TARDIS: Topological Algorithm for Robust DIscovery of Singularities

Julius von Rohrscheidt

April 26, 2024

HELMHOLTZ MUNICH

Technical University of Munich

TARDIS: Topological Algorithm for Robust DIscovery of **Singularities**

PMLR **Proceedings of Machine Learning Research**

Volume 202 JMLR DMLR TMLR MLOSS FAQ Submission Format RSS

[edit]

Topological Singularity Detection at Multiple Scales

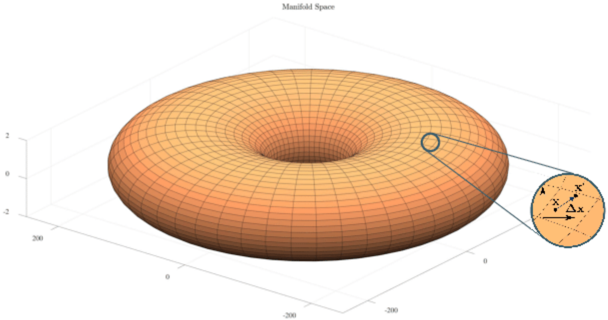
Julius Von Rohrscheidt, Bastian Rieck *Proceedings of the 40th International Conference on Machine Learning*, PMLR 202:35175-35197, 2023.

Abstract

The manifold hypothesis, which assumes that data lies on or close to an unknown manifold of low intrinsic dimension, is a staple of modern machine learning research. However, recent work has shown that real-world data exhibits distinct non-manifold structures, i.e. singularities, that can lead to erroneous findings. Detecting such singularities is therefore crucial as a precursor to interpolation and inference tasks. We address this issue by developing a topological framework that (i) quantifies the local intrinsic dimension, and (ii) yields a Euclidicity score for assessing the 'manifoldness' of a point along multiple scales. Our approach identifies singularities of complex spaces, while also capturing singular structures and local geometric complexity in image data.

`https://proceedings.mlr.press/v202/von-rohrscheidt23a.html`

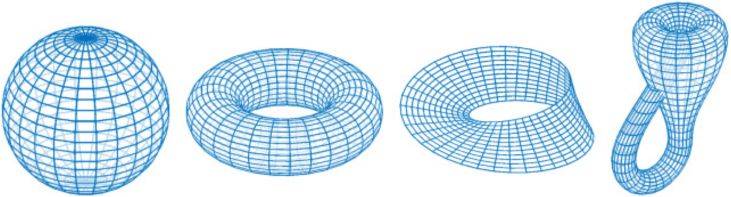
Manifolds in a nutshell



A **manifold** is a space that resembles Euclidean space locally, i.e. *every* point admits a neighbourhood that looks like a Euclidean ball.

Manifolds in a nutshell

Examples of manifolds:



(a) Sphere (b) Torus (c) Möbius strip (d) Klein bottle

- Manifolds can encode complex global behaviour
- However, locally they look ‘trivial‘.

Why manifolds?

- Manifolds are widely studied objects in mathematics
- In Data Science, most non-linear dimensionality reduction techniques (**UMAP, t-SNE, ...**) make use of the manifold hypothesis:

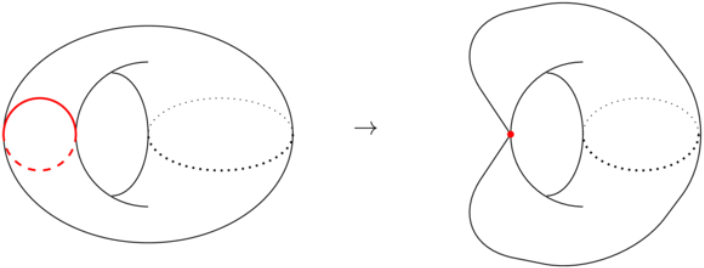
The **manifold hypothesis** assumes that the given data lies on a lower dimensional manifold.

- Performance of these algorithms depends on the correctness of the manifold hyothesis.

Singularities

A **singularity** is a point in a space that violates the assumption of being *locally Euclidean*.

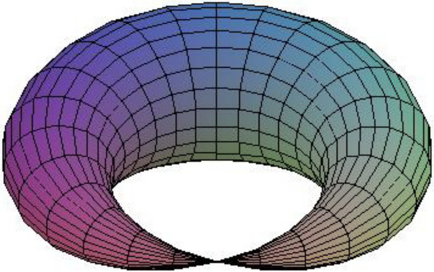
A **singular space** is a space that may admit singularities.



Singularities

A **singularity** is a point in a space that violates the assumption of being *locally Euclidean*.

A **singular space** is a space that may admit singularities.



Why singularities?

