Understanding the shape of data: a brief introduction to Topological Data Analysis

Frédéric Chazal
DataShape team
INRIA Saclay - Ile-de-France
frederic.chazal@inria.fr
What is topological structure of data?

Modern data carry complex, but important, geometric/topological structure!
What is topological structure of data?

A non obvious problem:
→ no direct access to topological/geometric information: need of intermediate constructions (simplicial complexes);
→ distinguish topological “signal” from noise;
→ topological information may be multiscale;
→ statistical analysis of topological information.

Topological Data Analysis (TDA)
Persistent homology!
What is topological structure of data?

A non obvious problem:

→ no direct access to topological/geometric information: need of intermediate constructions (simplicial complexes);
→ distinguish topological “signal” from noise;
→ topological information may be multiscale;
→ statistical analysis of topological information.

Topological Data Analysis (TDA)
Persistent homology!
What is Topological Data Analysis (TDA)?

Topological Data Analysis (TDA) is a recent field whose aim is to:

- infer relevant topological and geometric features from complex data,
- take advantage of topological/geometric information for further Data Analysis, Machine Learning and AI tasks.
For what kind of data is TDA useful?

- Complex data!
For what kind of data is TDA useful?

- Complex data!
- Examples (where TDA brings real added value):
  - Force fields in granular media
  - Nanomaterial design
  - (Chaotic) time-dependent data - see later in the talk
The classical TDA pipeline

1. Build a multiscale topol. structure on top of data: filtrations.
2. Compute multiscale topol. signatures: persistent homology
3. Take advantage of the signature for further Machine Learning and AI tasks: Statistical aspects and representations of persistence
Persistent homology
The theory of persistence

A recent theory that is subject to intense research activities:

- **from the mathematical perspective:**
  - general algebraic framework (persistence modules) and general stability results.
  - extensions and generalizations of persistence (zig-zag persistence, multi-persistence, etc...)
  - Statistical analysis of persistence.

- **from the algorithmic and computational perspective:**
  - efficient algorithms to compute persistence and some of its variants.
  - efficient software libraries (in particular, Gudhi: [https://project.inria.fr/gudhi/](https://project.inria.fr/gudhi/)).

- **from the data science perspective:**
  - representations of persistence that are suitable for Machine Learning
  - Topological/geometric information in combination with other features

A whole machinery at the crossing of mathematics and computer science!
Tracking and encoding the evolution of the connected components (0-dimensional homology) of the sublevel sets of a function.
Tracking and encoding the evolution of the connected components (0-dimensional homology) and cycles (1-dimensional homology) of the sublevel sets.

Homology: an algebraic way to rigorously formalize the notion of $k$-dimensional cycles through a vector space (or a group), the homology group whose dimension is the number of "independent" cycles (the Betti number).
Stability properties

What if $f$ is slightly perturbed?
What if $f$ is slightly perturbed?

**Theorem (Stability):**
For any *tame* functions $f, g : X \to \mathbb{R}$, $d_B(D_f, D_g) \leq \|f - g\|_\infty$.

[Cohen-Steiner, Edelsbrunner, Harer 05], [C., Cohen-Steiner, Glisse, Guibas, Oudot - SoCG 09], [C., de Silva, Glisse, Oudot 12]
The bottleneck distance between two diagrams $D_1$ and $D_2$ is

$$d_B(D_1, D_2) = \inf_{\gamma \in \Gamma} \sup_{p \in D_1} \|p - \gamma(p)\|_\infty$$

where $\Gamma$ is the set of all the bijections between $D_1$ and $D_2$ and $\|p - q\|_\infty = \max(|x_p - x_q|, |y_p - y_q|)$. 

The **bottleneck distance** between two diagrams $D_1$ and $D_2$ is
Some applications (illustrations)

- Persistence-based clustering  [C., Guibas, Oudot, Skraba - J. ACM 2013]

- Analysis of force fields in granular media  [Kramar, Mischaikow et al.]
Some applications (illustrations)

- Hand gesture recognition [Li, Ovsjanikov, C. - CVPR'14]

- Persistence-based pooling for shape recognition [Bonis, Ovsjanikov, Oudot, C. 2016]
Filtrations allow to construct “shapes” representing the data in a multiscale way.

