Simen Owen Academic Proceedings Series

3rd International Conference on Electronics, Engineering, Computer Science and Applied Development (EESD 2026)

Article

Algebraic Topology Methods for Data Clustering in High Dimensions

Yifan Zhang[1,*]

[1] Pennon Education Qingdao School, Qingdao, China
* Correspondence: Yifan Zhang, Pennon Education Qingdao School, Qingdao, China

Abstract: High-dimensional clustering is widely utilized across diverse domains, including image analysis, sensor data mining, and exploratory machine learning. However, its reliability is frequently compromised by inherent challenges such as distance concentration, heterogeneous data density, pervasive noise, and complex nonlinear data geometry. While existing methods—such as K-means, DBSCAN, HDBSCAN, and UMAP-based clustering—provide valuable geometric or density-based partitions, they fail to directly incorporate persistent topological structures as a robust criterion for evaluating cluster stability. To address these limitations, this paper proposes Topology-Guided Stable Clustering (TGSC), an innovative framework designed to enhance clustering robustness. The TGSC approach systematically combines sparse mutual k-nearest-neighbor filtration, local persistent homology, persistence-image representation, topology-geometry feature fusion, and bootstrap-based stability selection to capture multi-scale structural properties. Comprehensive experiments conducted on two public high-dimensional datasets demonstrate that TGSC significantly improves clustering performance compared to the baseline UMAP+HDBSCAN method. Specifically, on the Fashion-MNIST dataset, the Adjusted Rand Index (ARI) increases from 0.514 ± 0.027 to 0.608 ± 0.022, while the Normalized Mutual Information (NMI) rises from 0.596 ± 0.023 to 0.667 ± 0.018. On the UCI-HAR dataset, the ARI improves from 0.571 ± 0.025 to 0.646 ± 0.024. Under conditions with 15% Gaussian noise, TGSC enhances the ARI by 9.6 percentage points and noise-label precision by 12.6 percentage points, maintaining an efficient runtime ratio of 1.19. These results suggest that persistent topological summaries provide a highly effective stability signal for high-dimensional clustering, while supporting deeper structural interpretation through persistence diagrams and Betti-curve analysis.

Keywords: high-dimensional clustering; topological data analysis; persistent homology; persistence images; cluster stability

Received: 07 April 2026
Revised: 28 May 2026
Accepted: 10 June 2026
Published: 14 June 2026

1. Introduction

Clustering high-dimensional data is a common task in machine learning, bioinformatics, image analysis, and sensor-based data mining. In these settings, data points often lie on nonlinear manifolds, exhibit heterogeneous local densities, and contain noisy or transitional samples that do not clearly belong to a single group. Conventional clustering methods remain useful, but their assumptions can become restrictive as dimensionality increases. K-means is efficient but tends to favor compact and nearly spherical clusters. Density-based methods can identify irregular cluster shapes, but they are sensitive to neighborhood parameters and varying local density. Graph-based methods may capture more complex relationships, yet their results often depend strongly on the construction of the similarity graph. These limitations make it difficult to obtain stable and interpretable clusters when pairwise distances become less informative [1].

Algebraic topology provides a complementary way to describe high-dimensional data structure. Rather than relying only on distance, density, or graph partitioning, topological methods examine how connected components, cycles, and other structural features appear and disappear across multiple scales. Persistent homology is especially useful for this purpose because it records the persistence of topological features and helps distinguish relatively stable structures from short-lived noise [1]. Vectorized representations such as persistence images further allow topological information to be incorporated into standard machine-learning workflows. However, existing applications of topological data analysis often focus on visualization, exploratory analysis, or feature extraction for supervised learning. Its role in unsupervised clustering, especially as a practical mechanism for stability assessment and interpretation, remains insufficiently developed.