- Recently, Brown et. al.[1] found evidence that popular datasets (MNIST, FashionMNIST, ...) do not satisfy the manifold hypothesis.
- Moreover, Perea et. al.[2] showed empirically that manifold learning algorithms fail in general, when the underlying data does not stem from a manifold.

$\Rightarrow$ Let's test the manifold hypothesis!

[1] Brown, Bradley CA, et al. "The Union of Manifolds Hypothesis and its Implications for Deep Generative Modelling." arXiv preprint arXiv:2207.02862 (2022).

[2] Mike, Joshua Lee, and Jose Perea. "TALLEM: Topological Assembly of Locally Euclidean Models." 2022 Spring Western Sectional Meeting. AMS, 2022.

Cone of a topological space

For a topological space X, **the cone** of X is given by
$c^\circ X := X \times (0,1]/X \times \{1\}$

Cone of a topological space

For a topological space X, **the cone** of X is given by $c^\circ X := X \times (0,1]/X \times \{1\}$

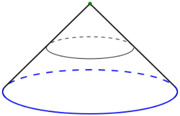
Examples:

- $c^\circ pt. = (0,1]$

Cone of a topological space

For a topological space X, **the cone** of X is given by
$c^\circ X := X \times (0,1]/X \times \{1\}$

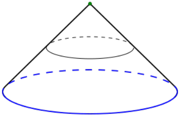
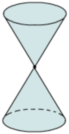
Examples:

- $c^\circ pt. = (0,1]$
- $c^\circ S^1 \cong D^2$ (2-dimensional disk) 

Cone of a topological space

For a topological space X, **the cone** of X is given by
$c^\circ X := X \times (0,1]/X \times \{1\}$

Examples:

- $c^\circ pt. = (0,1]$
- $c^\circ S^1 \cong D^2$ (2-dimensional disk) 
- $c^\circ(S^1 \sqcup S^1) \cong$ double cone 

Singular spaces

A 0-dimensional stratified pseudomanifold is a countable set of points with the discrete topology.

Singular spaces

A 0-dimensional stratified pseudomanifold is a countable set of points with the discrete topology.

An n-dimensional (PL) stratified pseudomanifold is a (PL) space X of dimension n, together with a filtration by closed (PL) subspaces $X = X_n \supset X_{n-1} = X_{n-2} \supset \cdots \supset X_0 \supset X_{-1} = \emptyset$ such that

Singular spaces

A 0-dimensional stratified pseudomanifold is a countable set of points with the discrete topology.

An n-dimensional (PL) stratified pseudomanifold is a (PL) space X of dimension n, together with a filtration by closed (PL) subspaces $X = X_n \supset X_{n-1} = X_{n-2} \supset \cdots \supset X_0 \supset X_{-1} = \emptyset$ such that

- Every non-empty $X_{n-k} - X_{n-k-1}$ is a (PL) manifold of dimension $n - k$.

Singular spaces

A 0-dimensional stratified pseudomanifold is a countable set of points with the discrete topology.

An n-dimensional (PL) stratified pseudomanifold is a (PL) space X of dimension n, together with a filtration by closed (PL) subspaces $X = X_n \supset X_{n-1} = X_{n-2} \supset \cdots \supset X_0 \supset X_{-1} = \emptyset$ such that

- Every non-empty $X_{n-k} - X_{n-k-1}$ is a (PL) manifold of dimension $n-k$.
- $X - X_{n-2}$ is dense in X.

Singular spaces

A 0-dimensional stratified pseudomanifold is a countable set of points with the discrete topology.

An n-dimensional (PL) stratified pseudomanifold is a (PL) space X of dimension n, together with a filtration by closed (PL) subspaces $X = X_n \supset X_{n-1} = X_{n-2} \supset \cdots \supset X_0 \supset X_{-1} = \emptyset$ such that

- Every non-empty $X_{n-k} - X_{n-k-1}$ is a (PL) manifold of dimension $n-k$.
- $X - X_{n-2}$ is dense in X.
- For each point $x \in X_{n-k} - X_{n-k-1}$, there exists an open neighborhood U of x in X and a compact (PL) stratified pseudomanifold L of dimension $k-1$ and a (PL) homeomorphism

$$\phi : U \xrightarrow{\simeq} \mathbb{R}^{n-k} \times c^{\circ}L$$

Singular spaces

A 0-dimensional stratified pseudomanifold is a countable set of points with the discrete topology.

