A statistical framework for analyzing shape in a time series of random geometric objects[1]

Anne van Delft[1]

(Joint work with Andrew J. Blumberg[2])

[1]Department of Statistics, Columbia University, NY, USA

[2]Department of Mathematics/Irving Institute for Cancer Dynamics, Columbia University, NY, USA

February 13, 2024

[1]Supported by NSF grant DMS-2311338

Outline

Overview: what is this talk about?

Modern data sets: complex mathematical structures

- Measurements from processes that vary over a continuum
  - sequentially collected;
  - sampled almost continuously on domain;
  - exhibit nonstationary behavior.

Overview: what is this talk about?

Modern data sets: complex mathematical structures

- Measurements from processes that vary over a continuum
  - sequentially collected;
  - sampled almost continuously on domain;
  - exhibit nonstationary behavior.

- In various applications: paramount interest in shape
  - Dimensionality reduction followed by clustering is ubiquitous

  $\hookrightarrow$ Clustering captures very coarse shape information.

  $\hookrightarrow$ Standard approaches to dimensionality reduction often assume linearity; e.g., PCA, compressed sensing, NMF, or a contractible smooth manifold (manifold learning).

Overview: what is this talk about?

Modern data sets: complex mathematical structures

- Measurements from processes that vary over a continuum
  - sequentially collected;
  - sampled almost continuously on domain;
  - exhibit nonstationary behavior.
- In various applications: paramount interest in shape
  - Dimensionality reduction followed by clustering is ubiquitous

  $\hookrightarrow$ Clustering captures very coarse shape information.

  $\hookrightarrow$ Standard approaches to dimensionality reduction often assume linearity; e.g., PCA, compressed sensing, NMF, or a contractible smooth manifold (manifold learning).

Idea: use more refined mathematical shape descriptors.

What is this talk about? (continued)

We are particularly interested in studying the evolution of shape over time:
$\hookrightarrow$ time series of geometric objects.

1. **foundational questions:** how to do statistical inference?
2. **algorithmic problems:** can we implement inference methods efficiently?
3. and **applications:** we focus on genomics

Application: cell differentation in development

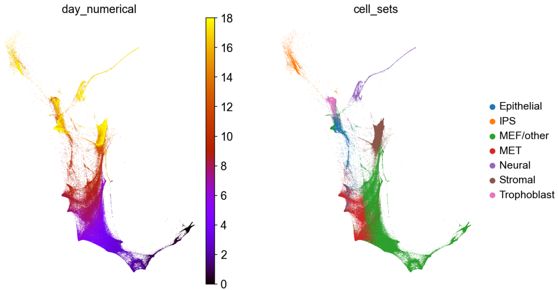


Figure: Developmental trajectories (color indicates time).

Goal: detect and capture changes in shape

Mathematical context

We are interested in data sampled from objects with complicated geometry, low-dimensional geometry.

- Nonlinear manifolds: for example, the circle.
- Things close to manifolds: spaces with corners (singularities), unions of manifolds of differing dimension.
- Non-manifold spaces, with a notion of local metric geometry (relevant in genomics).

Representative of data of interest: sequence of finite metric spaces
(e.g., time series of point clouds)

Outline

Topological Data Analysis

Topological data analysis applies invariants of algebraic topology to discrete data.

- Algebraic topology assigns algebraic objects (e.g., numbers, vector spaces) to geometric objects.

- These invariants are global and qualitative; e.g., the *kth* homology groups $H_k(X)$ of a space X count the number of k-dimensional holes.

$\hookrightarrow$ homology detects the connected components, tunnels, voids, etc., of a topological space.

$\hookrightarrow$ Insensitive to deformation.

$\hookrightarrow$ Generalization of clustering: H_0 counts the number of components.

