2017-01-04

Michael Lesnick, Universality of the Homotopy Interleaving Distance

Homotopy interleaving is a homotopy-invariant analogues of interleavings. Whereas ordinary interleavings can be interpreted as pairs of "approximate isomorphisms" between filtered spaces, homotopy interleavings can be viewed as pairs of "approximate weak equivalences". Homotopy interleaving distance d_{HI} is an universal pseudometric satisfying natural stability and homotopy invariance axioms. Furthermore, it can strengthen Sheehy's approximation results for sparse Rips filtrations to topological level, formulate a conjectural persistent analogue of the Whitehead Theorem, and a space-level formulation of Sheehy's homological persistent nerve theorem.

Peter Bubenik, Discovering Geometry using Topological Data Analysis

He gave an introduction on how to analyze data using landscapes. He ran analysis on two dataset: constant metric space and Alzheimers Disease Neuroimaging Initiative (ADNI).

When space with constant curvature -1(hyperbolic), 0(euclidean), 1(spherical) are considered, triangles from hyperbolic are thinner than from euclidean, and triangles from spherical are fatter than from euclidean. Hence, when $death/birth$ are compared, $hyperbolic < euclidean < spherical$ holds. Landscapes combined with SVM on 100 samples showed good classification rate.

Alzheimers Disease Neuroimaging Initiative (ADNI) data is a surface data in 3d. He filtered each hippocampus in 144 directions, calculated persistence landscape, and applied SVM. Classification rate was 73%.

He also mentioned our TDA package.

Matthew L Wright, Multidimensional Persistence: A Practical Approach

Paper: Interactive Visualization of 2-D Persistence Modules, [arXiv]

Video: Matthew Wright, Visualizing 2-Dimensional Persistent Homology [video]

Two-dimensional (2-D) persistence allow us to work with data indexed by two parameters, such as distance and density. 2-D persistence module M is a collection of k-vector modules $\{M_u\}_{u \in \mathbb{R}^2}$.

We can visualize 2-D persistence in three ways:

1. the dimension of each homology vector space $M_{i,j}$.

2. The rank invariant:

- For $u \leq v$, $rank(u, v)$ is the dimension of homology at u that also exists at v.

- Let L be the line through u and v.

- The restriction of M to L is a 1-D persistence module M^L, and we can visualize this.

3. The bigraded Betti numbers.

RIVET is Rank Invariant Visualization and Exploration Tool. We can get RIVET at http://rivet.online

Thomas Wanner, Topological Microstructure Analysis Using Persistence Landscapes

Model for Phase Separation

Quenching of homogeneous alloys may lead to phase separation generating complicated microstructures. Averaged persistence landscapes can be used to recover central system information in the Cahn-Hilliard theory of phase separation.

Genki Kusano, Kernel method for persistence diagrams

Paper: Genki Kusano, Kenji Fukumizu, Yasuaki Hiraoka, Persistence weighted Gaussian kernel for topological data analysis [arXiv]

Let Ω be a set, and let $k : \Omega \times \Omega \to \mathbb{R}$ a nice function (positive definite kernel). Gram matrix $(k(x_i, x_j))$ plays an important role for statistical analysis on Ω. $k : \Omega \times \Omega \to \mathbb{R}$ is called a positive definite kernel when $k(x, y) = k(y, x)$ holds and $(k(x_i, x_j))$ is positive semidefinite. Then Moore-Aroszajn theorem says that a positive definite kernel k uniquely defines the Hilbert space $\mathcal{H}_k$, where $k(\cdot, x) : \Omega \to \mathbb{R}$ is an element in $\mathcal{H}_k$.

Let $M_b(\Omega)$ be a finite signed Radon measure on Ω. Then we can define a mapping from $M_b(\Omega)$ to $\mathcal{H}_k$ by $\mu \mapsto E_k(\mu) := \int k(\cdot, x) d\mu(x)$.