An n-dimensional (PL) stratified pseudomanifold is a (PL) space X of dimension n, together with a filtration by closed (PL) subspaces $X = X_n \supset X_{n-1} = X_{n-2} \supset \cdots \supset X_0 \supset X_{-1} = \emptyset$ such that

- Every non-empty $X_{n-k} - X_{n-k-1}$ is a (PL) manifold of dimension $n-k$.
- $X - X_{n-2}$ is dense in X.
- For each point $x \in X_{n-k} - X_{n-k-1}$, there exists an open neighborhood U of x in X and a compact (PL) stratified pseudomanifold L of dimension $k-1$ and a (PL) homeomorphism

$$\phi : U \xrightarrow{\sim} \mathbb{R}^{n-k} \times c^\circ L$$

  (which is stratum-preserving.)

Local homology

- For a point $x \in X$, its i-th local homology $H_i(X, X - x)$ captures homological information of an infinitesimal small neighborhood of x, relative to an infinitesimal punctured neighbourhood of x (in X).

Local homology

- For a point $x \in X$, its i-th local homology $H_i(X, X - x)$ captures homological information of an infinitesimal small neighborhood of x, relative to an infinitesimal punctured neighbourhood of x (in X).
- Let X be a (stratified) pseudomanifold and $x \in X$. Then x has a distinguished neighborhood $U \cong \mathbb{R}^k \times c^\circ L$, where L is called the **link** of x.

Local homology

- For a point $x \in X$, its i-th local homology $H_i(X, X - x)$ captures homological information of an infinitesimal small neighborhood of x, relative to an infinitesimal punctured neighbourhood of x (in X).
- Let X be a (stratified) pseudomanifold and $x \in X$. Then x has a distinguished neighborhood $U \cong \mathbb{R}^k \times c^\circ L$, where L is called the **link** of x.
- The local homology of x will generally depend on the homology of L.

Local homology

- For a point $x \in X$, its i-th local homology $H_i(X, X - x)$ captures homological information of an infinitesimal small neighborhood of x, relative to an infinitesimal punctured neighbourhood of x (in X).
- Let X be a (stratified) pseudomanifold and $x \in X$. Then x has a distinguished neighborhood $U \cong \mathbb{R}^k \times c^\circ L$, where L is called the **link** of x.
- The local homology of x will generally depend on the homology of L.
- The motivation to use local homology for singularity detection stems from the following fact:

Local homology

- For a point $x \in X$, its i-th local homology $H_i(X, X - x)$ captures homological information of an infinitesimal small neighborhood of x, relative to an infinitesimal punctured neighbourhood of x (in X).
- Let X be a (stratified) pseudomanifold and $x \in X$. Then x has a distinguished neighborhood $U \cong \mathbb{R}^k \times c^\circ L$, where L is called the **link** of x.
- The local homology of x will generally depend on the homology of L.
- The motivation to use local homology for singularity detection stems from the following fact:
- If $U \cong c^\circ L$, one can show that

$$H_i(X, X - x) = \tilde{H}_{i-1}(L)$$

for all $i \geq 0$.

Local homology

- Let X be a (stratified) pseudomanifold and $x \in X$. Then x has a distinguished neighborhood $U \cong \mathbb{R}^k \times c^\circ L$, where L is called the **link** of x.
- The local homology of x will generally depend on the homology of L.
- The motivation to use local homology for singularity detection stems from the following fact:
- If $U \cong c^\circ L$, one can show that

$$H_i(X, X - x) = \tilde{H}_{i-1}(L)$$

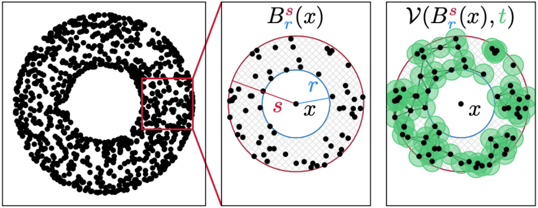
  for all $i \geq 0$.
- In particular, if $X = M$ is a manifold of dimension n, one obtains

$$H_i(M, M - x) = \tilde{H}_{i-1}(S^{n-1}) = \begin{cases} \mathbb{Z}, i = n \\ 0, i \neq n \end{cases}$$

How to test the manifold hypothesis?

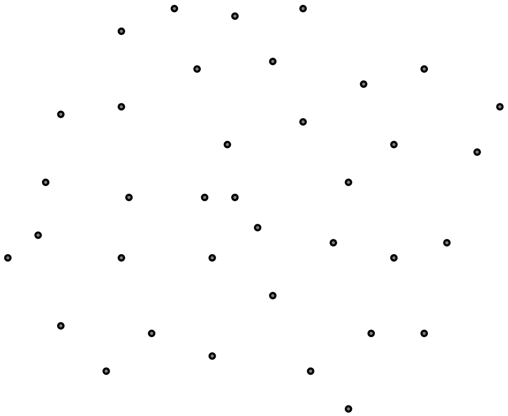
- As we have already seen, manifolds are characterised by a local property.
- **Idea:** Test the ‘manifoldness‘ of each point in the data space, individually.

