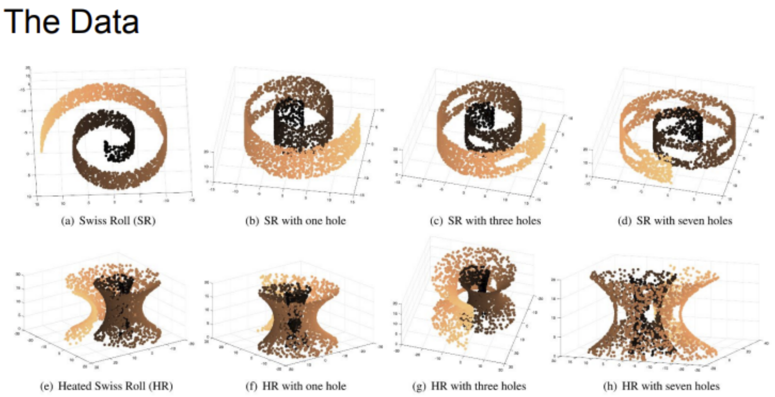
Topological Dimensionality Reduction

Anton Yang

May 2024

1 Introduction

Dimensionality Reduction plays a vital role in visualizing and interpreting high-dimensional data. A commonly used technique for this is Principal Component Analysis (PCA), which is a form of unsupervised learning. PCA transforms the data into a set of principal components that capture the most variance with fewer dimensions, potentially capturing data structure without losing too much information. While this approach can help reduce complexity, it's important to carefully determine the number of components to retain so that critical information isn't lost. However, PCA is weak at treating circular and nonlinear data, where PCA is defined as an orthogonal linear transformation. One dimensional reduction method is topological dimension reduction. This is a class of techniques in data analysis that aims to reduce high-dimensional data into lower-dimensional representations while preserving topological features or patterns in the data. These methods emphasize understanding the shape or structure of data rather than just minimizing information loss. The key characteristics are topology preservation, non-linear relationships, and flexibility in handling data.



(a) Swiss Roll (SR) (b) SR with one hole (c) SR with three holes (d) SR with seven holes

(e) Heated Swiss Roll (HR) (f) HR with one hole (g) HR with three holes (h) HR with seven holes

Figure 1: fig:Data Shapes

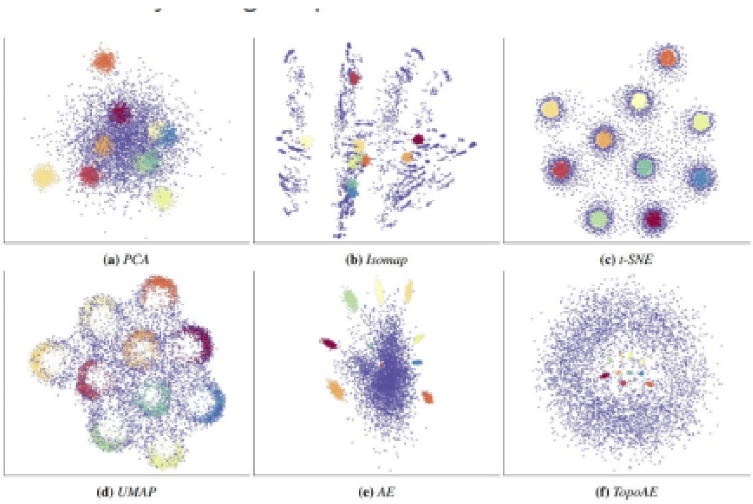

(a) PCA (b) Isomap (c) t-SNE
(d) UMAP (e) AE (f) TopoAE

Figure 2: fig:Spheres Dataset

2 Method

In the sequel, $||\cdot||$ is the euclidean norm in $\mathbb{R}^d$, for some $d \geqslant 3$. In the sequel, topological spaces $\mathcal{M}$ will considered to be second countable Hausdorff. Therefore, every pair of distinct points has a corresponding pair of disjoint neighborhoods. Its topology has a countable basis of open set. This assumption is satisfied in most topological spaces.[1]

Definition 1. A topological space $\mathcal{M}$ is called a (topological) manifold if, locally, it resembles a real n-dimensional Euclidean space, that is, there exists $n \in \mathbb{N}$ such that for all $x \in \mathcal{M}$, there exists a neighborhood $\mathcal{U}_x$ of x and a homoeomorphism $f : \mathcal{U}_x \to \mathbb{R}^n$. The pair $(\mathcal{U}_x, f)$ is referred to as a chart on $\mathcal{M}$ and f is alled a parametrization at x.

Definition 2. Let $\mathcal{M}$ be a manifold. $\mathcal{M}$ is said to be smooth if given $x \in \mathcal{M}$, the parametrization f and x has a smooth or continuous partial derivatives of any order and can be extended to a smooth function $g : \mathcal{M} \to \mathbb{R}^n$ such that $g|_{\mathcal{M} \cap \mathcal{U}_x} = f$.

