CLC _____ Number _____

UDC _____ Available for reference ☐Yes ☐No

SUSTech Southern University of Science and Technology

Undergraduate Thesis

Thesis Title: **Persistent Homology in Finance and Machine Learning: From Theory to Practice**

Student Name: **Yanche Wu**

Student ID: **12210614**

Department: **Department of Mathematics**

Program: **Mathematics and Applied Mathematics**

Thesis Advisor: **Yifei Zhu**

分类号 _____ 编号_____

U D C_____ 密级_____

南方科技大学
SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

本科生毕业设计（论文）

题　　目：金融与机器学习中的持续同调：从理论到实践

姓　　名：伍言澈

学　　号：12210614

院　　系：数学系

专　　业：数学与应用数学

指导教师：朱一飞

COMMITMENT OF HONESTY

1. I solemnly promise that the paper presented comes from my independent research work under my supervisor's supervision. All statistics and images are real and reliable.

2. Except for the annotated reference, the paper contents no other published work or achievement by person or group. All people making important contributions to the study of the paper have been indicated clearly in the paper.

3. I promise that I did not plagiarize other people's research achievement or forge related data in the process of designing topic and research content.

4. If there is violation of any intellectual property right, I will take legal responsibility myself.

Signature: 

Date: 2026.5.28

COMMITMENT OF HONESTY

1. 本人郑重承诺所呈交的毕业设计（论文），是在导师的指导下，独立进行研究工作所取得的成果，所有数据、图片资料均真实可靠。

2. 除文中已经注明引用的内容外，本论文不包含任何其他人或集体已经发表或撰写过的作品或成果。对本论文的研究作出重要贡献的个人和集体，均已在文中以明确的方式标明。

3. 本人承诺在毕业论文（设计）选题和研究内容过程中没有抄袭他人研究成果和伪造相关数据等行为。

4. 在毕业论文（设计）中对侵犯任何方面知识产权的行为，由本人承担相应的法律责任。

作者签名: 

2026 年 5 月 28 日

Persistent Homology in Finance and Machine Learning: From Theory to Practice

[ABSTRACT]: In this paper, we discuss a key issue in the application of topological data analysis, namely, persistence diagrams are not directly compatible with many standard statistical and machine learning methods. To address this, persistence landscapes and persistence images are introduced to convert topological information into appropriate mathematical structure.

The thesis consists of a theoretical review and two applications. In the theoretical review, we introduce the notions of simplex, complex, filtration, persistent homology, persistence diagrams, and vectorizations. In the first application, we reproduce a financial market analysis based on persistence landscapes and their L^q-type norms. In the second application, we reproduce a machine learning experiment in which persistence images are used for clustering synthetic geometric shapes.

[Keywords]: Persistent Homology, Persistence Diagram, Persistence Landscape, Persistence Image

[摘要]：在本文中，我们探讨了拓扑数据分析应用中的一个关键问题，即持续同调图与许多标准统计和机器学习方法并不直接兼容。为了解决这个问题，持续景观和持续图像被引入，旨在将拓扑信息转化入更合适的数学结构中。

本论文由理论综述和两个应用组成。在理论综述中，我们介绍了单形，复形，过滤，持续同调以及向量化方法。第一个应用中，我们复现了一项基于持续景观及其 L^q 型范数的金融市场分析。第二个应用中，我们复现了一个机器学习实验，其中持续图像被用于合成几何形状的聚类。

[关键词]：持续同调, 持续同调图, 持续景观, 持续图像

Contents

1. Introduction

Topological data analysis (TDA) is a set of tools providing a framework for studying data through its shape, connectivity, and other topological structures [4]. Persistent homology is particularly important, it tracks the birth and death of homological features across a filtraction [8] by persistence diagrams [6, 15]. However, a persistence diagram is a multiset of points instead of vectors, and distances between diagrams are usually defined through metrics such as the bottleneck or Wasserstein distance [5]. This makes persistence diagrams difficult to use directly in many standard statistical and machine learning pipelines. So several vectorization methods have been developed to solve this problem. In this thesis, we focus on two representative methods: persistence landscapes [3] and persistence images [1]. These methods transform persistence diagrams into functional or finite-dimensional representations.

This thesis organize contents from theory to application. The theoretical part introduces the mathematical background of simplex, complex, filtrations, persistent homology, persistence diagrams, and two vectorizations.

We apply the preceding theory in two case studies, each reproducing selected results from an existing paper.

The first application [9] concerns financial market analysis. In this setting, persistence landscapes and their L^q-type norms are used to summarize topological features of point clouds constructed from market return data. The goal is to examine whether these topological summaries exhibit visible changes around major unstable periods of markets. This analysis should be understood as exploratory rather than as a formal forecasting model: elevated topological signals may be associated with market crisis, but predictive early-warning claims require additional research.

The second application [1] is a machine learning task. Persistence images are used as features for clustering synthetic geometric point clouds. This experiment provides simple benchmarks for evaluating whether vectorized persistent homology can preserve shape in-

formation under noise.

The remainder of this thesis is organized as follows. Section 2 introduces the mathematical background. Section 3 presents two vectorization methods. Section 4 reproduces the financial market application. Section 5 reproduces the machine learning application. Section 6 concludes the thesis and gives suggestions on future work.

2. Mathematical Foundations of Persistent Homology

Persistent homology studies how homological features appear and disappear along a filtration. In this thesis, all simplicial complexes in the theoretical development are treated as finite abstract simplicial complexes. This formulation is standard in computational topology: chain groups, boundary maps, homology groups, filtrations, persistent homology groups, and persistence intervals can all be defined from finite abstract simplicial complex [2, 15].

This formulation is natural for our application sections, since Vietoris–Rips complexes are defined by declaring finite subsets of a metric space to be simplices by using a pairwise-distance condition.

Throughout this section, homology is taken with coefficients in a fixed field $\mathbb{F}$ (In computations, we mainly use $\mathbb{Z}/11\mathbb{Z}$ or $\mathbb{Z}/2\mathbb{Z}$). This convention makes all chain groups and homology groups vector spaces.

2.1 Abstract Simplicial Complexes

In this thesis, simplices are always nonempty, while the empty complex is allowed. Thus the empty complex contains no simplices.

Definition 2.1 (Abstract simplex)**.** Let V be a finite set. An **abstract k-simplex** in V is a nonempty subset $\sigma = \{v_0, v_1, \ldots, v_k\} \subseteq V$ with cardinality $k+1$. The elements of σ are called **vertices**.

Definition 2.2 (Face)**.** Let σ be an abstract simplex. A **face** of σ is a nonempty subset $\tau \subseteq \sigma$. If $|\tau| = \ell + 1$, then τ is an ℓ-simplex.

Definition 2.3 (Abstract simplicial complex)**.** An **abstract simplicial complex** K on a finite

vertex set V is a collection of nonempty finite subsets of V such that whenever

$$\sigma \in K \quad \text{and} \quad \emptyset \neq \tau \subseteq \sigma,$$

we also have $\tau \in K$. The elements of K are called simplices. Equivalently, under the convention used in this thesis, an abstract simplicial complex is closed under taking nonempty faces. The empty complex is allowed, but the empty set is not regarded as a simplex.

Definition 2.4 (Dimension)**.** The **dimension** of a simplex σ is $\dim \sigma = |\sigma| - 1$. If $K \neq \emptyset$, the dimension of a finite abstract simplicial complex K is $\dim K = \max_{\sigma \in K} \dim \sigma$. For the empty complex, we use the convention $\dim \emptyset = -1$.

The definition records only which finite subsets of vertices are declared to be simplices. This is sufficient for defining chain groups, boundary operators, homology, and persistent homology [2, 15].

2.2 Chain Groups and Homology

Definition 2.5 (Oriented simplex)**.** Let $\sigma = \{v_0, \ldots, v_p\}$ be an abstract p-simplex. An **orientation** of σ is an equivalence class of orderings $[v_0, \ldots, v_p]$, where two orderings are equivalent if they differ by an even permutation. Reversing the orientation changes the sign:

$$[v_{\pi(0)}, \ldots, v_{\pi(p)}] = \operatorname{sgn}(\pi)[v_0, \ldots, v_p].$$

Definition 2.6 (p-chain group)**.** Let K be a finite abstract simplicial complex. For $p \geq 0$, the p**-chain group** $C_p(K)$ is the vector space over $\mathbb{F}$ generated by the oriented p-simplices of K. An element of $C_p(K)$ is called a p**-chain** and can be written as $c = \sum_i a_i \sigma_i$, where $a_i \in \mathbb{F}$ and each σ_i is an oriented p-simplex of K.