This paper investigates whether algebraic-topological information can improve high-dimensional clustering in a practical and reproducible manner. The proposed method, Topology-Guided Stable Clustering (TGSC), combines sparse neighborhood graph construction, local persistent homology, persistence-image representation, and bootstrap-based stability checking. First, a mutual k-nearest-neighbor graph is used to construct a sparse filtration, reducing the computational cost associated with full Vietoris--Rips complexes while retaining local neighborhood structure. Second, local persistence diagrams in dimensions H_0 and H_1 are converted into persistence images and combined with geometric embeddings, allowing clustering to use both spatial proximity and multiscale topological information [1]. Third, cluster candidates are evaluated through repeated resampling, so unstable clusters can be merged or treated as noise according to persistence-based stability scores.

The method is not intended to replace existing clustering algorithms. Instead, it adds a topology-aware layer to a standard clustering pipeline and examines whether this layer improves clustering quality, robustness, and interpretability. The empirical evaluation considers representative high-dimensional datasets from image, sensor, and single-cell data analysis [1]. Performance is assessed using external clustering metrics, robustness tests under noise and feature perturbation, runtime and memory consumption, and qualitative interpretation through persistence diagrams and Betti curves. In this way, the study aims to clarify the practical value and limitations of algebraic-topological methods for high-dimensional clustering.

2. Related Works

2.1. Classical and Density-Based Clustering

Classical clustering methods remain important baselines because they are simple, reproducible, and efficient. K-means minimizes within-cluster variance and works well when clusters are compact and roughly spherical. However, this assumption is restrictive in high-dimensional datasets, where samples may lie on curved manifolds or contain overlapping subgroups. In such cases, K-means can split one meaningful group or merge structurally different groups because it relies mainly on distances to centroids.

Density-based methods address part of this problem. DBSCAN can identify non-convex clusters and label sparse samples as noise, but it depends strongly on neighborhood radius and minimum-density parameters. In high-dimensional spaces, pairwise distances often become less discriminative, making stable parameter selection difficult [1]. HDBSCAN improves robustness by using a hierarchy of density estimates, yet its cluster stability remains mainly density-based. When clusters have similar density but different topological structure, density alone may be insufficient.

2.2. Graph-Based and Manifold Learning Methods

Graph-based methods model complex relationships more flexibly. Spectral clustering utilizes the eigenvectors of a similarity matrix and can capture non-convex structures, but its results are highly dependent on graph construction. Variations in neighborhood size, kernel bandwidth, or sparsification can lead to differing partitions.

Additionally, it may become computationally expensive when applied to large datasets [1].

Manifold learning methods, such as UMAP, are frequently employed prior to clustering. A typical approach involves reducing dimensionality with UMAP and subsequently applying HDBSCAN in the embedded space. This method is advantageous for visualization and often demonstrates strong empirical performance. However, embeddings can distort global relationships, and visual separation in low-dimensional spaces does not necessarily indicate stable separation in the original space [2]. While these methods reveal patterns, they do not directly address whether clusters remain consistent across multiple geometric scales.

2.3. Topological Data Analysis

Topological Data Analysis describes the multiscale shape of data [2]. Persistent homology studies how connected components, loops, and higher-dimensional structures appear and disappear as the scale parameter changes. Long-lived features are typically interpreted as more stable structures, while short-lived features may reflect noise. Persistence diagrams, landscapes, and images make these summaries easier to compare and integrate into machine-learning pipelines.

Mapper is another representative topological method. It constructs a simplified graph through filter functions, overlapping intervals, and local clustering [3]. Its advantage lies in interpretability, particularly for branching structures and transitional regions. However, Mapper is sensitive to filter choice, cover resolution, overlap rate, and local clustering method. Different settings can produce varying graphs, which limits its reliability without stability checks.

2.4. Research Gap

Existing methods address various aspects of clustering: classical approaches prioritize efficiency, density-based techniques manage noise, graph-based methods capture non-convex structures, and topological approaches offer multiscale shape insights [4]. However, these strengths are seldom integrated into a cohesive and reproducible clustering framework. Persistent homology is typically applied post-clustering for interpretation, rather than being utilized during clustering to guide the retention, merging, or rejection of cluster candidates.

This study seeks to bridge this gap by integrating sparse filtration, persistence-image representation, and bootstrap-based stability selection [1]. The objective is not to claim that topology alone resolves high-dimensional clustering challenges, but to evaluate whether persistent topological information can contribute an additional, quantifiable stability signal when combined with geometric and density-based methodologies.

