A Family of Density-Scaled Filtered Complexes

Paper Author: *Abigail Hickok*

Presented By **Yikun Zhang**

Department of Statistics,
University of Washington

May 28, 2024

Department of
STATISTICS

UNIVERSITY *of* WASHINGTON

Introduction

High-Dimensional Data Challenges

Nowadays, high-dimensional point cloud data are ubiquitous.

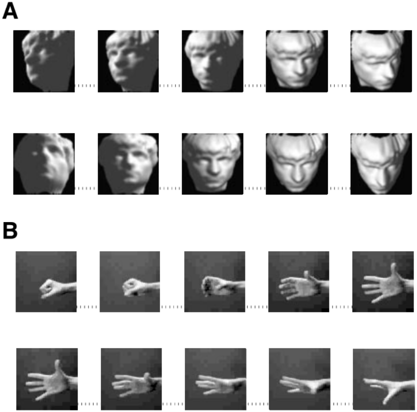


Figure 1: Images with many pixels (Tenenbaum et al., 2000).

High-Dimensional Data Challenges

Nowadays, high-dimensional point cloud data are ubiquitous.

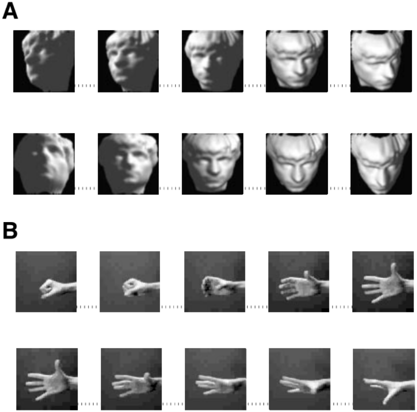


Figure 1: Images with many pixels (Tenenbaum et al., 2000).

- **Challenges:** Analyzing high-dimensional data is statistically and computationally challenging.

- **Manifold Hypothesis (Fefferman et al., 2016):**

High-dimensional data tend to lie in the vicinity of a low dimensional manifold.

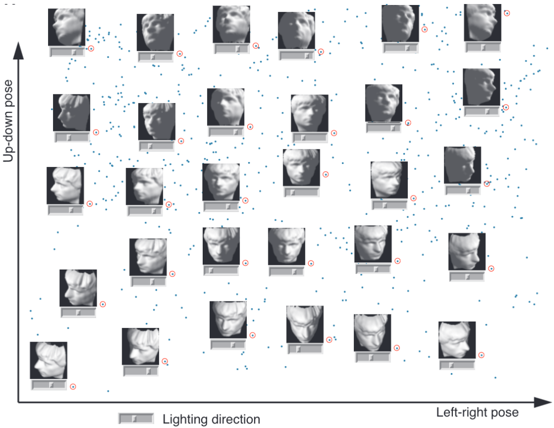


Figure 2: Two-dimensional parameterization of images (Tenenbaum et al., 2000).

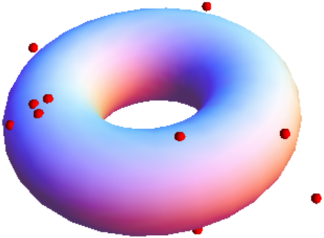
- **Goal:** Infer the homology of the underlying manifold M around which the point cloud $X = \{x_i\}_{i=1}^{N}$ lie.

[1] A simple example of a Calabi–Yau manifold is given by

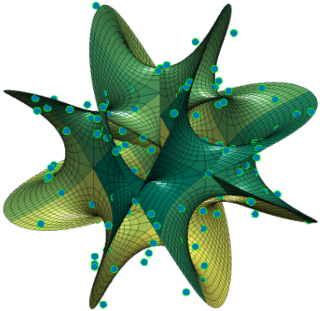
$$x^2 + y^2 + z^2 + w^2 = 0 \quad \text{with} \quad (x, y, z, w) \text{ from the complex projective 3-space.}$$

Homology Inference on Manifolds

- **Goal:** Infer the homology of the underlying manifold M around which the point cloud $X = \{x_i\}_{i=1}^N$ lie.
  - It can distinguish M from other manifolds with different homology.



(a) Data around a two dimensional torus (Fefferman et al., 2016).



(b) Data around the 3D projection of the Calabi-Yau manifold[1] (Yao et al., 2023).

[1] A simple example of a Calabi–Yau manifold is given by

$$x^2 + y^2 + z^2 + w^2 = 0 \quad \text{with} \quad (x, y, z, w) \text{ from the complex projective 3-space.}$$

- **Question:** How can infer the homology of a manifold M with dimension n from the point cloud $X = \{x_i\}_{i=1}^N$?

▶ **Question:** How can infer the homology of a manifold M with dimension n from the point cloud $X = \{x_i\}_{i=1}^N$?

1. Approximate M through a *filtered complex*, which is a collection of simplicial complexes $\{\mathcal{K}_r\}_{r\in\mathbb{R}}$ such that $\mathcal{K}_s \subseteq \mathcal{K}_r$ for all $s \leq r$.

▶ **Question:** How can infer the homology of a manifold M with dimension n from the point cloud $X = \{x_i\}_{i=1}^N$?

1. Approximate M through a *filtered complex*, which is a collection of simplicial complexes $\{\mathcal{K}_r\}_{r\in\mathbb{R}}$ such that $\mathcal{K}_s \subseteq \mathcal{K}_r$ for all $s \leq r$.

   - *Čech Complex* $\check{C}(X) \equiv \check{C}(M, d, X)$: The set of simplices in $\check{C}(M, d, X)_r$ at filtration level r is

$$\left\{ x_J : \bigcap_{j\in J} B(x_j, r) \neq \emptyset \text{ and } J \subseteq \{1, ..., N\} \right\},$$

     where (M, d) is a metric space, x_J denotes the simplex with vertices x_j for all $j \in J$, and $B(x, r) := \{y \in M : d(x, y) \leq r\}$.

▶ **Question:** How can infer the homology of a manifold M with dimension n from the point cloud $X = \{x_i\}_{i=1}^{N}$?

1. Approximate M through a *filtered complex*, which is a collection of simplicial complexes $\{\mathcal{K}_r\}_{r\in\mathbb{R}}$ such that $\mathcal{K}_s \subseteq \mathcal{K}_r$ for all $s \leq r$.

   - *Čech Complex* $\check{C}(X) \equiv \check{C}(M, d, X)$*:* The set of simplices in $\check{C}(M, d, X)_r$ at filtration level r is

   $$\left\{ x_J : \bigcap_{j\in J} B(x_j, r) \neq \emptyset \text{ and } J \subseteq \{1, ..., N\} \right\},$$

   where (M, d) is a metric space, x_J denotes the simplex with vertices x_j for all $j \in J$, and $B(x, r) := \{y \in M : d(x, y) \leq r\}$.

   - *Vietoris-Rips Complex* $VR(X) = VR(M, d, X)$*:* The set of simplices in $VR(M, d, X)_r$ at filtration level r is

   $$\left\{ x_J : d(x_i, x_j) \leq 2r \text{ for all } i, j \in J \text{ and } J \subseteq \{1, ..., N\} \right\}.$$

► **Question:** How can infer the homology of a manifold M with dimension n from the point cloud $X = \{x_i\}_{i=1}^N$?

2. Construct the *persistent homology* $H(\check{C}(X))$ of the Čech complex.

[2]Loosely speaking, it is k-dimensional hole with $k \leq n$.

▶ **Question:** How can infer the homology of a manifold M with dimension n from the point cloud $X = \{x_i\}_{i=1}^{N}$?

2. Construct the *persistent homology* $H(\check{C}(X))$ of the Čech complex.

3. Summarize $H(\check{C}(X))$ by a *persistent diagram*, which is a multiset of points in $[0, \infty]^2$ that records the birth and death of each homology class [2].