Definition 2.7 (Boundary operator)**.** The **boundary operator** $\partial_p : C_p(K) \to C_{p-1}(K)$ is defined on an oriented p-simplex $[v_0, \ldots, v_p]$ by

$$\partial_p[v_0, \ldots, v_p] = \sum_{i=0}^{p} (-1)^i [v_0, \ldots, \widehat{v_i}, \ldots, v_p],$$

where $\widehat{v_i}$ means that v_i is omitted. The definition is extended linearly to all p-chains. For $p = 0$, we set $\partial_0 = 0$.

Proposition 2.8 (Boundary of a boundary)**.** *For every $p \geq 1$, $\partial_{p-1} \circ \partial_p = 0$.*

Proof. It is enough to verify the identity on an oriented p-simplex $[v_0, \ldots, v_p]$. Applying the boundary operator twice produces a sum of oriented $(p-2)$-simplices. Each term is obtained by deleting two vertices. If the two deleted vertices are v_i and v_j with $i < j$, then the corresponding face appears twice: once by deleting v_i first and then v_j, and once by deleting v_j first and then v_i. These two occurrences have opposite signs. Hence all terms cancel in pairs, and therefore $\partial_{p-1}\partial_p[v_0, \ldots, v_p] = 0$. The identity holds for every p-chain by linearity. □

Definition 2.9 (Chain complex)**.** The sequence of vector spaces and boundary maps

$$\cdots \longrightarrow C_{p+1}(K) \xrightarrow{\partial_{p+1}} C_p(K) \xrightarrow{\partial_p} C_{p-1}(K) \longrightarrow \cdots$$

is called the **simplicial chain complex** of K.

Definition 2.10 (Cycle group and boundary group)**.** The **p-cycle group** of K is $Z_p(K) = \ker \partial_p$. The **p-boundary group** of K is $B_p(K) = \operatorname{im} \partial_{p+1}$.

Proposition 2.11 (Boundaries are cycles)**.** *For every $p \geq 0$, $B_p(K) \subseteq Z_p(K)$.*

Proof. Let $b \in B_p(K)$. Then there exists a $(p+1)$-chain $c \in C_{p+1}(K)$ such that $b = \partial_{p+1}c$. Applying ∂_p gives

$$\partial_p b = \partial_p \partial_{p+1} c = 0.$$

Hence $b \in \ker \partial_p = Z_p(K)$. Therefore $B_p(K) \subseteq Z_p(K)$. □

Definition 2.12 (Homology group)**.** The **p-th homology group** of K is the quotient vector space $H_p(K) = Z_p(K)/B_p(K)$. Elements of $H_p(K)$ are homology classes of p-cycles modulo p-boundaries.

Definition 2.13 (Betti number)**.** The p**-th Betti number** of K is $\beta_p(K) = \dim H_p(K)$.

The first Betti numbers have standard interpretations. The number β_0 counts connected components, β_1 counts independent one-dimensional loops, and β_2 counts two-dimensional voids. In the abstract setting, these terms should be understood as the usual topological interpretations of the corresponding homology groups.

2.3 Filtrations of Abstract Simplicial Complexes

Definition 2.14 (Subcomplex)**.** Let K be an abstract simplicial complex. A **subcomplex** of K is an abstract simplicial complex L such that $L \subseteq K$.

Definition 2.15 (Filtration)**.** A **filtration** of a finite abstract simplicial complex K is a nested sequence of subcomplexes

$$K^0 \subseteq K^1 \subseteq \cdots \subseteq K^m = K.$$

The index i is called the **filtration index**. If a simplex $\sigma \in K$ first appears in K^i, then i is called the filtration value of σ.

As the filtration grows, new simplices are added. Consequently, homology classes may be born, persist for several filtration levels, and die at later levels. This is the basic mechanism recorded by persistent homology [6, 8].

Definition 2.16 (Filtration value)**.** Let $K^\bullet$ be a filtration

$$\emptyset = K^0 \subseteq K^1 \subseteq \cdots \subseteq K^m = K.$$

The **filtration value** of a simplex $\sigma \in K$ is the smallest index i such that $\sigma \in K^i$.

2.4 Persistent Homology Groups

Let

$$K^\bullet : \quad K^0 \subseteq K^1 \subseteq \cdots \subseteq K^m$$

be a filtration of finite abstract simplicial complexes.

Proposition 2.17 (Inclusion-induced map on homology)**.** *For every pair of indices $i \leq j$, the inclusion $\iota^{i,j} : K^i \hookrightarrow K^j$ induces a linear map on homology $f_p^{i,j} : H_p(K^i) \to H_p(K^j)$.*

Proof. The inclusion $\iota^{i,j} : K^i \hookrightarrow K^j$ sends every simplex of K^i to the same simplex regarded as a simplex of K^j. Therefore it induces a linear map on chain groups $\iota_\#^{i,j} : C_p(K^i) \to C_p(K^j)$. This chain-level map commutes with the boundary operators: $\partial_p \iota_\#^{i,j} = \iota_\#^{i,j} \partial_p$. Hence cycles are mapped to cycles and boundaries are mapped to boundaries. Therefore, if two cycles in K^i differ by a boundary, their images in K^j also differ by a boundary. Thus the chain-level inclusion induces a well-defined linear map on homology, $f_p^{i,j} : H_p(K^i) \to H_p(K^j)$. □

Definition 2.18 (Persistent homology group)**.** For $i \leq j$, the **p-dimensional persistent homology group** from level i to level j is

$$H_p^{i,j} = \operatorname{im} f_p^{i,j}.$$

Thus, $H_p^{i,j}$ consists of the p-dimensional homology classes that are already present at level i and remain nontrivial at level j [6, 15].

Proposition 2.19 (Chain-level description of persistent homology)**.** *Let*

$$Z_p^i = Z_p(K^i), \qquad B_p^j = B_p(K^j).$$

Then

$$H_p^{i,j} \cong Z_p^i / (B_p^j \cap Z_p^i).$$

Proof. Define a map $\varphi : Z_p^i \to H_p(K^j)$ by sending each cycle $z \in Z_p^i$ to its homology class $[z]$ in $H_p(K^j)$. Since $K^i \subseteq K^j$, every cycle in K^i can also be regarded as a cycle in K^j.

The image of φ is exactly the image of the induced homology map $f_p^{i,j} : H_p(K^i) \to H_p(K^j)$. Hence $\operatorname{im} \varphi = H_p^{i,j}$. The kernel of φ consists of those cycles in Z_p^i that become boundaries in K^j, namely $\ker \varphi = B_p^j \cap Z_p^i$. By the first isomorphism theorem,

$$Z_p^i / (B_p^j \cap Z_p^i) \cong H_p^{i,j}.$$

□

Definition 2.20 (Persistent Betti number)**.** The p**-dimensional persistent Betti number** from level i to level j is

$$\beta_p^{i,j} = \dim H_p^{i,j}.$$

The number $\beta_p^{i,j}$ counts independent p-dimensional homology classes that persist from K^i to K^j.

2.5 Birth, Death, and Persistence

Definition 2.21 (Birth)**.** A p-dimensional homology class is **born** at level i if it appears in $H_p(K^i)$ but is not in the image of the induced map $H_p(K^{i-1}) \to H_p(K^i)$. For $i = 0$, every nonzero class in $H_p(K^0)$ is regarded as born at level 0.

Definition 2.22 (Death)**.** A p-dimensional homology class born at level i **dies entering** level j if it stops to represent an independent homology class when passing from K^{j-1} to K^j. Equivalently, it either becomes a boundary in K^j or merges with an older homology class under the standard persistence pairing convention.

The standard pairing convention says that when two homology classes merge, the class born earlier is regarded as surviving, while the class born later is regarded as dying [6, 15].

Definition 2.23 (Birth–death pair)**.** If a p-dimensional homology class is born at level i and dies entering level j, then the corresponding **birth–death pair** is $(i, j), i < j$. If the filtration is indexed by real parameters $\alpha_0 < \alpha_1 < \cdots < \alpha_m$, then the corresponding birth–death pair is written as (α_i, α_j).

Definition 2.24 (Persistence)**.** The **persistence** of a birth–death pair (b, d) is $\text{pers}(b, d) = d - b$.

A feature with large persistence survives over a long range of filtration values and is often interpreted as more significant. A feature with small persistence lies close to the diagonal in the birth–death plane and is often interpreted as less robust.

2.6 Persistence Diagrams

For a finite filtration, birth–death multiplicities can be described through persistent Betti numbers. Since the filtration is finite and homology is taken over a field, the associated persistence module decomposes into interval modules. The persistent Betti numbers therefore determine the interval multiplicities by the following rank-invariant formula.

Theorem 2.25 (Multiplicity formula)**.** *Let $i < j$. The number of p-dimensional homology classes born at level i and dying entering level j is*

$$\mu_p^{i,j} = \left(\beta_p^{i,j-1} - \beta_p^{i,j}\right) - \left(\beta_p^{i-1,j-1} - \beta_p^{i-1,j}\right),$$

with the convention that terms involving $i-1$ are zero when $i = 0$.