How to test the manifold hypothesis?

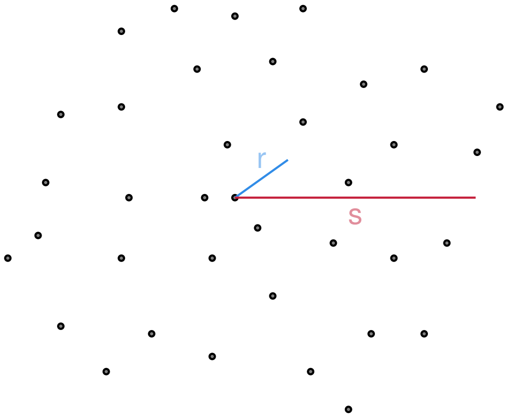


- From your given dataset $\mathbb{X}$, choose a point $x \in \mathbb{X}$.
- For two fixed radius parameters $r < s$, let $B_r^s(x)$ denote the set of data points with distance to x at least r, and at most s.
- Let $\mathcal{V}(B_r^s(x), t)$ denote the *Vietoris-Rips* construction w.r.t. $B_r^s(x)$ at filtration step t.

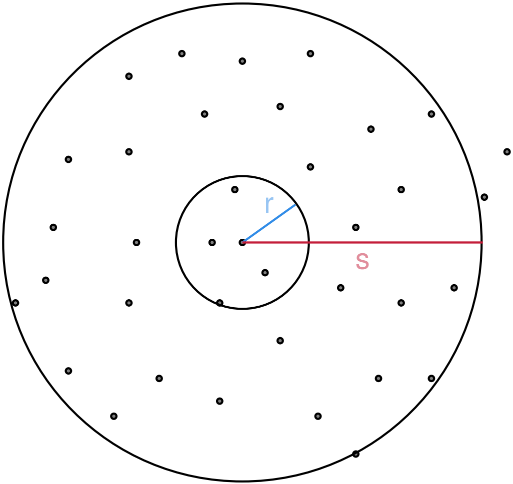
How to test the manifold hypothesis?



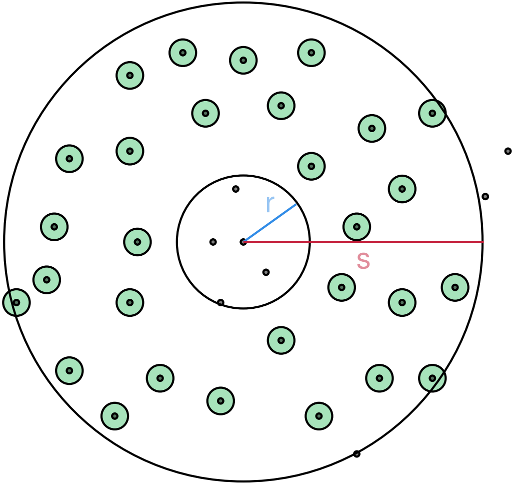
How to test the manifold hypothesis?



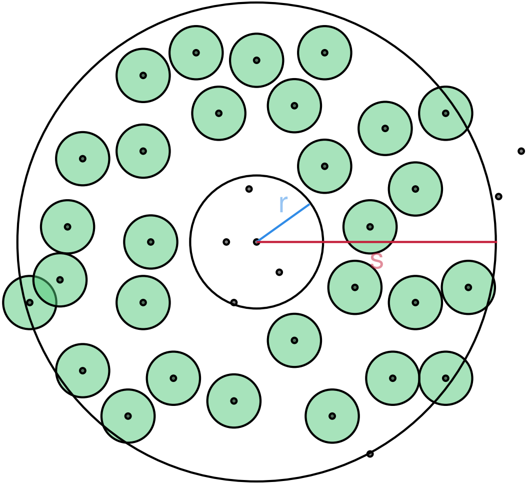
How to test the manifold hypothesis?



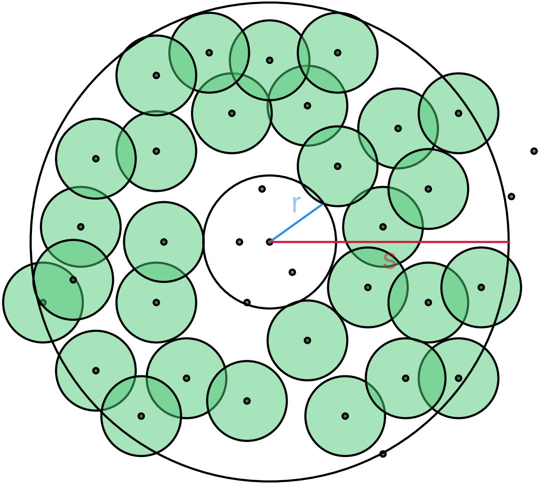
How to test the manifold hypothesis?



