An introduction to persistent homology

Bastian Rieck

Interdisciplinary Center for Scientific Computing
Heidelberg University

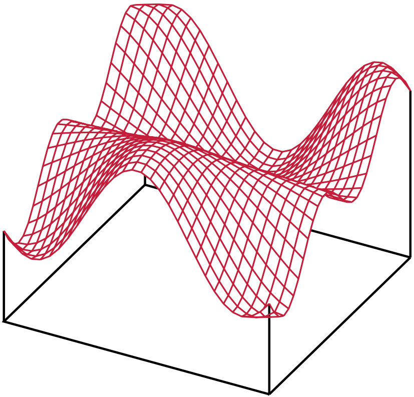
UNIVERSITÄT
HEIDELBERG
ZUKUNFT
SEIT 1386

HGS
MathComp

IWR

Motivation

What is the ‘shape’ of data?



Agenda

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

Part I

Theory: Algebraic topology

Algebraic topology

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

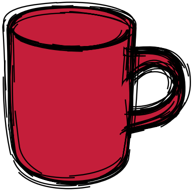
Topological spaces

Manifolds

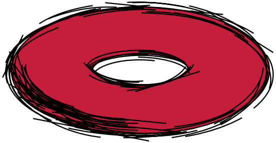
A d-dimensional Riemannian manifold $\mathbb{M}$ in some $\mathbb{R}^{n}$, with $d \ll n$, is a space where every point $p \in \mathbb{M}$ has a neighbourhood that 'locally looks' like $\mathbb{R}^{d}$.

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* the two spaces.

Algebraic invariants

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

Examples

1. *Dimension*: $\mathbb{R}^2 \neq \mathbb{R}^3$ because $2 \neq 3$.
2. *Determinant*: If matrices A and B are similar, their determinants are equal.

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.

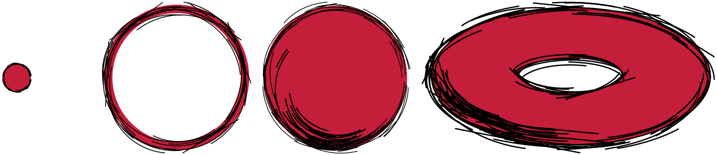
Betti numbers

A topological invariant

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

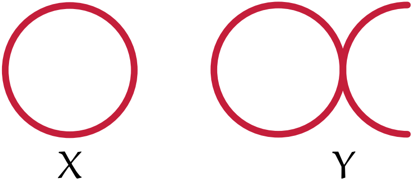
β_0	Connected components
β_1	Tunnels
β_2	Voids
$\vdots$	$\vdots$

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 $X \not\simeq Y$. The converse is *not* true, unfortunately:



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 $X \not\simeq Y$.

But to be completely honest, the second object is technically not a manifold. This is only meant as an illustration of the issue.

Calculating Betti numbers

The kth Betti number β_k is the rank of the kth *homology group* $H_k(X)$ of the topological space X.

Technically, I should write *simplicial homology group* everytime. I am not going to do this. Instead, let's first talk about *simplicial complexes*.

Simplicial complexes

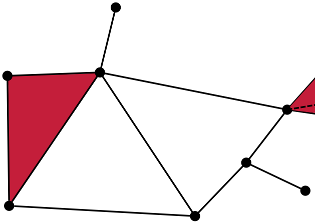
A family of sets K with a collection of subsets S is called an *abstract simplicial complex* if:

1. $\{v\} \in S$ for all $v \in K$.
2. If $\sigma \in S$ and $\tau \subseteq \sigma$, then $\tau \in K$.

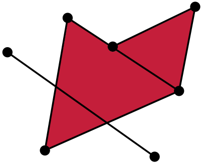
The elements of a simplicial complex are called *simplices*. A k-simplex consists of $k + 1$ indices.

Simplicial complexes

Example



Valid



Invalid

Chain groups

Given a simplicial complex K, the pth chain group C_p of K contains all linear combinations of p-simplices in the complex. Coefficients are in $\mathbb{Z}_2$, hence all elements of C_p are of the form $\sum_j \sigma_j$, for $\sigma_j \in K$. The group operation is addition with $\mathbb{Z}_2$ coefficients.