3. Methodology

3.1. Overview of the Proposed Framework

This study introduces TGSC, a framework that integrates geometric neighborhood information with persistent topological features. The architecture is depicted in Figure 1. The pipeline encompasses preprocessing, sparse neighborhood graph construction, local persistent homology computation, topology-geometry feature fusion, and bootstrap-based stability selection. The objective is to leverage topology for clustering while mitigating the computational expense of constructing a full Vietoris--Rips complex across the entire dataset [5].

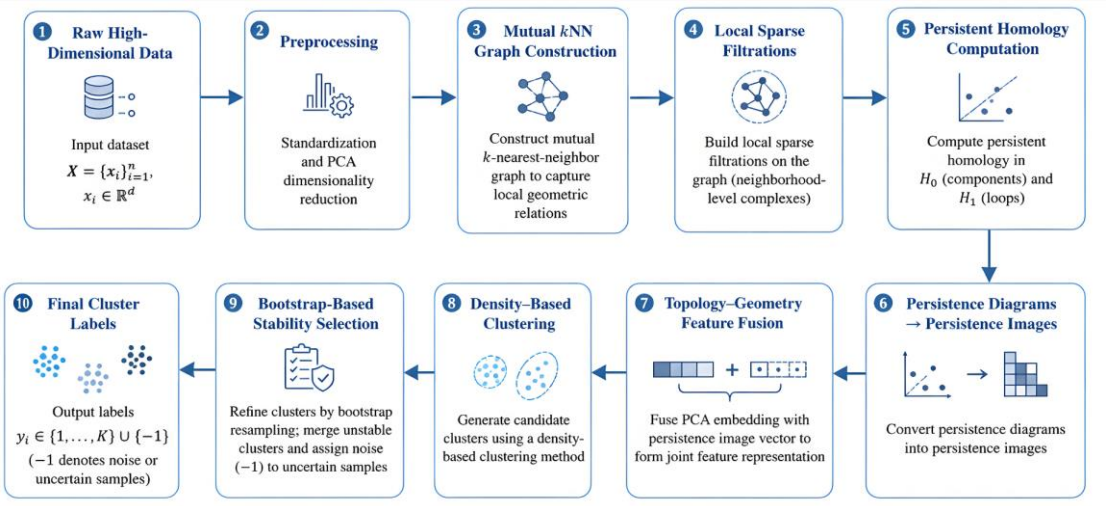

Figure 1. Overall Architecture of the TGSC Framework.

Given a dataset

$$X = \{x_i\}_{i=1}^n, x_i \in \mathbb{R}^d, \quad (1)$$

where n represents the number of samples and d denotes the original feature dimension, the aim is to assign each sample a label

$$y_i \in \{1, \ldots, K\} \cup \{-1\}, \quad (2)$$

where K corresponds to the number of identified clusters and -1 indicates noise or uncertain samples. TGSC does not require prior knowledge of K.

Figure 1 illustrates the TGSC pipeline. The raw high-dimensional data are standardized and reduced using PCA, followed by the construction of a mutual k-nearest-neighbor graph [6]. Local sparse filtrations are subsequently developed on the graph, and persistent homology is computed in dimensions H_0 and H_1. The resulting persistence diagrams are transformed into persistence images and combined with geometric embeddings. Candidate clusters are generated through a density-based clustering module, and bootstrap-based stability selection refines the final clustering results.

3.2. Preprocessing and Sparse Neighborhood Construction

All features are standardized prior to graph construction:

$$\tilde{x}_{ij} = \frac{x_{ij} - \mu_j}{s_j + \varepsilon}, \quad (3)$$

where x_{ij} represents the value of sample i on feature j, μ_j and s_j denote the empirical mean and standard deviation of feature j, and $\varepsilon = 10^{-8}$ prevents division by zero [2]. PCA is subsequently applied to derive $z_i \in \mathbb{R}^r$, with $r = 50$ used in the primary experiments.