Proof. By the interval decomposition theorem, the persistent Betti number $\beta_p^{i,j}$ counts the number of p-dimensional persistence intervals whose birth index is at most i and whose death index is greater than j. Therefore, the difference $\beta_p^{i,j-1} - \beta_p^{i,j}$ counts the number of p-dimensional intervals whose birth index is at most i and whose death index is exactly j. Similarly, $\beta_p^{i-1,j-1} - \beta_p^{i-1,j}$ counts the number of such intervals whose birth index is at most $i-1$ and whose death index is exactly j. Subtracting the second quantity from the first removes the intervals born before level i. Hence the remaining number is exactly the number of p-dimensional homology classes born at level i and dying entering level j. □

Definition 2.26 (Persistence diagram)**.** Let

$$K_\bullet : K^0 \subseteq K^1 \subseteq \cdots \subseteq K^m$$

be a filtration indexed by real parameters $\alpha_0 < \alpha_1 < \cdots < \alpha_m$. The p-dimensional persistence diagram of $K_\bullet$, denoted by $\mathrm{Dgm}_p(K_\bullet)$, is the multiset in the extended birth–death plane $\mathbb{R} \times (\mathbb{R} \cup \{\infty\})$ defined as follows.

For every finite birth–death pair with $i < j$, the point (α_i, α_j) is included with multiplicity $\mu_p^{i,j}$.

If a p-dimensional homology class is born at level i and does not die within the filtration, then the point (α_i, ∞) is included with the corresponding multiplicity. Equivalently, the multiplicity of the infinite point born at level i is $\mu_p^{i,\infty} = \beta_p^{i,m} - \beta_p^{i-1,m}$, with the convention that terms involving $i-1$ are zero when $i = 0$.

Each point in $\mathrm{Dgm}_p(K_\bullet)$ records the birth time and death time, finite or infinite, of a p-dimensional homology class. Points are counted with multiplicity.

Definition 2.27 (Diagonal)**.** The **diagonal** in the birth–death plane is $\Delta = \{(x, x) : x \in \mathbb{R}\}$.

A point far from the diagonal corresponds to a long-lived homology class, while a point close to the diagonal corresponds to a short-lived class.

Remark 2.28 (Finite and infinite intervals)*.* The persistence diagram $\mathrm{Dgm}_p(K_\bullet)$ may contain both finite points $(b, d), b < d < \infty$, and infinite points (b, ∞). Infinite points correspond to homology classes that are born during the filtration but do not die before the final complex K^m.

In the vectorization methods used later in this thesis, we restrict attention to finite off-diagonal points

$$(b, d) \in \mathrm{Dgm}_p(K_\bullet), \qquad b < d < \infty.$$

These finite birth–death pairs are the input for persistence landscapes and persistence images.

A persistence diagram is a multiset of points rather than a vector in a Euclidean space. Therefore, although persistence diagrams summarize the evolution of homology across a filtration, they are not directly compatible with many standard statistical and machine learning methods. This is the motivation of the vectorized representations introduced in the next section.

2.7 Vietoris–Rips Filtrations

We now describe the main way in which filtrations are obtained from point-cloud data in this thesis. The Vietoris–Rips construction is naturally formulated in the language of abstract simplicial complexes: it declares certain finite subsets of a metric space to be simplices

according to a pairwise-distance rule [2, 15].

Definition 2.29 (Vietoris–Rips complex)**.** Let (X, d) be a finite metric space and let $\varepsilon \geq 0$. The **Vietoris–Rips complex** of X at scale ε is the abstract simplicial complex

$$\mathrm{VR}_\varepsilon(X) = \{\emptyset \neq \sigma \subseteq X : d(x, y) \leq \varepsilon \text{ for all } x, y \in \sigma\}.$$

Proposition 2.30 (Vietoris–Rips complexes are abstract simplicial complexes)**.** *For every finite metric space (X, d) and every $\varepsilon \geq 0$, $\mathrm{VR}_\varepsilon(X)$ is an abstract simplicial complex.*

Proof. Let $\sigma \in \mathrm{VR}_\varepsilon(X)$. Then for every pair of vertices $x, y \in \sigma$, $d(x, y) \leq \varepsilon$. Let $\emptyset \neq \tau \subseteq \sigma$. Every pair of vertices of τ is also a pair of vertices of σ. Therefore, for every $x, y \in \tau$, $d(x, y) \leq \varepsilon$. Hence $\tau \in \mathrm{VR}_\varepsilon(X)$. Thus every nonempty face of every simplex in $\mathrm{VR}_\varepsilon(X)$ also belongs to $\mathrm{VR}_\varepsilon(X)$. Therefore $\mathrm{VR}_\varepsilon(X)$ is an abstract simplicial complex under the convention used in this thesis. □

Proposition 2.31 (Vietoris–Rips filtration)**.** *Let $0 \leq \varepsilon_0 < \varepsilon_1 < \cdots < \varepsilon_m$. Then*

$$\mathrm{VR}_{\varepsilon_0}(X) \subseteq \mathrm{VR}_{\varepsilon_1}(X) \subseteq \cdots \subseteq \mathrm{VR}_{\varepsilon_m}(X).$$

Therefore, the Vietoris–Rips construction gives a filtration.

Proof. Suppose $\sigma \in \mathrm{VR}_{\varepsilon_i}(X)$. Then for every pair of vertices $x, y \in \sigma$, $d(x, y) \leq \varepsilon_i$. If $i \leq j$, then $\varepsilon_i \leq \varepsilon_j$. Therefore $d(x, y) \leq \varepsilon_j$ for every pair $x, y \in \sigma$. Hence $\sigma \in \mathrm{VR}_{\varepsilon_j}(X)$. Thus

$$\mathrm{VR}_{\varepsilon_i}(X) \subseteq \mathrm{VR}_{\varepsilon_j}(X)$$

whenever $i \leq j$. This proves that the Vietoris–Rips complexes form a filtration. □

Given a finite metric space (X, d) and an increasing sequence of scale parameters, the Vietoris–Rips filtration produces a filtration of abstract simplicial complexes. Therefore, the persistent homology and persistence diagrams defined above apply directly to Vietoris–Rips filtrations.

3. Vectorization of Persistence Diagrams

As we mentioned before, persistence diagrams are not directly suitable for many standard statistical and machine learning methods since they are not vectors in a Euclidean space, and distances between diagrams, such as the bottleneck and Wasserstein distances, are based on matching problems [5]. Therefore, persistence diagrams are often transformed into vectorized representations.

This section introduces two representative vectorization methods: persistence landscapes [3] and persistence images [1]. The persistence landscape maps a diagram to a sequence of functions, while the persistence image maps a diagram to a finite-dimensional matrix.

3.1 Persistence Landscape

As discussed before, the p-dimensional persistent homology of a finite filtration can be represented by a persistence diagram $\mathrm{Dgm}_p(K_\bullet)$. When we do computations, the points of this diagram may be obtained by reducing the boundary matrix of the filtration.

In the following definition, we use only finite off-diagonal birth–death pairs $(b, d) \in \mathrm{Dgm}_p(K_\bullet)$ with $b < d < \infty$. Intervals with infinite death time are excluded from the landscape vectorization in this thesis.

For each finite point $(b, d) \in \mathrm{Dgm}_p(K_\bullet)$ with $b < d < \infty$, define the associated tent function

$$\Lambda_{(b,d)}(t) = \max\{0, \min(t - b, d - t)\}.$$

Equivalently,

$$\Lambda_{(b,d)}(t) = \begin{cases} t - b, & b \le t \le \frac{b+d}{2}, \\ d - t, & \frac{b+d}{2} \le t \le d, \\ 0, & \text{otherwise.} \end{cases}$$

This function is supported on $[b, d]$, reaches its maximum $\frac{d-b}{2}$ at the midpoint $\frac{b+d}{2}$, and is zero outside $[b, d]$.

For each $k \in \mathbb{N}$, the k-th persistence landscape function $\lambda_k : \mathbb{R} \to \mathbb{R}$ is defined by

$$\lambda_k(t) = k\text{- max}\left\{\Lambda_{(b,d)}(t) : (b,d) \in \text{Dgm}_p(K_\bullet)\right\},$$

where k- max denotes the k-th largest value in the multiset above. Diagram points are counted according to their multiplicities. If fewer than k tent functions are nonzero at t, then $\lambda_k(t) = 0$.

The persistence landscape associated with $\text{Dgm}_p(K_\bullet)$ is the sequence of functions $\lambda = (\lambda_1, \lambda_2, \lambda_3, \ldots)$. Thus, a persistence landscape should be viewed not as a single function, but as a sequence of functions obtained by ordering the tent functions pointwise.