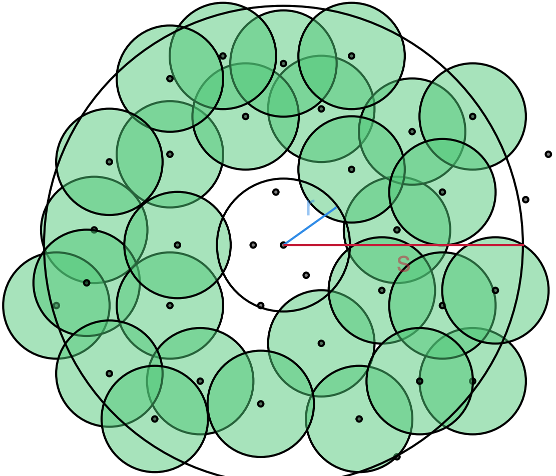
How to test the manifold hypothesis?



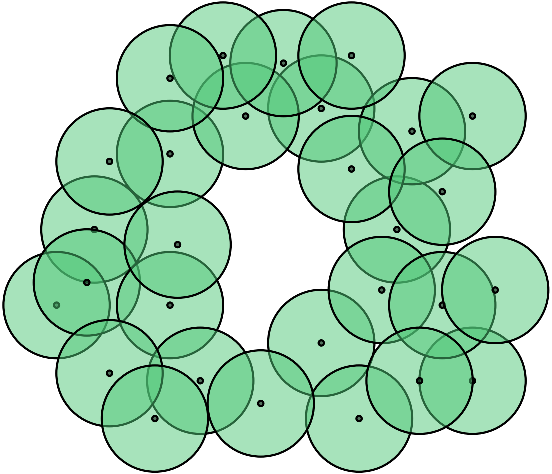
How to test the manifold hypothesis?



How to test the manifold hypothesis?



How to test the manifold hypothesis?



Back to data: Persistent homology

- Given a finite metric space $(\mathbb{X}, d)$, the *Vietoris–Rips complex* at step t is defined as the abstract simplicial complex $\mathcal{V}(\mathbb{X}, t)$, in which an abstract k-simplex $(x_0, \ldots, x_k)$ of points in $\mathbb{X}$ is spanned if and only if $d(x_i, x_j) \leq t$ for all $0 \leq i \leq j \leq k$.

Back to data: Persistent homology

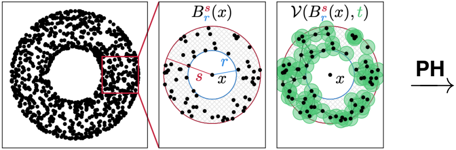
- Given a finite metric space $(\mathbb{X}, d)$, the *Vietoris–Rips complex* at step t is defined as the abstract simplicial complex $\mathcal{V}(\mathbb{X}, t)$, in which an abstract k-simplex $(x_0, \ldots, x_k)$ of points in $\mathbb{X}$ is spanned if and only if $d(x_i, x_j) \leq t$ for all $0 \leq i \leq j \leq k$.
- For $t_1 \leq t_2$, the inclusions $\mathcal{V}(\mathbb{X}, t_1) \hookrightarrow \mathcal{V}(\mathbb{X}, t_2)$ yield a filtration which we denote by $\mathcal{V}(\mathbb{X}, \bullet)$.

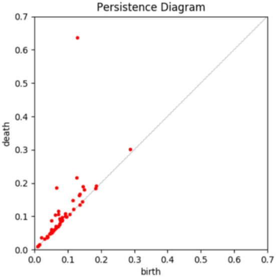
Back to data: Persistent homology

- Given a finite metric space $(\mathbb{X}, d)$, the *Vietoris–Rips complex* at step t is defined as the abstract simplicial complex $\mathcal{V}(\mathbb{X}, t)$, in which an abstract k-simplex $(x_0, \ldots, x_k)$ of points in $\mathbb{X}$ is spanned if and only if $d(x_i, x_j) \leq t$ for all $0 \leq i \leq j \leq k$.
- For $t_1 \leq t_2$, the inclusions $\mathcal{V}(\mathbb{X}, t_1) \hookrightarrow \mathcal{V}(\mathbb{X}, t_2)$ yield a filtration which we denote by $\mathcal{V}(\mathbb{X}, \bullet)$.
- This leads to $H_i(\mathcal{V}(\mathbb{X}, t_1)) \to H_i(\mathcal{V}(\mathbb{X}, t_2))$ for any $t_1 \leq t_2$

