Persistent homology for multivariate data visualization

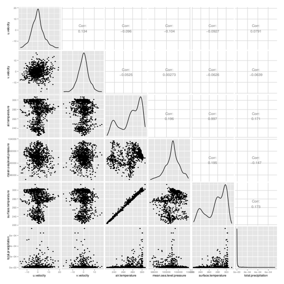
Bastian Rieck

Interdisciplinary Center for Scientific Computing
Heidelberg University

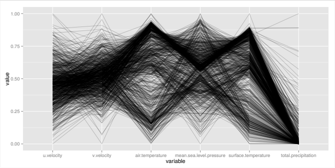
Motivation

Understanding the 'shape' of data

Unstructured data



Scatterplot matrix



Parallel coordinates

Agenda

1. Theory: Algebraic topology
2. Theory: Persistent homology
3. Applications

Part I

Theory: Algebraic topology

Algebraic topology

Algebraic topology is the branch of mathematics that uses tools from abstract algebra to study manifolds. The basic goal is to find algebraic invariants that classify topological spaces up to homeomorphism.

Adapted from `https://en.wikipedia.org/wiki/Algebraic_topology`.

Manifolds

A d-dimensional Riemannian manifold $\mathbb{M}$ in some $\mathbb{R}^n$, with $d \ll n$, is a space where every point $p \in \mathbb{M}$ has a neighbourhood that 'locally looks' like $\mathbb{R}^d$.

*

A 2-dimensional manifold

Homeomorphisms

A homeomorphism between two spaces X and Y is a continuous function $f\colon X \to Y$ whose inverse $f^{-1}\colon Y \to X$ exists and is continuous as well.

$\cong$

Intuitively, we may *stretch*, *bend*—but not *tear* and *glue* the two spaces.

Algebraic invariants

An invariant is a property of an object that remains unchanged upon transformations such as scaling or rotations.

Example

Dimension is a simple invariant: $\mathbb{R}^2 \neq \mathbb{R}^3$ because $2 \neq 3$.

In general

Let $\mathcal{M}$ be the family of manifolds. An invariant permits us to define a function $f\colon \mathcal{M} \times \mathcal{M} \to \{0, 1\}$ that tells us whether two manifolds are different or 'equal' (with respect to that invariant).

No invariant is *perfect*—there will be objects that have the same invariant even though they are different.

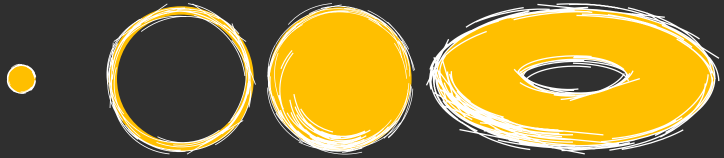
Betti numbers

A useful topological invariant

Informally, they count the number of holes in different dimensions that occur in a data set.

β_0	Connected components
β_1	Tunnels
β_2	Voids
$\vdots$	$\vdots$

Space	β_0	β_1	β_2
Point	1	0	0
Circle	1	1	0
Sphere	1	0	1
Torus	1	2	1



Signature property

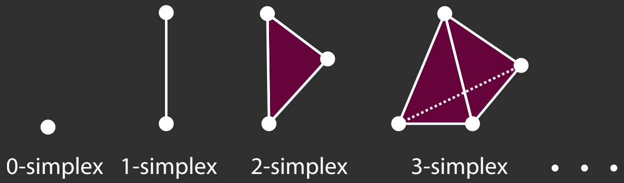
If $\beta_i^X \neq \beta_i^Y$, we know that $X \not\cong Y$. The converse is *not* true, unfortunately:

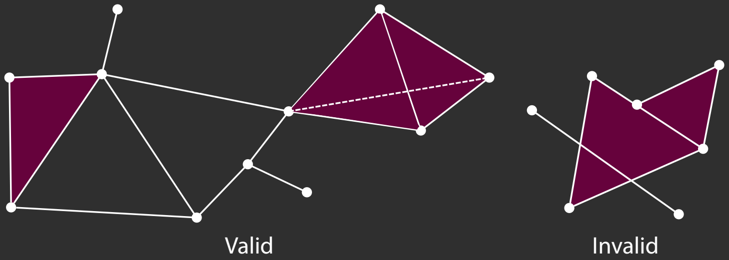
X Y

Space	β_0	β_1
X	1	1
Y	1	1

We have $\beta_0 = 1$ and $\beta_1 = 1$ for X and Y, but still $X \not\cong Y$.

