A Statistical Approach to Topological Data Analysis

Bertrand MICHEL

Laboratoire de Statistique Théorique et Appliquée
Université Pierre et Marie Curie

Stochastic Geometry Conference
Nantes, April 6 2016
I - Introduction: Statistics and Topological Data Analysis
Topological data analysis and topological inference

- Geometric inference, algebraic topology tools and computational topology have recently witnessed important developments with regards to data analysis, giving birth to the field of topological data analysis (TDA).

- The aim of TDA is to infer relevant, qualitative and quantitative topological structures (clusters, holes ...) directly from the data.

- The two popular methods in TDA: **Mapper algorithm** [Singh et al., 2007] and **persistent homology** [Edelsbrunner et al., 2002].

- Topological inference methods aim to infer topological properties of an unknown topological space $X$, typically from a point cloud $X_n$ “close” to $X$. 
Application fields of TDA methods

[distribution of galaxies]

[Sensor Data]

Magnetometer Data (walking)

Magnetometer Data (cross trainer)

[3D shape database]
Topological data analysis methods can be used:

- For **exploratory analysis**, visualization:

- For **feature extraction** in supervised settings (prediction):

[Chazal et al., 2014b]

[Chazal et al., 2015a]
Statistics and TDA

Until very recently, TDA and topological inference mostly relied on deterministic approaches. Alternatively, a statistical approach to TDA means that:

- we consider data as generated from an unknown distribution
- the inferred topological features by TDA methods are seen as estimators of topological quantities describing an underlying object.

Non exhaustive list of questions for a statistical approach to TDA:

- proving consistency of TDA methods.
- providing confidence regions for topological features and discussing the significance of the estimated topological quantities.
- selecting relevant scales at which the topological phenomenon should be considered.
- dealing with outliers and providing robust methods for TDA.
- ...
II- Homology and Persistent homology
Point clouds in themselves do not carry any non trivial topological or geometric structure.

For a point cloud $\mathbb{X}_n$ in $\mathbb{R}^d$ (or in a metric space), the $\alpha$-offset of $\mathbb{X}_n$ is defined by

$$\mathbb{X}_n^\alpha = \bigcup_{x \in \mathbb{X}_n} B(x, \alpha).$$

More generally, for any compact set $\mathbb{X}$,

$$\mathbb{X}^\alpha := \bigcup_{x \in \mathbb{X}} B(x, \alpha) = d_{\mathbb{X}}^{-1}([0, \alpha])$$

where the distance function $d_{\mathbb{X}}$ to $\mathbb{X}$ is

$$d_{\mathbb{X}}(y) = \inf_{x \in \mathbb{X}} \|x - y\| \quad \text{(in } \mathbb{R}^d)$$

General idea: deduce from $(\mathbb{X}_n^\alpha)_{r>0}$ some topological and geometric information of an underlying object.
Non-discrete sets such as offsets, and also continuous mathematical shapes like curves, surfaces cannot easily be encoded as finite discrete structures.

A geometric simplicial complex \( C \) is a set of simplices such that:

- Any face of a simplex from \( C \) is also in \( C \).
- The intersection of any two simplices \( s_1, s_2 \in C \) is either a face of both \( s_1 \) and \( s_2 \), or empty.
Basic tools for TDA: Offsets and Simplicial Complexes

Examples:

- A simplex \([x_0, x_1, \cdots, x_k]\) is in the Čech complex \(\check{\text{Cech}}_{\alpha}(X_n)\) if and only if \(\bigcap_{j=0}^{k} B(x_j, \alpha) \neq \emptyset\).

- A simplex \([x_0, x_1, \cdots, x_k]\) is in the Rips complex \(\text{Rips}_{\alpha}(X_n)\) if and only if \(\|x_j - x_{j'}\| \leq \alpha\) for all \(j, j' \in \{1, \ldots, k\}\).

Can be also defined for a set of points in any metric space or for any compact metric space.