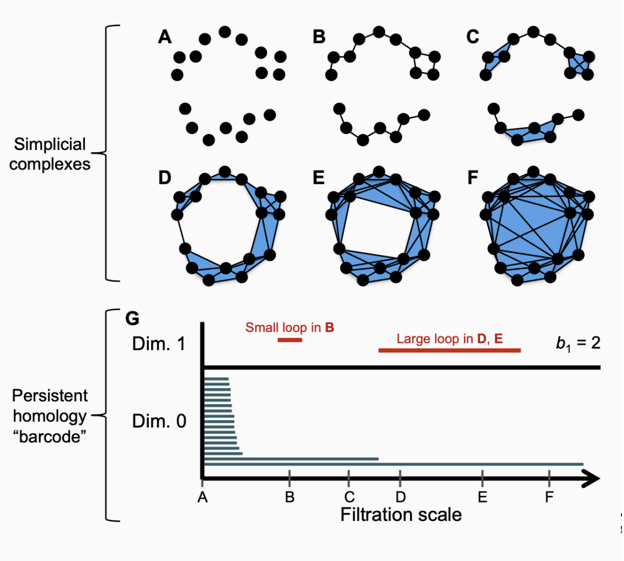
Persistent homology

A core topological invariant in TDA is persistent homology: Describes multi-scale topological features of a point cloud (i.e., a finite metric space)

1. Involves construction of a sequence of simplicial complexes from (X, ∂_X)
2. Associates to these simplices topological invariants such as homology
3. Assigns a birth and death value to each topological feature

This creates a filtered vector space for each k which can be represented as a multiset of intervals (a, b) referred to as a barcode or persistence diagram $PH_k(X)$.

Persistent homology captures information at different scales



Evolution of shape over time

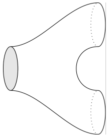


Figure: A surface evolving in time; time increases along the x-axis from left to right.

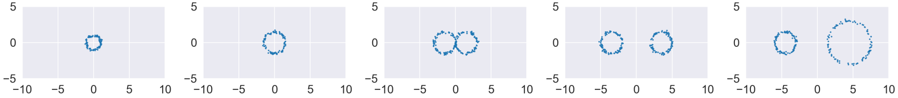


Figure: Samples from slices at fixed times from the evolving surface as time increases.

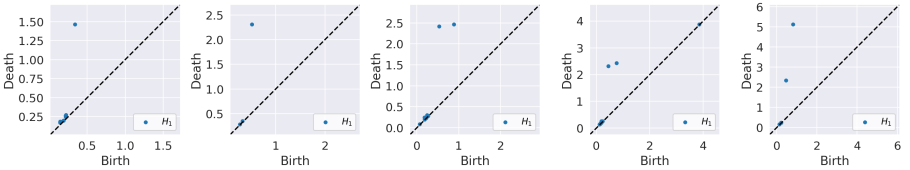


Figure: Persistence diagrams of the samples from the slices. The points away from the line $x = y$ represent the circles.

Main theorems of persistent homology

There are two foundational theoretical results that speak for it:

1. There are comparatively efficient algorithms to compute persistent homology
2. Persistent homology is stable[1]: for compact metric spaces

$$d_{\mathscr{B}}(PH_k(X), PH_k(Y)) \leqslant d_{GH}((X, \partial_X), (Y, \partial_Y))$$

where $d_{GH}(X, Y)$ denotes the Gromov-Hausdorff distance

$$d_{GH}(X, Y) = \frac{1}{2} \inf \left\{ \mathrm{dist}(\mathcal{R}) \mid \mathcal{R} \text{ correspondence between } X \text{ and } Y \right\}$$

where $\mathrm{dist}(\mathcal{R}) = \sup_{(x,y),(x',y') \in \mathcal{R}} |\partial_X(x, x') - \partial_Y(y, y')|$.

[1] original version due to D. Cohen-Steiner, H. Edelsbrunner, H. and J. Harer, 2007

The ‘barcode’ process

Given an ensemble of point clouds $(X_t^n)_{t=1}^T$ we can track and analyze the features via the process $(PH_k(X_t^n))_{t=1}^T$.