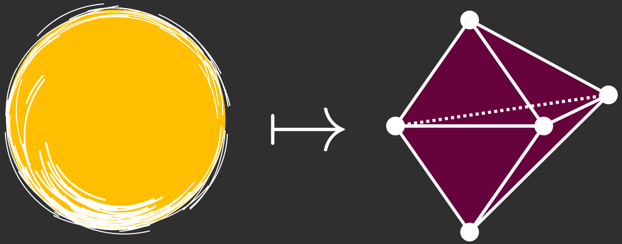
Simplicial complexes





Simplicial complexes

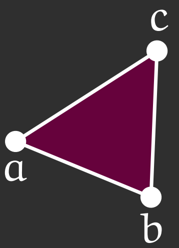
Example



The simplicial complex representation is compact and permits the calculation of the Betti numbers using an efficient matrix reduction scheme.

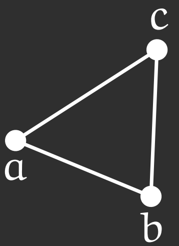
Basic idea

Calculating boundaries



The boundary of the triangle is:

$$\partial_2\{a, b, c\} = \{b, c\} + \{a, c\} + \{a, b\}$$



The set of edges does *not* have boundary:

$$\begin{aligned}\partial_1 & (\{b, c\} + \{a, c\} + \{a, b\}) \\ &= \{c\} + \{b\} + \{c\} + \{a\} + \{b\} + \{a\} \\ &= 0\end{aligned}$$

Fundamental lemma

For all p, we have $\partial_{p-1} \circ \partial_p = 0$: *Boundaries do not have a boundary themselves*.

This permits us to calculate Betti numbers of simplicial complexes by reducing a *boundary matrix* to its *Smith Normal Form* using Gaussian elimination.

Summary

- We want to differentiate between different objects.

Summary

- We want to differentiate between different objects.
- This endeavour requires algebraic invariants.

Summary

- We want to differentiate between different objects.
- This endeavour requires algebraic invariants.
- One invariant, the *Betti numbers*, measures intuitive aspects of our data.

Summary

- We want to differentiate between different objects.
- This endeavour requires algebraic invariants.
- One invariant, the *Betti numbers*, measures intuitive aspects of our data.
- Their calculation requires a *simplicial complex* and a *boundary operator*.

Summary

- We want to differentiate between different objects.
- This endeavour requires algebraic invariants.
- One invariant, the *Betti numbers*, measures intuitive aspects of our data.
- Their calculation requires a *simplicial complex* and a *boundary operator*.

Part II

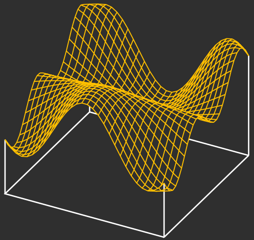
Theory: Persistent homology

Real-world multivariate data

- Unstructured point clouds
- n items with D attributes; $n \times D$ matrix
- Non-random sample from $\mathbb{R}^D$

Manifold hypothesis

There is an unknown d-dimensional manifold $\mathbb{M} \subseteq \mathbb{R}^D$, with $d \ll D$, from which our data have been sampled.



*

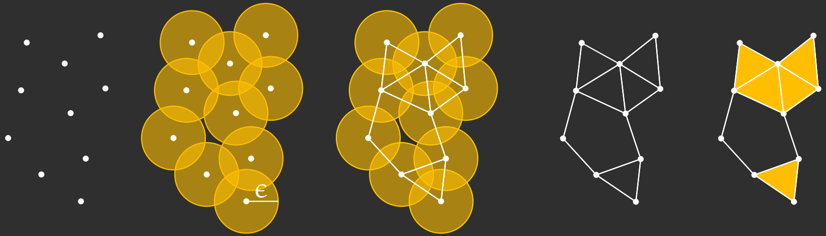
2-manifold in $\mathbb{R}^3$

Agenda

1. Convert our input data into a simplicial complex K.
2. Calculate the Betti numbers of K.
3. Use the Betti numbers to compare data sets.

