Les données ont-elles une forme?
Une petite introduction à l’Analyse Topologique des Données

Frédéric Chazal
DataShape team
INRIA Saclay - Ile-de-France
frederic.chazal@inria.fr
What is Topological Data Analysis (TDA)?

Modern data carry complex, but important, geometric/topological structure!
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:
  - using topological features in ML pipelines,
  - taking advantage of topological information to improve ML pipelines.
Challenges and goals

Problem(s):
- how to infer the topological structure of data?
- how to compare topological properties (invariants) of close shapes/data sets?

• Challenges and goals:
  → 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.
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

- 90′s: size theory (P. Frosini et al), framed Morse complex and stability (S.A. Barannikov).
- important mathematical and practical developments since the 2000′s.
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.
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.
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.
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.
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.

Persistence barcode

Persistence diagram
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 $K$. 
Abstract simplicial complexes

Let $P = \{p_1, \cdots p_n\}$ 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 $\{e_1, \cdots e_n\}$ a basis of $\mathbb{R}^n$. “The” geometric realization of $K$ is the (geometric) subcomplex $|K|$ of the simplex spanned by $e_1, \cdots e_n$ such that:

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

$|K|$ is a topological space (subspace of an Euclidean space)!
Let $P = \{p_1, \cdots, p_n\}$ 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 $\mathcal{S}$ built on top of a set $X$ is a family $(S_a \mid a \in \mathbb{R})$ of subcomplexes of some fixed simplicial complex $\overline{\mathcal{S}}$ with vertex set $X$ such that $S_a \subseteq S_b$ for any $a \leq b$. 
A filtered simplicial complex $S$ built on top of a set $X$ is a family $(S_a | a \in \mathbb{R})$ of subcomplexes of some fixed simplicial complex $\overline{S}$ with vertex set $X$ s. t. $S_a \subseteq S_b$ for any $a \leq b$.

**Examples:** Let $(X, d_X)$ be a metric space.

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

  $$[x_0, x_1, \cdots, x_k] \in \text{Rips}(X, a) \iff d_X(x_i, x_j) \leq a, \text{ for all } i, j.$$  

- Čech complex: $\check{\text{Cech}}(X, a)$ is the complex with vertex set $X$ s.t.

  $$[x_0, x_1, \cdots, x_k] \in \check{\text{Cech}}(X, a) \iff \cap_{i=0}^{k} B(x_i, a) \neq \emptyset$$
Comparing persistence diagrams

A matching between two diagrams $dgm_1$ and $dgm_2$ is a subset

$$m \subseteq (dgm_1 \cup \Delta) \times (dgm_2 \cup \Delta)$$

such that every points in $dgm_1 \setminus \Delta$ and $dgm_2 \setminus \Delta$ appears exactly once in $m$.

The Bottleneck distance between $dgm_1$ and $dgm_2$ is then defined by

$$d_B(dgm_1, dgm_2) = \inf_{\text{matching } m} \max_{(p,q) \in m} \|p - q\|_\infty.$$
Stability properties

“Stability theorem”: Close spaces/data sets have close persistence diagrams!
[C., de Silva, Oudot - Geom. Dedicata 2013]

If $X$ and $Y$ are pre-compact metric spaces, then

$$d_b(dgm(Rips(X)), dgm(Rips(Y))) \leq d_{GH}(X, Y).$$

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).
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 $d(b, A) = \sup_{a \in A} d(b, a)$. 
Application: non rigid shape classification

MDS using bottleneck distance.

- 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.
Statistical setting and “linear representations”

\( \mathbf{X} \) is now a random point cloud (in some metric space)

Filt is a deterministic filtration (e.g. Rips)

\( D[\text{Filt}(\mathbf{X})] \) becomes random
Statistical setting and “linear representations”

\(X\) is now a random point cloud (in some metric space)

\(\text{Filt}\) is a deterministic filtration (e.g. Rips)

\(D[\text{Filt}(X)]\) becomes random

What can be said about the distribution of diagrams \(D[\text{Filt}(X)]\)?
X is now a random point cloud (in some metric space)

Filt is a deterministic filtration (e.g. Rips)

\( D[\text{Filt}(X)] \) becomes random

What can be said about the distribution of diagrams \( D[\text{Filt}(X)] \)?