- $(PH_k(X_t^n))_{t=1}^T$ takes values in the set of barcodes $\mathscr{B}$
- $\mathscr{B}$ forms a metric space under various metrics, specifically the bottleneck distance $d_{\mathscr{B}}$
- $(\overline{\mathscr{B}}, d_{\mathscr{B}})$ is a complete separable metric space.

Inference?

Not straightforward:

- Only access to point clouds of latent process.
- The space of barcodes is Polish.
- Statistical inference on topological invariants not well-developed.

Outline

The process of interest (latent)

- let (M, ∂_M) and $(M', \partial_{M'})$ be compact metric spaces
- We consider stochastic processes $(\mathbb{X}_t : t \in \mathbb{Z})$ defined by

$$\mathbb{X}_t : \Omega \to C(M, M')$$

The process of interest (latent)

- let (M, ∂_M) and $(M', \partial_{M'})$ be compact metric spaces
- We consider stochastic processes $(\mathbb{X}_t : t \in \mathbb{Z})$ defined by

$$\mathbb{X}_t : \Omega \to C(M, M')$$

- Then the process $(\tilde{\mathbb{X}}_t(m) : t \in \mathbb{Z}, m \in M)$ defined by

$$\tilde{\mathbb{X}}_t(m) := e_m \circ \mathbb{X}_t$$

  with $e_m : C(M, M') \to M', \xi \mapsto \xi(m)$ takes values in M'.

Interpretation: we think of M as a parameter space and the images in M' as representing the geometric object of interest

Locally stationary metric space-valued SP

Need asymptotic theory under nonstationarity;

Let $(\mathbb{X}_{t,T} : t \in \mathbb{Z}, T \in \mathbb{N})$ be an (S, ∂_S)-valued stochastic process.

Definition 1

$(\mathbb{X}_{t,T} : t \in \mathbb{Z}, T \in \mathbb{N})$ is locally stationary if, for all $u = t/T \in [0,1]$, $\exists$ an (S, ∂_S)-valued stationary process $(\mathbb{X}_t(u) : t \in \mathbb{Z}, u \in [0,1])$ such that

$$\partial_S\big(\mathbb{X}_{t,T}, \mathbb{X}_t(\frac{t}{T})\big) = O_p(T^{-1}) \quad \text{and} \quad \partial_S(\mathbb{X}_t(u), \mathbb{X}_t(v)) = O_p(|u-v|)$$

uniformly in $t = 1, \ldots, T$ and $u, v \in [0,1]$.

The observed process

- An n-dimensional point cloud for the function $\mathbb{X}_t$ is given by

$$\tilde{\mathbb{X}}^n_{t,T} := e_{m_1,\ldots,m_n} \circ \mathbb{X}_{t,T}$$

- The data arises as an ensemble of point clouds $(\tilde{\mathbb{X}}^n_{t,T})_{t=1}^T$ where $n = n(T)$.
- For the data to be representative of the latent process, we must have conditions such that

$$\lim_{n\to\infty} \tilde{\mathbb{X}}^n_{t,T} \approx \mathbb{X}_{t,T}(M) = \tilde{\mathbb{X}}_{t,T}$$

in an appropriate sense.

Outline

Overview: What is this talk about?

Topological Data Analysis

A statistical framework

Topological invariants to characterize (S, ∂_S)-valued processes

Application of methodology: testing for topological change

A weak invariance principle

Extension of Gromov's characterization to ergodic MMPDS

- Gromov's characterization: a metric measure space (S, ∂_S, ν_S) is up to isometry determined by the infinite-dimensional distance matrix distribution
$$\{\partial_S(s_i, s_j)\}_{(i,j)\in\mathbb{N}\times\mathbb{N}}$$
where $\{s_i\} \in S$ is an iid sequence with common distribution ν_S.

Extension of Gromov's characterization to ergodic MMPDS

- Gromov's characterization: a metric measure space (S, ∂_S, ν_S) is up to isometry determined by the infinite-dimensional distance matrix distribution
$$\{\partial_S(s_i, s_j)\}_{(i,j)\in\mathbb{N}\times\mathbb{N}}$$
where $\{s_i\} \in S$ is an iid sequence with common distribution ν_S.

