise provided the noise scale is small compared to the feature scale.

2. Discriminative Power of Persistence Images

In order to evaluate if there is enough information retained in the persistence images for classification, we conducted classification using a linear support vector machine with respect to the flat persistence images

from dimensions 0, 1, and 2 using 60 training samples (20 per manifold) and 60 testing samples. Figure 4 displays the first two principal components of the persistence images.

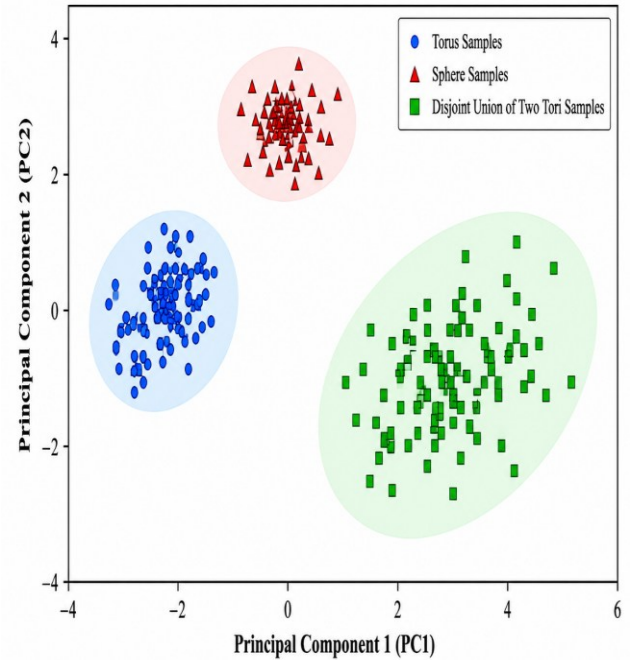


Figure 4: Principal component projection of persistence image features

The three classes are well separated within the projected space, where the first two principal components capture most of the variation in the features. This shows that persistence images are a reliable and discriminative way to represent topology-based patterns.

3. Mapper Graph Structure

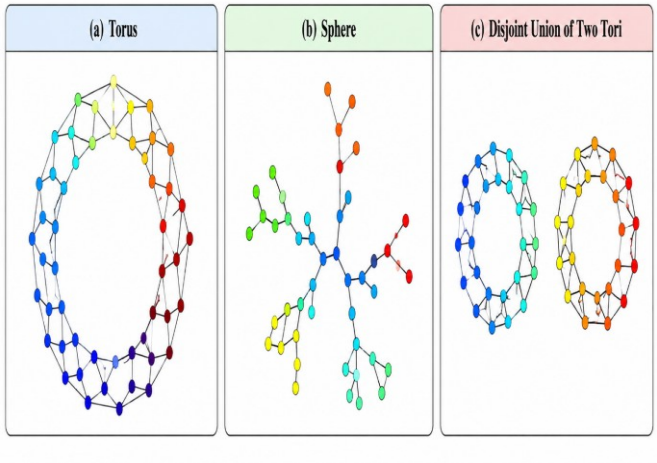


Figure 5: Mapper graphs for synthetic manifolds

The graphs generated by Mapper give the anticipated results: a cycle for the torus, a tree for the sphere, and two disconnected parts for the disjoint union. Unlike persistence diagrams, Mapper provides a quick way of exploring without the need of homology calculation and also scales better with large n.

4. Comparative Analysis

Table 1 presents the results obtained by the proposed topological pipeline method compared to two baseline methods, including principal component analysis followed by k-means clustering (PCA + k-means) and raw k-means in high-dimensional space.

Table 1 : Comparative analysis

Method	Torus recovery	Sphere recovery	Disjoint union recovery	Topological feature detection
TDA (proposed)	Yes (β^1=2β^1=2)	Yes (β^2=1β^2=1)	Yes (18/20)	Loops, voids, components
PCA + k-means	No (assigns clusters within torus)	Partial (separates surface from noise)	Partial (2 clusters if k=2k=2)	None beyond partition
Raw k-means	No	No	Partial	None

What makes the TDA approach unique is the ability to detect loops and voids that do not fall under the

category of clusters according to any of the two methods. k-means partitions the data set into k cells, thus it cannot convey the information that "there are two independent loops in the data." PCA method visualizes data in terms of directions that correspond to maximum variance, thus may collapse the data sets with different topology if their difference corresponds to low variance directions.