[2] Loosely speaking, it is k-dimensional hole with $k \leq n$.

► **Question:** How can infer the homology of a manifold M with dimension n from the point cloud $X = \{x_i\}_{i=1}^N$?

2. Construct the *persistent homology* $H(\check{C}(X))$ of the Čech complex.

3. Summarize $H(\check{C}(X))$ by a *persistent diagram*, which is a multiset of points in $[0, \infty]^2$ that records the birth and death of each homology class [2].

Theorem (Nerve Theorem in Borsuk 1948)

If $\cap_{j \in J} B(x_j, r)$ is either contractible or empty for all $J \subseteq \{1, ..., N\}$, then $\check{C}(M, d, X)_r$ is homotopy-equivalent to $\cap_{i=1}^N B(x_i, r)$.

[2] Loosely speaking, it is k-dimensional hole with $k \leq n$.

► **Question:** How can infer the homology of a manifold M with dimension n from the point cloud $X = \{x_i\}_{i=1}^N$?

2. Construct the *persistent homology* $H(\check{C}(X))$ of the Čech complex.

3. Summarize $H(\check{C}(X))$ by a *persistent diagram*, which is a multiset of points in $[0, \infty]^2$ that records the birth and death of each homology class [2].

Theorem (Nerve Theorem in Borsuk 1948)

If $\cap_{j \in J} B(x_j, r)$ is either contractible or empty for all $J \subseteq \{1, ..., N\}$, then $\check{C}(M, d, X)_r$ is homotopy-equivalent to $\cap_{i=1}^N B(x_i, r)$.

► **Caveat:** On a general Riemannian manifold, $\cap_{j \in J} B(x_j, r)$ is contractible only when r is sufficiently small.

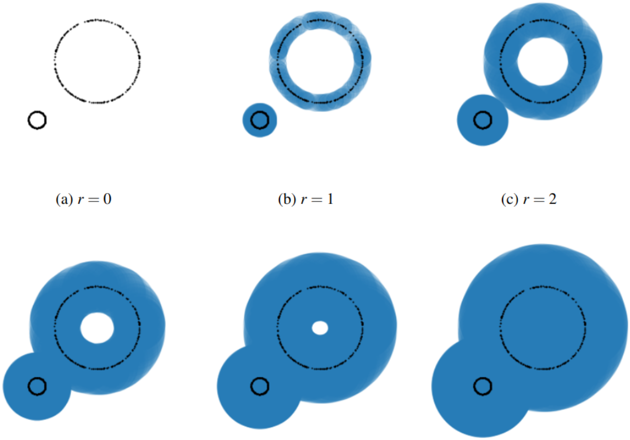
[2] Loosely speaking, it is k-dimensional hole with $k \leq n$.

Motivating Example: Two Circle With Different Radii

- **Conventional Wisdom:** Homology classes with the long lifetimes are true topological features, while those with the short lifetimes are noises.

Motivating Example: Two Circle With Different Radii

- **Conventional Wisdom:** Homology classes with the long lifetimes are true topological features, while those with the short lifetimes are noises.



(a) $r = 0$ (b) $r = 1$ (c) $r = 2$

(d) $r = 3$ (e) $r = 4$ (f) $r = 5$

The smaller circle has a homology class with a much shorter lifetime, but both homology classes are true topological features!!

Proposition (Proposition 3.1 in Niyogi et al. 2008)

Let the closure of

$$\{x \in \mathbb{R}^m : \exists \text{ distinct } y, z \in M \text{ s.t. } d(x, M) = d(x, y) = d(x, z)\}$$

be the ***medial axis*** *of a submanifold M in $\mathbb{R}^m$ and $\sigma(x)$ be the distance from $x \in M$ to the medial axis. Denote the* ***condition number*** *of M by $1/\tau$, where*

$$\tau = \inf_{x \in M} \sigma(x).$$

Then, $\check{C}(X)_r$ is homotopy-equivalent to M when $r < \sqrt{\frac{3}{5}}\,\tau$ and X is sufficiently dense in M.

Proposition (Proposition 3.1 in Niyogi et al. 2008)

Let the closure of

$$\{x \in \mathbb{R}^m : \exists \text{ distinct } y, z \in M \text{ s.t. } d(x, M) = d(x, y) = d(x, z)\}$$

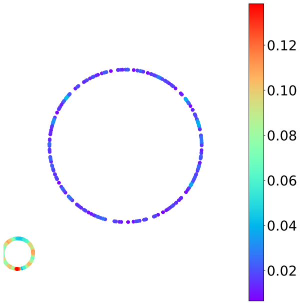
be the ***medial axis*** *of a submanifold M in $\mathbb{R}^m$ and $\sigma(x)$ be the distance from $x \in M$ to the medial axis. Denote the* ***condition number*** *of M by $1/\tau$, where*

$$\tau = \inf_{x \in M} \sigma(x).$$

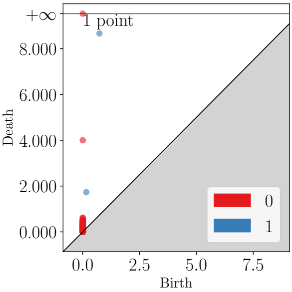
Then, $\check{C}(X)_r$ is homotopy-equivalent to M when $r < \sqrt{\frac{3}{5}}\,\tau$ and X is sufficiently dense in M.

- In the previous two-circle example, τ is equal to the radius of each circle when we view them separately.

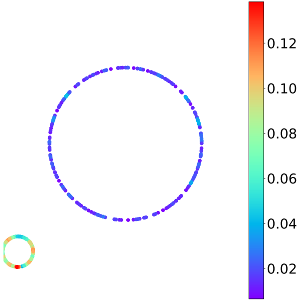
  $\Longrightarrow$ $\check{C}(X)$ may only be homotopy-equivalent to M for a very small range of filtration values r.



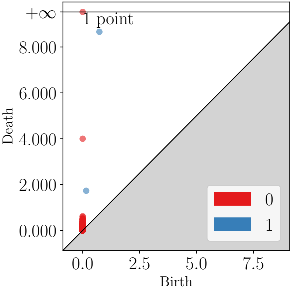
(a) The point cloud density that is inversely proportional to the local feature size.



(b) Persistent diagram of standard Vietoris-Rips complexes.



(a) The point cloud density that is inversely proportional to the local feature size.



(b) Persistent diagram of standard Vietoris-Rips complexes.

- Standard distance-based filtered complexes are not invariant under homeomorphism.

$\Longrightarrow$ Their corresponding persistent homology is not topologically invariant.

Main Contributions of This Paper

Let (M, g) be an n-dimensional Riemannian manifold and $X = \{x_i\}_{i=1}^N$ be points sampled from a smooth density function $f : M \to (0, \infty)$.

► **Key Insight:** Modify the metric g to construct a family of "density-scaled filtered complexes".

Main Contributions of This Paper

Let (M, g) be an n-dimensional Riemannian manifold and $X = \{x_i\}_{i=1}^N$ be points sampled from a smooth density function $f : M \to (0, \infty)$.

► **Key Insight:** Modify the metric g to construct a family of "density-scaled filtered complexes".

- Define a conformally equivalent metric $\widetilde{g} := \sqrt[n]{f^2 \alpha(N)^2} \cdot g$, where

$$\alpha(N) := \begin{cases} \frac{N}{(\log N + (n-1) \log \log N) \log N}, & N > 1, \\ 1, & N = 1. \end{cases}$$

Main Contributions of This Paper

Let (M, g) be an n-dimensional Riemannian manifold and $X = \{x_i\}_{i=1}^N$ be points sampled from a smooth density function $f : M \to (0, \infty)$.

▶ **Key Insight:** Modify the metric g to construct a family of "density-scaled filtered complexes".