- Stability properties \( \Rightarrow \) asymptotic properties, confidence bands, Wasserstein stability, ...
- Other representation of persistence that are well-suited for Machine Learning (landscapes, Betti curves, pers. images, kernels, ...)

Statistical setting and “linear representations”
Statistical setting

$$(M, \rho)$$ metric space

$\mu$ a probability measure with compact support $X_\mu$.

Sample $m$ points according to $\mu$.

**Example:**

- $\text{Filt}(\hat{X}_m) = \text{Rips}_\alpha(\hat{X}_m)$

**Questions:**

- Statistical properties of $\text{dgm}(\text{Filt}(\hat{X}_m))$? $\text{dgm}(\text{Filt}(\hat{X}_m)) \to ?$ as $m \to +\infty$?
Statistical setting

\((\mathcal{M}, \rho)\) metric space

\(\mu\) a probability measure with compact support \(X_{\mu}\).

Sample \(m\) points according to \(\mu\).

Example:
- \(\text{Filt}(\hat{X}_m) = \text{Rips}_\alpha(\hat{X}_m)\)

Questions:

- Statistical properties of \(\text{dgm}(\text{Filt}(\hat{X}_m))\) → as \(m \to +\infty\)?
- Can we do more statistics with persistence diagrams? What can be said about distributions of diagrams?
Statistical setting

$(\mathbb{M}, \rho)$ metric space

$\mu$ a probability measure with compact support $X_\mu$.

Sample $m$ points according to $\mu$.

Example:
- $\text{Filt}(\hat{X}_m) = \text{Rips}_\alpha(\hat{X}_m)$

Stability thm: $d_b(d_{\text{dgm}}(\text{Filt}(X_\mu)), d_{\text{dgm}}(\text{Filt}(\hat{X}_m))) \leq 2d_{GH}(X_\mu, \hat{X}_m)$

So, for any $\varepsilon > 0$,

$$\mathbb{P}\left(d_b\left(d_{\text{dgm}}(\text{Filt}(X_\mu)), d_{\text{dgm}}(\text{Filt}(\hat{X}_m))\right) > \varepsilon\right) \leq \mathbb{P}\left(d_{GH}(X_\mu, \hat{X}_m) > \frac{\varepsilon}{2}\right)$$
For $a, b > 0$, $\mu$ satisfies the $(a, b)$-standard assumption if for any $x \in \mathbb{X}_\mu$ and any $r > 0$, we have $\mu(B(x, r)) \geq \min(ar^b, 1)$. 
For $a, b > 0$, $\mu$ satisfies the $(a, b)$-standard assumption if for any $x \in \mathbb{X}_\mu$ and any $r > 0$, we have $\mu(B(x,r)) \geq \min(ar^b,1)$.

**Theorem:** If $\mu$ satisfies the $(a, b)$-standard assumption, then for any $\varepsilon > 0$:

$$\mathbb{P} \left( d_b \left( \text{dgm}(\text{Filt}(\mathbb{X}_\mu)), \text{dgm}(\text{Filt}(\hat{\mathbb{X}}_m)) \right) > \varepsilon \right) \leq \min \left( \frac{8^b}{a\varepsilon^b} \exp(-ma\varepsilon^b), 1 \right).$$
Deviation inequality and rate of convergence

For $a, b > 0$, $\mu$ satisfies the $(a, b)$-standard assumption if for any $x \in X_\mu$ and any $r > 0$, we have $\mu(B(x, r)) \geq \min(ar^b, 1)$.

**Theorem:** If $\mu$ satisfies the $(a, b)$-standard assumption, then for any $\varepsilon > 0$:

$$
P\left(d_b\left(d_{gm}(\text{Filt}(X_\mu)), d_{gm}(\text{Filt}(\hat{X}_m))\right) > \varepsilon\right) \leq \min\left(\frac{8b}{a\varepsilon^b} \exp(-ma\varepsilon^b), 1\right).
$$