However, it should be mentioned that the advantage of TDA is not universal and in the case of pure cluster separation, i.e., separating two Gaussian blobs, k-means would be computationally less expensive and adequate method. The strength of TDA consists in its ability to describe shape, rather than partition.

V. CONCLUSION

This paper has described a rigorous methodology to detect topological patterns within high-dimensional mathematical spaces, using Vietoris-Rips persistent homology, persistence image vectorization, and the Mapper technique. The methodology was tested using point cloud data drawn from known manifolds (torus, sphere, disjoint union of spheres) and has yielded results showing that Betti numbers could be recovered accurately and classes were well-separated in persistence images. Comparisons have shown that the topological patterns captured structural information (loops, voids, connectivity) which is not captured by partition and projection-based methodologies such as PCA and k-means.

Implications for pattern detection in mathematical spaces are profound. Mathematical objects typically exhibit topology that is analytically well-understood yet challenging to discover from the sample: the genus of a surface, the number of connected components of a variety, whether the fundamental group is non-trivial. TDA serves as a computational link between sampled representations of mathematical objects and their topological properties. The process outlined above does not require any knowledge about the manifolds in question and yields quantified measures (Betti numbers, persistence entropy, persistence images) that are comparable among different datasets, parameters, or models.

A number of limitations need to be pointed out as well. First, the complexity of persistent homology analysis increases exponentially with the size of the dataset and the homology dimension used, which means that for particularly large datasets, some approximations have to be made. Second, the filtration selected (Vietoris-Rips in our case) makes certain assumptions regarding the scale of structures that matter; the multiparametric approach, though theoretically interesting, is yet hard to compute. Third, our synthetic manifolds have low intrinsic dimensionality, despite being embedded in higher dimensional space; it remains to be seen how well the methods would apply to manifolds with high-dimensional homology (like high-genus surfaces).

Three directions should be pursued in future research. First, the application of TDA to mathematically defined spaces, such as samples from moduli spaces, solutions to polynomial equations, or configuration spaces, would extend the applicability of the methodology beyond synthetically generated manifolds. Second, the incorporation of topological summaries into parametric modeling frameworks (such as using persistence images as covariates for regression on mathematical parameters) would allow topological sensitivity analysis to be performed. Third, the construction of TDA descriptors specifically designed for mathematical problems (for example, homology with coefficients from a field suitable to the problem or filtrations defined algebraically and not metrically) remains almost completely unstudied. With further advancement in TDA software packages and the increasing prevalence of high-dimensional data in the mathematical sciences, the framework described above provides a principled and effective way to get from points to topology.

REFERENCES

1. M. Uray, B. Giunti, M. Kerber, and S. Huber, “Topological data analysis in smart manufacturing processes—A survey on the state of the art,” arXiv preprint arXiv:2310.09319, 2023.
2. B. G. Hem, “Poset functor cocalculus and applications to topological data analysis,” arXiv preprint arXiv:2501.05996, 2025.
3. Y. Singh and E. Quaia, “Unraveling the invisible: Topological data analysis as the new frontier in

radiology's diagnostic arsenal," Tomography, vol. 11, no. 3, 2025.
4. P. Debnath and K. G. Mills, "Pynerve: Topological data analysis for Python and C++," PyPI package version 1.0.0, 2026.
5. "Topological data analysis and topological deep learning beyond persistent homology: A review," Artificial Intelligence Review, Springer, 2025.
6. "Topological Data Analysis Reveals a Subgroup of Luminal B Breast Cancer," IEEE Journal of Translational Engineering in Health and Medicine, pp. 465–471, May 2025.
7. "Persistent homology and Euler characteristic profiles in topological data analysis," EMS Magazine, no. 132, 2024.
8. "Topology and simplicial complexes in topological data analysis," Panamerican Mathematical Journal, vol. 35, no. 2s, 2025.
9. "Mapping firms' locations over time using the Mapper algorithm," Zenodo preprint, 2023.
10. American Mathematical Society, "Special Session on Advances in Applied Topology and Topological Data Analysis," 2025 Fall Southeastern Sectional Meeting, Oct. 2025.