- Define a conformally equivalent metric $\widetilde{g} := \sqrt[n]{f^2 \alpha(N)^2} \cdot g$, where

$$\alpha(N) := \begin{cases} \frac{N}{(\log N + (n-1)\log\log N)\log N}, & N > 1, \\ 1, & N = 1. \end{cases}$$

$\widetilde{g}$ helps shrink the distances between points in sparse regions of the manifold and enlarge the distances in dense regions.

Let (M, g) be an n-dimensional Riemannian manifold and $X = \{x_i\}_{i=1}^N$ be points sampled from a smooth density function $f : M \to (0, \infty)$.

▶ **Key Insight:** Modify the metric g to construct a family of "density-scaled filtered complexes".

- Define a conformally equivalent metric $\widetilde{g} := \sqrt[n]{f^2\alpha(N)^2} \cdot g$, where

$$\alpha(N) := \begin{cases} \frac{N}{(\log N + (n-1)\log\log N)\log N}, & N > 1, \\ 1, & N = 1. \end{cases}$$

$\widetilde{g}$ helps shrink the distances between points in sparse regions of the manifold and enlarge the distances in dense regions.

- The density-scaled Čech complex is homotopy-equivalent to M for filtration values in (r_1, r_2) with $r_1 \to 0, r_2 \to \infty$ in probability as $N \to \infty$ and is invariant under conformal transformations.

Methodology and Theoretical Properties

Related Works

1. *k-Nearest Neighbor Filtration:* At filtration level k, the set of simplices is

$$\left\{ x_J : \|x_i - x_j\| \leq \left\|x_i - x_{N_i^k}\right\| \text{ or } \left\|x_j - x_{N_j^k}\right\| \text{ for all } i, j \in J \text{ and all } J \subseteq \{1, \ldots, N\} \right\},$$

where N_i^k is the index of the k-th nearest neighbor of x_i.

- Fail in regions that are close in Euclidean but far in Riemannian distance.

Related Works

1. *k-Nearest Neighbor Filtration:* At filtration level k, the set of simplices is

$$\left\{ x_J : ||x_i - x_j|| \leq \left|\left|x_i - x_{N_i^k}\right|\right| \text{ or } \left|\left|x_j - x_{N_j^k}\right|\right| \text{ for all } i,j \in J \text{ and all } J \subseteq \{1,\ldots,N\} \right\},$$

where N_i^k is the index of the k-th nearest neighbor of x_i.

- Fail in regions that are close in Euclidean but far in Riemannian distance.
- However, the modified continuous kNN graph (Berry and Sauer, 2019) induced a similar density-scaled metric $\sqrt[n]{f^2} \cdot g$.

Related Works

1. *k-Nearest Neighbor Filtration:* At filtration level k, the set of simplices is

$$\left\{x_J : \left|\left|x_i - x_j\right|\right| \leq \left|\left|x_i - x_{N_i^k}\right|\right| \text{ or } \left|\left|x_j - x_{N_j^k}\right|\right| \text{ for all } i, j \in J \text{ and all } J \subseteq \{1, ..., N\}\right\},$$

   where N_i^k is the index of the k-th nearest neighbor of x_i.
   - Fail in regions that are close in Euclidean but far in Riemannian distance.
   - However, the modified continuous kNN graph (Berry and Sauer, 2019) induced a similar density-scaled metric $\sqrt[n]{f^2} \cdot g$.

2. *Fermat Distance:* Fernández et al. (2023) considered a different density-scaled metric $\frac{1}{\sqrt[n]{f^{2(p-1)}}} \cdot g$ for some parameter $p > 1$.
   - It did the opposite as what this paper proposed.

Related Works

1. *k-Nearest Neighbor Filtration:* At filtration level k, the set of simplices is

$$\left\{ x_J : \left|\left|x_i - x_j\right|\right| \leq \left|\left|x_i - x_{N_i^k}\right|\right| \text{ or } \left|\left|x_j - x_{N_j^k}\right|\right| \text{ for all } i, j \in J \text{ and all } J \subseteq \{1, ..., N\} \right\},$$

   where N_i^k is the index of the k-th nearest neighbor of x_i.
   - Fail in regions that are close in Euclidean but far in Riemannian distance.
   - However, the modified continuous kNN graph (Berry and Sauer, 2019) induced a similar density-scaled metric $\sqrt[n]{f^2} \cdot g$.

2. *Fermat Distance:* Fernández et al. (2023) considered a different density-scaled metric $\frac{1}{\sqrt[n]{f^{2(p-1)}}} \cdot g$ for some parameter $p > 1$.
   - It did the opposite as what this paper proposed.

3. *Density-Weighted Complex:* Define the radius of a ball at point x as $r_x(t) := \frac{t}{\sqrt[n]{\alpha(N) \cdot f(x)}}$ of a given filtration parameter t.
   - The proposed density-scaled complexes are more robust than the above density-weighted complexes.

Density-Scaled Filtered Complex

Let (M, g) be an n-dimensional Riemannian manifold and $X = \{x_i\}_{i=1}^N$ be points sampled from a smooth density function $f : M \to (0, \infty)$.

- Recall that the density-scaled Riemannian metric is $\widetilde{g} := \sqrt[n]{f^2 \alpha(N)^2} \cdot g$, where

$$\alpha(N) := \begin{cases} \frac{N}{(\log N + (n-1)\log\log N)\log N}, & N > 1, \\ 1, & N = 1. \end{cases}$$

Density-Scaled Filtered Complex

Let (M, g) be an n-dimensional Riemannian manifold and $X = \{x_i\}_{i=1}^N$ be points sampled from a smooth density function $f : M \to (0, \infty)$.

- Recall that the density-scaled Riemannian metric is $\widetilde{g} := \sqrt[n]{f^2\alpha(N)^2} \cdot g$, where
$$\alpha(N) := \begin{cases} \frac{N}{(\log N + (n-1)\log\log N)\log N}, & N > 1, \\ 1, & N = 1. \end{cases}$$

- The uniform probability measure on $(M, \widetilde{g})$ is $\mathbb{P}(A) = \int_A \frac{1}{\widetilde{\mu}(M)}\, d\widetilde{V}$ for any Borel set $A \subseteq M$, where
$$d\widetilde{V} = \sqrt{|\widetilde{g}|}\, dx^1 \wedge \cdots \wedge dx^n = \alpha(N) f \sqrt{|g|}\, dx^1 \wedge \cdots \wedge dx^n = \alpha(N) f\, dV$$
is the volume form on $(M, \widetilde{g})$ and $\widetilde{\mu}(M)$ is the volume of $(M, \widetilde{g})$.
$$\implies \frac{1}{\widetilde{\mu}(M)}\, d\widetilde{V} = f\, dV.$$

Density-Scaled Filtered Complex

Let (M, g) be an n-dimensional Riemannian manifold and $X = \{x_i\}_{i=1}^N$ be points sampled from a smooth density function $f : M \to (0, \infty)$.

- Recall that the density-scaled Riemannian metric is $\widetilde{g} := \sqrt[n]{f^2 \alpha(N)^2} \cdot g$, where

$$\alpha(N) := \begin{cases} \frac{N}{(\log N + (n-1) \log\log N) \log N}, & N > 1, \\ 1, & N = 1. \end{cases}$$

- The uniform probability measure on $(M, \widetilde{g})$ is $\mathbb{P}(A) = \int_A \frac{1}{\widetilde{\mu}(M)} \, d\widetilde{V}$ for any Borel set $A \subseteq M$, where

$$d\widetilde{V} = \sqrt{|\widetilde{g}|} \, dx^1 \wedge \cdots \wedge dx^n = \alpha(N) f \sqrt{|g|} \, dx^1 \wedge \cdots \wedge dx^n = \alpha(N) f \, dV$$

  is the volume form on $(M, \widetilde{g})$ and $\widetilde{\mu}(M)$ is the volume of $(M, \widetilde{g})$.

$$\implies \frac{1}{\widetilde{\mu}(M)} \, d\widetilde{V} = f \, dV.$$

Sampling points from (M, g) with probability density function f is equivalent to sampling points uniformly at random from $(M, \widetilde{g})$!!