(Fair warning: It won’t be so simple)

Converting unstructured data into a simplicial complex

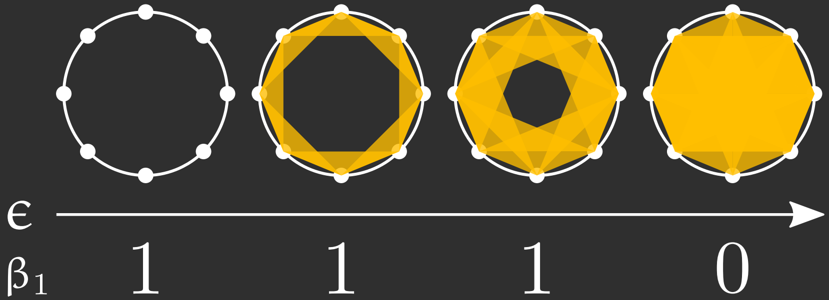


Require: Distance measure (e.g. Euclidean distance), maximum scale threshold ϵ.

Construct the Vietoris–Rips complex $\mathcal{V}_\epsilon$ by adding a k-simplex whenever all of its $(k-1)$-dimensional faces are present.

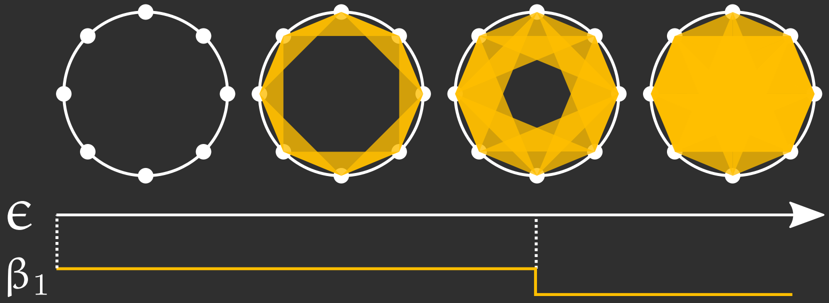
Calculating Betti numbers directly from $\mathcal{V}_\epsilon$

Unstable behaviour



Calculating *persistent* Betti numbers

Persistent homology



How does this work in practice?

An example for β_0

- Have a function $f\colon \mathcal{V}_\epsilon \to \mathbb{R}$ on the *vertices* of the Vietoris–Rips complex.

How does this work in practice?

An example for β_0

- Have a function $f\colon \mathcal{V}_\epsilon \to \mathbb{R}$ on the *vertices* of the Vietoris–Rips complex.
- Extend it to a function on the whole simplicial complex by setting $f(\sigma) = \max\{f(v) \mid v \in \sigma\}$.

How does this work in practice?

An example for β_0

- Have a function $f\colon \mathcal{V}_\epsilon \to \mathbb{R}$ on the *vertices* of the Vietoris–Rips complex.
- Extend it to a function on the whole simplicial complex by setting $f(\sigma) = \max\{f(v) \mid v \in \sigma\}$.
- Analyse the connectivity changes in the sublevel sets of f, i.e. sets of the form

$$L_\alpha^-(f) = \{v \mid f(v) \leqslant \alpha\}.$$

How does this work in practice?

An example for β_0

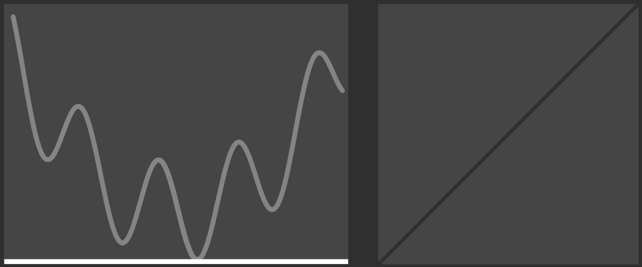
- Have a function $f\colon \mathcal{V}_\epsilon \to \mathbb{R}$ on the *vertices* of the Vietoris–Rips complex.
- Extend it to a function on the whole simplicial complex by setting $f(\sigma) = \max\{f(v) \mid v \in \sigma\}$.
- Analyse the connectivity changes in the sublevel sets of f, i.e. sets of the form

$$L_\alpha^-(f) = \{v \mid f(v) \leqslant \alpha\}.$$

- This can be done by traversing the values of f in increasing order and stopping at 'critical points'.