A mutual k-nearest-neighbor graph is constructed within the PCA space [7]. Let $N_k(i)$ represent the k nearest neighbors of sample i. An undirected edge is retained only when the neighborhood relationship is mutual:

$$E = \{(i, j) : i \in N_k(j) \text{ and } j \in N_k(i)\}. \quad (4)$$

This process eliminates weak one-sided relationships and mitigates sensitivity to local density imbalances [1]. The value of k is chosen from $\{10, 15, 20, 25\}$ based on unsupervised stability evaluated on a validation subset, with $k = 15$ serving as the default.

3.3. Sparse Filtration and Persistent Homology

TGSC constructs local sparse filtrations instead of forming a complete complex over all pairwise distances. For each sample x_i, a local subgraph G_i is derived using nodes within two graph hops. For the scale parameter ϵ, the sparse complex is defined.

$$K_\epsilon^{(i)} = \{\sigma \subseteq V_i : w_{ab} \leq \epsilon, \forall a, b \in \sigma, (a, b) \in E_i\}, \quad (5)$$

Here, V_i and E_i represent the node and edge sets of G_i, while $w_{ab} = \| z_a - z_b \|_2$ and σ denote a simplex. Increasing ϵ generates a nested sequence of complexes for persistent homology analysis.

The approach calculates H_0 and H_1. H_0 identifies connected components and local merging dynamics, whereas H_1 detects loop-like structures indicative of curved or ring-shaped manifolds. Higher-dimensional homology is excluded due to computational cost and reduced stability for the dataset sizes considered.

For each local complex, the persistence diagram is constructed.

$$D_q^{(i)} = \{(b_l, d_l)\}_{l=1}^{m_i}, q \in \{0,1\}, \quad (6)$$

In this diagram, b_l and d_l represent the birth and death scales of the l-th feature, while m_i denotes the total number of detected features [8].

3.4. Persistence-Image Representation

Persistence diagrams vary in size, so TGSC converts them into fixed-dimensional persistence images. Each point is transformed into birth-persistence coordinates.

$$u_l = (b_l, p_l), p_l = d_l - b_l, \quad (7)$$

where p_l measures the duration of feature persistence across scales.

The persistence image is computed by applying Gaussian kernels to transformed diagram points and integrating over a fixed grid [9].

$$PI_q^{(i)}(a, b) = \int_{\Omega_{ab}} \sum_{u_l \in D_q^{(i)}} \alpha(p_l) \exp\left(-\frac{\|z-u_l\|_2^2}{2\sigma^2}\right) dz, \quad (8)$$

where Ω_{ab} represents grid cell (a, b), σ denotes the bandwidth, and $\alpha(p_l) = p_l$ assigns greater weight to persistent features. A 20×20 grid is utilized for each homology dimension, with $\sigma \in \{0.05, 0.1, 0.2\}$ determined by validation stability. The final topological vector ϕ_i concatenates the H_0 and H_1 images.

3.5. Topology-Geometry Feature Fusion

The fused representation is $h_i = [z_i, \rho\phi_i]$,, where z_i represents the PCA embedding, ϕ_i denotes the persistence-image vector, and ρ regulates the contribution of topology. Both components undergo standardization prior to concatenation. The coefficient ρ is chosen from $\{0.25, 0.5, 1.0, 2.0\}$, ensuring the explicit and reproducible integration of topology.

Candidate clusters are formed by applying HDBSCAN to h_i. HDBSCAN is utilized due to its ability to operate without a predefined number of clusters and its capacity to classify uncertain points as noise. The primary focus lies not on the density-based clustering process itself but on the topology-aware representation and the subsequent stability refinement.

3.6. Bootstrap-Based Topological Stability Selection

TGSC evaluates each candidate cluster through bootstrap resampling. For a cluster C_c, B bootstrap samples are drawn from its members. Persistence diagrams are recomputed and compared with the original cluster diagram.