- The *density-scaled Čech complex* is defined as:

$$D\check{C}(M, g, f, X) := \check{C}\left(M, d_{M,\widetilde{g}}, X\right).$$

$\iff$ The set of simplices in $D\check{C}(M, g, f, X)_r$ at filtration level r is

$$\left\{x_J : \bigcap_{j \in J} B(x_j, r) \neq \emptyset \text{ and } J \subseteq \{1, ..., N\}\right\} \text{ with } B(x, r) := \left\{y \in M : d_{M,\widetilde{g}}(x, y) \leq r\right\}.$$

- The *density-scaled Čech complex* is defined as:

$$D\check{C}(M, g, f, X) := \check{C}\left(M, d_{M,\widetilde{g}}, X\right).$$

$\iff$ The set of simplices in $D\check{C}(M, g, f, X)_r$ at filtration level r is

$$\left\{ x_J : \bigcap_{j\in J} B(x_j, r) \neq \emptyset \text{ and } J \subseteq \{1, ..., N\} \right\} \text{ with } B(x, r) := \left\{ y \in M : d_{M,\widetilde{g}}(x, y) \leq r \right\}.$$

- The *density-scaled Vietoris-Rips complex* is defined as:

$$DVR(M, g, f, X) := VR\left(M, d_{M,\widetilde{g}}, X\right).$$

$\iff$ The set of simplices in $DVR(M, g, f, X)_r$ at filtration level r is

$$\left\{ x_J : d_{M,\widetilde{g}}(x_i, x_j) \leq 2r \text{ for all } i, j \in J \text{ and } J \subseteq \{1, ..., N\} \right\}.$$

► **Notes:** More generally, one can define a density-scaled version of any filtered complex via $d_{M,\widetilde{g}}$.

Convergence Property of Density-Scaled Filtered Complex

- The *convexity radius* of a Riemannian manifold (M, g) is

  $r^{\text{convex}} := \sup \{r : B(x,s) \text{ is geodesically convex for all } x \in M \text{ and all } 0 \leq s < r\}$,

  where $B(x, s) := \left\{y \in M : d_{M,g}(x, y) \leq r\right\}$.

- Let r_N^{convex} be the convexity radius of $(M, \widetilde{g}_N)$, where $\widetilde{g}_N$ denotes the density-scaled Riemannian metric with N points.

Convergence Property of Density-Scaled Filtered Complex

- The *convexity radius* of a Riemannian manifold (M, g) is
  $r^{\text{convex}} := \sup \{r : B(x, s) \text{ is geodesically convex for all } x \in M \text{ and all } 0 \leq s < r\}$,
  where $B(x, s) := \left\{y \in M : d_{M,g}(x, y) \leq r\right\}$.

- Let r_N^{convex} be the convexity radius of $(M, \widetilde{g}_N)$, where $\widetilde{g}_N$ denotes the density-scaled Riemannian metric with N points.

- The *coverage radius* of a point cloud X on (M, g) is defined as:

$$r^{\text{cover}} := \inf \left\{ r : M \subseteq \bigcup_{x \in X} B(x, r) \right\}.$$

- Let r_N^{cover} be the coverage radius of a point cloud X on $(M, \widetilde{g}_N)$.

- The *convexity radius* of a Riemannian manifold (M, g) is $r^{\text{convex}} := \sup \{r : B(x, s) \text{ is geodesically convex for all } x \in M \text{ and all } 0 \leq s < r\}$, where $B(x, s) := \{y \in M : d_{M,g}(x, y) \leq r\}$.
- Let r_N^{convex} be the convexity radius of $(M, \widetilde{g}_N)$, where $\widetilde{g}_N$ denotes the density-scaled Riemannian metric with N points.
- The *coverage radius* of a point cloud X on (M, g) is defined as:

$$r^{\text{cover}} := \inf \left\{ r : M \subseteq \bigcup_{x \in X} B(x, r) \right\}.$$

- Let r_N^{cover} be the coverage radius of a point cloud X on $(M, \widetilde{g}_N)$.

Theorem (Theorem 3 in Hickok 2021)

If $r_N^{cover} < r < r_N^{convex}$, then $D\check{C}(M, g, f, X)$ is homotopy-equivalent to M.

- *If M is compact, then $r_N^{convex} \to \infty$ as $N \to \infty$.*
- *If M is compact and connected, then $r_N^{cover} \xrightarrow{P} 0$ as $N \to \infty$.*

- Let (M_1, g_1) and (M_2, g_2) be Riemannian manifolds.
- Let $F : (M_2, g_2) \to (M_1, g_1)$ be a conformal transformation (or specifically, a diffeomorphism).
- Let $f_1 : M_1 \to (0, \infty)$ be a smooth density function.

Conformal Invariance of Density-Scaled Filtered Complex

- Let (M_1, g_1) and (M_2, g_2) be Riemannian manifolds.
- Let $F : (M_2, g_2) \to (M_1, g_1)$ be a conformal transformation (or specifically, a diffeomorphism).
- Let $f_1 : M_1 \to (0, \infty)$ be a smooth density function.
- The function $f_2 : M_2 \to (0, \infty)$ is the *pullback* of f_1 under F, *i.e.*, $f_2 dV_2 = F^*(f_1 dV_1)$.

Sampling a point cloud Y from f_2 is equivalent to sampling a point cloud X from f_1 and setting $Y = F^{-1}(X)$.

- Let (M_1, g_1) and (M_2, g_2) be Riemannian manifolds.
- Let $F : (M_2, g_2) \to (M_1, g_1)$ be a conformal transformation (or specifically, a diffeomorphism).
- Let $f_1 : M_1 \to (0, \infty)$ be a smooth density function.
- The function $f_2 : M_2 \to (0, \infty)$ is the *pullback* of f_1 under F, *i.e.*, $f_2 dV_2 = F^*(f_1 dV_1)$.

Sampling a point cloud Y from f_2 is equivalent to sampling a point cloud X from f_1 and setting $Y = F^{-1}(X)$.

Theorem (Theorem 5 in Hickok 2021)

Let $\Sigma(M, d, X)$ be a distance-based filtered complex that is invariant under global isometry. Then, the density-scaled filtered complex $D\Sigma$ is invariant under all conformal transformations.

- $D\Sigma(M_1, g_1, f_1, X)$ is isomorphic to $D\Sigma\left(M_2, g_2, f_2, F^{-1}(X)\right)$.

Implementation of Density-Scaled Filtered Complex

Estimate f by kernel density estimation on Riemannian manifold (Pelletier, 2005; Ozakin and Gray, 2009):

$$\widehat{f}_N(y) := \frac{1}{N} \sum_{x \in X} \frac{1}{h_N^n} K\left(\frac{||y - x||}{h_N} \right),$$

where $K : \mathbb{R} \to [0, \infty)$ is a kernel function such that

$$K(-x) = K(x),\ \int_{||z|| \leq 1} K(||z||)\, d^n z = 1,\ \text{and } K(x) = 0 \text{ for } x \notin (-1, 1).$$

Implementation of Density-Scaled Filtered Complex

Estimate f by kernel density estimation on Riemannian manifold (Pelletier, 2005; Ozakin and Gray, 2009):

$$\widehat{f}_N(y) := \frac{1}{N}\sum_{x\in X}\frac{1}{h_N^n}K\left(\frac{||y-x||}{h_N}\right),$$

where $K:\mathbb{R}\to[0,\infty)$ is a kernel function such that

$$K(-x)=K(x),\ \int_{||z||\leq 1}K(||z||)\,d^n z=1,\ \text{and } K(x)=0 \text{ for } x\notin(-1,1).$$