Persistent homology: encode the evolution of the topology across the scales → multi-scale topological signatures.

Persistent homology for point cloud data
• Filtrations allow to construct “shapes” representing the data in a multiscale way.

• **Persistent homology:** encode the evolution of the topology across the scales → multi-scale topological signatures.
• Filtrations allow to construct “shapes” representing the data in a multiscale way.

• **Persistent homology:** encode the evolution of the topology across the scales → multi-scale topological signatures.

**Persistent homology for point cloud data**
• Filtrations allow to construct “shapes” representing the data in a multiscale way.

• **Persistent homology**: encode the evolution of the topology across the scales → multi-scale topological signatures.
Filtrations allow to construct “shapes” representing the data in a multiscale way.

Persistent homology: encode the evolution of the topology across the scales → multi-scale topological signatures.

- **Perspective homology**: encode the evolution of the topology across the scales → multi-scale topological signatures.
Given a set $P = \{p_0, \ldots, p_k\} \subset \mathbb{R}^d$ of $k + 1$ affinely independent points, the $k$-dimensional simplex $\sigma$, or $k$-simplex for short, spanned by $P$ is the set of convex combinations

$$
\sum_{i=0}^{k} \lambda_i p_i, \quad \text{with} \quad \sum_{i=0}^{k} \lambda_i = 1 \quad \text{and} \quad \lambda_i \geq 0.
$$

The points $p_0, \ldots, p_k$ are called the vertices of $\sigma$. 

0-simplex: vertex
1-simplex: edge
2-simplex: triangle
3-simplex: tetrahedron

etc...
A (finite) simplicial complex $K$ in $\mathbb{R}^d$ is a (finite) collection of simplices such that:

1. any face of a simplex of $K$ is a simplex of $K$,
2. the intersection of any two simplices of $K$ is either empty or a common face of both.

The underlying space of $K$, denoted by $|K| \subset \mathbb{R}^d$ is the union of the simplices of $K$. 

Simplicial complexes
Abstract simplicial complexes

Let \( P = \{p_1, \ldots, p_n\} \) be a (finite) set. An abstract simplicial complex \( K \) with vertex set \( P \) is a set of subsets of \( P \) satisfying the two conditions:

1. The elements of \( P \) belong to \( K \).
2. If \( \tau \in K \) and \( \sigma \subseteq \tau \), then \( \sigma \in K \).

The elements of \( K \) are the simplices.

Let \( \{e_1, \ldots, e_n\} \) a basis of \( \mathbb{R}^n \). “The” geometric realization of \( K \) is the (geometric) subcomplex \( |K| \) of the simplex spanned by \( e_1, \ldots, e_n \) such that:

\[
[e_{i_0} \ldots e_{i_k}] \in |K| \text{ iff } \{p_{i_0}, \ldots, p_{i_k}\} \in K
\]

\( |K| \) is a topological space (subspace of an Euclidean space)!
Let $P = \{p_1, \ldots, p_n\}$ be a (finite) set. An abstract simplicial complex $K$ with vertex set $P$ is a set of subsets of $P$ satisfying the two conditions:

1. The elements of $P$ belong to $K$.
2. If $\tau \in K$ and $\sigma \subseteq \tau$, then $\sigma \in K$.

The elements of $K$ are the simplices.

**IMPORTANT**

Simplicial complexes can be seen at the same time as geometric/topological spaces (good for top./geom. inference) and as combinatorial objects (abstract simplicial complexes, good for computations).
A filtered simplicial complex (or a filtration) $\mathcal{S}$ built on top of a set $X$ is a family $(S_a \mid a \in \mathbb{R})$ of subcomplexes of some fixed simplicial complex $\mathcal{S}$ with vertex set $X$ s. t. $S_a \subseteq S_b$ for any $a \leq b$.