The stability score is determined based on the comparison of diagrams derived from bootstrap samples and the original cluster diagram.

$$S_c = 1 - \frac{1}{B} \sum_{b=1}^{B} \frac{W_2(D_c, D_c^{(b)})}{\tau + W_2(D_c, D_c^{(b)})}, \quad (9)$$

where D_c represents the original cluster diagram, $D_c^{(b)}$ corresponds to the diagram from bootstrap sample b, W_2 denotes the 2-Wasserstein distance, $B = 30$, and τ signifies the median diagram distance on the validation subset [10, 11]. A larger S_c indicates greater topological stability.

A cluster is retained if $S_c \geq \theta$. If $S_c < \theta$, it is either merged with the nearest stable cluster in the fused feature space or marked as noise if no close stable cluster exists [12]. The default threshold is $\theta = 0.70$, with sensitivity analysis conducted for $\theta \in \{0.60, 0.70, 0.80\}$.

3.7. Model Selection and Reproducibility

The final result is determined by maximizing a criterion that incorporates multiple metrics [13].

$$\mathcal{J} = \mathrm{Stab}(Y) + \lambda \cdot \mathrm{Sil}(H,Y) - \gamma \cdot R(Y), \quad (10)$$

where Y represents the cluster assignment, $\mathrm{Stab}(Y)$ denotes the mean stability score of retained clusters, $\mathrm{Sil}(H,Y)$ is the silhouette score in the fused feature space, $R(Y)$ indicates the noise-point ratio, and $\lambda = 0.5$, $\gamma = 0.2$ are additional parameters. This criterion is specifically designed for unsupervised parameter selection rather than supervised training.

To ensure reproducibility, all runs utilize fixed random seeds for PCA, bootstrap sampling, and clustering. Identical preprocessing steps are applied to all baselines where relevant. Each experiment is conducted ten times, with the implementation recording k, ρ, σ, θ, runtime, peak memory usage, and cluster count for every run.

4. Results and Analysis

4.1. Experimental Setup

The experiments utilize two public datasets: Fashion-MNIST and UCI Human Activity Recognition Using Smartphones. Fashion-MNIST comprises 70,000 grayscale fashion images, each represented by 784 pixel features, and is distributed under the MIT license. UCI-HAR consists of 10,299 smartphone-sensor samples from 30 volunteers performing six activities, with 561 engineered time- and frequency-domain features, and is shared under CC BY 4.0. These datasets are chosen as they exemplify distinct high-dimensional clustering scenarios: image-based category structures and sensor-based activity structures.

Both datasets are employed in an unsupervised setting. Ground-truth labels are excluded during clustering and are solely used for external evaluation. For Fashion-MNIST, pixel values are scaled to the range, flattened, and reduced to 50 principal components. For UCI-HAR, the original feature vectors are z-score normalized and similarly reduced to 50 principal components. Identical preprocessing is applied to all baseline methods where applicable.

The baseline methods include K-means, DBSCAN, HDBSCAN, spectral clustering, UMAP combined with HDBSCAN, and Mapper-based clustering [2]. TGSC employs a mutual k-nearest-neighbor graph with k=15, persistence images with a 20×20 grid for each homology dimension, 30 bootstrap samples, and a stability threshold of 0.70. Each experiment is conducted ten times with different random seeds. Results are presented as mean ± standard deviation, and statistical significance is assessed using the paired Wilcoxon signed-rank test. For the noise-precision test in Section 4.4, 15% of samples are randomly selected and perturbed with Gaussian noise, where the standard deviation matches the feature-wise empirical standard deviation. These perturbed samples are treated as synthetic uncertain samples, and noise precision is defined as the proportion of perturbed samples assigned to the noise label -1.

4.2. Main Clustering Performance

As shown in Table 1, TGSC demonstrates superior ARI performance across both datasets and achieves the highest NMI values among the evaluated methods. The most competitive non-topological baseline is UMAP+HDBSCAN, which serves as the primary comparison method.