Nerve Theorem: the offsets \(X_n^\alpha\) of a point cloud \(X_n\) in \(\mathbb{R}^d\) are homotopy equivalent to the Čech complex \(\check{\text{Cech}}_{\alpha}(X_n)\)
Given a point cloud $X_n$ in $\mathbb{R}^d$, we generally define a filtration of (nested simplicial) complexes by considering all the possible scale parameters $\alpha : (C_\alpha)_{\alpha \in \mathcal{A}}$. 

\[ C_{\alpha_1} \quad C_{\alpha_2} \quad \cdots \quad \cdots \quad \cdots \]
Homology inference

- **Singular homology** provides a algebraic description of “holes” in a geometric shape (connected components, loops, etc ...)

- **Betti number** $\beta_k$ is the rank of the $k$-th homology group.

- **Computational Topology** : Betti numbers can be computed on simplicial complexes.

---

**Homology inference** [Niyogi et al., 2008 and 2011] [Balakrishnan et al., 2012] : The Betti number (actually the homotopy type) of Riemannian manifolds with positive reach can be recovered with high probability from offsets of a sample on (or close to) the manifold.
Persistent homology

Starting from a point cloud $X_n$, let $\text{Filt} = (C_\alpha)_{\alpha \in \mathcal{A}}$ be a filtration of nested simplicial complexes.

Persistent homology: identification of “persistent” topological features along the filtration.

- multiscale information ;
- more stable and more robust ;
- (but does not answer the scale selection problem...)
Barecodes and Persistence Diagrams

Filtration of simplicial complexes $\text{Filt}(X_n)$

Offsets

Barecode
Barecodes and Persistence Diagrams

Filtration of simplicial complexes $\text{Filt}(X_n)$

Offsets

Barecode

Dgm($\text{Filt}(X_n)$)

Persistence diagram of the filtration $\text{Filt}(X_n)$ built on $X_n$. 
The bottleneck distance between two diagrams $Dgm_1$ and $Dgm_2$ is

$$d_b(Dgm_1, Dgm_2) = \inf_{\gamma \in \Gamma} \sup_{p \in Dgm_1} \|p - \gamma(p)\|_\infty$$

where $\Gamma$ is the set of all the bijections between $Dgm_1$ and $Dgm_2$ and

$$\|p - q\|_\infty = \max(|x_p - x_q|, |y_p - y_q|).$$
Distance between persistence diagrams and stability

Theorem [Chazal et al., 2012]: For any compact metric spaces \((X, \rho)\) and \((Y, \rho')\),

\[
d_b \left( \text{Dgm}(\text{Filt}(X)), \text{Dgm}(\text{Filt}(Y)) \right) \leq 2 \ d_{GH} \ (X, Y).
\]

Consequently, if \(X\) and \(Y\) are embedded in the same metric space \((\mathbb{M}, \rho)\) then

\[
d_b \left( \text{Dgm}(\text{Filt}(X)), \text{Dgm}(\text{Filt}(Y)) \right) \leq 2 \ d_H \ (X, Y).
\]
III - Statistics and Persistent homology
Persistence diagram inference [Chazal et al., 2014b]
Joint work with F. Chazal, M. Glisse and C. Labruère.

$(\mathbb{M}, \rho)$ metric space
$X$ compact set in $\mathbb{M}$.

$n$ points sampled in $X$ according to $\mu$

$\hat{X}_n$

well defined for any compact metric space [Chazal et al., 2012]

$\hat{X}_n$

Estimator of $\text{Dgm}(\text{Filt}(K))$

$\text{Dgm}(\text{Filt}(X))$

Convergence ???

$\text{Dgm}(\text{Filt}(\hat{X}_n))$
For $a, b > 0$, $\mu$ satisfies the $(a, b)$-standard assumption on its support $X_\mu$ if for any $x \in X_\mu$ and any $r > 0$:

$$\mu(B(x, r)) \geq \min(ar^b, 1).$$

$\mathcal{P}(a, b, M)$ : set of all the probability measures satisfying the $(a, b)$-standard assumption on the metric space $(M, \rho)$.

**Theorem:** For $a, b > 0$:

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

where $C$ only depends on $a$ and $b$.