Now we use kernel on persistence diagram. By appropriate weight function $w : \mathbb{R}^2 \to \mathbb{R}$, a persistence diagram D is represented as a weighted measure $\mu_D^w = \sum_{x \in D} w(x)\delta_x$. Then we can sequentially embed as $D \mapsto \mu_D^w \mapsto E_k(\mu_D^w)$. Practically, we propose to use Gaussian kernel $k_G(x, y) = \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right)$ and weight function $w_{ave}(x) = \arctan(C\text{pers}(x)^p)$.

Stability theorem [Fukumizu, Hiraoka, 2016] implies that $\|E_k(\mu^w_{D_q(X)}) - E_k(\mu^w_{D_q(Y)})\|_{\mathcal{H}} \leq L(M) d_H(D_q(X), D_q(Y))$.

Their application was about which index is the change point between liquid and glass?

Matthew Kahle, Maximally persistent cycles in random geometric complexes

Paper: Omer Bobrowski, Matthew Kahle, Primoz Skraba, Maximally Persistent Cycles in Random Geometric Complexes [arXiv]

Book: Mathew Penrose, Random Geometric Graphs

The random geometric graph $G(n, r)$ has its vertices n points chosen i.i.d. in $\mathbb{R}^d$. Then two vertices are adjacent if they are within distance r. Then we have following proposition:

Proposition. *Let G_1 be the number of vertices in the largest connected component of $G(n,r)$*

There exists a constant c^* with following property:

If $r \le \frac{c}{\sqrt{n}}$ and $c < c^*$, then w.h.p $G_1 = O(\log n)$.

If $r \ge \frac{c}{\sqrt{n}}$ and $c > c^*$, then w.h.p $G_1 = \Omega(n)$.

Random geometric graph extends to random geometric complexes, where $C(n,r)$ is the Cech complex. Then we have following proposition:

Proposition. *Consider n points chosen i.i.d. uniformly in the unit square $[0,1]^2$.*

If $r \ll n^{-3/4}$ then w.h.p $H_1(C(n,r)) = 0$.

If $r \gg \sqrt{\log n/n}$ then w.h.p. $H_1(C(n,r)) = 0$.

If $n^{-3/4} \ll r \ll \sqrt{\log n/n}$ then w.h.p. $H_1(C(n,r)) \neq 0$.

For persistent homology, they considered multiplicative persistence rather than additive persistence.

Definition. $p(\sigma) = d(\sigma)/b(\sigma)$ rather than $p(\sigma) = d(\sigma) - b(\sigma)$.

Advantages of the multiplicative definition are as follows:

1. In the random setting, many cycles σ satisfy $d(\sigma) - b(\sigma) \approx d(\sigma)$.
2. This makes dimensionless.
3. The relationship between Cech and Rips is a multiplicative one.

Then we have following theorem:

Theorem. *(Two-dimensional case.) Consider n points chosen i.i.d. uniformly in the unit square $[0,1]^2$. Then w.h.p. the maximal persistence in degree one homology is of order $\max_{\sigma} p(\sigma) \asymp \frac{\log n}{\log\log n}$.*

where main tool is from isoperimetric inequality.

Theorem. *(Federer-Fleming, 1960) If σ is a k-cycle in $\mathbb{R}^d$ of k-dimensional volume V, then the filling radius R satisfies $R = O(V^{1/k})$.*

There are several questions.

1. $\lim \frac{\max_\sigma p(\sigma)}{\log n/\log\log n} =?$
2. other distributions? e.g. multivariate normal

2017-01-05
Seth Sullivant, Introduction to Algebraic Statistics
See his webpage for book draft [link]

Megan Owen, Means and a Central Limit Theorem in tree space.

The space of metric phylogenetic trees introduced by Billera, Holmes, and Vogtmann "Geometry of the Space of Phylogenetic Trees" (2001) [pdf] is a polyhedral cone complex. It is also non-positively curved or CAT(0), so there is a unique shortest path (geodesic) between any two trees and and a well-defined notion of a mean tree for a given set of trees. Also, the calculation is fast.

Mean tree is sticky: mean tends to be pulled towards lower-dimensional data.