This sequence can be regarded as an element of a suitable L^q-type function space. For $1 \leq q < \infty$, its landscape norm is commonly written as

$$\|\lambda\|_q = \left(\sum_{k=1}^{\infty} \int_{\mathbb{R}} |\lambda_k(t)|^q \, dt\right)^{1/q}.$$

Equivalently,

$$\|\lambda\|_q = \left(\sum_{k=1}^{\infty} \|\lambda_k\|_{L^q(\mathbb{R})}^q\right)^{1/q}.$$

This notation emphasizes that the norm is taken over the entire landscape sequence, rather than over a single landscape function.

3.2 Persistence Image

Persistence images provide another vectorized representation of a persistence diagram. They transform a diagram into a finite-dimensional representation that can be used in statistical and machine learning algorithms.

Let

$$D_p := \text{Dgm}_p(K_\bullet) = \{(b_\ell, d_\ell)\}_{\ell \in I}$$

denote the p-dimensional persistence diagram. As in the persistence-landscape construction, we restrict attention to finite off-diagonal points.

The first step is to transform each birth–death pair into birth–persistence coordinates:

$T(b_\ell, d_\ell) = (b_\ell, d_\ell - b_\ell)$. Write $(u_\ell, v_\ell) = T(b_\ell, d_\ell)$, so that $u_\ell = b_\ell, v_\ell = d_\ell - b_\ell$. Thus, v_ℓ is the persistence of the corresponding homological feature.

Let $w : \mathbb{R} \times \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ be a nonnegative weighting function on the birth–persistence plane. In practice, w is usually chosen to increase with persistence and to vanish, or become small, near $v = 0$, thereby reducing the contribution of short-lived features close to the diagonal.

For each transformed point (u_ℓ, v_ℓ), we associate a smoothing kernel centered at (u_ℓ, v_ℓ). A common choice is the two-dimensional Gaussian kernel

$$\phi_{\sigma,(u_\ell,v_\ell)}(x, y) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{(x - u_\ell)^2 + (y - v_\ell)^2}{2\sigma^2}\right),$$

where $\sigma > 0$ is a smoothing parameter.

The persistence surface associated with D_p is defined by

$$\rho_{D_p}(x, y) = \sum_{\ell \in I} w(u_\ell, v_\ell)\phi_{(u_\ell,v_\ell)}(x, y).$$

This surface is a smoothed and weighted representation of the transformed persistence diagram in the birth–persistence plane.

Let $\{P_{rs}\}_{1\leq r\leq m,\, 1\leq s\leq n}$ be a finite $m \times n$ grid of pixels covering a prescribed region of the birth–persistence plane. The persistence image is the matrix whose (r, s)-entry is

$$\mathrm{PI}_{rs}(D_p) = \iint_{P_{rs}} \rho_{D_p}(x, y)\, dx\, dy.$$

Thus, the persistence image is obtained by integrating the weighted persistence surface over each pixel of the grid. In the numerical implementation, the pixel integral is approximated by evaluating the persistence surface at the pixel center and multiplying by the pixel area.

After vectorizing the resulting matrix, the persistence image becomes an ordinary finite-dimensional feature vector. The grid resolution, smoothing parameter, and weighting function all affect the final representation, so these choices are important in applications.

4. Application I: Financial Market Analysis

4.1 Problem Setting

This section reproduces the financial market application of Gidea and Katz [9], who applied topological data analysis to daily returns of four major US stock market indices. Their method constructs rolling point clouds from multivariate return data, computes one-dimensional persistent homology, and summarizes the resulting persistence landscapes by their L^q-type norms.

The purpose of this reproduction is to examine whether the main empirical patterns reported in the original study can be recovered under a individual pipeline. We mainly focus on whether the persistence landscape norms display obvious changes around the 2000 technology crash and the 2008 global financial crisis. The analysis should be understood as retrospective and exploratory: it studies selected historical crisis windows and does not by itself establish an out-of-sample forecasting rule.

The reproduction focuses on:

- the long-term evolution of the normalized L^1 and L^2 norms of persistence landscapes over the sample period;
- the localized behavior of the L^1 norm in the 1,000 trading days preceding the 2000 and 2008 crisis dates;
- secondary rolling indicators based on the variance, low-frequency spectral behavior, and lag-1 autocorrelation of the L^1 norm.

4.2 Methodology

The persistent homology in this reproduction were computed by using the GUDHI [12] library. The overall workflow follows the financial-market pipeline of Gidea and Katz [9]: daily market returns are converted into a sequence of rolling point clouds, persistent homology is computed on each point cloud, and persistence landscapes are summarized by their L^q-type norms.

Data Preprocessing and Log-returns. We consider $d = 4$ daily time series of adjusted closing prices $P_{i,j}$ for the S&P 500, DJIA, NASDAQ, and Russell 2000 indices. In the original paper, the data were downloaded from Yahoo Finance. In this reproduction, the historical price series covering the period from 1988 to 2016 were assembled from Stooq, The Wall Street Journal, and the Federal Reserve Economic Data (FRED) database [13, 14, 7] . This difference in data source may contribute to discrepancies between the reproduced figures and the original results.

To reduce non-stationarity and focus on relative price changes, we compute the daily log-returns $r_{i,j} = \log\left(\frac{P_{i,j}}{P_{i-1,j}}\right)$, where i denotes the trading day and $j \in \{1, \ldots, 4\}$ identifies the index. Each day i is therefore represented by a vector $x_i = (r_{i,1}, r_{i,2}, r_{i,3}, r_{i,4}) \in \mathbb{R}^4$. The resulting four-dimensional return vectors form the input time series for the topological analysis.

Sliding Window and Point Cloud Construction. To capture the evolution of the market state, we apply a sliding window of size w. For each time index n, we collect w consecutive return vectors and form a point cloud $X_n = \{x_n, x_{n+1}, \ldots, x_{n+w-1}\} \subset \mathbb{R}^4$. We use a window size of $w = 50$ trading days with a sliding step of one day. The topological statistic computed from X_n is time-stamped at the end of the window, namely at trading day $n + w - 1$.

Persistent Homology and Vietoris–Rips Filtration. For each rolling point cloud X_n, we construct a Vietoris–Rips filtration using the Euclidean distance in $\mathbb{R}^4$. At scale ε, the Vietoris–Rips complex $\mathrm{VR}_\varepsilon(X_n)$ includes a k-simplex whenever all pairwise distances among its $k + 1$ vertices are at most ε. As ε increases, this produces a nested family of simplicial complexes:

$$\mathrm{VR}_{\varepsilon_1}(X_n) \subseteq \mathrm{VR}_{\varepsilon_2}(X_n) \subseteq \cdots \subseteq \mathrm{VR}_{\varepsilon_m}(X_n), \quad \varepsilon_1 < \varepsilon_2 < \cdots < \varepsilon_m.$$

For each rolling point cloud, we compute the persistence diagram of the first homology group H_1. In the implementation, the Vietoris–Rips complex is constructed up to dimension 2,

which is sufficient for computing H_1.

Persistence Landscapes and Landscape Norms. Given a finite birth–death pair (b_i, d_i) in a persistence diagram, we define the associated triangular tent function $\Lambda_i : \mathbb{R} \to [0, \infty)$ by $\Lambda_i(t) = \max\{0, \min(t - b_i, d_i - t)\}$. This function is centered at $(b_i + d_i)/2$ and reaches height $(d_i - b_i)/2$.

The persistence landscape is then defined as a sequence of functions $\lambda = \{\lambda_k\}_{k\in\mathbb{N}}$, where $\lambda_k(t)$ is the k-th largest value among the collection $\{\Lambda_i(t)\}_i$ for each $t \in \mathbb{R}$: $\lambda_k(t) = k\text{-}\max\{\Lambda_i(t) : i = 1, \ldots, N\}$. To summarize the strength of persistent topological features by a scalar quantity, we compute the L^q-norm of the persistence landscape:

$$\|\lambda\|_q = \left(\sum_{k=1}^{\infty} \int_{\mathbb{R}} |\lambda_k(t)|^q \, dt \right)^{1/q}, \qquad 1 \le q < \infty.$$

In this study, we focus on the L^1 and L^2 norms of the persistence landscapes derived from H_1. These norms provide scalar summaries of persistent loop-like features in the four-dimensional return space. In the numerical implementation, the landscape representation is discretized, and the corresponding integrals are approximated numerically. If a rolling point cloud has no finite H_1 interval, its landscape norm is set to zero.

Normalization for Visualization. For visualization, the computed landscape-norm series are rescaled by min–max normalization, $\widetilde{x}_t = \frac{x_t - \min_s x_s}{\max_s x_s - \min_s x_s}$. The normalization range depends on the reproduced figure. In the long-term plot, the minimum and maximum are computed over the whole sample period. In the localized plots, the time window is first extracted and the normalization is then applied within that window.