Under additional technical hypotheses, for any estimator $\hat{\text{Dgm}}_n$ of $\text{Dgm}(\text{Filt}(X_\mu))$:

$$\liminf_{n \to \infty} \sup_{\mu \in \mathcal{P}(a, b, M)} \mathbb{E}
\left[ d_b(\text{Dgm}(\text{Filt}(X_\mu)), \hat{\text{Dgm}}_n) \right] \geq C' n^{-1/b}$$

where $C'$ is an absolute constant.
Confidence sets for persistence diagrams [Fasy et al., 2014]

\[ P \left( \text{Dgm}(\text{Filt}(K)) \in \hat{R} \right) \geq 1 - \alpha \]
Confidence sets for persistence diagrams [Fasy et al., 2014]

Using the Hausdorff stability, we can define confidence sets for persistence diagrams:

\[ d_b \left( \text{Dgm}(\text{Filt}(K)), \text{Dgm}(\text{Filt}(X_n)) \right) \leq d_H(K, X_n). \]

It is sufficient to find \( c_n \) such that

\[ \limsup_{n \to \infty} \left( d_H(K, X_n) > c_n \right) \leq \alpha. \]
IV - Robust distance functions for TDA and geometric inference
Standard TDA methods are not robust to outliers

\[ X^r := \bigcup_{x \in X} B(x, r) = d_X^{-1}([0, r]) \]

where the distance function \( d_X \) to \( X \) is

\[ d_X(y) = \inf_{x \in X} \| x - y \| \]
Standard TDA methods are not robust to outliers

\[ X^r := \bigcup_{x \in X} B(x, r) \]
\[ = d_X^{-1}([0, r]) \]

where the distance function \( d_X \) to \( X \) is

\[ d_X(y) = \inf_{x \in X} \| x - y \| \]
We would like to consider the sub levels of an alternative distance function related to the sampling measure, which support is $\mathbb{X}$, or close to $\mathbb{X}$. 

Robust TDA with an alternative distance function?
Preliminary distance function to a measure $P$:
Let $u \in ]0, 1[$ be a positive mass, and $P$ a probability measure on $\mathbb{R}^d$:

$$\delta_{P,u}(x) = \inf \{ r > 0 : P(B(x,r)) \geq u \}$$

$\delta_{P,u}$ is the smallest distance needed to capture a mass of at least $u$.

$\delta_{P,u}$ is the quantile function at $u$ of the r.v. $\|x - X\|$ where $X \sim P$. 
Preliminary distance function to a measure \( P \):

Let \( u \in ]0, 1[ \) be a positive mass, and \( P \) a probability measure on \( \mathbb{R}^d \):

\[
\delta_{P,u}(x) = \inf \{ r > 0 : P(B(x,r)) \geq u \}
\]

**Definition:** Given a probability measure \( P \) on \( \mathbb{R}^d \) and \( m > 0 \), the distance function to the measure \( P \) (DTM) is defined by

\[
d_{P,m} : x \in \mathbb{R}^d \mapsto \left( \frac{1}{m} \int_0^m \delta_{P,u}(x) \, du \right)^{1/2}
\]
Properties of the DTM:

- Stability under Wasserstein perturbations:

\[ \|d_{P,m} - d_{Q,m}\|_{\infty} \leq \frac{1}{\sqrt{m}} W_2(P,Q) \]

- The function \( x \mapsto d_{P,m}^2(x) \) is semiconcave, this is ensuring strong regularity properties on the geometry of its sublevel sets.

- Consequently, if \( \tilde{P} \) is a probability distribution close to \( P \) for Wasserstein distance \( W_2 \), then the sublevel sets of \( d_{\tilde{P},m} \) provide a topologically correct approximation of the support of \( P \).
Distance to The Empirical Measure (DTEM)

Let $X_1, \ldots, X_n$ sample according to $P$ and let $P_n$ be the empirical measure. Then

$$d^2_{P_n, \frac{k}{n}}(x) = \frac{n}{k} \sum_{i=1}^{k} \|x - X(i)\|^2$$

where $\|X(1) - x\| \geq \|X(2) - x\| \geq \cdots \geq \|X(k) - x\| \cdots \geq \|X(n) - x\|$
Estimation of the DTM via the empirical DTM

[Chazal et al., 2014b] and [Chazal et al., 2015b]

Quantity of interest:

\[ d_{P_n,k/n}^2(x) - d_{P,k/n}^2(x) \]

- Observe that

\[ d_{P,m}^2(x) = \frac{1}{m} \int_0^m F_x^{-1}(u) du \]

where \( F_x \) is the cdf of \( \|x - X\|^2 \) with \( X \sim P \).