- We show a similar result holds for ergodic metric measure-preserving dynamical systems, i.e., tuples $(S, \partial_S, \mu_S, \theta_S)$ where θ_S is a measure-preserving function that is ergodic under the measure μ_S.

- Thus, the infinite-dimensional distance matrix distribution

$$\phi(X) = (\partial_S(X_t, X_s) : t, s \in \mathbb{Z})$$

is a complete invariant of a stationary ergodic Polish-valued stochastic process $X = (X_t \colon t \in \mathbb{Z})$.
- The ball volumes are fully determined by the infinite-dimensional matrix distribution $\phi(X)$

Conditions such that the ball volumes characterize $\phi(X)$, and thus μ_X?

$(\overline{\mathcal{B}}, d_{\mathscr{B}})$-valued proc. are determined by their values on balls[1]

Theorem 4.1

The space $\mathscr{B}_\alpha^N$ of N-point bounded barcodes has the property that any Borel measure is determined by its values on balls.

Corollary 2

The pushforward of any Borel measure on point clouds under the persistent homology functor PH_k is determined by its values on balls.

Corollary 3

The pushforward of any Borel measure on point clouds under the zigzag persistent homology functor is determined by its values on balls.

[1] Van Delft & Blumberg (2024) arXiv:2401.11125: "Measures determined by their values on balls and Gromov-Wasserstein convergence."

Consequence of the preceding theorem

Characterizing the geometry of a Polish-valued process X via the ball volume processes of the fidis
$\rightarrow$ convenient for inference as it reduces to analyzing U-processes.

To see this, note for example that

$$\mu_J^X\big(B(\pi_J \circ X, r)\big) = \mathbb{E}_{(X')\sim\mu_J}\big[\prod_{j\in J} 1_{\partial_S(X_j, X'_j)\leqslant r}\big].$$

where

$$B(s,r) = (s'_j : \max_{1\leqslant j\leqslant |J|} \partial_S(s, s'_j) \leqslant r)$$

denotes the ball volume on the $|J|$-dimensional product metric space $S^{|J|}$.

Outline

Detecting nonstationary behavior in the marginals

- Let $\nu_t := \mathbb{P} \circ (PH_k(\tilde{\mathbb{X}}_t))^{-1}$ denote the marginal distribution at time t.
- The process

$$\Big(\varphi_t(r) := \nu_t(B(PH_k(\tilde{\mathbb{X}}_t), r)) : r \geqslant 0\Big) \quad PH_k(\tilde{\mathbb{X}}_t) \sim \nu_t$$

characterizes the measure ν_t up to isometry.

Detecting nonstationary behavior in the marginals

- Let $\nu_t := \mathbb{P} \circ (PH_k(\tilde{\mathbb{X}}_t))^{-1}$ denote the marginal distribution at time t.
- The process

$$\left(\varphi_t(r) := \nu_t(B(PH_k(\tilde{\mathbb{X}}_t), r)) : r \geqslant 0\right) \quad PH_k(\tilde{\mathbb{X}}_t) \sim \nu_t$$

  characterizes the measure ν_t up to isometry.

Then we are interested in testing pair of hypotheses

$$H_0 : \mathbb{E}\varphi_t(r) = \mathbb{E}\varphi(r) \quad \forall t \in \mathbb{Z}, r \in [0, \mathscr{R}]$$

versus

$$H_A : \mathbb{E}\varphi_t(r) \neq \mathbb{E}\varphi(r) \text{ for some } t \in \mathbb{Z}, r \in [0, \mathscr{R}].$$

Outline

Test statistic

- Given we observe an ensemble of point clouds $\left(\tilde{\mathbb{X}}_t^{n(T)}\right)_{t=1}^T$
- Create the barcode sample $\left(PH_k(\tilde{\mathbb{X}}_t^{n(T)})\right)_{t=1}^T$.