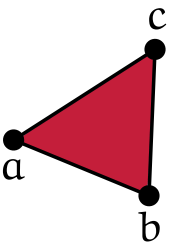
Example

$$\{a, b, c\} + \{a, b, c, d, e\}$$
$$\{a, b, c\} + \{a, b, c\} + \{a, b\} = \{a, b\}$$

We need chain groups to algebraically express the concept of a *boundary*.

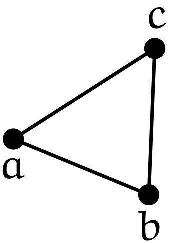
Basic idea

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:

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

Boundary homomorphism

Given a simplicial complex K, the pth boundary homomorphism is the homomorphism that assigns each simplex $\sigma = \{v_0, \ldots, v_p\} \in K$ to its boundary:

$$\partial_p \sigma = \sum_i \{v_0, \ldots, \hat{v}_i, \ldots, v_k\} \tag{1}$$

In the equation above, $\hat{v}_i$ indicates that the set does *not* contain the ith vertex. The function $\partial_p \colon C_p \to C_{p-1}$ is thus a homomorphism between the chain groups.

Fundamental lemma

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

Chain complex

$$0 \xrightarrow{\partial_{n+1}} C_n \xrightarrow{\partial_n} C_{n-1} \xrightarrow{\partial_{n-1}} \dots \xrightarrow{\partial_2} C_1 \xrightarrow{\partial_1} C_0 \xrightarrow{\partial_0} 0 \qquad (2)$$

Cycle and boundary groups

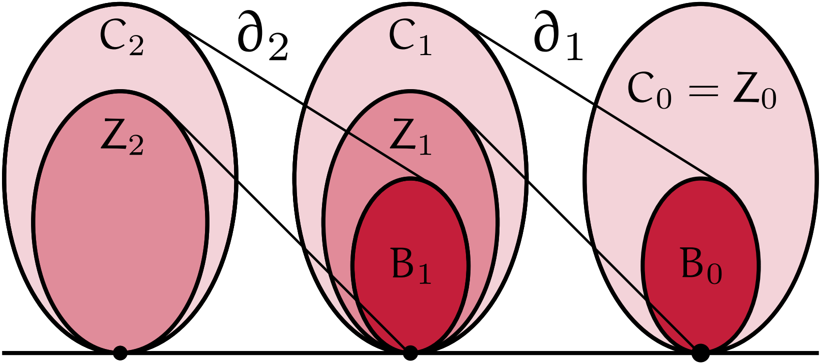
$$\text{Cycle group } Z_p = \ker \partial_p \tag{3}$$

$$\text{Boundary group } B_p = \operatorname{im} \partial_{p+1} \tag{4}$$

We have $B_p \subseteq Z_p$ in the group-theoretical sense. In other words, every boundary is also a cycle.

Illustration of the nesting relations

Following Zomorodian, Edelsbrunner, and many more…



Homology groups & Betti numbers

The pth homology group H_p is a quotient group, defined by 'removing' cycles that are boundaries from a higher dimension:

$$H_p = Z_p / B_p = \operatorname{ker} \partial_p / \operatorname{im} \partial_{p+1}, \tag{5}$$

With this definition, we may finally calculate the pth Betti number:

$$\beta_p = \operatorname{rank} H_p \tag{6}$$

Intuitively: Calculate all boundaries, remove the boundaries that come from higher-dimensional objects, and count what's left.

Summary

- We want to differentiate between different objects.

Summary

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

Summary

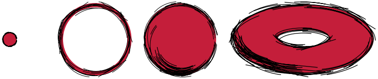
- We want to differentiate between different objects.
- This endeavour requires algebraic invariants.
- One invariant, the *Betti numbers*, measures intuitive aspects of our data.