- The default kernel in this paper is the biweight kernel

$$K(x)=\frac{\bar{K}(x)}{|\mathbb{S}^{n-1}|\int_0^1\bar{K}(r)r^{n-1}\,dr}\quad\text{with}\quad\bar{K}(x):=\frac{15}{16}(1-x^2)^2\mathbb{1}_{\{x\in(-1,1)\}}.$$

Estimate f by kernel density estimation on Riemannian manifold (Pelletier, 2005; Ozakin and Gray, 2009):

$$\widehat{f}_N(y) := \frac{1}{N} \sum_{x \in X} \frac{1}{h_N^n} K\left(\frac{||y-x||}{h_N}\right),$$

where $K : \mathbb{R} \to [0, \infty)$ is a kernel function such that

$$K(-x) = K(x), \int_{||z|| \leq 1} K(||z||) \, d^n z = 1, \text{ and } K(x) = 0 \text{ for } x \notin (-1, 1).$$

- The default kernel in this paper is the biweight kernel

$$K(x) = \frac{\bar{K}(x)}{|\mathbb{S}^{n-1}| \int_0^1 \bar{K}(r) r^{n-1} \, dr} \quad \text{with} \quad \bar{K}(x) := \frac{15}{16}(1-x^2)^2 \mathbb{1}_{\{x \in (-1,1)\}}.$$

- The bandwidth parameter is chosen as $h_N = N^{-\frac{1}{n+4}}$ in this paper, because when $h_N \propto N^{-\frac{1}{n+4}}$,

$$\mathbb{E}\left[\left(\widehat{f}_N(y) - f(y)\right)^2\right] = O\left(N^{-\frac{4}{n+4}}\right) \text{ is optimal.}$$

Implementation of Density-Scaled Filtered Complex

Estimate f by kernel density estimation on Riemannian manifold (Pelletier, 2005; Ozakin and Gray, 2009):

$$\widehat{f}_N(y) := \frac{1}{N} \sum_{x \in X} \frac{1}{h_N^n} K\left(\frac{||y-x||}{h_N}\right),$$

where $K : \mathbb{R} \to [0, \infty)$ is a kernel function such that

$$K(-x) = K(x), \int_{||z|| \le 1} K(||z||)\, d^n z = 1, \text{ and } K(x) = 0 \text{ for } x \notin (-1,1).$$

- The default kernel in this paper is the biweight kernel

$$K(x) = \frac{\bar{K}(x)}{|\mathbb{S}^{n-1}| \int_0^1 \bar{K}(r) r^{n-1}\, dr} \quad \text{with} \quad \bar{K}(x) := \frac{15}{16}(1-x^2)^2 \mathbb{1}_{\{x \in (-1,1)\}}.$$

- The bandwidth parameter is chosen as $h_N = N^{-\frac{1}{n+4}}$ in this paper, because when $h_N \propto N^{-\frac{1}{n+4}}$,

$$\mathbb{E}\left[\left(\widehat{f}_N(y) - f(y)\right)^2\right] = O\left(N^{-\frac{4}{n+4}}\right) \text{ is optimal.}$$

- Require prior knowledge of the manifold dimension n.

Estimate the Riemannian distance $d_{M,\widetilde{g}}$ as follows:

[3] Choose k to be the first k for which the number of connected components in $G_{k'NN}(X)$ is equal to those in $G_{kNN}(X)$ for all $k' \in \{k-5, ..., k\}$.

Implementation of Density-Scaled Filtered Complex

Estimate the Riemannian distance $d_{M,\widetilde{g}}$ as follows:

1. Construct the k-nearest neighbor[3] graph $G_{kNN}(X)$.

[3] Choose k to be the first k for which the number of connected components in $G_{k'NN}(X)$ is equal to those in $G_{kNN}(X)$ for all $k' \in \{k-5, \ldots, k\}$.

Estimate the Riemannian distance $d_{M,\widetilde{g}}$ as follows:

1. Construct the k-nearest neighbor[3] graph $G_{kNN}(X)$.
2. Define the weight of an edge $(x_i, x_j) \in G_{kNN}(X)$ to

$$w(x_i, x_j) := \sqrt[n]{\alpha(N) \max\left\{\widehat{f}_N(x_i), \widehat{f}_N(x_j)\right\}} \, ||x_i - x_j|| \, .$$

[3] Choose k to be the first k for which the number of connected components in $G_{k'NN}(X)$ is equal to those in $G_{kNN}(X)$ for all $k' \in \{k-5, ..., k\}$.

Estimate the Riemannian distance $d_{M,\widetilde{g}}$ as follows:

1. Construct the k-nearest neighbor[3] graph $G_{kNN}(X)$.
2. Define the weight of an edge $(x_i, x_j) \in G_{kNN}(X)$ to

$$w(x_i, x_j) := \sqrt[n]{\alpha(N) \max\left\{\widehat{f}_N(x_i), \widehat{f}_N(x_j)\right\}}\; ||x_i - x_j||\,.$$

3. The estimate $\widehat{d_{M,\widetilde{g}}}(x_i, x_j)$ is the length of the shortest weighted path in $G_{kNN}(X)$ from x_i to x_j. Set $\widehat{d_{M,\widetilde{g}}}(x_i, x_j) = \infty$ if x_i, x_j are not connected.

The approximate density-scaled Vietoris-Rips complex $\widehat{DVR}(n, k, X)$ at filtration level t is

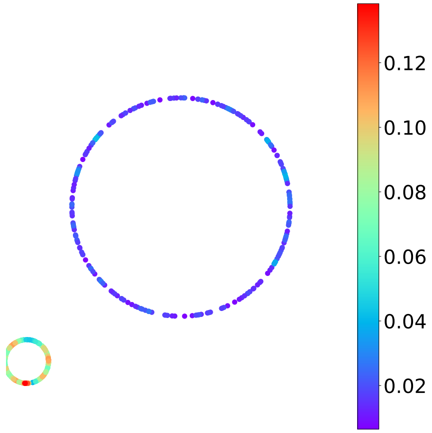
$$\left\{x_J : \widehat{d_{M,\widetilde{g}}}(x_i, x_j) \leq 2t \text{ for all } i, j \in J \text{ and all } J \subseteq \{1, ..., N\}\right\}.$$

[3] Choose k to be the first k for which the number of connected components in $G_{k'NN}(X)$ is equal to those in $G_{kNN}(X)$ for all $k' \in \{k - 5, ..., k\}$.

Experimental Results

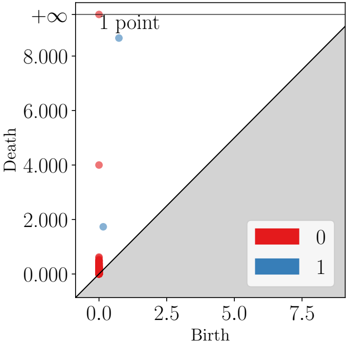
Two-Circle of Different Radii

- Sample $X = \{x_i\}_{i=1}^{500}$ points from two disjoint circles C_1 and C_2 of radii $R_1 = 1$ and $R_2 = 5$, respectively.
- The density function is given by $f(x) = \begin{cases} \frac{1}{4\pi R_1}, & x \in C_1, \\ \frac{1}{4\pi R_2}, & x \in C_2. \end{cases}$

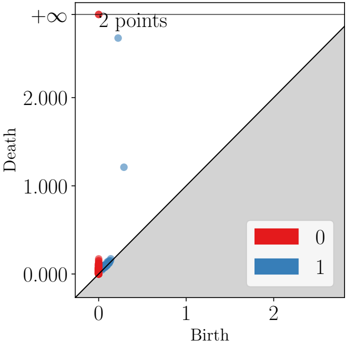


Two-Circle of Different Radii

- Sample $X = \{x_i\}_{i=1}^{500}$ points from two disjoint circles C_1 and C_2 of radii $R_1 = 1$ and $R_2 = 5$, respectively.
- The density function is given by $f(x) = \begin{cases} \frac{1}{4\pi R_1}, & x \in C_1, \\ \frac{1}{4\pi R_2}, & x \in C_2. \end{cases}$



(a) Persistent diagram for $H(VR(X))$.