Statistical Indicators for Early-Warning Analysis. To investigate possible pre-crisis behavior of the topological signal, we analyze the raw L^1-norm series using a secondary rolling window of 500 trading days. This secondary window is also trailing: for each date, the corresponding indicator is computed from the 500 observations ending at that date.

Within each 500-day window, we compute the sample variance, the average spectral density at low frequencies, and the lag-1 autocorrelation. The low-frequency spectral indicator is obtained from a spectral estimate of the demeaned L^1-norm series, and is summarized by averaging over the lowest positive frequency range. These derived indicators are used as exploratory summaries of the topological signal rather than as a calibrated forecasting model.

4.3 Long-term Empirical Results: Reproduction of Figure 9 [9]

Following the procedure described above, we computed the daily time series of the normalized L^1 and L^2 norms of persistence landscapes. This analysis provides a macroscopic view of the evolution of the topological signal over nearly three decades.

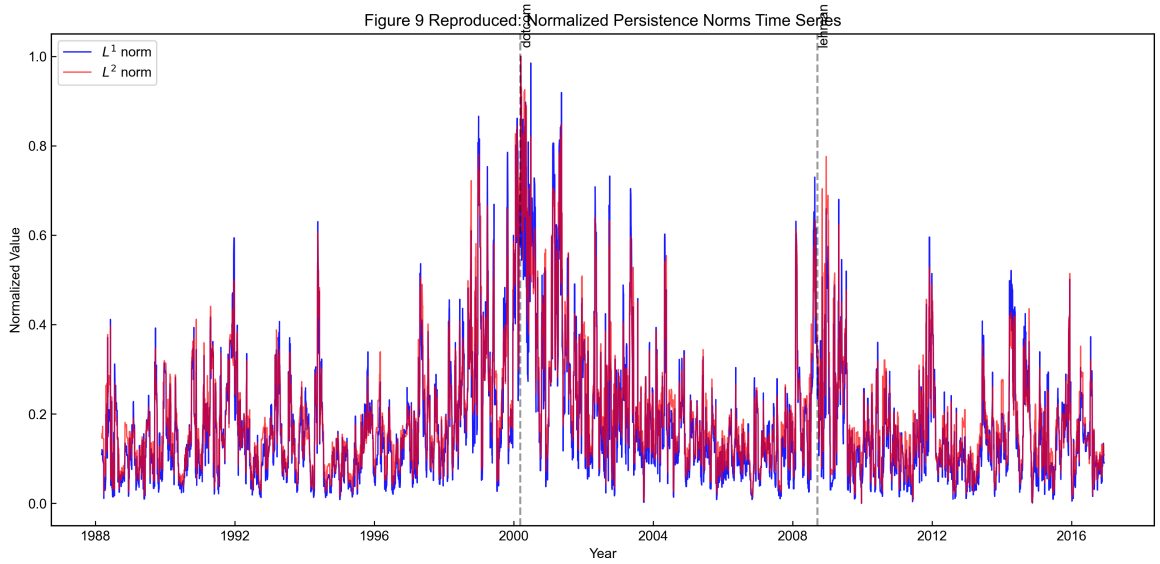

Figure 1. Time series of normalized L^1 (blue) and L^2 (red) norms of persistence landscapes for the four US stock market indices (1988–2016). Vertical dashed lines indicate the Dot-com crash (2000) and the Lehman Brothers bankruptcy (2008).

The reproduced Figure 1 is qualitatively consistent with the main pattern reported in the original study. The normalized L^1 and L^2 norms display obvious spikes during periods of severe market turbulence. In particular:

- **Dot-com crash (2000):** One of the most prominent peaks in the entire sample occurs around March 2000. This indicates that the persistence landscape becomes substantially larger during the collapse of the technology bubble, reflecting a stronger H_1

features.

- **Global Financial Crisis (2008):** A second major cluster of elevated values appears around the Lehman Brothers bankruptcy in September 2008.

The close co-movement between the L^1 and L^2 norms suggests that the increase in persistent topological structure is not specific to one particular choice of q.

4.4 Localized Results: Reproduction of Figure 10 [9]

To better assess the potential of TDA as an early-warning tool, we next examine the 1,000 trading days preceding the technology crash of 2000 and the Lehman Brothers bankruptcy of 2008. Figure 2 and Figure 3 compare the normalized S&P 500 index with the corresponding L^1 norm of the persistence landscape.

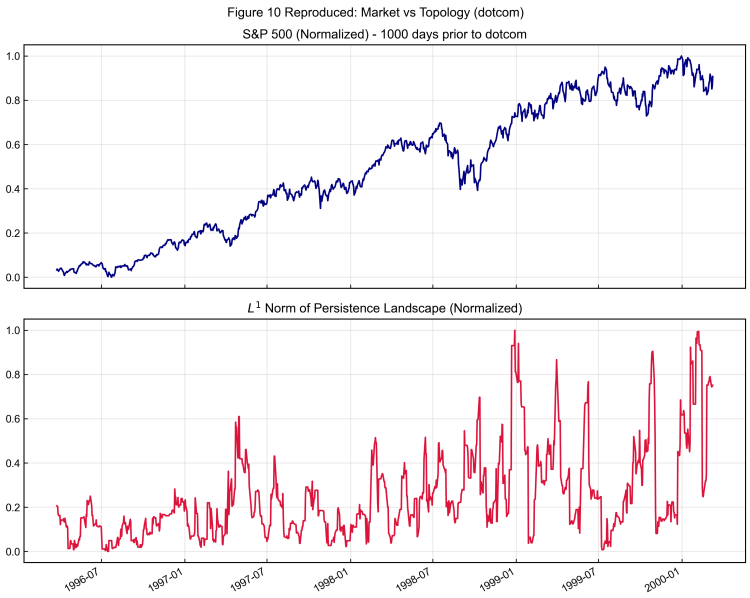

Figure 2. S&P 500 index (top) and L^1 norm (bottom) for the 1,000 trading days prior to March 2000.

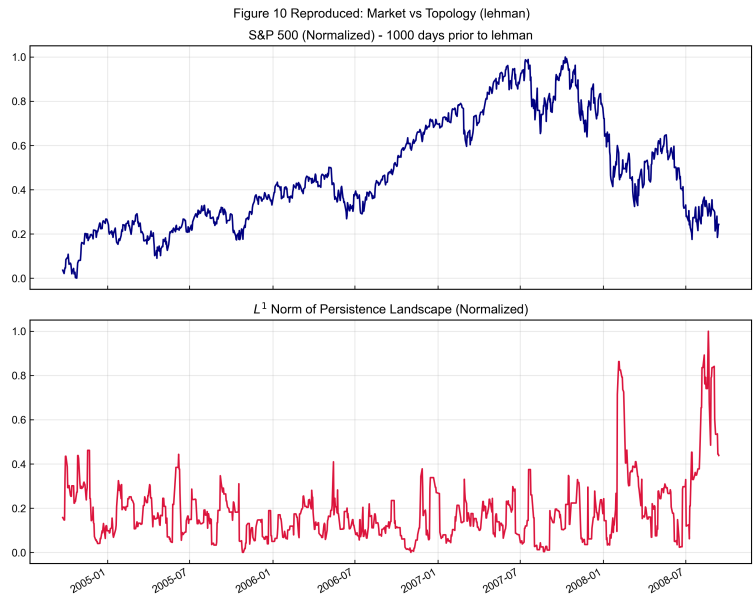

Figure 3. S&P 500 index (top) and L^1 norm (bottom) for the 1,000 trading days prior to September 2008.

In Figure 2, the S&P 500 exhibits a strong but volatile upward movement from 1996 through early 2000. The topological signal, however, reveals a more irregular pattern of growing instability:

- beginning as early as 1997, the L^1 norm displays several spikes that gradually increase in frequency and magnitude;
- a major surge appears in early 1999, well before the primary market peak.

Figure 3 shows the lead-up to the 2008 crisis. In this case, the S&P 500 peaks in late 2007 and then begins to weaken. The corresponding topological signal remains comparatively decreasing through 2005 and 2006, but rises substantially once financial stress becomes more visible:

- starting in late 2007, the L^1 norm begins to increase;
- a major spike appears in early 2008, followed by another sharp rise close to the Lehman bankruptcy in September 2008.

In general, these localized plots suggest that, in the two reproduced crisis windows, the topological signal becomes elevated before or around the most dramatic phase of the market

breakdown. This is consistent with the exploratory interpretation of persistence landscape norms as potential early-warning indicators, but it does not establish a general forecasting rule.