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

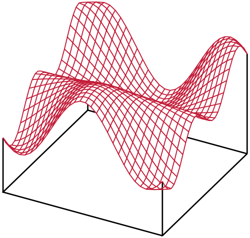
Theory: Persistent homology

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 $\mathbb{R}^3$

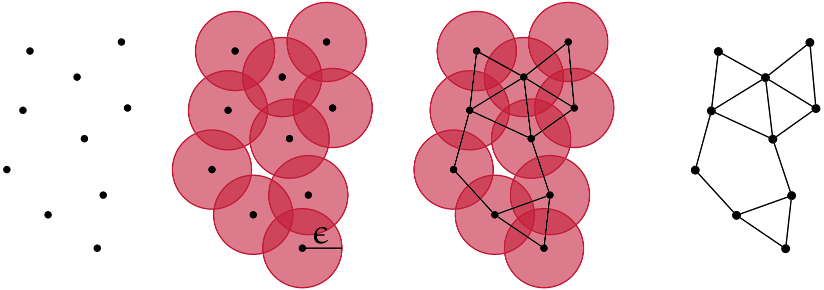
Agenda

1. Convert our input data into a simplicial complex $\mathbb{K}$.
2. Calculate simplicial homology of $\mathbb{K}$.
3. Use the Betti numbers to distinguish between different data sets.

Fair warning: It won’t be so simple, of course...

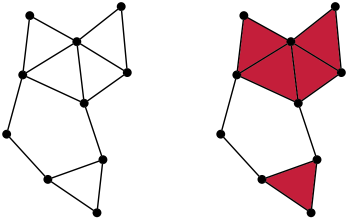
Converting unstructured data into a simplicial complex

Rips graph $\mathcal{R}_\epsilon$



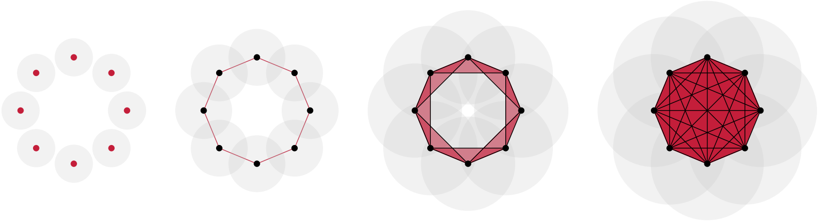
How to get a simplicial complex from $\mathcal{R}_\epsilon$?

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

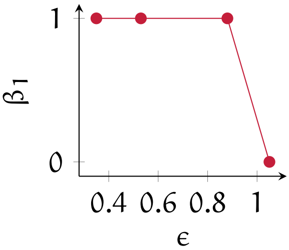


Calculating Betti numbers directly from $\mathcal{V}_\epsilon$

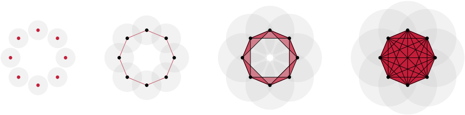
Unstable behaviour



$\epsilon = 0.35$ $\epsilon = 0.53$ $\epsilon = 0.88$ $\epsilon = 1.05$



Solution: Persistent homology



Exploit the nesting properties of the Rips graph and the Vietoris–Rips complex. For $\epsilon \leqslant \epsilon'$, we have:

$$\mathcal{R}_\epsilon \subseteq \mathcal{R}_{\epsilon'} \tag{7}$$

$$\mathcal{V}_\epsilon \subseteq \mathcal{V}_{\epsilon'} \tag{8}$$

Hence: Calculate ‘multi-scale Betti numbers’—observe how the Betti numbers change with a varying distance threshold.

Technical details

A *filtration* is a sequence of sets

$$\emptyset = K_0 \subseteq K_1 \subseteq \cdots \subseteq K_{n-1} \subseteq K_n = K \tag{9}$$

such that each K_i is a valid simplicial subcomplex of K. It turns out that we can reduce a *single* large boundary matrix of a complex in filtration order to get persistent homology!

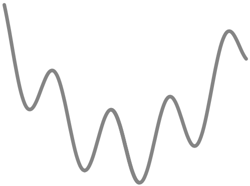
Central idea: A simplex may either *increase* the Betti number in a certain dimension, *decrease* it, or not change it at all.

Further details: *Computing Persistent Homology* (Afra Zomorodian and Gunnar Carlsson), 2005.

Persistent homology & persistence diagrams

One-dimensional example

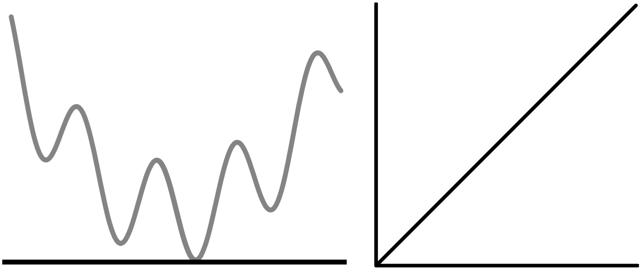
The simplicial complex is implicitly given by connecting points that are ‘adjacent’ on the function.