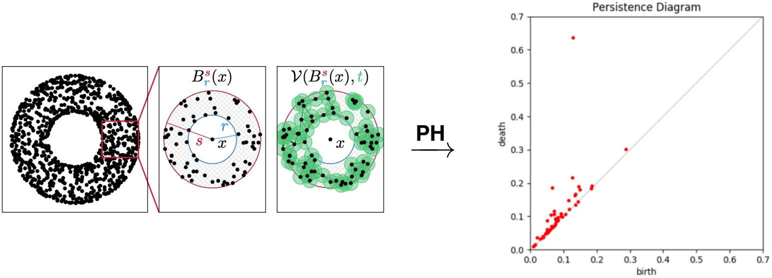
The i-th **persistent homology (PH)** of $\mathbb{X}$ with respect to the Vietoris-Rips construction is defined to be the collection of all these i-th homology groups, together with the respective induced maps between them, and denoted by $\mathrm{PH}_i(\mathcal{V}(\mathbb{X}, \bullet))$

Back to data: Euclidicity





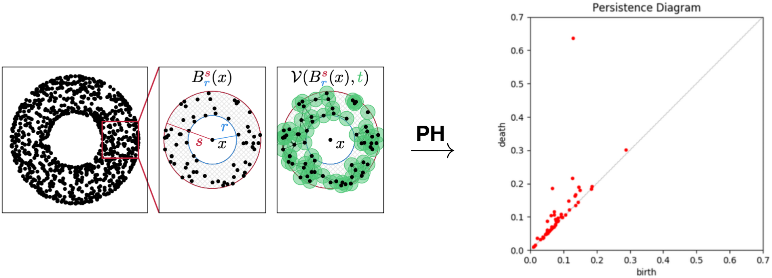
Back to data: Euclidicity



- We denote the resulting persistence information by $\mathrm{PH}(\mathcal{V}(B_r^s(x), \bullet))$.
- The idea is now to compare the topological information of $B_r^s(x)$ with the one of a known Euclidean model space $EucB_r^s(x)$:
$\mathrm{d_B}^{r,s} := \mathrm{d_B}\left[\mathrm{PH}(\mathcal{V}(B_r^s(x), \bullet)), \mathrm{PH}(\mathcal{V}(EucB_r^s(x), \bullet))\right]$
- Finally, we vary r and s and take the average of these distances: $\mathfrak{E}(x) := \frac{1}{C}\sum_{(r,s)} \mathrm{d_B}^{r,s}$

$\mathfrak{E}(x)$ is called the **Euclidicity** of x (w.r.t. the ambient data).

Euclidicity enjoys theoretical guarantees



- $\mathrm{d_B}{}^{r,s} := \mathrm{d_B}\left[\mathrm{PH}(\mathcal{V}(B_r^s(x), \bullet)), \mathrm{PH}(\mathcal{V}(EucB_r^s(x), \bullet))\right]$
- $\mathfrak{E}(x) := \frac{1}{C}\sum_{(r,s)} \mathrm{d_B}{}^{r,s}$

$\mathfrak{E}(x)$ is called the **Euclidicity** of x (w.r.t. the given data $\mathbb{X}$).

When the dataset $\mathbb{X}$ is sampled from a manifold, $\mathfrak{E}(x)$ will be small, for any point x.

Euclidicity tends to zero for 'manifold points'

Theorem

Let $M \subset \mathbb{R}^N$ be a smooth n-dimensional manifold and let $\mathbb{X} \subset M$ be a finite sample of size $S := |\mathbb{X}|$. For a given $\epsilon > 0$, sufficiently large S and a point $x \in \mathbb{X}$, there exists $s_\epsilon > 0$ that (up to a constant) only depends on ϵ, such that $\mathfrak{E}(x)$ is bounded above by ϵ, for any radius configuration with maximum outer radius at most s_ϵ.

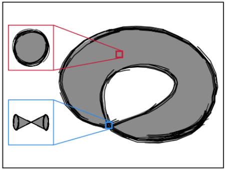
Euclidicity tends to zero for 'manifold points'

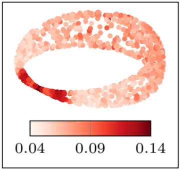
Theorem

Let $M \subset \mathbb{R}^N$ be a smooth n-dimensional manifold and let $\mathbb{X} \subset M$ be a finite sample of size $S := |\mathbb{X}|$. For a given $\epsilon > 0$, sufficiently large S and a point $x \in \mathbb{X}$, there exists $s_\epsilon > 0$ that (up to a constant) only depends on ϵ, such that $\mathfrak{E}(x)$ is bounded above by ϵ, for any radius configuration with maximum outer radius at most s_ϵ.

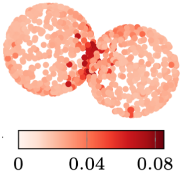
However, $\mathfrak{E}(x)$ will usually *not* tend to zero when x is a singularity! (Homology of the link of x is usually different to the homology of a sphere.)