(b) Persistent diagram of $H(\widehat{DVR}(X))$.

Number of Connected Components in $G_{kNN}(X)$

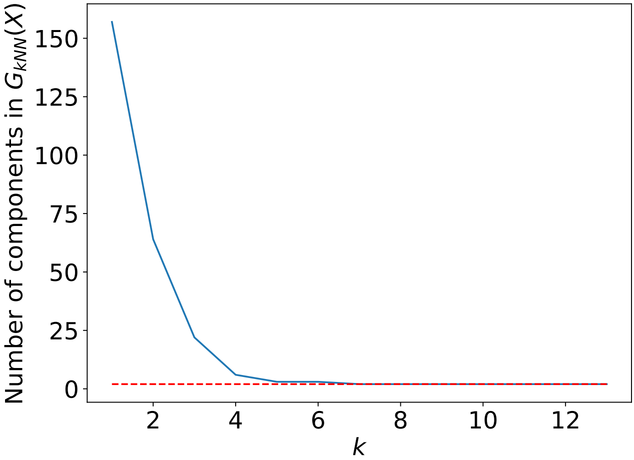


Figure 6: The number of connected components in $G_{kNN}(X)$ for the two-circle point cloud example. For $k \in \{5, \ldots, 74\}$, the number of components is the true value 2.

Robustness to the Choices of k and Kernel Functions

Table 1: Comparison of Kernel Functions and k Values

k	Kernel function	$\frac{\text{Lifetime of second-most persistent 1D homology class}}{\text{Lifetime of most-persistent 1D homology class}}$	Number of infinite 0D homology classes
10	Biweight	.659	2
10	Epanechnikov	.604	2
10	Triweight	.678	2
5	Biweight	.011	2
15	Biweight	.442	2

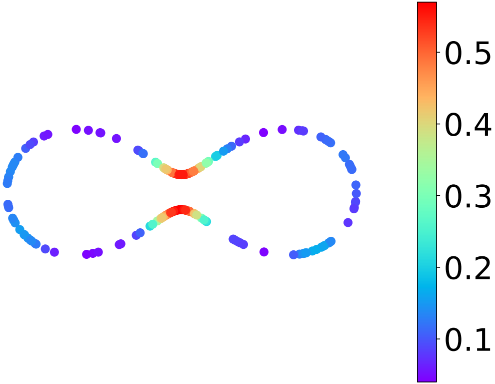
- The triweight kernel with $k = 10$ yields the highest ratio (0.678), slightly higher than the ratio for the biweight kernel with $k = 10$.

- The biweight kernel with $k = 5$ leads to very poor results (a ratio of 0.011), because $k = 5$ is too low for all of the adjacent points in X on the largest circle to be connected by an edge in $G_{kNN}(X)$.

- Sample $X = \{x_i\}_{i=1}^{200}$ points from a Cassini curve

$$r^4 - 2r^2\cos(2\theta) = e^4 - 1,$$

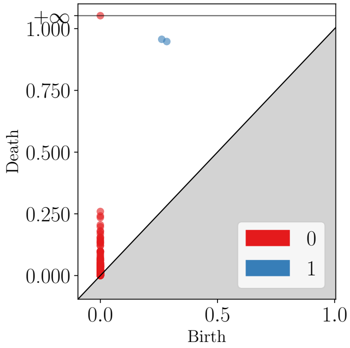
where $e = 1.01$ is the eccentricity and $\theta \sim \text{Uniform}[0, 2\pi)$.



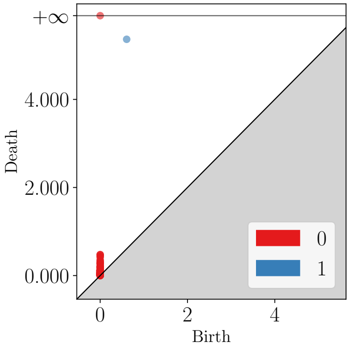
- Sample $X = \{x_i\}_{i=1}^{200}$ points from a Cassini curve

$$r^4 - 2r^2\cos(2\theta) = e^4 - 1,$$

where $e = 1.01$ is the eccentricity and $\theta \sim \text{Uniform}[0, 2\pi)$.



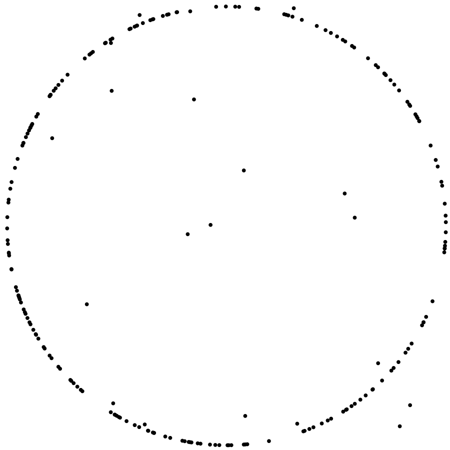
(a) Persistent diagram for $H(VR(X))$.



(b) Persistent diagram of $H(\widehat{DVR}(X))$.

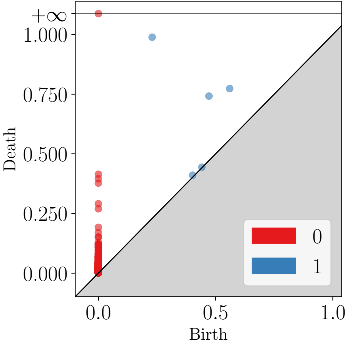
Point Cloud With Outliers

- Sample 200 points uniformly at random from $\mathbb{S}^1$ and 20 points[4] uniformly from the square $[-1,1]^2$.

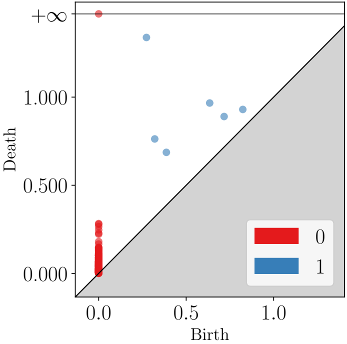


[4]In the paper, the author only sampled 10 outliers from $[-1,1]^2$.

- Sample 200 points uniformly at random from $\mathbb{S}^1$ and 20 points[4] uniformly from the square $[-1,1]^2$.



(a) Persistent diagram for $H(VR(X))$.

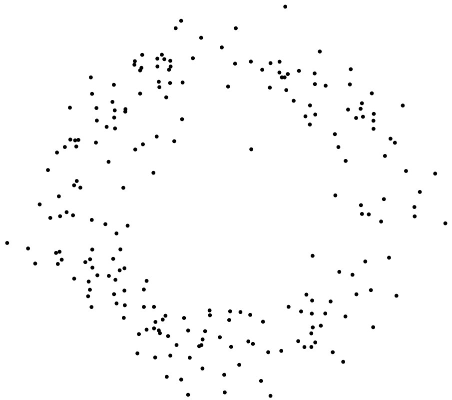


(b) Persistent diagram of $H(\widehat{DVR}(X))$.

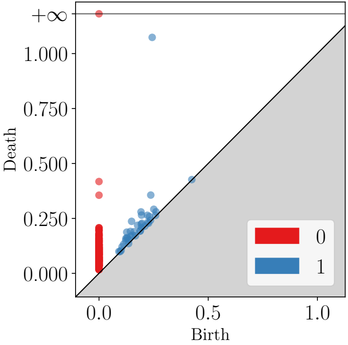
[4] In the paper, the author only sampled 10 outliers from $[-1,1]^2$.

