An introduction to persistent homology

Bastian Rieck

Interdisciplinary Center for Scientific Computing
Heidelberg University
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 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 [...]
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 homeomorphism between two spaces $X$ and $Y$ is a continuous function $f : X \rightarrow Y$ whose inverse $f^{-1} : Y \rightarrow X$ exists and is continuous as well.

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 $f: \mathcal{M} \times \mathcal{M} \to \{0, 1\}$ 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.

<table>
<thead>
<tr>
<th>$\beta_0$</th>
<th>Connected components</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_1$</td>
<td>Tunnels</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>Voids</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Space</th>
<th>$\beta_0$</th>
<th>$\beta_1$</th>
<th>$\beta_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Point</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Circle</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Sphere</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Torus</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>
Signature property

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

<table>
<thead>
<tr>
<th>Space</th>
<th>$\beta_0$</th>
<th>$\beta_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X$</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$Y$</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

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

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 $k$th Betti number $\beta_k$ is the rank of the $k$th 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.
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 $p$th 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 \textit{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:

\[
\partial_1 (\{b, c\} + \{a, c\} + \{a, b\}) \\
= \{c\} + \{b\} + \{c\} + \{a\} + \{b\} + \{a\} \\
= 0
\]
Boundary homomorphism

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

$$\partial_p \sigma = \sum_{i} \{v_0, \ldots, \hat{v}_i, \ldots, v_k\}$$  \hspace{1cm} (1)

In the equation above, $\hat{v}_i$ indicates that the set does not contain the $i$th vertex. The function $\partial_p : C_p \to C_{p-1}$ 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

\[ 0 \xrightarrow{\partial_{n+1}} C_n \xrightarrow{\partial_n} C_{n-1} \xrightarrow{\partial_{n-1}} \ldots \xrightarrow{\partial_2} C_1 \xrightarrow{\partial_1} C_0 \xrightarrow{\partial_0} 0 \quad (2) \]
Cycle and boundary groups

Cycle group $Z_p = \ker \partial_p$ \hspace{1cm} (3)
Boundary group $B_p = \text{im} \partial_{p+1}$ \hspace{1cm} (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 $p$th homology group $H_p$ is a quotient group, defined by ‘removing’ cycles that are boundaries from a higher dimension:

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

(5)

With this definition, we may finally calculate the $p$th Betti number:

$$
\beta_p = \text{rank } H_p
$$

(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 \( M \subseteq \mathbb{R}^D \), with \( d \ll D \), from which our data have been sampled.
Agenda

1. Convert our input data into a simplicial complex $K$.
2. Calculate simplicial homology of $K$.
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 $\mathcal{R}_\varepsilon$
How to get a simplicial complex from $\mathcal{R}_\varepsilon$?

Construct the Vietoris–Rips complex $\mathcal{V}_\varepsilon$ 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

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

\[ \beta_1 \]

\[ \epsilon \]

0 0.4 0.6 0.8 1
Solution: Persistent homology

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

$$R_\epsilon \subseteq R_{\epsilon'} \quad (7)$$
$$V_\epsilon \subseteq V_{\epsilon'} \quad (8)$$

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

\[ \emptyset = K_0 \subseteq K_1 \subseteq \cdots \subseteq K_{n-1} \subseteq K_n = K \]  

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.

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 $\mathbb{L}_c^{-}(f) = \{x \mid f(x) \leq c\}$ of the function.
Persistent homology & persistence diagrams

One-dimensional example
Persistent homology & persistence diagrams

One-dimensional example
Persistent homology & persistence diagrams

One-dimensional example

[Diagram showing a 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

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

Distance calculations

\[ W^2(X, Y) = \inf_{\eta: X \rightarrow Y} \sum_{x \in X} \| x - \eta(x) \|_\infty^2 \]
Uses for persistence diagrams

Distance calculations

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

\[
W_2(X, Y) \leq C \| f - g \|_\infty^{1 - \frac{k}{2}},
\]

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

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

Total persistence:
\[ \text{pers}(\mathcal{D})_p := \sum_{(c,d) \in \mathcal{D}} (c - d)^p \] (13)

In essence, these are topological summary statistics.
Total persistence

\( p = 2 \)
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!