Definition 3. Let $\mathcal{M}$ and $\mathcal{N}$ be differentiable manifolds. A function $\psi : \mathcal{M} \to \mathcal{N}$ is an embedding if ψ is an injective function.

Definition 4. Consider a Hausdorff topological manifold $\mathcal{M}$ homeomorphic to an open subset of the half-euclidean space $\mathbb{R}^n_+$. Let the interior $\text{int}(\mathcal{M})$ of $\mathcal{M}$ be the subspace of $\mathcal{M}$ formed by all points s that have a neighborhood homeomorphic to $\mathbb{R}^n$. Then, the boundary of $\mathcal{M}$ is defined as a complement of int($\mathcal{M}$) in $\mathcal{M}$, that is, $\mathcal{M}\setminus$ int($\mathcal{M}$), which is an n-1-dimensional topological manifold.

These are the initial considerations and assumptions for every model of topological dimensional reduction.

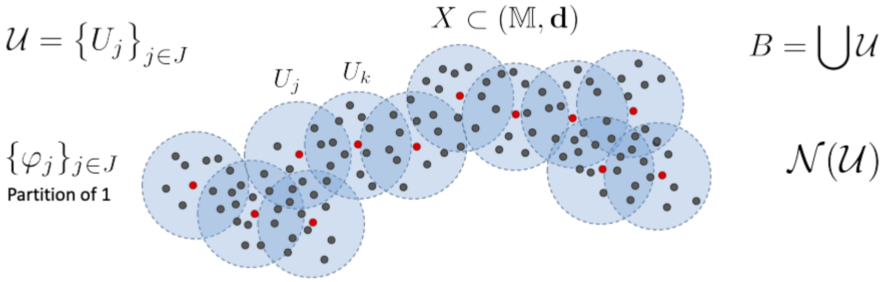

$\mathcal{U} = \{U_j\}_{j\in J}$
$X \subset (\mathbb{M}, \mathbf{d})$
$B = \bigcup \mathcal{U}$
U_j
U_k
$\{\varphi_j\}_{j\in J}$
Partition of 1
$\mathcal{N}(\mathcal{U})$

Figure 3: fig:Partition of Unity

3 Gradient-Free 3D Topology Optimization

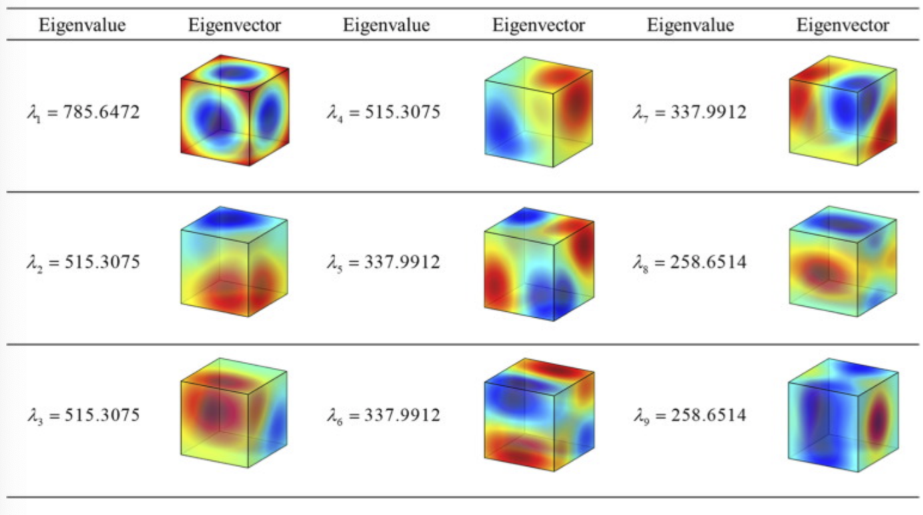
Researchers are studying based machine learning for efficient gradient-free 3D topology optimization, which a technique to integrate the material-field series expansion topological parameterization and the deep neural networks is proposed. [2] This optimization method reduces the computational time by 1–4 orders of magnitude compared with the coarse-mesh-based gradient-free methods and can handle up to 200 design parameters in design dimension reduction model. Several designs with high performances are obtained with the present method. Continuum topology optimization can be explained as finding the optimal distribution of materials in a structural design:

$$\min f(\mathcal{X}, \mathbf{u}) \text{ such that } G(\mathcal{X}, \mathbf{u}) \text{ with } g_k(\mathcal{X}, \mathbf{u}) \leqslant 0, k = 1, 2, ...$$

where $\mathcal{X}$ denotes the material distribution function, $f(\mathcal{X}, \mathbf{u})$ is the performance function to be optimized. This can be given in the form of compliance, structural stress, band gap, etc. $\mathbf{u}$ is the structural response obtained by solving the equilibrium function $G(\mathcal{X}, \mathbf{u}) = 0$, and g_k is the inequality constraint in topology optimization, which contains a volume or a mass constraint.
First, researchers assumed that the correlation is expressed as an exponential distance function:

$$c(\mathbf{x}_i, \mathbf{x}_j) = exp(-||\mathbf{x}_i - \mathbf{x}_j||^2/l_c^2$$

where $c(\mathbf{x}_i, \mathbf{x_j})$ represents the correlation of the material field function value of the two points $\mathbf{x}_i$ and $\mathbf{x}_j$, and l_c represents the correlation length. They utilized the efficient gradient-free 3D topology optimization to better understand the expansion with eigenvector λ_k, and they illustrate the correlation matrix for a cube design domain. The different orders of eigenvectors show different basis functions, and the eigenvalues decrease rapidly with increased order, which indicated that the major topology configuration can be approximately represented by the first several orders of eigenvectors.