- The distance to the empirical measure is the empirical counterpart of the distance to \( P \):

\[ d_{P_n,m}^2(x) = \frac{1}{m} \int_0^m F_{x,n}^{-1}(u) du \]

where \( F_{x,n} \) is the cdf of \( \|x - X\|^2 \) with \( X \sim P_n \).

- Finally we get that

\[ d_{P_n,k/n}^2(x) - d_{P,k/n}^2(x) = \frac{1}{m} \int_0^m \{ F_{x,n}^{-1}(u) - F_x^{-1}(u) \} du \]
Estimation of the DTM via the empirical DTM

[Chazal et al., 2014b] and [Chazal et al., 2015b]

Quantity of interest:

\[ d_{P, \frac{k}{n}}^2(x) - d_{P, \frac{k}{n}}^2(x) \]

Two complementary approaches of the problem:

- Asymptotic approach: \( \frac{k}{n} = m \) is fixed and \( n \) tends to infinity.

- Non asymptotic approach: \( n \) is fixed, and we want a tight control over the fluctuations of the empirical DTM, in function of \( k \), which can be taken very small.

We do not use Wasserstein stability for either of the two approaches. Wasserstein rates of convergence [Fournier and Guillin, 2013; Dereich et al., 2013] do not provide tight rates for the DTM in this context.
Theorem: Let $P$ be a measure on $\mathbb{R}^d$ with compact support. Let $D$ be a compact domain on $\mathbb{R}^d$ and $m \in (0,1)$. Assume that there exists a uniform upper bound $\omega_D$ on the modulus of continuity for the family $(F_x^{-1})_{x \in D}$ satisfying

$$\lim_{u \to 0} \omega_D(u) = \omega_D(0) = 0.$$ 

Then $\sqrt{n}(d_{P_n,m}^2 - d_{P,m}^2)$ converges in distribution to $B$ on $D$, where $B$ is a centered Gaussian process with covariance kernel

$$\kappa(x, y) = \frac{1}{m^2} \int_0^{F_x^{-1}(m)} \int_0^{F_y^{-1}(m)} \left( \mathbb{P} \left[ B(x, \sqrt{t}) \cap B(y, \sqrt{s}) \right] - F_x(t)F_y(s) \right) ds \, dt.$$
Theorem: Let $x$ be a fixed observation point in $\mathbb{R}^d$. Assume that $\omega_x : (0, 1] \rightarrow \mathbb{R}^+$ is an upper bound on the modulus of continuity of $F^{-1}_x$. Let $k < \frac{n}{2}$. For any $\lambda > 0$:

$$P \left( \left| d_{P_n, \frac{k}{n}}^2(x) - d_{P, \frac{k}{n}}^2(x) \right| \geq \lambda \right) \leq 2 \exp \left( -\frac{n}{8} \frac{k}{n} \omega_x \left( \frac{k}{n} \right)^2 \lambda^2 \right) + ...$$

Assume moreover that $\omega_x(u)/u$ is a non increasing function, then

$$\mathbb{E} \left( \left| d_{P_n, \frac{k}{n}}^2(x) - d_{P, \frac{k}{n}}^2(x) \right| \right) \leq \frac{C}{\sqrt{n}} \sqrt{\frac{k}{n}} \omega_x \left( \frac{k}{n} \right).$$

renormalization by the mass proportion  
localization at the origin  
parametric rate of convergence  
statistical complexity of the problem
Fluctuations of the DTEM [Chazal et al., 2015b]

The quantile function $F_x^{-1}$ carries some geometric information. For instance $\omega(0^+) = 0$ means that the support of $dF_x$ is a closed interval.
**Aim**: studying the persistent homology of the sub-levels of the DTM and providing confidence regions.