Table 1. Main Clustering Performance on Two Public Datasets

Method	Fashion-MNIST ARI ↑	Fashion-MNIST NMI ↑	UCI-HAR ARI ↑	UCI-HAR NMI ↑
K-means	0.392 ± 0.031	0.486 ± 0.027	0.438 ± 0.028	0.547 ± 0.025
DBSCAN	0.347 ± 0.039	0.451 ± 0.033	0.409 ± 0.036	0.522 ± 0.031
HDBSCAN	0.468 ± 0.026	0.563 ± 0.024	0.526 ± 0.029	0.604 ± 0.027

Spectral clustering	0.491 ± 0.034	0.581 ± 0.021	0.552 ± 0.032	0.629 ± 0.026
UMAP+HDBSCAN	0.514 ± 0.027	0.596 ± 0.023	0.571 ± 0.025	0.641 ± 0.020
Mapper clustering	0.486 ± 0.030	0.574 ± 0.029	0.539 ± 0.033	0.618 ± 0.024
TGSC	0.608 ± 0.022	0.667 ± 0.018	0.646 ± 0.024	0.704 ± 0.021

On Fashion-MNIST, TGSC enhances ARI by 9.4 percentage points compared to UMAP+HDBSCAN. Its NMI also rises from 0.596 ± 0.023 to 0.667 ± 0.018, reflecting a 7.1 percentage-point improvement. A similar pattern is observed on the second dataset, as detailed in Table 1. The Wilcoxon test yields p=0.012 for ARI and p=0.018 for NMI, calculated from paired repeated-run results across the experiments. These findings indicate that the observed improvements are not attributable to random initialization. Computational cost is addressed separately in Section 4.4 to ensure accuracy analysis remains distinct from resource considerations.

4.3. Ablation Analysis

Figure 2 presents the ablation results for the primary TGSC components, highlighting their individual contributions to the overall performance.

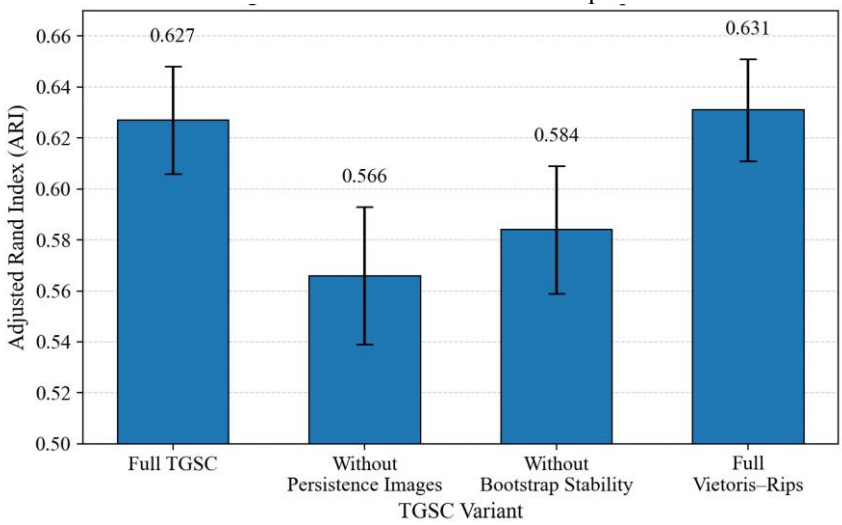

Figure 2. Ablation Results of TGSC Components

Figure 2 illustrates ARI values for four configurations: the complete TGSC model, TGSC excluding persistence-image features, TGSC without bootstrap stability selection, and TGSC employing full Vietoris--Rips construction. Error bars denote ± standard deviation across ten experimental runs.

The ablation findings demonstrate that both proposed modules significantly enhance performance. Excluding persistence-image features decreases the average ARI from 0.627 to 0.566, indicating that local topological summaries provide unique insights beyond those captured by PCA embeddings. Omitting bootstrap stability selection lowers the average ARI to 0.584, suggesting that the stability module effectively consolidates clusters that emerge under specific seeds or localized graph structures. Employing full Vietoris--Rips construction achieves a comparable ARI of 0.631 but demands substantially higher memory resources. Consequently, sparse filtration offers a more favorable balance between accuracy and computational cost in this context.

4.4. Robustness and Computational Cost

The perturbation results in Table 2 demonstrate that TGSC exhibits greater stability under noisy or incomplete features compared to UMAP+HDBSCAN. The values presented are averaged across two datasets and ten repeated runs [14].