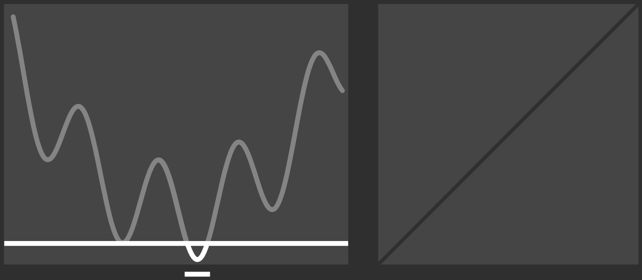
Persistent homology & persistence diagrams

One-dimensional example



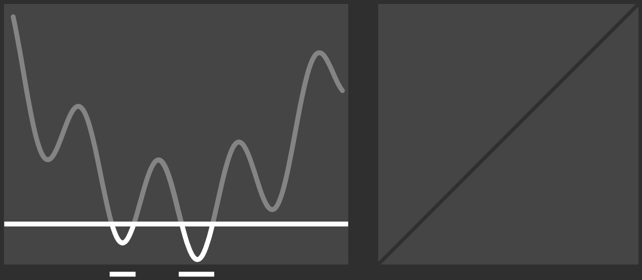
Persistent homology & persistence diagrams

One-dimensional example



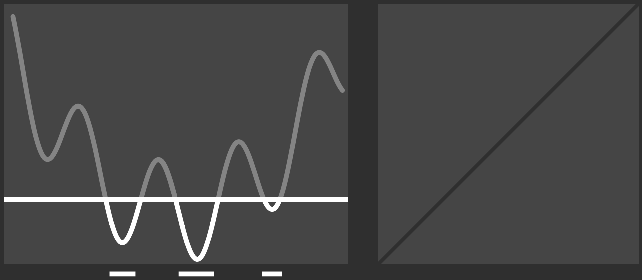
Persistent homology & persistence diagrams

One-dimensional example



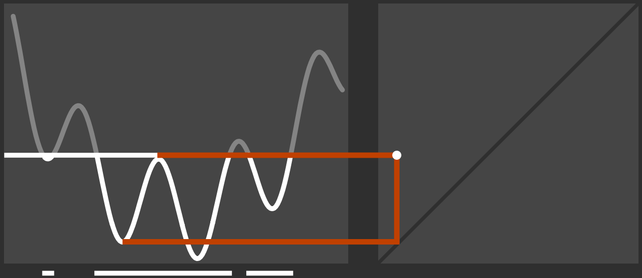
Persistent homology & persistence diagrams

One-dimensional example



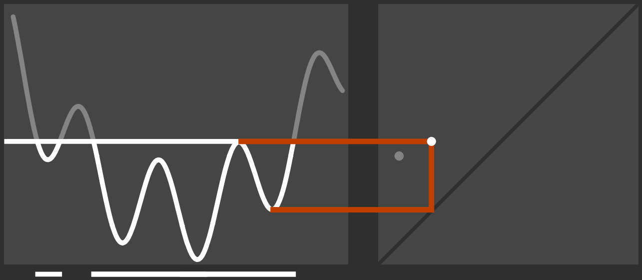
Persistent homology & persistence diagrams

One-dimensional example



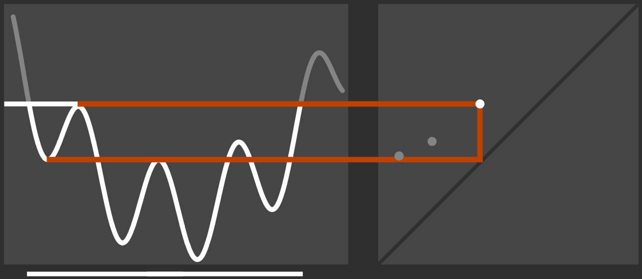
Persistent homology & persistence diagrams

One-dimensional example



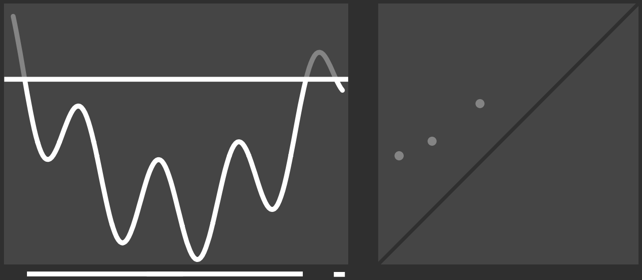
Persistent homology & persistence diagrams

One-dimensional example



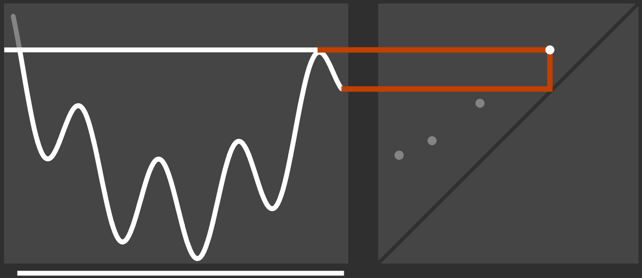
Persistent homology & persistence diagrams