Filtration order is given by traversing function values in ascending order. We shall observe changes in the *connected components* of the sublevel sets $L_c^-(f) = \{x \mid f(x) \leqslant c\}$ of the function.

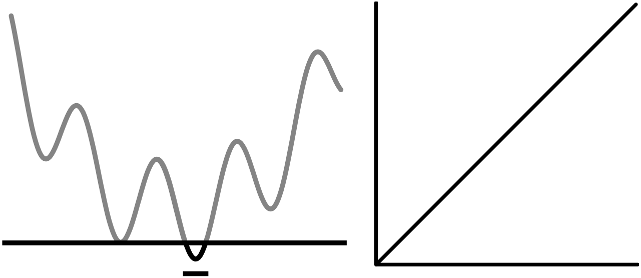
Persistent homology & persistence diagrams

One-dimensional example



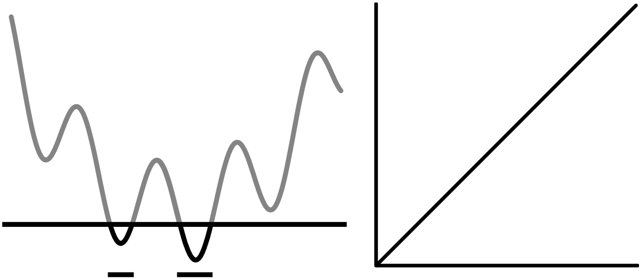
Persistent homology & persistence diagrams

One-dimensional example



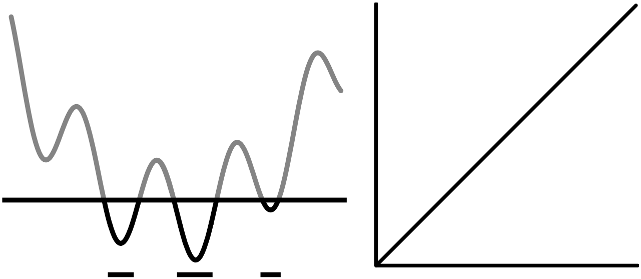
Persistent homology & persistence diagrams

One-dimensional example



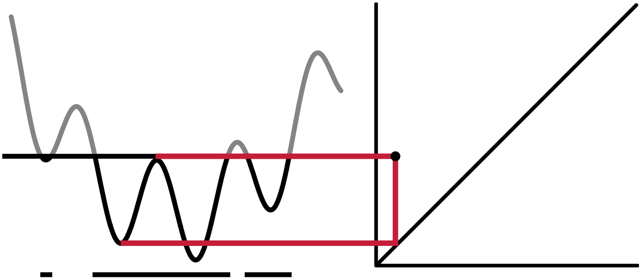
Persistent homology & persistence diagrams

One-dimensional example



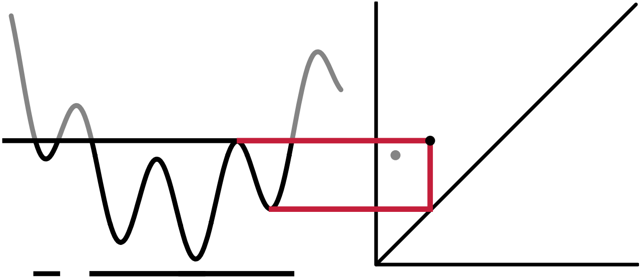
Persistent homology & persistence diagrams

One-dimensional example



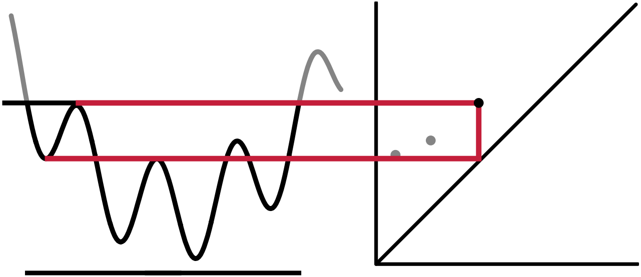
Persistent homology & persistence diagrams