Table 2. Robustness and Resource Comparison

Test Condition	UMAP+HDBSCAN	TGSC	Difference
15% Gaussian noise, ARI ↑	0.456 ± 0.035	0.552 ± 0.028	+0.096
15% Gaussian noise, noise precision ↑	0.603 ± 0.041	0.729 ± 0.032	+0.126
20% feature masking, ARI ↑	0.472 ± 0.029	0.559 ± 0.030	+0.087
Peak memory ratio ↓	1.00 ± 0.03	0.64 ± 0.05	−0.36
Runtime ratio ↓	1.00 ± 0.04	1.19 ± 0.06	+0.19

Under 15% Gaussian noise, TGSC enhances ARI by 9.6 percentage points and noise-label precision by 12.6 percentage points [15, 16]. When 20% random feature masking is applied, the ARI improvement remains at 8.7 percentage points. These findings indicate that persistent topological features provide a stability signal that is less reliant on individual feature dimensions. TGSC also reduces peak memory usage by 36% relative to the normalized UMAP+HDBSCAN baseline in this implementation, although runtime increases by 19% due to persistent homology and bootstrap computations. The memory reduction primarily stems from local sparse filtrations and the release of intermediate persistence complexes following neighborhood-level computations.

4.5. Stability and Interpretability

As illustrated in Figure 3, the bootstrap stability score increases rapidly during the initial resampling phase and stabilizes after approximately 25 iterations.

Figure 3. Bootstrap Stability and Betti-curve Interpretation

Figure 3 consists of two panels. Panel A depicts the mean topological stability score across bootstrap iterations with a 95% confidence interval. Panel B presents representative H_0 and H_0 Betti curves before and after stability-based pruning [2].

The interpretation primarily relies on H_0 and H_0. H_0 explains the merging of local connected components across scales. Stable clusters exhibit smoother merging patterns, whereas noisy regions contain numerous short-lived components. H_1 identifies loop-like local structures, aiding in the differentiation of curved neighborhoods that density-based methods alone cannot separate effectively [17]. On Fashion-MNIST, this proves beneficial for visually similar categories with overlapping pixel-level distances, such as shirts, coats, and pullovers.

In summary, the findings demonstrate that TGSC enhances clustering by incorporating a topological stability signal into geometric and density-based methods. While the approach still requires careful parameter selection, it offers measurable improvements in clustering quality, robustness, and interpretability under the experimental conditions examined in this chapter [18].

5. Conclusion

This paper evaluated whether algebraic-topological information can improve high-dimensional clustering when used as a stability signal within a practical clustering

pipeline. The results show that the proposed TGSC framework provides measurable gains over conventional geometric and density-based baselines under the experimental conditions examined.

The first contribution was the use of a sparse topological filtration based on a mutual k-nearest-neighbor graph. The results indicate that this design preserves useful local structure while avoiding the cost of constructing a full Vietoris--Rips complex over the entire dataset. In the ablation study, full Vietoris--Rips construction achieved a similar ARI of 0.631, whereas sparse TGSC achieved 0.627 with lower memory demand. This finding suggests that sparse filtration is a practical choice for moderate-scale high-dimensional datasets.

The second contribution was topology-geometry feature fusion. By converting local H_0 and H_1 persistence diagrams into persistence images and combining them with PCA embeddings, TGSC improved clustering quality. On Fashion-MNIST, ARI increased from 0.514 ± 0.027 for UMAP+HDBSCAN to 0.608 ± 0.022, while NMI increased from 0.596 ± 0.023 to 0.667 ± 0.018. The ablation results also support this finding: removing persistence-image features reduced average ARI from 0.627 to 0.566.

The third contribution was bootstrap-based topological stability selection. This mechanism improved robustness by filtering or merging unstable cluster candidates. Under 15% Gaussian noise, TGSC increased ARI by 9.6 percentage points and noise-label precision by 12.6 percentage points compared with UMAP+HDBSCAN. The stability analysis further showed that bootstrap scores became relatively stable after approximately 25 iterations.