One-dimensional example



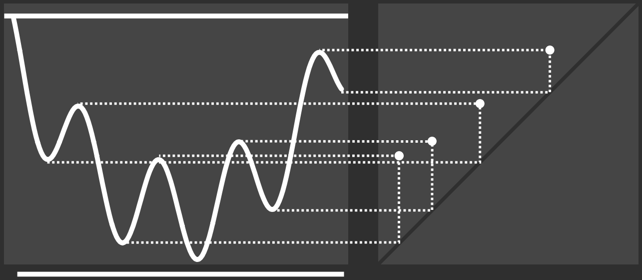
Persistent homology & persistence diagrams

One-dimensional example



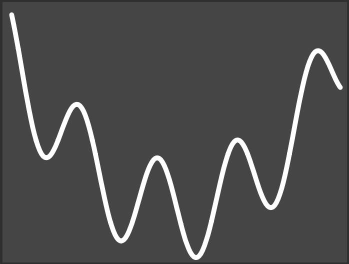
Persistent homology & persistence diagrams

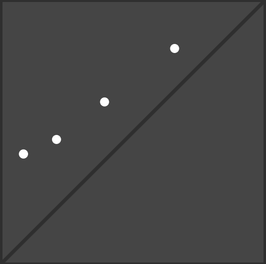
One-dimensional example



Persistent homology & persistence diagrams

One-dimensional example



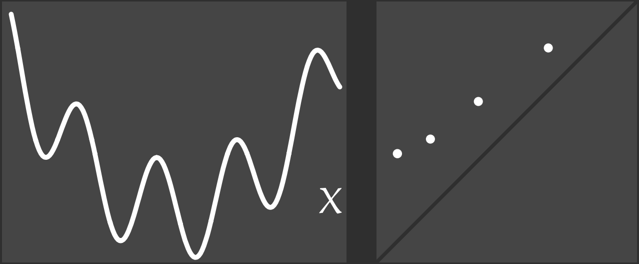


Uses for persistence diagrams

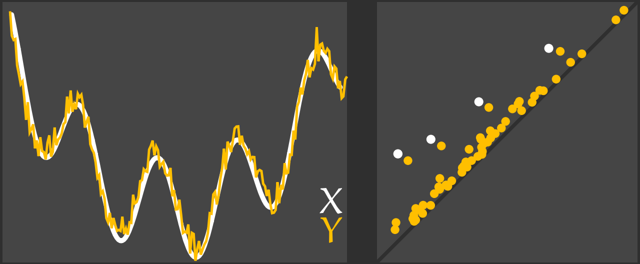
A persistence diagram is a multi-scale summary of topological activity in a data set. But the diagrams go well and beyond a simple comparison of Betti numbers!

> *L'algèbre est généreuse, elle donne souvent plus qu'on lui demande.*
>
> —D'Alembert

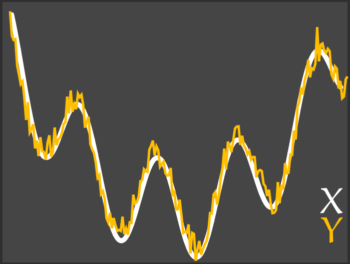
Distance calculations

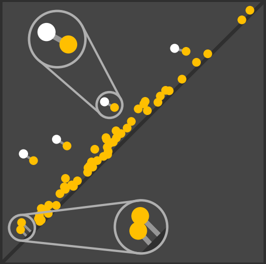


Distance calculations



Distance calculations

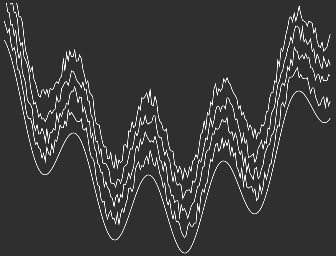




$$W_2(X, Y) = \sqrt{\inf_{\eta : X \to Y} \sum_{x \in X} \|x - \eta(x)\|_\infty^2}$$

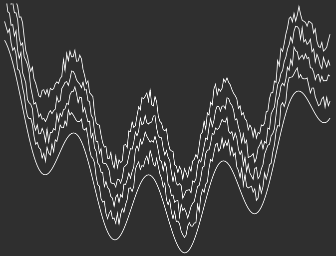
Sensitivity of distances

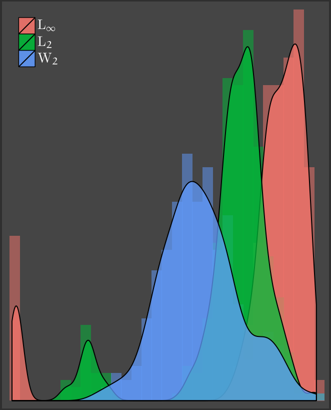
Wasserstein versus function space distances



Sensitivity of distances

Wasserstein versus function space distances



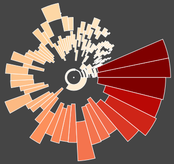


Only the Wasserstein distance does not distort the 'shape' of noise in the data.