One-dimensional example



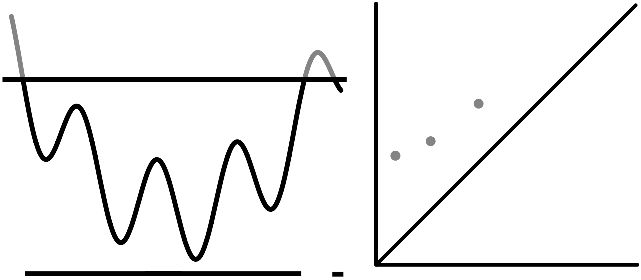
Persistent homology & persistence diagrams

One-dimensional example



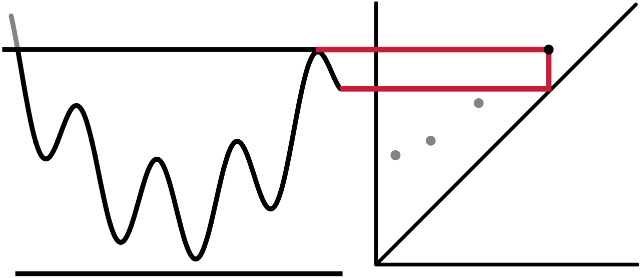
Persistent homology & persistence diagrams

One-dimensional example



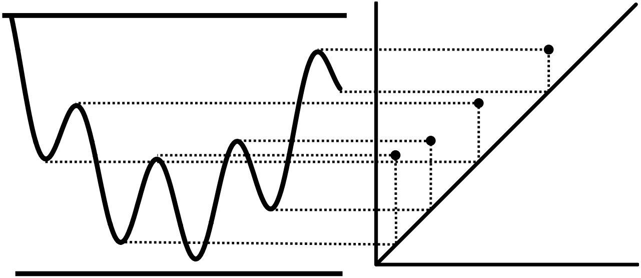
Persistent homology & persistence diagrams

One-dimensional example



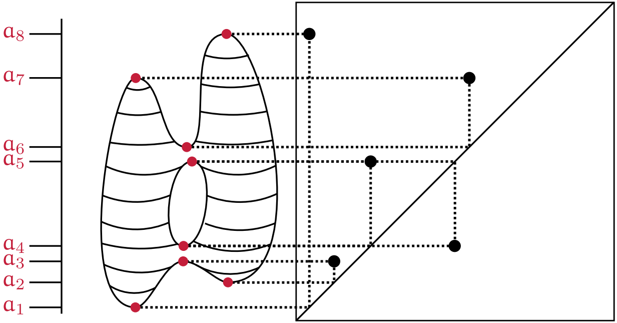
Persistent homology & persistence diagrams

One-dimensional example



Connections to Morse theory

2D manifolds

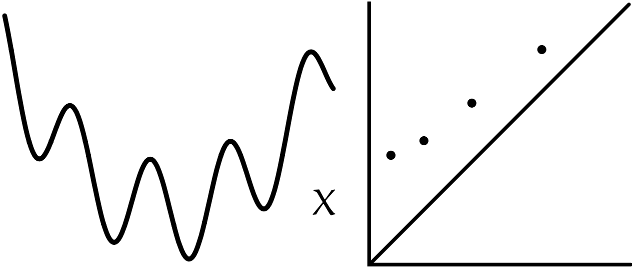


Critical points are in one-to-one correspondence with points in the persistence diagram:

- Minima create new connected components
- Maxima destroy connected components by merging them
- Saddle points either create holes or merge two connected components