Two alternative bootstrap methods:

- by bootstrapping the DTM
- Bottleneck Bootstrap
Bootstrap and significance of topological features
[Chazal et al., 2014b]

Bootstrapping the DTM
For $m \in (0, 1)$, define $c_{\alpha}$ by

$$\mathbb{P} \left( \sqrt{n} \left\| d_{P,m}^2 - d_{P_n,m}^2 \right\|_{\infty} > c_{\alpha} \right) = \alpha.$$ 

Let $X_1^*, \ldots, X_n^*$ be a sample from $P_n$, and let $P_n^*$ be the corresponding (bootstrap) empirical measure.

We consider the bootstrap quantity $d_{P_n^*,m}(x)$ of $d_{P_n,m}$.

The bootstrap estimate $\hat{c}_{\alpha}$ is defined by

$$\mathbb{P} \left( \sqrt{n} \left\| d_{P_n,m}^2 - d_{P_n^*,m}^2 \right\|_{\infty} > \hat{c}_{\alpha} \mid X_1, \ldots, X_n \right) = \alpha$$

where $\hat{c}_{\alpha}$ can be approximated by Monte Carlo.

**Theorem:** If $F_{x}^{-1}$ is regular enough, the distance to measure function is Hadamard differentiable at $P$. Consequently, the bootstrap method for the DTM is asymptotically valid.
Bootstrap and significance of topological features
[Chazal et al., 2014b]

**Bootstrapping the DTM**

$Dgm$: persistence diagram of the sub-levels of $d_{P,m}$

$\widehat{Dgm}$: persistence diagram of the sub-levels of $d_{P_n,m}$.

Let

$$C_n = \left\{ E \in \mathcal{D}iag : d_b(\widehat{Dgm}, E) \leq \frac{\hat{c}_\alpha}{\sqrt{n}} \right\},$$

where $\mathcal{D}iag$ is the set of all the persistence diagrams.

Then,

$$\mathbb{P}(Dgm \in C_n) = \mathbb{P}\left(d_b(Dgm, \widehat{Dgm}) \leq \frac{\hat{c}_\alpha}{\sqrt{n}} \right) \geq \mathbb{P}\left(\|d_{P,m}^2 - d_{P_n,m}^2\|_\infty \leq \frac{\hat{c}_\alpha}{\sqrt{n}} \right)$$

**Bootstrap estimate**
Bootstrap and significance of topological features
[Chazal et al., 2014b]

The Bottleneck Bootstrap

\( \hat{Dgm} : \) persistence diagram of the sub-levels of \( d_{P,m} \)
\( \hat{Dgm} : \) persistence diagram of the sub-levels of \( d_{P_n,m} \).
\( \hat{Dgm}^* : \) persistence diagram of the sub-levels of \( d_{P_n^*,m} \).

We directly bootstrap in the set of the persistence diagram by considering the random quantity \( d_b(\hat{Dgm}^*, \hat{Dgm}) \). We define \( \hat{t}_\alpha \) by

\[
\mathbb{P} \left( \sqrt{n} d_b(\hat{Dgm}^*, \hat{Dgm}) > \hat{t}_\alpha \mid X_1, \ldots, X_n \right) = \alpha.
\]

The quantile \( \hat{t}_\alpha \) can be estimated by Monte Carlo.
Bootstrap and significance of topological features
[Chazal et al., 2014b]

For both methods we can identify significant features by putting a band of size $2\hat{c}_\alpha$ or $2\hat{t}_\alpha$ around the diagonal:

In practice, the bottleneck bootstrap can lead to more precise inferences because in many cases the following stability result is not sharp

$$d_b(\hat{D}gm, Dgm) \leq \|d_{P,m}^2 - d_{P_n,m}^2\|_\infty.$$
Concluding remarks

• TDA methods focus on the topological properties (homology / persistent homology) of a shape.