Part III

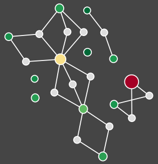
Applications

Published projects



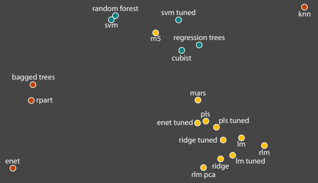
Persistence rings

Rieck, Mara, Leitte: *Multivariate data analysis using persistence-based filtering and topological signatures*



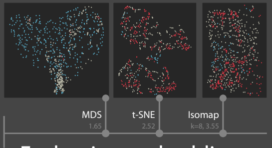
Simplicial chain graphs

Rieck, Leitte: *Structural analysis of multivariate point clouds using simplicial chains*



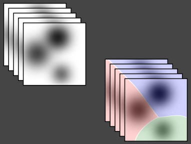
Model landscapes

Rieck, Leitte: *Enhancing comparative model analysis using persistent homology*



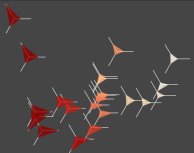
Evaluating embeddings

Rieck, Leitte: *Persistent homology for the evaluation of dimensionality reduction schemes*



Agreement analysis

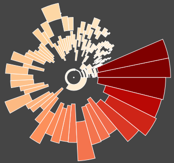
Rieck, Leitte: *Agreement analysis of quality measures for dimensionality reduction*



Data descriptor landscapes

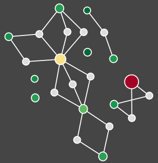
Rieck, Leitte: *Comparing dimensionality reduction methods using data descriptor landscapes*

Published projects



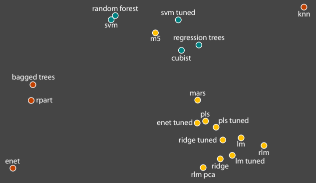
Persistence rings

Rieck, Mara, Leitte: *Multivariate data analysis using persistence-based filtering and topological signatures*



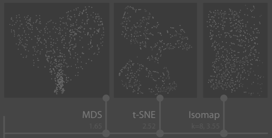
Simplicial chain graphs

Rieck, Leitte: *Structural analysis of multivariate point clouds using simplicial chains*



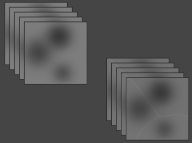
Model landscapes

Rieck, Leitte: *Enhancing comparative model analysis using persistent homology*



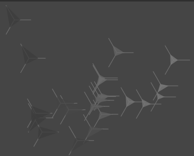
Evaluating embeddings

Rieck, Leitte: *Persistent homology for the evaluation of dimensionality reduction schemes*



Agreement analysis

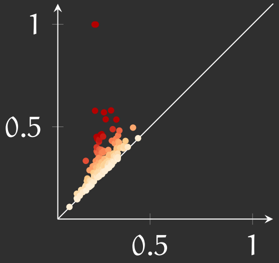
Rieck, Leitte: *Agreement analysis of quality measures for dimensionality reduction*



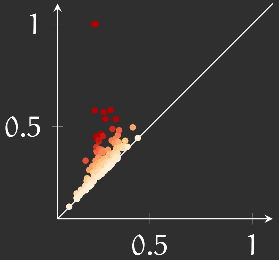
Data descriptor landscapes

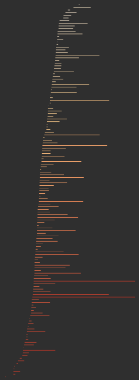
Rieck, Leitte: *Comparing dimensionality reduction methods using data descriptor landscapes*