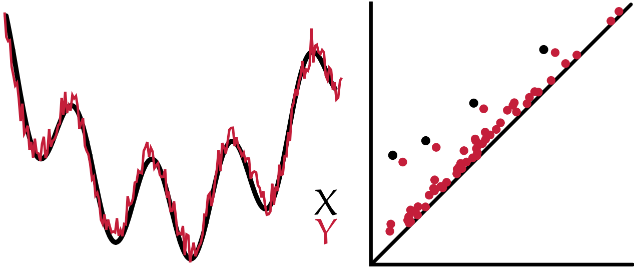
Uses for persistence diagrams

Distance calculations



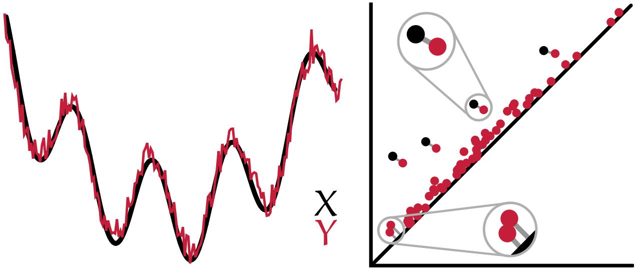
Uses for persistence diagrams

Distance calculations



Uses for persistence diagrams

Distance calculations



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

Stability

Theorem

Let f *and* g *be two Lipschitz-continuous functions. There are constants* k *and* C *that depend on the input space and on the Lipschitz constants of* f *and* g *such that*

$$W_2(X, Y) \leqslant C\|f - g\|_\infty^{1-\frac{k}{2}}, \tag{10}$$

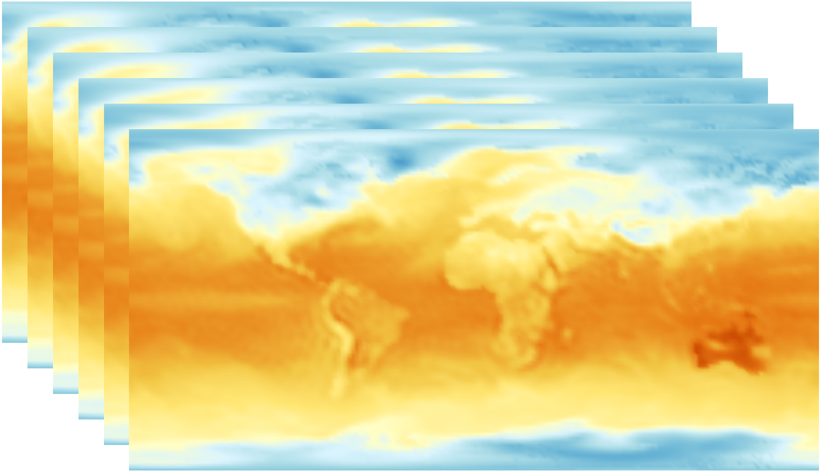
where X *and* Y *refer to the persistence diagrams of* f *and* g*.*

Part III

Examples

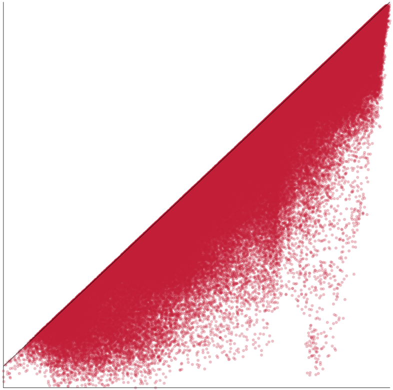
Scalar field analysis

Climate research



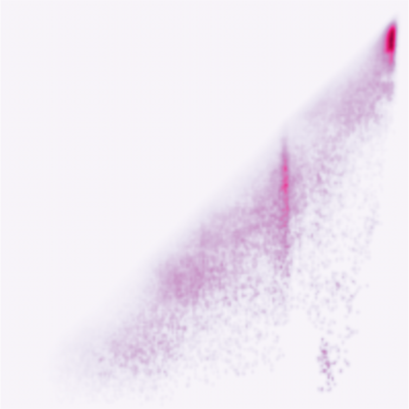
Combined persistence diagram

1460 time steps



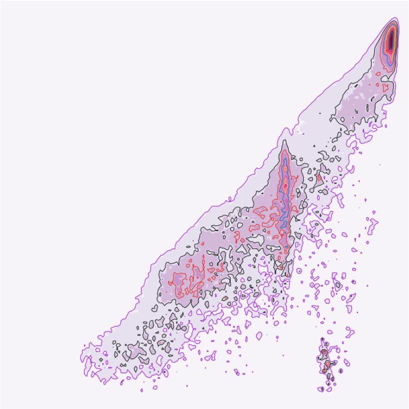
Combined persistence diagram

Kernel density estimates



Combined persistence diagram

Kernel density estimates



What kind of questions does this answer?

- What is the topology of an ‘average’ climate data scalar field?
- What time steps are outliers in the topological sense?
- Are two runs of a time-varying simulation similar?