**Corollary:** Let $\mathcal{P}(a, b, M)$ be the set of $(a, b)$-standard proba measures on $M$. Then:

$$
\sup_{\mu \in \mathcal{P}(a, b, M)} \mathbb{E} \left[d_b(d_{gm}(\text{Filt}(X_\mu)), d_{gm}(\text{Filt}(\hat{X}_m)))\right] \leq C \left(\frac{\ln m}{m}\right)^{1/b}
$$

where the constant $C$ only depends on $a$ and $b$ (not on $M$!). Moreover, the upper bound is tight (in a minimax sense)!
Persistence landscapes

$D = \{ \left( \frac{d_i+b_i}{2}, \frac{d_i+b_i}{2} \right) \}_{i \in I}$

For $p = \left( \frac{b+d}{2}, \frac{d-b}{2} \right) \in D$, 

$$\Lambda_p(t) = \begin{cases} 
  t - b & t \in [b, \frac{b+d}{2}] \\
  d - t & t \in (\frac{b+d}{2}, d] \\
  0 & \text{otherwise}.
\end{cases}$$

Persistence landscape [Bubenik 2012]: 

$$\lambda_D(k, t) = \max_{p \in \text{dgm}} \Lambda_p(t), \quad t \in \mathbb{R}, k \in \mathbb{N},$$

where $k_{\text{max}}$ is the $k$th largest value in the set.

Many other ways to “linearize” persistence diagrams: intensity functions, image persistence, kernels,...
Persistence landscapes [Bubenik 2012]:

\[ \lambda_D(k, t) = \max_{p \in dgm} A_p(t), \quad t \in \mathbb{R}, k \in \mathbb{N}, \]

Properties

- For any \( t \in \mathbb{R} \) and any \( k \in \mathbb{N} \), \( 0 \leq \lambda_D(k, t) \leq \lambda_D(k + 1, t) \).
- For any \( t \in \mathbb{R} \) and any \( k \in \mathbb{N} \), \( |\lambda_D(k, t) - \lambda_{D'}(k, t)| \leq d_B(D, D') \) where \( d_B(D, D') \) denotes the bottleneck distance between \( D \) and \( D' \).
Persistence landscapes

• Persistence encoded as an element of a functional space (vector space!).
• Expectation of distribution of landscapes is well-defined and can be approximated from average of sampled landscapes.
• process point of view: convergence results and convergence rates → confidence intervals can be computed using bootstrap.
Stability of expected landscapes

\[ \Lambda_{\mu,m}(t) = \mathbb{E}_{P_{\mu}}[\lambda(t)] \]

\[ |\lambda_n(t) - \Lambda_{\mu,m}(t)| \]

Stability w.r.t. \( \mu \)?
Stability of expected landscapes

Let $(\mathbb{M}, \rho, \mu)$ be a metric space and let $\mu, \nu$ be probability measures on $\mathbb{M}$ with compact supports. We have

$$\|\Lambda_{\mu,m} - \Lambda_{\nu,m}\|_\infty \leq m^{\frac{1}{p}} W_p(\mu, \nu)$$

where $W_p$ denotes the Wasserstein distance with cost function $\rho(x, y)^p$. 

Theorem: Let $(\mathbb{M}, \rho)$ be a metric space and let $\mu, \nu$ be probability measures on $\mathbb{M}$ with compact supports. We have

$$\|\Lambda_{\mu,m} - \Lambda_{\nu,m}\|_\infty \leq m^{\frac{1}{p}} W_p(\mu, \nu)$$

where $W_p$ denotes the Wasserstein distance with cost function $\rho(x, y)^p$. 

Repeat $n$ times: $\lambda_1(t), \cdots, \lambda_n(t) \rightarrow \bar{\Lambda}_n(t)$

Bootstrap

$|\bar{\Lambda}_n(t) - \Lambda_{\mu,m}(t)| = \mathbb{E}_{P_\mu}[\lambda(t)]$
TDA and Machine Learning: some illustrative examples on real applications
(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):

Persistent silhouette
[Chazal & al, 2013]

Persistent surface
[Adams & al, 2016]

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

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

<table>
<thead>
<tr>
<th>Class</th>
<th>Naive</th>
<th>Multi</th>
<th>FEA</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>99.8</td>
<td>99.9</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-Stand</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
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!

Software:
- The Gudhi library (C++/Python):
  https://project.inria.fr/gudhi/software/
- R package TDA