Qualitative visualizations: Persistence rings

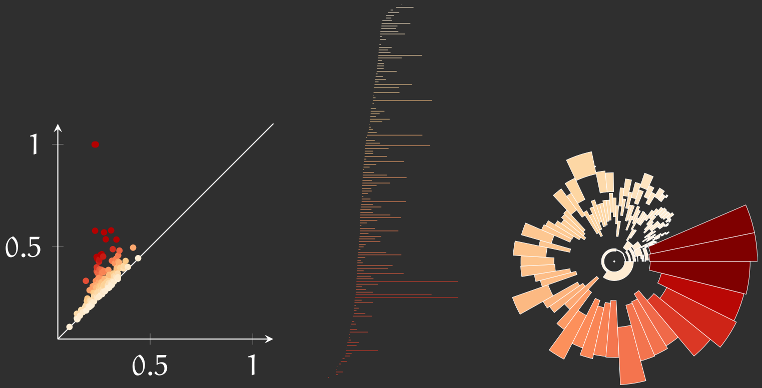


Qualitative visualizations: Persistence rings





Qualitative visualizations: Persistence rings



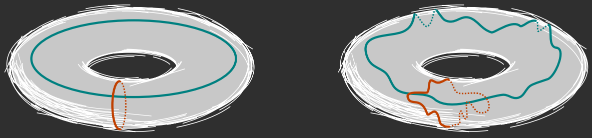
Qualitative visualizations: Simplicial chain graphs

Motivation

Analyse the connectivity of topological features for *ensembles* of data sets: Different runs of an experiment, different times at which measurements are being taken…

Central idea

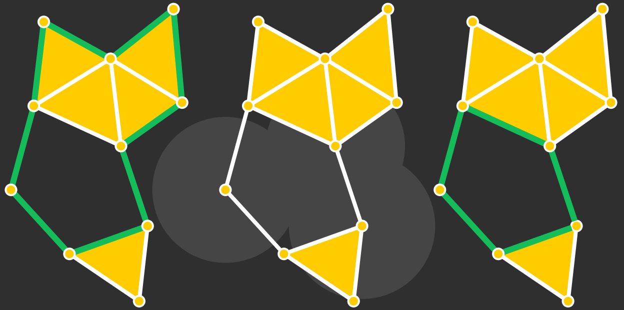
Obtain geometrical descriptions of topological features ('holes') while calculating persistent homology.



This is known as the 'localization problem' in persistent homology.

Solving the localization problem

1. Define a *geodesic ball* in a simplicial complex.
2. Solve all-pairs-shortest-paths problem to find possible sites.
3. Branch-and-bound strategy to improve performance.



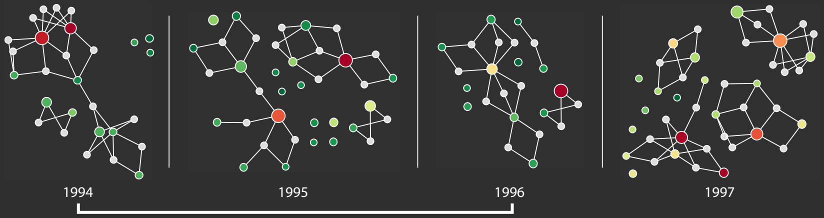
Simplicial chain graph

Basic example

Advantage: The space in which we localize the features usually has a high dimensions, but the graph will always be drawn in $\mathbb{R}^2$.

Results

Data: Tropical Atmosphere Ocean Array



Lessons learned

1. We can obtain features via persistent homology that permit a comparative analysis.
2. Visualizing these features becomes abstract very quickly.
3. Need more ‘quantitative’ topological visualizations.

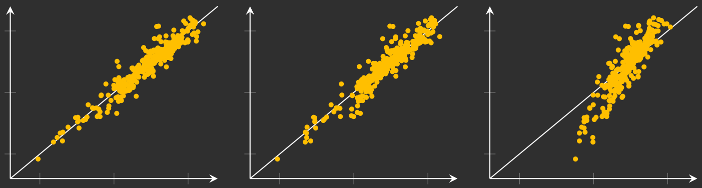
Quantitative visualizations: Model landscapes

Example: Solubility analysis

- 19 different mathematical models
- 1267 chemical compounds, described by 228-dimensional feature vectors
- Measured ground truth (solubility values)

Each model is a function $f\colon \mathbb{D} \to \mathbb{R}$. How to evaluate similarities & differences between the models?