Statistical view: Two-sample tests, clustering, …

Towards topological time-series analysis

Derived properties of persistence diagrams

∞-norm:

$$\|\mathcal{D}\|_\infty := \max_{(c,d)\in\mathcal{D}} |c-d| \tag{11}$$

p-norm:

$$\|\mathcal{D}\|_p := \left(\sum_{(c,d)\in\mathcal{D}} (c-d)^p\right)^{\frac{1}{p}} \tag{12}$$

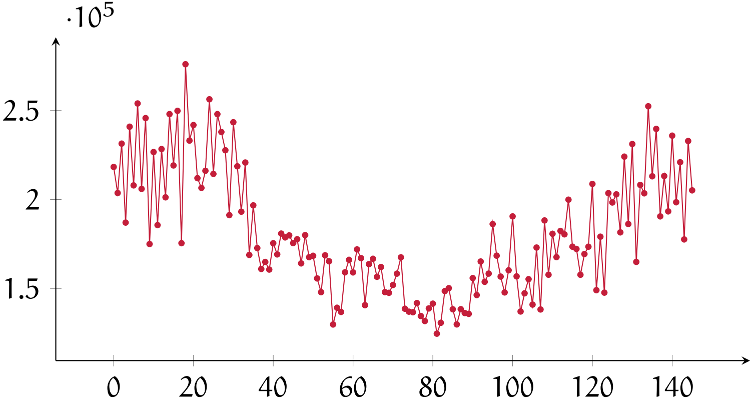
Total persistence:

$$\mathrm{pers}(\mathcal{D})_p := \sum_{(c,d)\in\mathcal{D}} (c-d)^p \tag{13}$$

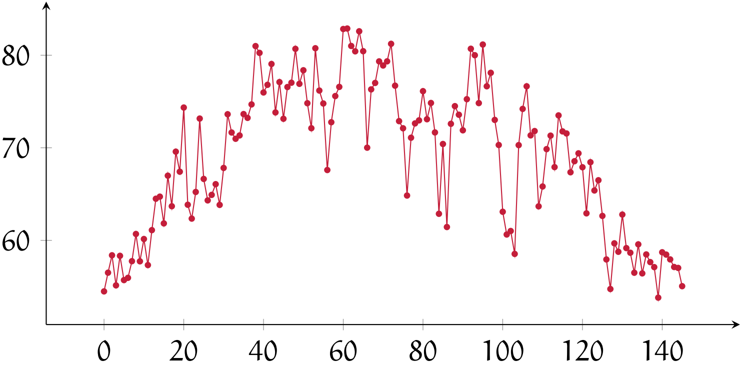
In essence, these are *topological summary statistics*.

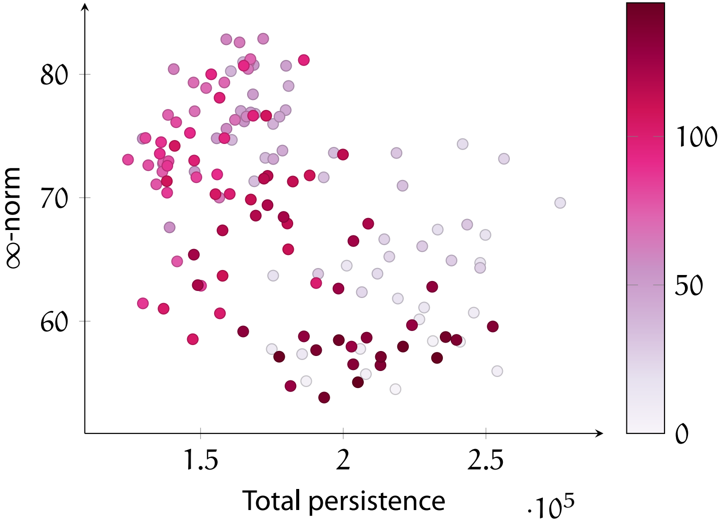
Total persistence

$p = 2$



∞-norm




80
70
60
∞-norm
1.5
2
2.5
Total persistence
$\cdot 10^5$
100
50
0

Conclusion

Take-away messages

1. Persistent homology is a new way of looking at complex data.
2. It has a rich mathematical theory and many desirable properties (robustness, invariance).
3. Lots of interesting applications!

Interested? Drop me a line at bastian.rieck@iwr.uni-heidelberg.de!