4.5 Derived Statistical Indicators: Reproduction of Figure 11 [9]

To further examine the pre-crash behavior of the topological signal, we compute three secondary indicators from the raw L^1-norm series using a 500-trading-day trailing rolling window: variance, average spectral density at low frequencies, computed from a spectral estimate over the lowest positive frequency range, and the lag-1 autocorrelation (ACF lag-1). Figure 4 and Figure 5 show these metrics for the 250 trading days preceding the March 2000 and September 2008 crashes.

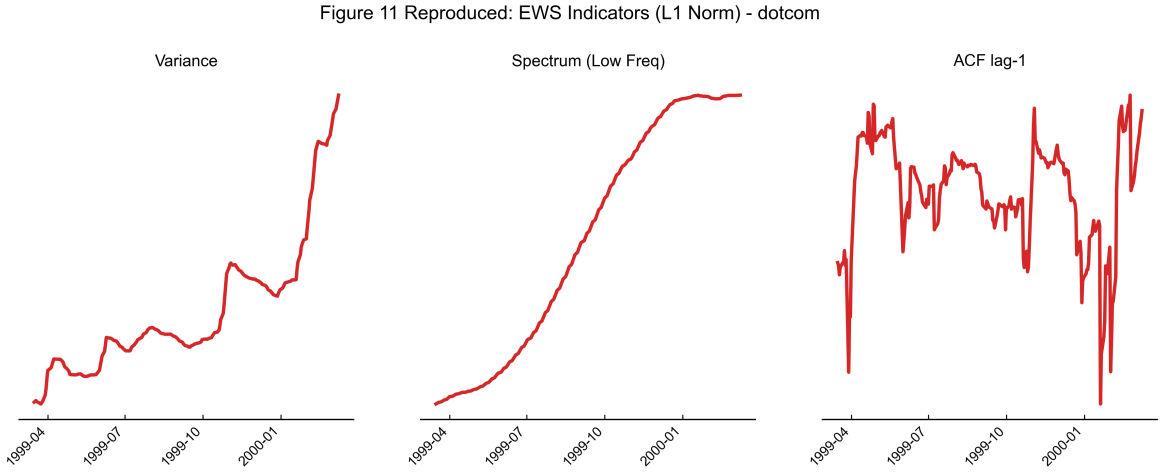

Figure 4. Reproduced early-warning indicators for the 250 trading days prior to the Dot-com crash (March 10, 2000). The panels show the variance, low-frequency spectral density, and ACF lag-1 of the L^1 norm.

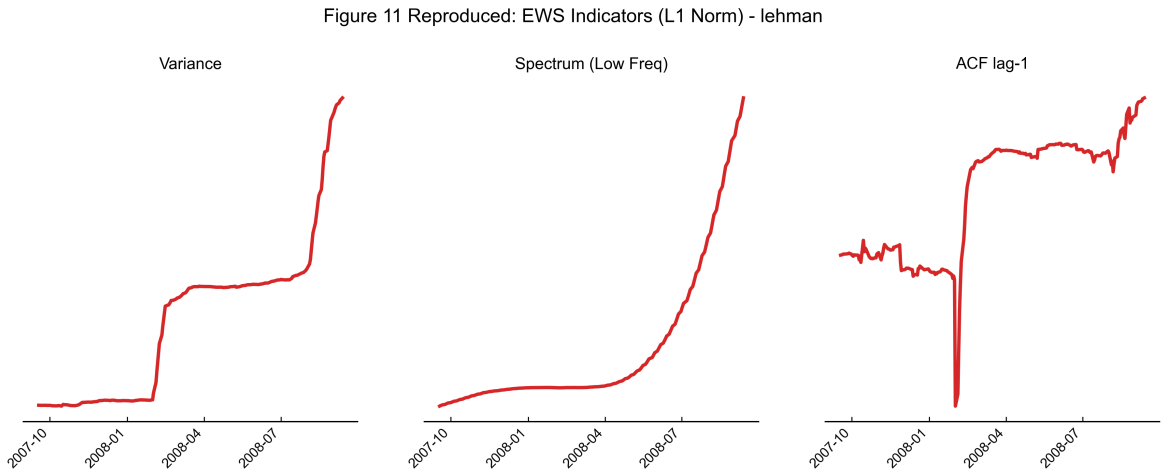

Figure 5. Reproduced early-warning indicators for the 250 trading days prior to the Lehman Brothers bankruptcy (September 15, 2008).

The statistical indicators reveal distinct patterns, although their stability differs across measures:

- **Variance:** In both crisis periods, the variance of the L^1 norm shows a clear upward trend.
- **Spectrum (low frequency):** In the 250 trading days preceding both crashes, the average spectral density at low frequencies generally increases. This indicates that the fluctuations of the topological signal become increasingly dominated by slower-moving components.

By contrast, the lag-1 autocorrelation is considerably less stable. In the original study, the ACF does not display a consistent upward trend prior to either crash, and our reproduction likewise shows substantial fluctuations, with occasional sharp drops, especially during the Lehman period. This suggests that, in this setting, the low-frequency spectrum is a more robust indicator than the raw lag-1 autocorrelation.

4.6 Summary and Reflection

Our reproduced indicators differ in some details from those reported in the original paper [9] . These discrepancies may arise from differences in data source, software implementation, numerical approximation, and plotting conventions. Nevertheless, the main qualitative

pattern is broadly consistent with the original study: in the two selected crisis windows, the persistence-landscape norms and their derived indicators become elevated before or around the most severe phase of market stress.

At the same time, we can see why this experiment should be interpreted as a retrospective exploratory reproduction rather than a formal forecasting test. For example, the crisis windows are chosen around known historical events. The conclusion is mainly deduced by visual patterns, descriptive comparison, lacks a complete out-of-sample forecasting validation with non-crisis comparison windows and standard financial baselines.

The current situation in academic field also fits my opinions, the financial interpretation of these topological signals remains partly unresolved. Explanations based on critical transitions, spectral reddening, or systemic stress should be understood as heuristic interpretations rather than established economic mechanisms.

We can see the potential of TDA in areas like quantitative trading, however, a full validation of predictive value would require further studies.

5. Application II: Machine Learning Application

5.1 Problem Setting

In this application, we reproduce part of the geometric-shape experiment from Adams et al. [1] and treat it as a concrete TDA workflow for unsupervised clustering. The dataset contains six synthetic shape classes: *solid cube*, *circle*, *sphere*, *three clusters*, *hierarchical clusters*, and *torus*.

The task is to evaluate whether persistence image features preserve shape information under noise. We compute persistence diagrams in dimensions H_0 and H_1, transform them into persistence images, build pairwise distance matrices with L^1, L^2, and L^∞ metrics, and perform K-medoids clustering [10] with $K = 6$. Clustering quality is reported as best match accuracy after optimal label alignment.

5.2 Methodology

Our pipeline is:

$$\begin{aligned}\text{Point Cloud} &\longrightarrow \text{Persistence Diagram} \longrightarrow \text{Persistence Image}\\ &\longrightarrow \text{Distance Matrix} \longrightarrow \text{K-medoids and Evaluation.}\end{aligned}$$

Throughout this application, let $\eta \in \{0.05, 0.10\}$ denote the noise level. For each fixed η, let $a = 1, \ldots, N_\eta$, $N_\eta = 150$, index the point-cloud samples in the same noise group. We use $k \in \{0, 1\}$ for the homological dimension.

Data Generation. Following the synthetic shape experiment in the original paper [1], we consider six shape classes: solid cube, circle, sphere, three clusters, hierarchical clusters, and torus. For each shape class, each noise level $\eta \in \{0.05, 0.10\}$, and each sample index a, we generate a clean point set $X_a = \{x_{a,i}\}_{i=1}^{500} \subset \mathbb{R}^3$ from the corresponding geometric model, then perturb it by isotropic Gaussian noise:

$$\widetilde{x}_{a,i}^{(\eta)} = x_{a,i} + \varepsilon_{a,i}^{(\eta)}, \qquad \varepsilon_{a,i}^{(\eta)} \sim \mathcal{N}(0, \eta^2 I_3), \qquad i = 1, \ldots, 500.$$

Where I_3 denotes the 3×3 identity matrix, so the noise is independent across coordinates with variance η^2 in each coordinate. This yields the noisy point cloud $\widetilde{X}_a^{(\eta)} = \{\widetilde{x}_{a,i}^{(\eta)}\}_{i=1}^{500}$. For each noise level, we generate 25 point clouds per class, giving 150 point clouds in total. In details:

Table 1. Mathematical definitions of the synthetic shape classes.