Extension: Point Cloud Near the Manifold

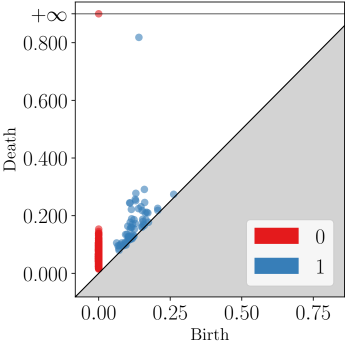
- Sample 220 points uniformly at random from $\mathbb{S}^1$ with radial noises $\mathcal{N}(0, 0.2^2)$.



- Sample 220 points uniformly at random from $\mathbb{S}^1$ with radial noises $\mathcal{N}(0, 0.2^2)$.



(a) Persistent diagram for $H(VR(X))$.



(b) Persistent diagram of $H(\widehat{DVR}(X))$.

Extension: Point Cloud Near the Manifold

- Sample 220 points uniformly at random from $\mathbb{S}^1$ with radial noises $\mathcal{N}(0, 0.2^2)$.

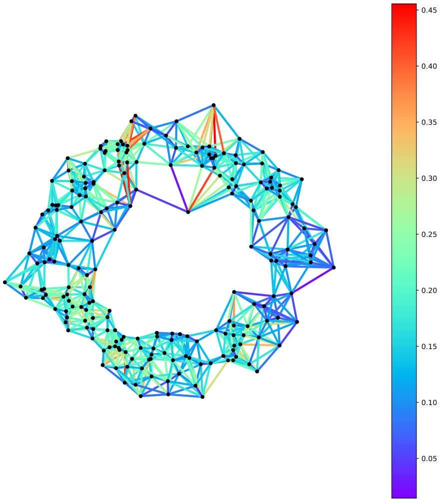


Figure 9: The weighted kNN graph $G_{kNN}(X)$ with $k = 9$.

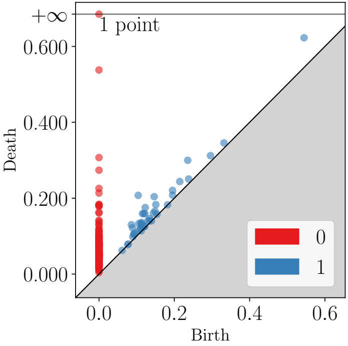
Application: Clustering

► **Goal:** Use $\widehat{DVR}$ to identify the number of clusters in a point cloud when clusters have different densities.

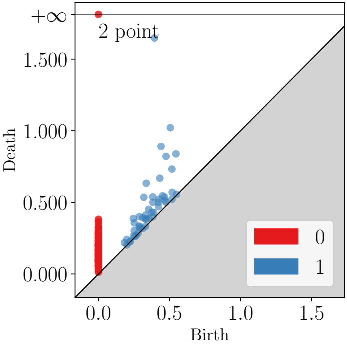
- Sample $N = 200$ points from the union of squares $[0,1]^2$ and $[1.5, 2.5] \times [0,1]$.

▶ **Goal:** Use $\widehat{DVR}$ to identify the number of clusters in a point cloud when clusters have different densities.

- Sample $N = 200$ points from the union of squares $[0,1]^2$ and $[1.5, 2.5] \times [0,1]$.



(a) Persistent diagram for $H(VR(X))$.



(b) Persistent diagram of $H(\widehat{DVR}(X))$.

- Apply $\widehat{DVR}$ to a point cloud generated from the Lorenz dynamical system (Lorenz, 1963):

$$\begin{cases} \frac{dx}{dt} & = \gamma(y - x), \\ \frac{dy}{dt} & = x(\rho - z) - y, \\ \frac{dz}{dt} & = xy - \beta z, \end{cases}$$

  where we set $\gamma = 10$, $\rho = 28$, and $\beta = \frac{8}{3}$.

- Set the initial condition to $(x_0, y_0, z_0) = (1, 1, 1)$ and solve the system from $t = 0$ to $t = 50$ using SciPy ODE solver (Virtanen et al., 2020).

Application: Lorenz System

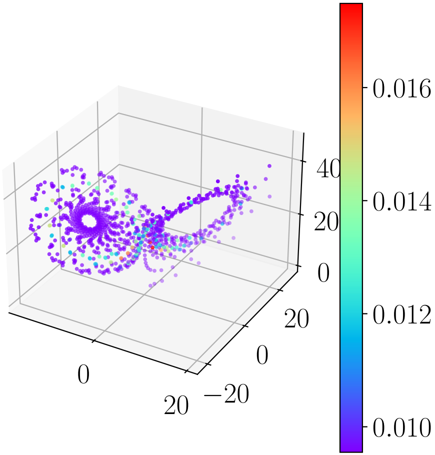
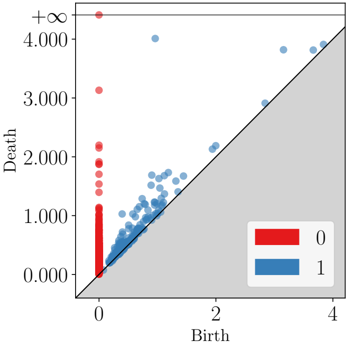
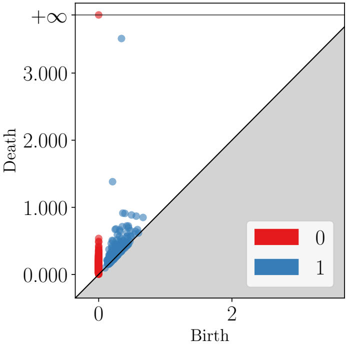


Figure 11: Collection of points $\{(x(t_i), y(t_i), z(t_i))\}_{i=1}^{1000}$ with time steps $t_i = 0.05i$.[5]

[5] In the paper, the point cloud is taken to be a 2-dim time-delay embedding of $x(t)$ with time lag $\tau = 0.05$.



(a) Persistent diagram for $H(VR(X))$.



(b) Persistent diagram of $H(\widehat{DVR}(X))$.

Discussion

Conclusion

► **Summary:** This paper (Hickok, 2021) proposed a family of density-scaled filtered complexes for inferring the homology of a manifold M.

► **Summary:** This paper (Hickok, 2021) proposed a family of density-scaled filtered complexes for inferring the homology of a manifold M.

► **Main Contribution:**

- The density-scaled Čech complex is homotopy-equivalent to M for a growing interval of filtration values as $N \to \infty$, regardless of the geometric properties of M.
- The density-scaled filtered complexes are invariant under conformal transformations.
- Introduce a practical algorithm for construct the density-scaled Vietoris-Rips complex $\widehat{DVR}$.

1. **Stability of $\widehat{DVR}$:** If two point clouds X, Y are close in Wasserstein distance $W_{\inf}(X, Y) := \inf_{\eta: X \to Y} \max_{x \in X} ||x - \eta(x)||$, then the bottleneck distance[5] between the persistence diagrams of $\widehat{DVR}(n, k, X)$ and $\widehat{DVR}(n, k, Y)$ are close.

[5] The bottleneck distance between two diagrams is

$$W_{\infty}\left(\mathrm{dgm}(\mathbb{V}), \mathrm{dgm}(\mathbb{U})\right) := \inf_{\eta} \sup_{x \in \mathrm{dgm}(\mathbb{V})} \left|\left|x - \eta(x)\right|\right|_{\infty},$$

and the infimum is taken over all bijections $\eta : \mathrm{dgm}(\mathbb{V}) \to \mathrm{dgm}(\mathbb{U})$.