State-of-the-art



- Existing measures (RMSE or R^2) only focus on *values* of a model.
- The structure/shape is not being used!
- Shortcomings: Sensitivity to noise, 'masking' the influence of outliers…

Our approach

1. Calculate Vietoris–Rips complex $\mathcal{V}_\epsilon$ on the molecular descriptors.

Our approach

1. Calculate Vietoris–Rips complex $\mathcal{V}_\epsilon$ on the molecular descriptors.
2. Use model values & ground truth to obtain a set of functions on $\mathcal{V}_\epsilon$.

Our approach

1. Calculate Vietoris–Rips complex $\mathcal{V}_\epsilon$ on the molecular descriptors.
2. Use model values & ground truth to obtain a set of functions on $\mathcal{V}_\epsilon$.
3. Calculate persistence diagrams for each function.

Our approach

1. Calculate Vietoris–Rips complex $\mathcal{V}_\epsilon$ on the molecular descriptors.
2. Use model values & ground truth to obtain a set of functions on $\mathcal{V}_\epsilon$.
3. Calculate persistence diagrams for each function.
4. Compare diagrams using the Wasserstein distance.

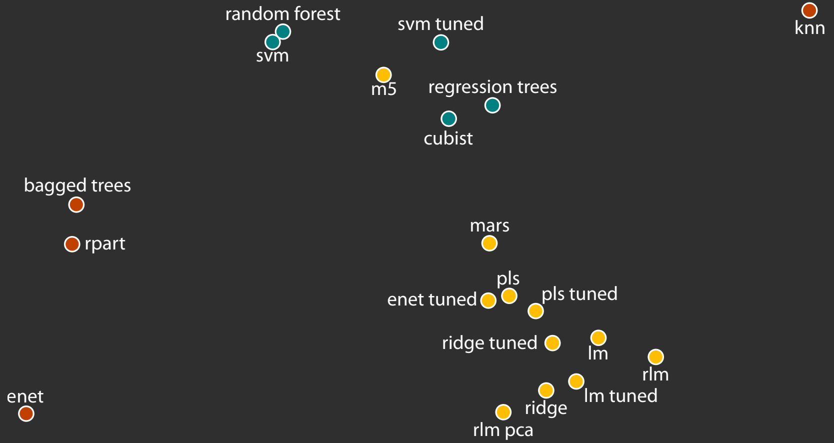
Our approach

1. Calculate Vietoris–Rips complex $\mathcal{V}_\epsilon$ on the molecular descriptors.
2. Use model values & ground truth to obtain a set of functions on $\mathcal{V}_\epsilon$.
3. Calculate persistence diagrams for each function.
4. Compare diagrams using the Wasserstein distance.
5. *Absolute* comparison with ground truth diagram.

Our approach

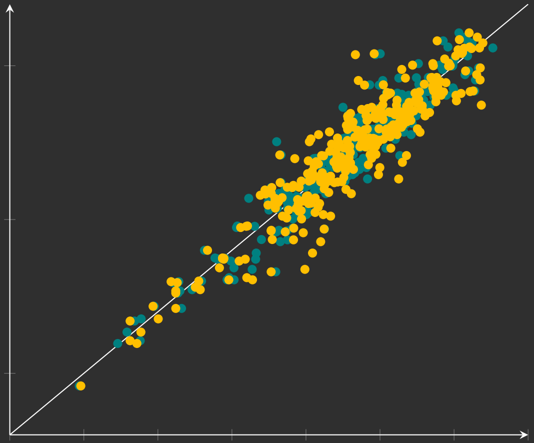
1. Calculate Vietoris–Rips complex $\mathcal{V}_\epsilon$ on the molecular descriptors.
2. Use model values & ground truth to obtain a set of functions on $\mathcal{V}_\epsilon$.
3. Calculate persistence diagrams for each function.
4. Compare diagrams using the Wasserstein distance.
5. *Absolute* comparison with ground truth diagram.
6. *Relative* comparison with all diagrams.

Visualization of relative model differences



Why is m5 rated differently?

Comparison with cubist



Summary

Topological methods yield quantitative and qualitative information about data sets—often, this goes well and beyond the scope of regular geometric approaches.

Summary

Topological methods yield quantitative and qualitative information about data sets—often, this goes well and beyond the scope of regular geometric approaches.

Future work

1. Performance improvements: Smaller complexes, other distance measures, …
2. Ensemble data & 'average' topological structures
3. Connection to geometric features in data

Summary

Topological methods yield quantitative and qualitative information about data sets—often, this goes well and beyond the scope of regular geometric approaches.

Future work

1. Performance improvements: Smaller complexes, other distance measures, …
2. Ensemble data & 'average' topological structures
3. Connection to geometric features in data

Thank you for your attention!