Shape class	Clean point-cloud generation
`solid_cube`	$x_i \sim \mathrm{Unif}([0,1]^3)$.
`circle`	$\theta_i \sim \mathrm{Unif}(0, 2\pi)$, $x_i = (\cos\theta_i, \sin\theta_i, 0)$.
`sphere`	$\phi_i \sim \mathrm{Unif}(0, 2\pi)$, $s_i \sim \mathrm{Unif}(-1, 1)$, and $x_i = (\sqrt{1-s_i^2}\cos\phi_i, \sqrt{1-s_i^2}\sin\phi_i, s_i)$.
`three_clusters`	Centers are $c_1 = (0,0,0)$, $c_2 = (2,2,2)$, and $c_3 = (4,0,4)$. Points are allocated equally among the three centers, and $x_i = c_j + \xi_i$, where $\xi_i \sim \mathcal{N}(0, 0.1^2 I_3)$.
`hierarchical_clusters`	Main centers are $c_1 = (0,0,0)$, $c_2 = (5,5,5)$, and $c_3 = (10,0,10)$. Offsets are $o_1 = (0.5,0,0)$, $o_2 = (0,0.5,0)$, and $o_3 = (0,0,0.5)$. Points are allocated equally among the nine combinations of main center and offset, and sampled as $x_i = c_j + o_h + \xi_i$, where $\xi_i \sim \mathcal{N}(0, 0.05^2 I_3)$.
`torus`	$\theta_i, \phi_i \sim \mathrm{Unif}(0, 2\pi)$, and $x_i = ((R_{\text{tor}} + r_{\text{tor}}\cos\theta_i)\cos\phi_i, (R_{\text{tor}} + r_{\text{tor}}\cos\theta_i)\sin\phi_i, r_{\text{tor}}\sin\theta_i)$, where $R_{\text{tor}} = 2.0$ and $r_{\text{tor}} = 0.6$.

Persistent Homology. Given the noisy point cloud $\widetilde{X}_a^{(\eta)}$ endowed with the ambient Euclidean metric, we construct the Vietoris–Rips filtration and compute persistent homology in dimensions 0 and 1. In the numerical implementation, the Vietoris–Rips complex is constructed up to dimension 2, which is sufficient for computing H_0 and H_1.

For each sample a, noise level η, and homological dimension $k \in \{0, 1\}$, this produces the persistence diagram

$$D_{a,k}^{(\eta),\text{raw}} := \mathrm{Dgm}_k\big(\mathrm{VR}(\widetilde{X}_a^{(\eta)})_\bullet\big).$$

Each point in $D_{a,k}^{(\eta),\text{raw}}$ is a birth–death pair (b_ℓ, d_ℓ), recording the birth and death values of a k-dimensional homological feature.

For the persistence-image construction, we retain only finite off-diagonal points and define

$$D_{a,k}^{(\eta)} := \left\{(b_\ell, d_\ell) \in D_{a,k}^{(\eta),\text{raw}} : d_\ell < \infty\right\},$$

where multiplicities of repeated birth–death pairs are preserved. In particular, for H_0, the essential class that never dies within the filtration is excluded from the vectorized representation. Thus each point-cloud sample is represented, for the purpose of persistence-image

vectorization, by the pair $\big(D_{a,0}^{(\eta)}, D_{a,1}^{(\eta)}\big)$.

Persistence Image Construction. Each finite birth–death point is first transformed into birth–persistence coordinates: $T : (b, d) \mapsto (u, v), u = b, v = d - b$. For each sample a, noise level η, and homological dimension $k \in \{0, 1\}$, write $\widehat{D}_{a,k}^{(\eta)} := T\big(D_{a,k}^{(\eta)}\big)$ for the corresponding multiset of transformed points. Thus $\widehat{D}_{a,k}^{(\eta)} = \{(u_\ell, v_\ell)\}_{\ell \in I_{a,k}^{(\eta)}}$.

For each fixed noise level and homological dimension, the maximum persistence over all samples in the same noise group is defined as

$$v_{*,\eta}^{(k)} := \max\left\{v_\ell : (u_\ell, v_\ell) \in \widehat{D}_{a,k}^{(\eta)} \text{ for some } a = 1, \ldots, N_\eta\right\}.$$

We then use the piecewise linear weight function $w_{\eta,k}(u, v) = \bar{w}_{\eta,k}(v)$, where

$$\bar{w}_{\eta,k}(v) = \begin{cases} 0, & v \le 0, \\ v/v_{*,\eta}^{(k)}, & 0 < v < v_{*,\eta}^{(k)}, \\ 1, & v \ge v_{*,\eta}^{(k)}. \end{cases}$$

Thus, the persistence-image weight and grid range are fitted separately for each noise level and each homological dimension. This is consistent with the fixed-dataset clustering setting of the original experiment.

The Gaussian smoothing kernel is parameterized by its variance. We set $\sigma^2 = 0.1$, and use persistence-image resolution $R_{\mathrm{PI}} = 20$. To keep the notation compatible with the persistence-image construction in Section 3.2, write the one-dimensional Gaussian density centered at v_ℓ as

$$\phi_{\sigma,v_\ell}^{(1)}(t) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(t - v_\ell)^2}{2\sigma^2}\right),$$

and the two-dimensional isotropic Gaussian density centered at (u_ℓ, v_ℓ) as

$$\phi_{\sigma,(u_\ell,v_\ell)}^{(2)}(x, y) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{(x - u_\ell)^2 + (y - v_\ell)^2}{2\sigma^2}\right).$$

For H_0, all connected components in the Vietoris–Rips filtration are born at filtration value zero under the usual convention. Therefore, the birth coordinate carries no additional

information for finite H_0 points, and the persistence coordinate is equivalent to the death coordinate. Following the convention described in the original paper [1], we use a one-dimensional persistence-image representation for H_0, based only on the persistence coordinate, rather than a full two-dimensional birth–persistence image. This is a one-dimensional analogue of the persistence-image construction in Section 3.2, specialized to the H_0 case where all finite components are born at zero.

For H_0, define the one-dimensional persistence surface

$$\rho_{a,0}^{(\eta)}(t) = \sum_{(0,v_\ell)\in \widehat{D}_{a,0}^{(\eta)}} \bar{w}_{\eta,0}(v_\ell)\phi_{\sigma,v_\ell}^{(1)}(t).$$

The H_0 image is a one-dimensional discretization with R_{PI} bins over $[0,\, v_{*,\eta}^{(0)} + 3\sigma]$. Let $\{Q_m^{(\eta,0)}\}_{m=1}^{R_{\mathrm{PI}}}$ be the corresponding one-dimensional grid of intervals, and let t_m be the center of $Q_m^{(\eta,0)}$. The m-th entry of the H_0 persistence image is

$$\mathrm{PI}_m^{(0)}\left(D_{a,0}^{(\eta)}\right) = \int_{Q_m^{(\eta,0)}} \rho_{a,0}^{(\eta)}(t)\, dt.$$

In the numerical implementation, this integral is approximated by $\mathrm{PI}_m^{(0)}\left(D_{a,0}^{(\eta)}\right) \approx \rho_{a,0}^{(\eta)}(t_m)\Delta t$, where Δt is the bin width. If a sample has no finite H_0 points after discarding the essential class, the corresponding H_0 vector is set to zero.

For H_1, we use the usual two-dimensional birth–persistence plane. Define

$$u_{*,\eta}^{(1)} := \max\left\{u_\ell : (u_\ell, v_\ell) \in \widehat{D}_{a,1}^{(\eta)} \text{ for some } a = 1, \ldots, N_\eta\right\}.$$

The H_1 image is constructed on an $R_{\mathrm{PI}} \times R_{\mathrm{PI}}$ grid over $[0,\, u_{*,\eta}^{(1)} + 3\sigma] \times [0,\, v_{*,\eta}^{(1)} + 3\sigma]$. Let $\{P_{rs}^{(\eta,1)}\}_{1\le r,s\le R_{\mathrm{PI}}}$ be the corresponding grid of pixels, and let (x_r, y_s) be the center of $P_{rs}^{(\eta,1)}$.

For H_1, define the persistence surface

$$\rho_{a,1}^{(\eta)}(x,y) = \sum_{(u_\ell,v_\ell)\in \widehat{D}_{a,1}^{(\eta)}} w_{\eta,1}(u_\ell, v_\ell)\phi_{\sigma,(u_\ell,v_\ell)}^{(2)}(x,y).$$

The (r, s)-entry of the H_1 persistence image is

$$\mathrm{PI}_{rs}\big(D_{a,1}^{(\eta)}\big) = \iint_{P_{rs}^{(\eta,1)}} \rho_{a,1}^{(\eta)}(x, y)\, dx\, dy.$$

In the numerical implementation, this integral is approximated by $\mathrm{PI}_{rs}\big(D_{a,1}^{(\eta)}\big) \approx \rho_{a,1}^{(\eta)}(x_r, y_s)\Delta x \Delta y$. Equivalently,

$$\mathrm{PI}_{rs}\big(D_{a,1}^{(\eta)}\big) \approx \sum_{(u_\ell, v_\ell) \in \widehat{D}_{a,1}^{(\eta)}} w_{\eta,1}(u_\ell, v_\ell)\phi^{(2)}_{\sigma,(u_\ell,v_\ell)}(x_r, y_s)\Delta x \Delta y.$$

If a sample has no finite H_1 points, the corresponding H_1 image is set to the zero vector.