Can prove a Central Limit Theorem if mean is in interior of a top-dimensional orthant:

Let $\hat{x}$: Frechet mean, $\hat{x}_k$: Frechet mean. Then $\sqrt{k}(\hat{x}_k - \hat{x}) \xrightarrow{L} N(0, A^\top V A)$. Also have a CLT for when mean in interior of codimension 1 boundary.

There is a work for confidence intervals: Amy Willis, Confidence sets for phylogenetic trees [arXiv].

Jeff Sommars, A Computer Algebra System for R: Macaulay2 and the m2r package

Macaulay2 is the language for algebraic geometry. And m2r is a socket between Macaulay2 and R.

2016-01-06

Donald Richards, Distance Correlation: A New Tool for Detecting Association and Measuring Correlation Between Data Sets

It is unwise to apply linear regression to percentage data. We can instead consider distance correlation: joint characteristic function can be defined as $\psi_{X,Y}(s,t) = \mathbb{E}\exp\left[i(sX + tY)\right]$. Then the distance covariance is defined as $\mathcal{V}(X,Y) = \frac{1}{\gamma_p \gamma_q}\int_{\mathbb{R}^{p+q}} \left|\psi_{X,Y}(s,t) - \psi_X(s)\psi_Y(t)\right|^2$. And then the distance correlation is defined as $\mathcal{R}(X,Y) = \frac{\mathcal{V}(X,Y)}{\sqrt{\mathcal{V}(X,X)}\sqrt{\mathcal{V}(Y,Y)}}$. Its empirical version is $\mathcal{R}_n(X,Y) = \frac{\mathcal{V}_n(X,Y)}{\sqrt{\mathcal{V}_n(X,X)}\sqrt{\mathcal{V}_n(Y,Y)}}$. Distance correlation has higher statistical power.

Peter Bubenik, An Introduction to Topological Data Analysis

He gave an brief introduction on the concept of persistent homology and landscapes, and how they can be used to analyze biological data.

Erica Flapan, Topological Complexity in Protein Structures

In proteins, knots prevent proteins from degrading. If Taylor's theory is correct, only knots appearing in protein has unknotting number 1.

Statistical proof and the problem of irreproducibility

Mumford, "Intelligent Design Found in the Sky with p $<$ 0.001". But we should consider the correction for multiplicity in the number of possible variables selected as significant using multiple hypotheses correction and FDR control.

2016-01-07

Miguel del Alamo, Variational Multiscale Estimators for Nonparametric Regression and Statistical Inverse Problems

Our Model: nonparametric regression

$y_n(x) = f(x) + \xi_n(x)$, $x \in \Gamma_n$ (equidistant grid)

Penelized estimation : $\min \mathcal{L}(g, y_n) + \lambda S(g)$

Smoothness-constrained estimator : $\min \mathcal{L}(g, y_n)$ s.t. $S(g) \leq \eta$

Data-friendly-constrained estimation : $\min S(g)$ s.t. $\mathcal{L}(g, y_n) \leq \gamma$

Multiresolution norm

$\|y\|_{\mathcal{B}} = \sup\limits_{B \in \mathcal{B}} \frac{1}{\sqrt{\#\Gamma_n \cap \mathcal{B}}} \left| \sum\limits_{s \in \Gamma_n \cap \mathcal{B}} y(x) \right|$ for $y \in \mathbb{R}^{\Gamma_n}$

Multiscale Nemirovski-Dantzig (MIND) estimator

$\min\limits_{g} S(g)$

s.t. $\|S_n g - y_n\|_{\mathcal{B}} \leq \gamma_n \sim \sqrt{\log n}$

Choose γ_n s.t. $\|S_n f - y_n\|_{\mathcal{B}} \leq \gamma_n$ with high probability, then f is admissible for the minimization problem.

TV regularization

$\hat{f} \in \arg\min |g|_{BV}$ s.t. $\|S_n g - y_n\|_{\mathcal{B}} \leq \gamma_n$

nearly minimax optimal