• TDA methods can be used
  – as an “exploratory method”, in particular when the point cloud is sampled on (close to) a real geometric object
  – as a “feature extraction” procedure, next these extracted features can be used for learning purposes.

• TDA is an emerging field, at the interface maths, computer sciences, statistics.

• Many topics about the statistical analysis of TDA

• Applications in many fields of sciences (medecine, biology, dynamic systems, astronomy, dynamical systems, physics ...)

• TDA methods need to bring together Geometric Inference, Computational Topology and Geometry, Statistics and Learning methods.
Thank you !
References


References


References


Topological invariants

How topological spaces can be compared from a topological point of view?

For comparing topological spaces, we consider topological invariants (preserved by homeomorphism): numbers, groups, polynomials.
Topological invariants

How topological spaces can be compared from a topological point of view?

For comparing topological spaces, we consider topological invariants (preserved by homeomorphism) : numbers, groups, polynomials.

Homotopy is weaker than homeomorphism but is preserves many topological invariants.

- Two continuous functions $f : X \to Y$ and $g : X \to Y$ are **homotopic** if there exists a continuous application $H : X \times [0, 1] \to Y$ such that $H(\cdot, 0) = f$ and $H(\cdot, 1) = g$.
- Two topological spaces $X$ and $Y$ are **homotopic** if there exists two continuous applications $f : X \to Y$ and $g : Y \to X$ such that
  - $g \circ f$ is homotopic to $\text{id}_X$;
  - $f \circ g$ is homotopic to $\text{id}_Y$;

Topological Stability and Regularity

Topological inference: under “regularity assumptions”, topological properties of $X$ can be recovered from (the off-sets) of a close enough object $Y$. 
Topological Stability and Regularity

Topological inference: under “regularity assumptions”, topological properties of $X$ can be recovered from (the off-sets) of a close enough object $Y$.

- The *local feature size* is a local notion of regularity:
  For $x \in X$, $\text{lfs}_X(x) := d(x, \mathcal{M}(X^c))$.

- The global version of the local feature size is the *reach* [Federer, 1959]:

  $$\kappa(X) = \inf_{x \in X^c} \text{lfs}_X(x).$$

  The reach is small if either $X$ is not smooth or if $X$ is close to being self-intersecting.

- Weak feature size and its extensions [Chazal and Lieutier, 2007] (by considering the critical values of $d_X$).
Topological Stability and Regularity

Topological inference: under “regularity assumptions”, topological properties of $X$ can be recovered from (the offsets) of a close enough object $Y$.

$$d_H(X, Y) = \inf \{ \alpha \geq 0 \mid X \subset Y^\alpha \text{ and } Y \subset X^\alpha \}$$

Example:

**Theorem** [Chazal and Lieutier, 2007]: Let $X$ and $Y$ be two compact sets in $\mathbb{R}^d$ and let $\varepsilon > 0$ be such that $d_H(X, Y) < \varepsilon$, $\text{wfs}(X) > 2\varepsilon$ and $\text{wfs}(Y) > 2\varepsilon$. Then for any $0 < \alpha < 2\varepsilon$, $X^\alpha$ and $Y^\beta$ are homotopy equivalent.
Topological Stability and Regularity

Topological inference: under “regularity assumptions”, topological properties of $X$ can be recovered from (the off-sets) of a close enough object $Y$.

$$d_H(X, Y) = \inf \{ \alpha \geq 0 \mid X \subset Y^\alpha \text{ and } Y \subset X^\alpha \}$$

Example:

**Theorem** [Chazal and Lieutier, 2007]: Let $X$ and $Y$ be two compact sets in $\mathbb{R}^d$ and let $\varepsilon > 0$ be such that $d_H(X, Y) < \varepsilon$, $\text{wfs}(X) > 2\varepsilon$ and $\text{wfs}(Y) > 2\varepsilon$. Then for any $0 < \alpha < 2\varepsilon$, $X^\alpha$ and $Y^\beta$ are homotopy equivalent.

Sampling conditions in Hausdorff metric.

Statistical analysis of homotopy inference can be deduced from support estimation of a distribution under the Hausdorff metric.