Eigenvalue	Eigenvector	Eigenvalue	Eigenvector	Eigenvalue	Eigenvector
$\lambda_1 = 785.6472$		$\lambda_4 = 515.3075$		$\lambda_7 = 337.9912$	
$\lambda_2 = 515.3075$		$\lambda_5 = 337.9912$		$\lambda_8 = 258.6514$	
$\lambda_3 = 515.3075$		$\lambda_6 = 337.9912$		$\lambda_9 = 258.6514$	

Figure 4: fig:The first nine orders of eigenvalues and eigenvectors of the correlation matrix

4 Single Cell RNA-seq Data Using Graph Autoencoder

The topology-preserving dimensionality was also used in Biology field. Dimensionality Reduction is crucial for the visualization and interpretation of the high-dimensional single-cell RNA sequencing (scRNA-seq) data. Preserving topological structure among cells to low dimensional space remains a challenge. A dimensionality reduction method that preserves topological structure in scRNA-seq, an ideal approach for investigating cell-cell variation. They utilized PCA and t-Distributed Stochastic Neighbor Embedding (t-SNE) to implement a scRNA-seq data for visualization, and this significantly increase the understanding of cellular heterogeneity and development progress. However, the data generated by parallel scRNA-seq are of high dropout and high noise with complex structure, which posed a series of challenges on dimensionality reduction. It is a big challenge to preserve the complex topological structure among cells. [3]

This research used deep learning, and it embedded the distinct cell features while ignoring the cell-cell relationships, which limited their ability to reveal the

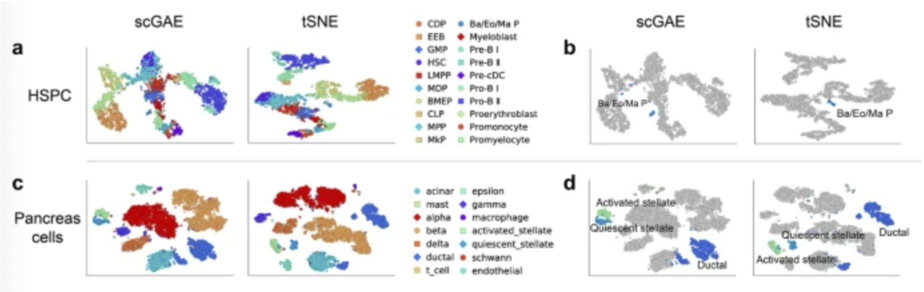


Figure 5: fig:Analyses of Two Real Datasets

complex topological structure among cells. According to figure 5, scGAE shows preserved topological structure among human pancreatic cells populations. The function of the pancreas hinges on complex interactions among distinct cell types and cell populations. They found that the distances and topological structures among cell types inferred by scGAE better fit our knowledge. Moreover, scGAE clearly separated other cell populations while SAUCIE, Ivid, and PHATE mixed some of the clusters Overall, scGAE preserved the topological structure among different cell populations, which greatly benefit the understanding of the cellular relationships.

5 Conclusion

Overall, Topological dimension reduction has become a valuable tool in a wide range of fields, from biology to engineering. These techniques allow researchers to explore and visualize complex high-dimensional datasets, revealing underlying structures, clusters, or patterns that might otherwise go unnoticed. By applying topology and non-linear relationships, topological methods offer a different perspective compared to traditional linear approaches like PCA, providing insights that are critical for decision-making and further analysis.

However, the application of topological dimension reduction is not without its challenges. Effective use of these techniques requires careful parameter tuning, and the results can be sensitive to these choices. Additionally, while they excel in capturing complex data relationships, they may require considerable computational resources and expertise to implement effectively. However, topological dimension reduction remains a powerful approach, enabling researchers to distill high-dimensional data, especially complex dataset, into more manageable forms without sacrificing essential information.

References

[1] E. Kwessi. Topological comparison of some dimension reduction methods using persistent homology on eeg data. *Axioms*, 12(7), 2023.

[2] Z. Sun, Y. Wang, P. Liu, and Y. Luo. Topological dimensionality reduction-based machine learning for efficient gradient-free 3d topology optimization. *Materials Design*, 220:110885, 2022.

[3] Z. Z. W. J. Zixiang Luo, Chenyu Xu. A topology-preserving dimensionality reduction method for single-cell rna-seq data using graph autoencoder. 11, 2021.