Discussion

1. **Stability of $\widehat{DVR}$:** If two point clouds X, Y are close in Wasserstein distance $W_{\inf}(X, Y) := \inf_{\eta: X \to Y} \max_{x \in X} ||x - \eta(x)||$, then the bottleneck distance[5] between the persistence diagrams of $\widehat{DVR}(n, k, X)$ and $\widehat{DVR}(n, k, Y)$ are close.
   - The results are stated in $\epsilon - \delta$ language, and we don't know the rate of convergence and its dependence on N, K, and n.

[5]The bottleneck distance between two diagrams is

$$W_\infty\left(\mathrm{dgm}(\mathbb{V}), \mathrm{dgm}(\mathbb{U})\right) := \inf_{\eta} \sup_{x \in \mathrm{dgm}(\mathbb{V})} ||x - \eta(x)||_\infty,$$

and the infimum is taken over all bijections $\eta : \mathrm{dgm}(\mathbb{V}) \to \mathrm{dgm}(\mathbb{U})$.

1. **Stability of $\widehat{DVR}$:** If two point clouds X, Y are close in Wasserstein distance $W_{\inf}(X,Y) := \inf_{\eta:X\to Y} \max_{x\in X} ||x-\eta(x)||$, then the bottleneck distance[5] between the persistence diagrams of $\widehat{DVR}(n,k,X)$ and $\widehat{DVR}(n,k,Y)$ are close.
   - The results are stated in $\epsilon - \delta$ language, and we don't know the rate of convergence and its dependence on N, K, and n.

2. **Bandwidth Selection:** Other bandwidth selection methods, such as least square cross-validation (Stone, 1984) and plug-in method (Sheather and Jones, 1991), are worth studying.

[5] The bottleneck distance between two diagrams is

$$W_\infty\left(\mathrm{dgm}(\mathbb{V}), \mathrm{dgm}(\mathbb{U})\right) := \inf_{\eta} \sup_{x\in\mathrm{dgm}(\mathbb{V})} ||x-\eta(x)||_\infty ,$$

and the infimum is taken over all bijections $\eta : \mathrm{dgm}(\mathbb{V}) \to \mathrm{dgm}(\mathbb{U})$.

Discussion

1. **Stability of $\widehat{DVR}$:** If two point clouds X, Y are close in Wasserstein distance $W_{\inf}(X, Y) := \inf_{\eta: X \to Y} \max_{x \in X} ||x - \eta(x)||$, then the bottleneck distance[5] between the persistence diagrams of $\widehat{DVR}(n, k, X)$ and $\widehat{DVR}(n, k, Y)$ are close.
   - The results are stated in $\epsilon - \delta$ language, and we don't know the rate of convergence and its dependence on N, K, and n.
2. **Bandwidth Selection:** Other bandwidth selection methods, such as least square cross-validation (Stone, 1984) and plug-in method (Sheather and Jones, 1991), are worth studying.
3. **Computational Efficiency:** Computing $\widehat{DVR}$ requires knowledge of the pairwise Euclidean distances between the points in X, which has at least $O(N^2)$ time and space complexity.

[5] The bottleneck distance between two diagrams is

$$W_\infty(\mathrm{dgm}(\mathbb{V}), \mathrm{dgm}(\mathbb{U})) := \inf_{\eta} \sup_{x \in \mathrm{dgm}(\mathbb{V})} ||x - \eta(x)||_\infty,$$

and the infimum is taken over all bijections $\eta : \mathrm{dgm}(\mathbb{V}) \to \mathrm{dgm}(\mathbb{U})$.

Thank you!

More details can be found in

Abigail Hickok. A Family of Density-Scaled Filtered Complexes. *arXiv preprint*, 2021. https://arxiv.org/abs/2112.03334.

Reference

M. Berger. *A panoramic view of Riemannian geometry*. Springer, 2003.

T. Berry and T. Sauer. Consistent manifold representation for topological data analysis. *Foundations of Data Science*, 1(1):1–38, 2019.

K. Borsuk. On the imbedding of systems of compacta in simplicial complexes. *Fundamenta Mathematicae*, 35(1):217–234, 1948.

C. Fefferman, S. Mitter, and H. Narayanan. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 29(4):983–1049, 2016.

X. Fernández, E. Borghini, G. Mindlin, and P. Groisman. Intrinsic persistent homology via density-based metric learning. *Journal of Machine Learning Research*, 24(75):1–42, 2023.

L. Flatto and D. J. Newman. Random coverings. *Acta Mathematica*, 138(none):241 – 264, 1977.

A. Hickok. A family of density-scaled filtered complexes. *arXiv preprint arXiv:2112.03334*, 2021.

E. N. Lorenz. Deterministic nonperiodic flow. *Journal of atmospheric sciences*, 20(2):130–141, 1963.

P. Niyogi, S. Smale, and S. Weinberger. Finding the homology of submanifolds with high confidence from random samples. *Discrete & Computational Geometry*, 39:419–441, 2008.

A. Ozakin and A. Gray. Submanifold density estimation. In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc., 2009.

B. Pelletier. Kernel density estimation on riemannian manifolds. *Statistics & probability letters*, 73(3): 297–304, 2005.

S. J. Sheather and M. C. Jones. A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society: Series B (Methodological)*, 53(3):683–690, 1991.

C. J. Stone. An asymptotically optimal window selection rule for kernel density estimates. *The Annals of Statistics*, pages 1285–1297, 1984.

Reference

J. B. Tenenbaum, V. d. Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, et al. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature methods*, 17(3):261–272, 2020.

Z. Yao, J. Su, and B. Li. Manifold fitting: An invitation to statistics. *arXiv preprint arXiv:2304.07680*, 2023.

Theorem (Theorem 3 in Hickok 2021)

If $r_N^{cover} < r < r_N^{convex}$, then $D\check{C}(M, g, f, X)$ is homotopy-equivalent to M.

- *If M is compact, then $r_N^{convex} \to \infty$ as $N \to \infty$.*
- *If M is compact and connected, then $r_N^{cover} \xrightarrow{P} 0$ as $N \to \infty$.*

Proof (Sketch). If $r < r_N^{\text{convex}}$, then the intersection $\cap_{j \in J} B(x_j, r)$ is convex for all $J \subseteq \{1, ..., N\}$, so it is either contractible or empty. If $r > r_N^{\text{cover}}$, then $\cup_i B(x_i, r) = M$. Then, apply the Nerve Theorem.

- The convexity radius of a compact manifold is positive; see Chapter 6.5.3 in Berger (2003). Thus, $r_1^{\text{convex}} > 0$, and $r_N^{\text{convex}} = \sqrt[n]{\alpha(N)} \cdot r_1^{\text{convex}}$.

Theorem (Theorem 3 in Hickok 2021)

If $r_N^{cover} < r < r_N^{convex}$, then $D\check{C}(M, g, f, X)$ is homotopy-equivalent to M.

- *If M is compact, then $r_N^{convex} \to \infty$ as $N \to \infty$.*
- *If M is compact and connected, then $r_N^{cover} \xrightarrow{P} 0$ as $N \to \infty$.*

Proof (Continued).

- The convergence of r_N^{cover} is controlled by the *filling factor* $\Lambda := \frac{N\nu_n r^n}{\mu(M)}$. Define

$$\Lambda_N = \log N + (n-1)\log\log N + w(N) \quad \text{and} \quad r_N = \sqrt[n]{\frac{\alpha(N)\Lambda_N}{N\nu_n}},$$

where $w(N)$ is a sequence with $w(N) \to \infty$ and $\frac{w(N)}{\log N} \to 0$ as $N \to \infty$, while v_n is the volume of a Euclidean unit n-ball.
For any $\epsilon > 0$, $r_N < \epsilon$ when N is sufficiently large, so

$$\mathbb{P}\left(r_N^{\text{convex}} > \epsilon\right) < \mathbb{P}\left(r_N^{\text{cover}} > r_N\right) \to 0 \quad \text{as} \quad N \to \infty$$

by Theorem 1.1 in Flatto and Newman (1977) and Corollary 1 in Hickok (2021). □