After vectorization, define

$$z_{a,0}^{(\eta)} := \mathrm{vec}\Big(\big(\mathrm{PI}_m^{(0)}(D_{a,0}^{(\eta)})\big)_{m=1}^{R_{\mathrm{PI}}}\Big) \in \mathbb{R}^{20},$$

and

$$z_{a,1}^{(\eta)} := \mathrm{vec}\Big(\big(\mathrm{PI}_{rs}(D_{a,1}^{(\eta)})\big)_{1\le r,s\le R_{\mathrm{PI}}}\Big) \in \mathbb{R}^{400}.$$

Although the H_0 and H_1 vectors can be concatenated into a 420-dimensional feature vector, the results reported in this reproduction evaluate the H_0 and H_1 features separately, matching the structure of the comparison table in the original experiment.

Distance Geometry and K-medoids Evaluation. For each noise level η, homological dimension $k \in \{0, 1\}$, and norm index $q \in \{1, 2, \infty\}$, we compute the pairwise distance matrix $\Delta^{(\eta,k,q)} = \big(\Delta_{\eta,k,q}(a, b)\big)_{1\le a,b\le N_\eta}$, where $\Delta_{\eta,k,q}(a, b) = \Big\| z_{a,k}^{(\eta)} - z_{b,k}^{(\eta)} \Big\|_q$. K-medoids clustering is then performed with $K = 6$ by solving

$$\min_{\mathcal{M}\subset\{1,\dots,N_\eta\},\,|\mathcal{M}|=K} \sum_{a=1}^{N_\eta} \min_{m\in\mathcal{M}} \Delta_{\eta,k,q}(a, m),$$

where $\mathcal{M}$ is the set of selected medoid indices.

In the reproduction run, we use $n_{\mathrm{init}} = 50$ random initializations and select the solution with the lowest unsupervised clustering objective. This initialization budget is smaller than

that used in the original paper [1], and may contribute to discrepancies in the reproduced clustering accuracies.

After clustering, the cluster labels are matched to the true shape labels using the Hungarian algorithm [11] applied to the contingency matrix. The reported clustering accuracy is the best-match accuracy after this optimal label alignment. The true class labels are used only in this final evaluation step.

5.3 Results

Table 2 compares our reproduction results with the PI entries reported in the original paper [1]. The comparison is made for the two noise levels $\eta = 0.05$ and $\eta = 0.10$, the two homological dimensions H_0 and H_1, and the three vector norms L^1, L^2, and L^∞.

Table 2. Comparison between reproduction and original PI results (accuracy).

Noise	Component	Metric	Reproduction	Original	Reproduction − Original
0.05	H_0	L^1	1.0000	0.9330	+0.0670
0.05	H_0	L^2	1.0000	0.9270	+0.0730
0.05	H_0	L^∞	1.0000	0.9400	+0.0600
0.05	H_1	L^1	1.0000	1.0000	+0.0000
0.05	H_1	L^2	1.0000	1.0000	+0.0000
0.05	H_1	L^∞	1.0000	1.0000	+0.0000
0.10	H_0	L^1	1.0000	0.9530	+0.0470
0.10	H_0	L^2	1.0000	0.9530	+0.0470
0.10	H_0	L^∞	1.0000	0.9600	+0.0400
0.10	H_1	L^1	0.7867	0.9530	-0.1663
0.10	H_1	L^2	0.9333	0.9600	-0.0267
0.10	H_1	L^∞	0.9400	0.9600	-0.0200

In general, the reproduction supports the main conclusion that persistence image features are effective for this shape-clustering task. At the lower noise level $\eta = 0.05$, both H_0 and H_1 PI features achieve perfect clustering accuracy in this reproduction run. At the higher noise level $\eta = 0.10$, the H_0-based features remain perfectly separated, while the H_1-based features show some degradation, especially under the L^1 metric.

5.4 Summary and Reflection

The reproduced results are broadly consistent with the original study in showing that persistence images provide effective finite-dimensional features for distinguishing the six synthetic shape classes.

The H_0 results should be interpreted carefully. Since the synthetic shapes are not rescaled to a common geometric size, H_0 persistence can reflect metric-geometric information such as scale, density, and cluster separation, rather than purely topological connectivity. This does not invalidate the clustering result, but it affects the interpretation of why the H_0 features separate the classes so well.

The differences between the reproduction and the original may arise from random sampling of the point clouds (the reported accuracies come from a single fixed random seed, with multiple K-medoids initializations inside that run, rather than from a multi-seed average), noise realizations, implementation details in the persistence image construction, and the K-medoids optimization procedure. In particular, this reproduction uses a smaller number of random K-medoids initializations than the original paper [1], which may contribute to the lower H_1 accuracy at $\eta = 0.10$.

A more complete extension would report mean and standard deviation across multiple random seeds, use a larger K-medoids initialization budget, and test the sensitivity of the results to shape normalization and persistence image parameters.

6. Conclusion

This thesis examined persistent homology and two vectorized representations of persistence diagrams, persistence landscapes and persistence images, in two reproductions: financial market analysis and machine learning. In general, our results suggest that these vectorized representations provide feasible ways to transform the information of persistence diagrams into mathematical structures that fits standard computational pipelines. Despite certain limitations, it has already demonstrated potential for application in fields such as finance and machine learning.

References

[1] H. Adams, T. Emerson, M. Kirby, R. Neville, C. Peterson, P. Shipman, S. Chepushtanova, E. Hanson, F. Motta, and L. Ziegelmeier. Persistence images: a stable vector representation of persistent homology. *Journal of Machine Learning Research*, 18(8):1–35, 2017. URL: https://jmlr.org/papers/v18/16-337.html.

[2] J.-D. Boissonnat, F. Chazal, and M. Yvinec. *Geometric and Topological Inference*. Cambridge University Press, Cambridge, 2018.

[3] P. Bubenik. Statistical topological data analysis using persistence landscapes. *Journal of Machine Learning Research*, 16(3):77–102, 2015. URL: https://jmlr.org/papers/v16/bubenik15a.html.

[4] G. Carlsson. Topology and data. *Bulletin of the American Mathematical Society*, 46(2):255–308, 2009. DOI: 10.1090/S0273-0979-09-01249-X.

[5] D. Cohen-Steiner, H. Edelsbrunner, and J. Harer. Stability of persistence diagrams. *Discrete & Computational Geometry*, 37(1):103–120, 2007. DOI: 10.1007/s00454-006-1276-5.

[6] H. Edelsbrunner, D. Letscher, and A. Zomorodian. Topological persistence and simplification. *Discrete & Computational Geometry*, 28(4):511–533, 2002. DOI: 10.1007/s00454-002-2885-2.

[7] Federal Reserve Bank of St. Louis. Federal reserve economic data (FRED). https://fred.stlouisfed.org/, 2026. Accessed: 2026-04-28.

[8] R. Ghrist. Barcodes: the persistent topology of data. *Bulletin of the American Mathematical Society*, 45(1):61–75, 2008. DOI: 10.1090/S0273-0979-07-01191-3.

[9] M. Gidea and Y. Katz. Topological data analysis of financial time series: landscapes of crashes. *Physica A: Statistical Mechanics and its Applications*, 491:820–834, 2018. DOI: 10.1016/j.physa.2017.09.028.

[10] L. Kaufman and P. J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley & Sons, New York, 1990.

[11] H. W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1–2):83–97, Mar. 1955. DOI: 10.1002/nav.3800020109.

[12] C. Maria, J.-D. Boissonnat, M. Glisse, and M. Yvinec. The GUDHI library: simplicial complexes and persistent homology. In *Mathematical Software – ICMS 2014*, volume 8592 of *Lecture Notes in Computer Science*, pages 167–174, Berlin, Heidelberg. Springer, 2014. DOI: 10.1007/978-3-662-44199-2_28.

[13] Stooq. Stooq historical market data. https://stooq.com/db/h/, 2026. Accessed: 2026-04-28.

[14] The Wall Street Journal. Market data. https://www.wsj.com/market-data, 2026. Accessed: 2026-04-28.

[15] A. Zomorodian and G. Carlsson. Computing persistent homology. *Discrete & Computational Geometry*, 33(2):249–274, 2005. DOI: 10.1007/s00454-004-1146-y.

Acknowledgments

I am profoundly grateful to Professor Yifei Zhu for being my guide in topology. His courses provided me with the necessary expertise ranging from point-set topology to persistent homology. My thanks also go to Ms. Xiaoxue Ren and Ms. Yuan Qu for their helpful consultation and administrative support throughout the thesis preparation process. Their assistance made the journey much smoother.