More generally, filtration $=$ nested family of spaces.
Filtrations of simplicial complexes

- A filtered simplicial complex (or a filtration) $\mathcal{S}$ built on top of a set $\mathbb{X}$ is a family $(\mathcal{S}_a | a \in \mathbb{R})$ of subcomplexes of some fixed simplicial complex $\overline{\mathcal{S}}$ with vertex set $X$ s. t. $\mathcal{S}_a \subseteq \mathcal{S}_b$ for any $a \leq b$.

- More generally, filtration = nested family of spaces.

**Example:** Let $(\mathbb{X}, d_{\mathbb{X}})$ be a metric space.

- The Vietoris-Rips filtration is the filtered simplicial complex defined by: for $a \in \mathbb{R}$,

  \[ [x_0, x_1, \cdots, x_k] \in \text{Rips}(\mathbb{X}, a) \iff d_{\mathbb{X}}(x_i, x_j) \leq a, \text{ for all } i, j. \]
Filtrations of simplicial complexes

• A filtered simplicial complex (or a filtration) $S$ built on top of a set $X$ is a family $(S_a \mid a \in \mathbb{R})$ of subcomplexes of some fixed simplicial complex $\bar{S}$ with vertex set $X$ s. t. $S_a \subseteq S_b$ for any $a \leq b$.

• More generally, filtration $=$ nested family of spaces.

Many other examples and ways to design filtrations depending on the application and targeted objectives : sublevel and upperlevel sets, Čech complex,...
“Stability theorem”: Close spaces/data sets have close persistence diagrams!

If \( X \) and \( Y \) are pre-compact metric spaces, then

\[
\text{d}_b(\text{dgm}(\text{Rips}(X)), \text{dgm}(\text{Rips}(Y))) \leq d_{GH}(X, Y).
\]

Bottleneck distance

Gromov-Hausdorff distance

\[
d_{GH}(X, Y) := \inf_{Z, \gamma_1, \gamma_2} d_H(\gamma_1(X), \gamma_2(X))
\]

\( Z \) metric space, \( \gamma_1 : X \to Z \) and \( \gamma_2 : Y \to Z \)

isometric embeddings.

Rem: This result also holds for other families of filtrations (particular case of a more general thm).
“Stability theorem”: Close spaces/data sets have close persistence diagrams!

If \( X \) and \( Y \) are pre-compact metric spaces, then

\[
d_b\left(\text{dgm}(\text{Rips}(X)), \text{dgm}(\text{Rips}(Y))\right) \leq d_{GH}(X, Y).
\]

\( d_{GH}(X, Y) := \inf_{Z, \gamma_1, \gamma_2} d_H(\gamma_1(X), \gamma_2(X)) \)

\( Z \) metric space, \( \gamma_1 : X \to Z \) and \( \gamma_2 : Y \to Z \) isometric embeddings.

Rem: This result also holds for other families of filtrations (particular case of a more general thm).

From a statistical perspective, when \( X \) is a random point cloud, such result links the study of statistical properties of persistence diagrams to support estimation problems.
Let $A, B \subset M$ be two compact subsets of a metric space $(M, d)$

$$d_H(A, B) = \max \{ \sup_{b \in B} d(b, A), \sup_{a \in A} d(a, B) \}$$

where $d(b, A) = \sup_{a \in A} d(b, a)$. 
Non rigid shapes in a same class are almost isometric, but computing Gromov-Hausdorff distance between shapes is extremely expensive.

Compare diagrams of sampled shapes instead of shapes themselves.
Persistent homology with the GUDHI library

GUDHI:

- a C++/Python open source software library for TDA,
- a developers team, an editorial board, open to external contributions,
- provides state-of-the-art TDA data structures and algorithms: design of filtrations, computation of pre-defined filtrations, persistence diagrams,...
- part of GUDHI is interfaced to R through the TDA package.

http://gudhi.gforge.inria.fr/
TDA and Machine Learning: some illustrative examples on real applications
TDA and Machine Learning for sensor data

(Multivariate) time-dependent data can be converted into point clouds: sliding window, time-delay embedding,...
TDA and Machine Learning for sensor data

TDA pipeline

GUDHI software

Topol. signatures

Feature engineering

Representations of persistence (linearization):

ML/AI

Features extraction
Random forests
Deep learning
Etc...
combined with other features!