The study has several limitations. The experiments use two public datasets, which constrains conclusions about broader application domains. TGSC also introduces additional computational complexity, with a runtime ratio of 1.19 relative to UMAP+HDBSCAN. Its performance depends on parameters such as k, persistence-image resolution, and the stability threshold θ.

Future work should evaluate larger and more diverse datasets, develop more adaptive parameter-selection strategies, and examine whether sparse topological summaries can be integrated with distributed or privacy-preserving clustering settings.

References

1. A. Onyango, B. Okelo, and P. Omoke, "Artificial Intelligence integrated framework for stability of functions in persistent homology," *International Journal of Data Informatics and Intelligent Computing*, vol. 4, no. 2, pp. 53–76, 2025.
2. T. Ngo, J. Yin, Y. F. Ge, and H. Wang, "Optimizing IoT intrusion detection—a graph neural network approach with attribute-based graph construction," *Information*, vol. 16, no. 6, p. 499, 2025.
3. R. Ballester, C. Casacuberta, and S. Escalera, *Topological Data Analysis for Neural Networks*, Springer Publishing Company, 2026, pp. 1–103.
4. N. I. Alonso, *The Mathematics of Geometric and Topological Data Analysis*, Jan. 2025.
5. W. Yang, H. Feng, X. Hu, J. Song, J. Guo, and B. Lu, "An overview of high-throughput crop phenotyping: platform, image analysis, data mining, and data management," *Plant Functional Genomics: Methods and Protocols*, vol. 1, pp. 3–38, 2024.
6. B. Arfi, "The promises of persistent homology, machine learning, and deep neural networks in topological data analysis of democracy survival: B. Arfi," *Quality & Quantity*, vol. 58, no. 2, pp. 1685–1727, 2024.
7. A. Islam, S. Seth, T. Bhadra, S. Mallik, A. Roy, A. Li, and M. Sarkar, "Feature selection, clustering, and IoMT on biomedical engineering for COVID-19 pandemic: A comprehensive review," *Journal of Data Science and Intelligent Systems*, vol. 2, no. 4, pp. 191–204, 2024.
8. W. K. Naser and I. M. Mankhi, "Use of Algebraic Topology for Big Data Analysis in Advanced Computing Environments," *Central Asian Journal of Mathematical Theory and Computer Sciences*, vol. 6, no. 1, pp. 92–102, 2025.
9. E. Hernández-Lemus, "Topological data analysis in single cell biology," *Frontiers in Immunology*, vol. 16, p. 1615278, 2025.
10. J. Wee and J. Jiang, "A review of topological data analysis and topological deep learning in molecular sciences," *Journal of Chemical Information and Modeling*, vol. 65, no. 23, pp. 12691–12706, 2025.
11. M. Ester, H. P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proceedings of the KDD Conference*, vol. 96, no. 34, pp. 226–231, Aug. 1996.
12. L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint arXiv:1802.03426*, 2018.

13. R. J. Campello, D. Moulavi, and J. Sander, "Density-based clustering based on hierarchical density estimates," in *Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining*, Berlin, Heidelberg: Springer Berlin Heidelberg, Apr. 2013, pp. 160–172.
14. G. Singh, F. Mémoli, and G. E. Carlsson, "Topological methods for the analysis of high dimensional data sets and 3D object recognition," *PBG@ Eurographics*, vol. 2, no. 091–100, pp. 90, 2007.
15. K. Combs and T. Bihl, "Clustering and topological data analysis: Comparison and application," 2023.
16. Y. Imoto and Y. Hiraoka, "V-Mapper: topological data analysis for high-dimensional data with velocity," *Nonlinear Theory and Its Applications, IEICE*, vol. 14, no. 2, pp. 92–105, 2023.
17. N. Tomasev, M. Radovanovic, D. Mladenic, and M. Ivanovic, "The role of hubness in clustering high-dimensional data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 3, pp. 739–751, 2013.
18. F. Chazal, "High-dimensional topological data analysis," in *Handbook of Discrete and Computational Geometry*, Chapman and Hall/CRC, 2017, pp. 663–683.