Euclidicity detects singularities

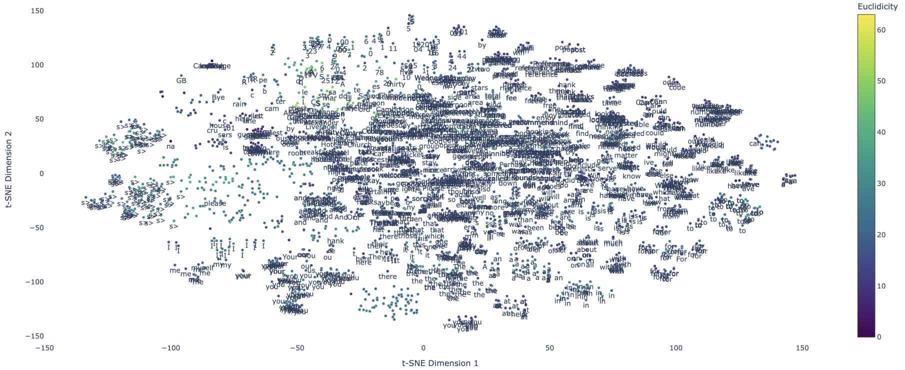




Euclidicity scores of *singular* points are higher than for *non-singular* points.



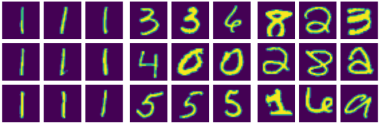
- The following are embeddings of tokens of a Large Language Model (RoBERTa)



Euclidicity detects non-linearities in image datasets

By flattening images, we obtain point cloud representations of image datasets in order to calculate Euclidicity scores.

It turns out that high Euclidicity values correspond to images that possess a high degree of geometric complexity **inside of** the image.



(a) MNIST

(b) FASHIONMNIST

Figure 6: Left to right: samples images exhibiting low, median, and high Euclidicity, respectively.

Misclassified samples admit higher Euclidicity scores

We trained a simple neural network to analyse the Euclidicity scores of misclassified vs. correctly classified samples.

Misclassified samples admit significantly **higher** Euclidicity scores than correctly classified samples.

Acknowledgement: This experiment was conducted together with Francesco Conti (Università di Pisa)

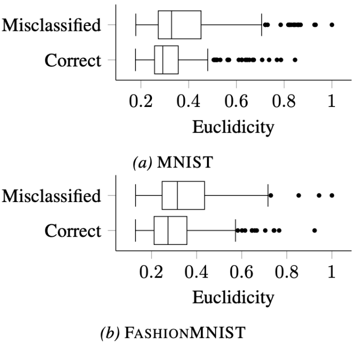


Figure 8: A comparison of Euclidicity scores for misclassified and correctly classified samples in two image data sets.

Persistent intrinsic dimension (PID)

- We have already seen that if $X = M$ is a manifold of dimension n and $x \in M$, its local homology reads

$$H_i(M, M - x) = \tilde{H}_{i-1}(S^{n-1}) = \begin{cases} \mathbb{Z}, i = n \\ 0, i \neq n \end{cases}$$

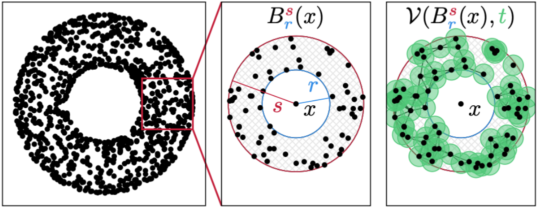
Persistent intrinsic dimension (PID)

- We have already seen that if $X = M$ is a manifold of dimension n and $x \in M$, its local homology reads

$$H_i(M, M - x) = \tilde{H}_{i-1}(S^{n-1}) = \begin{cases} \mathbb{Z}, i = n \\ 0, i \neq n \end{cases}$$

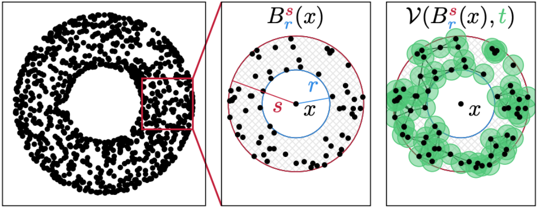
- This means that we can deduce the intrinsic dimension of M, by looking at its local homology!