Test statistic

- Given we observe an ensemble of point clouds $\big(\tilde{\mathbb{X}}_t^{n(T)}\big)_{t=1}^T$
- Create the barcode sample $\big(PH_k(\tilde{\mathbb{X}}_t^{n(T)})\big)_{t=1}^T$.
- Define the partial sum process

$$S_T(u,r) = \frac{1}{T^2}\sum_{s,t=1}^{\lfloor uT \rfloor} h(\tilde{\mathbb{X}}_t^{n(T)}, \tilde{\mathbb{X}}_s^{n(T)}, r) \quad r \in [0,\mathscr{R}], u \in [0,1].$$

where, for a compact metric space (R, ∂_R), the kernel $h : R \times R \times [0,\mathscr{R}] \to \mathbb{R}$ is given by

$$h(x,x',r) = 1\big\{d_{\mathcal{B}}\big(PH_k(x), PH_k(x')\big) \leqslant r\big\}, \quad x, x' \in R.$$

Test statistic

- Given we observe an ensemble of point clouds $\left(\tilde{\mathbb{X}}_t^{n(T)}\right)_{t=1}^T$
- Create the barcode sample $\left(PH_k(\tilde{\mathbb{X}}_t^{n(T)})\right)_{t=1}^T$.
- Define the partial sum process

$$S_T(u,r) = \frac{1}{T^2} \sum_{s,t=1}^{\lfloor uT \rfloor} h(\tilde{\mathbb{X}}_t^{n(T)}, \tilde{\mathbb{X}}_s^{n(T)}, r) \quad r \in [0, \mathscr{R}], u \in [0,1].$$

where, for a compact metric space (R, ∂_R), the kernel $h : R \times R \times [0, \mathscr{R}] \to \mathbb{R}$ is given by

$$h(x, x', r) = 1\{d_{\mathcal{B}}\big(PH_k(x), PH_k(x')\big) \leqslant r\}, \quad x, x' \in R.$$

Then let

$$U_T(u,r) = S_T(u,r) - u^2 S_T(1,r).$$

Test statistic

- Given we observe an ensemble of point clouds $\left(\tilde{\mathbb{X}}_t^{n(T)}\right)_{t=1}^T$
- Create the barcode sample $\left(PH_k(\tilde{\mathbb{X}}_t^{n(T)})\right)_{t=1}^T$.
- Define the partial sum process

$$S_T(u,r) = \frac{1}{T^2}\sum_{s,t=1}^{\lfloor uT\rfloor} h(\tilde{\mathbb{X}}_t^{n(T)}, \tilde{\mathbb{X}}_s^{n(T)}, r) \quad r \in [0,\mathscr{R}], u \in [0,1].$$

where, for a compact metric space (R, ∂_R), the kernel $h : R \times R \times [0,\mathscr{R}] \to \mathbb{R}$ is given by

$$h(x,x',r) = 1\left\{d_{\mathcal{B}}\left(PH_k(x), PH_k(x')\right) \leqslant r\right\}, \quad x, x' \in R.$$

Then let

$$U_T(u,r) = S_T(u,r) - u^2 S_T(1,r).$$

$\rightarrow$ Under H_0, $\mathbb{E}U_T(u,r) = 0$.

Test statistic (continued)

We consider suitably self-normalized versions of

$$\sup_{r\in[0,\mathscr{R}]}\sup_{u\in[0,1]}\sqrt{T}\left|U_T(u,r)\right|$$

and of

$$T\int_0^{\mathscr{R}}\int_0^1\big(U_T(u,r)\big)^2\,dudr.$$

weak invariance principle in $D([0,1]\times[0,\mathscr{R}])$

Theorem

Under the regularity assumptions

$$\left\{T^{1/2}\Big(S_T(u,r)-\mathbb{E}S_T(u,r)\Big)\right\}_{u\in[0,1],r\in[0,\mathscr{R}]} \overset{D}{\rightsquigarrow} \left\{\mathbb{G}(u,u,r)\right\}_{u\in[0,1],r\in[0,\mathscr{R}]}.$$