Persistences diagram

Persistent silhouette
[Chazal & al, 2013]

Persistent surface
[Adams & al, 2016]
With landscapes: patient monitoring

A joint industrial research project between

A French SME with innovating technology to reconstruct pedestrian trajectories from inertial sensors (ActiMyo)

Objective: precise analysis of movements and activities of pedestrians.

Applications: personal healthcare; medical studies; defense.
With landscapes: patient monitoring

**Example:** Dyskinesia crisis detection and activity recognition:

- Data collected in non controlled environments (home) are very chaotic.
- Data registration (uncertainty in sensors orientation/position).
- Reliable and robust information is mandatory.
- Events of interest are often rare and difficult to characterize.

<table>
<thead>
<tr>
<th>Class</th>
<th>Naive</th>
<th>Multi</th>
<th>FEAT</th>
<th>QUA</th>
<th>TDA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Walking</td>
<td>97.6</td>
<td>98.4</td>
<td>99.3</td>
<td>99.0</td>
<td>99.5</td>
</tr>
<tr>
<td>Upstairs</td>
<td>97.2</td>
<td>99.8</td>
<td>97.8</td>
<td>98.0</td>
<td>97.7</td>
</tr>
<tr>
<td>Downstairs</td>
<td>99.6</td>
<td>99.7</td>
<td>99.0</td>
<td>98.4</td>
<td>98.3</td>
</tr>
<tr>
<td>Sitting</td>
<td>87.1</td>
<td>93.1</td>
<td>89.7</td>
<td>91.8</td>
<td>96.5</td>
</tr>
<tr>
<td>Standing</td>
<td>87.0</td>
<td>97.7</td>
<td>97.2</td>
<td>97.2</td>
<td>98.1</td>
</tr>
<tr>
<td>Laying</td>
<td>92.4</td>
<td>100.</td>
<td>99.8</td>
<td>99.9</td>
<td>100.</td>
</tr>
<tr>
<td>Stand-Sit</td>
<td>90.8</td>
<td>95.6</td>
<td>89.1</td>
<td>91.3</td>
<td>93.4</td>
</tr>
<tr>
<td>Sit-Stand</td>
<td>100.</td>
<td>99.9</td>
<td>100.</td>
<td>100.</td>
<td>100.</td>
</tr>
<tr>
<td>Sit-Lie</td>
<td>87.1</td>
<td>81.1</td>
<td>84.2</td>
<td>90.0</td>
<td>95.1</td>
</tr>
<tr>
<td>Lie-Sit</td>
<td>81.4</td>
<td>81.8</td>
<td>85.9</td>
<td>91.8</td>
<td>87.9</td>
</tr>
<tr>
<td>Stand-Lie</td>
<td>74.2</td>
<td>87.6</td>
<td>86.5</td>
<td>87.4</td>
<td>81.5</td>
</tr>
<tr>
<td>Lie-stand</td>
<td>80.4</td>
<td>72.1</td>
<td>83.2</td>
<td>77.7</td>
<td>83.2</td>
</tr>
</tbody>
</table>

Multi-channels CNN $+$ TDA neural network

Results on publicly available data set (HAPT) - improve the state-of-the-art.

---

**sysnav**

**informatics** mathematics
TDA-DL pipeline for arrhythmia detection

Objective: Arrhythmia detection from ECG data.

- Improvement over state-of-the-art.
- Better generalization.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Accuracy[%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>UCLA (2018)</td>
<td>93.4</td>
</tr>
<tr>
<td>Li et al. (2016)</td>
<td>94.6</td>
</tr>
<tr>
<td>Inria-Fujitsu (2018)*</td>
<td>98.6</td>
</tr>
</tbody>
</table>
Thank you for your attention!