Persistent intrinsic dimension (PID)



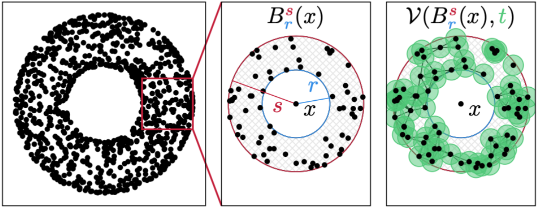
- **Idea for data that is sampled from a manifold:** Same construction as before (in order to approximate the link) and look at the maximum degree homology generators.

Persistent intrinsic dimension (PID)



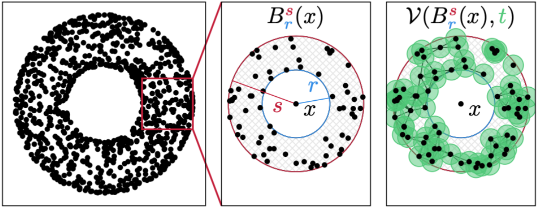
- **Idea for data that is sampled from a manifold:** Same construction as before (in order to approximate the link) and look at the maximum degree homology generators.
- This maximum homology degree is *n*.

Persistent intrinsic dimension (PID)



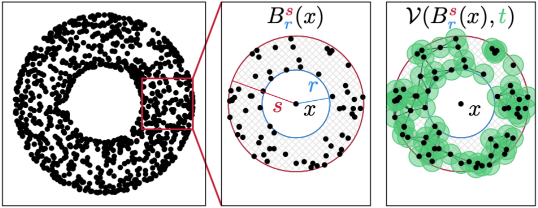
- **Idea for data that is sampled from a manifold:** Same construction as before (in order to approximate the link) and look at the maximum degree homology generators.
- This maximum homology degree is *n*.
- In practice, data may be noisy. We therefore only consider homology generators that exceed a certain persistence threshold.

Persistent intrinsic dimension (PID)



- **Idea for data that is sampled from a manifold:** Same construction as before (in order to approximate the link) and look at the maximum degree homology generators.
- This maximum homology degree is *n*.
- In practice, data may be noisy. We therefore only consider homology generators that exceed a certain persistence threshold.
- Finally, we vary *r* and *s*, and average the resulting dimension estimates.

Persistent intrinsic dimension (PID)



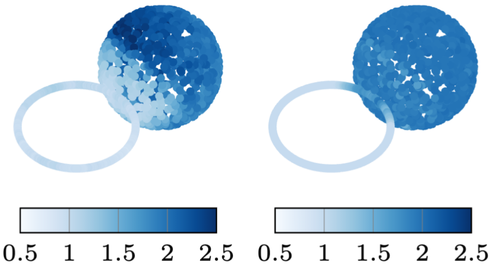
- **Idea for data that is sampled from a manifold:** Same construction as before (in order to approximate the link) and look at the maximum degree homology generators.
- This maximum homology degree is *n*.
- In practice, data may be noisy. We therefore only consider homology generators that exceed a certain persistence threshold.
- Finally, we vary *r* and *s*, and average the resulting dimension estimates. This is called the **persistent intrinsic dimension (PID)** of *x*.

Persistent intrinsic dimension (PID)

Theorem

Let $M \subset \mathbb{R}^N$ be an n-dimensional compact smooth manifold and let $\mathbb{X} := \{x_1, \ldots, x_S\}$ be a collection of uniform samples from M. For a sufficiently large S, PID calculates the correct intrinsic dimension of M in a small neighbourhood around x, for any $x \in M$. Moreover, this neighbourhood can be chosen arbitrarily small by increasing S.

Persistent intrinsic dimension (PID)



(a) twoNN *(b)* PID

- Dimensionality estimates: twoNN vs. PID.
- PID is more nuanced in capturing changes in dimensionality, assigning 1 to almost all points of the circle, i.e. S^1, while highlighting that points closer to S^2 exhibit an increase in dimensionality.

Conclusion and outlook

- Real-world data is often far from being sampled from manifolds.

Conclusion and outlook

- Real-world data is often far from being sampled from manifolds.
- We proposed a framework to assess if a given data point should be considered to lie on a manifold, or not.

Conclusion and outlook

- Real-world data is often far from being sampled from manifolds.
- We proposed a framework to assess if a given data point should be considered to lie on a manifold, or not.
- The given framework can be used to estimate the intrinsic dimension around the data point, locally.

Conclusion and outlook

- Real-world data is often far from being sampled from manifolds.
- We proposed a framework to assess if a given data point should be considered to lie on a manifold, or not.
- The given framework can be used to estimate the intrinsic dimension around the data point, locally.
- Experiments suggest that singularities have meaning: can we regularise for singularities, how?

Code

TARDIS: Topological Algorithms for Robust DIscovery of Singularities

arXiv 2210.00069 | maintainability A | contributors 2 | license BSD-3-Clause | docs passing

`https://github.com/aidos-lab/TARDIS`