in $D([0,1]\times[0,\mathscr{R}])$ w.r.t. the Skorokhod topology as $T\to\infty$, where $\{\mathbb{G}(u,v,r)\}_{v\leqslant u\in[0,1],r\in[0,\mathscr{R}]}$ is a zero-mean Gaussian process with covariance structure

$$Cov(\mathbb{G}(u_1,v_1,r_1),\mathbb{G}(u_2,v_2,r_2))=\int_0^{\min(u_1,u_2)}\sigma(\eta,v_1,r_1)\sigma(\eta,v_2,r_2)d\eta.$$

Corollary 4

Under the previous conditions

$$\left\{T^{1/2}\Big(U_T(u,r)-\mathbb{E}(U_T(u,r))\Big)\right\}_{u\in[0,1],r\in[0,\mathscr{R}]}$$

$$\underset{T\to\infty}{\rightsquigarrow}\quad \left\{\int_0^u \sigma(\eta,u,r)d\mathbb{B}(\eta)-u^2\int_0^1 \sigma(\eta,1,r)d\mathbb{B}(\eta)\right\}_{u\in[0,1],r\in[0,\mathscr{R}]}.$$

in $D[0,1]$ w.r.t. Skorokhod topology.

Under H_0, this reduces to

$$\left\{T^{1/2}U_T(u,r)\right\}_{u\in[0,1],r\in[0,\mathscr{R}]}\quad\underset{T\to\infty}{\rightsquigarrow}\quad\left\{u\sigma(r)\big(\mathbb{B}(u)-u\mathbb{B}(1)\big)\right\}_{u\in[0,1],r\in[0,\mathscr{R}]}.$$

Self-normalized (focus on max-type)

Define the range

$$V_T(r) = \max_{1\leqslant k\leqslant T}\big(U(k/T,r) - \mathbb{E}[U_T(k/T,r]\big) - \min_{1\leqslant k\leqslant T}\big(U(k/T,r) - \mathbb{E}[U_T(k/T,r)]\big)$$

And consider the empirical distance

$$\mathbb{D}_T^{\max} := \max_r \max_k \frac{|U(k/T,r)|}{V_T(r)}$$

Then, under H_0,

$$\mathbb{D}_T^{\max} \underset{T\to\infty}{\Longrightarrow} \sup_u \frac{\cancel{\sigma(r)}|u\mathbb{B}(u) - u^2\mathbb{B}(1)|}{\cancel{\sigma(r)}\sup_u\big(u\mathbb{B}(u) - u^2\mathbb{B}(1)\big) - \inf_u\big(u\mathbb{B}(u) - u^2\mathbb{B}(1)\big)} =: \mathbb{D}^{\max}$$

- RHS are pivotal and quantiles can be easily simulated
- Used to construct asymptotic level-α tests for the hypotheses of interest.

Summary

- Comprehensive theoretical framework based on FTS developed to infer on the evolving geometric features
- Naturally incorporates:
  - Nonstationary temporal and spatial dependence
  - Irregular and noise corrupted sampling
  - Analysis of convergence rate and non-asymptotic error bounds
- Simulation results: see arXiv.
- Applied to: developmental trajectories in single cell RNA-seq.

Outlook

Progress:

- We are exploiting theorem 4.1 without assuming a doubling measure.
- Estimation of break locations (to appear soon)
- Extensions mathematical shape descriptors

References

A. van Delft, A. & A. J. Blumberg.
A statistical framework for analyzing shape in a time series of random geometric objects
Available: arXiv:2304.01984

A. van Delft & A. J. Blumberg.
Measures determined by their values on balls and Gromov-Wasserstein convergence.
Available: arXiv:2401.11125

D. Cohen-Steiner, H. Edelsbrunner, H. and J. Harer.
Stability of Persistence Diagrams.
Discrete Comput Geom, 37:103–120, 2007.