Statistical Inference For Geometric Data

Jisu KIM

March 31, 2017

Carnegie Mellon University

Thesis Committee:
Larry Wasserman (Co-Chair)
Alessandro Rinaldo (Co-Chair)
Sivaraman Balakrishnan
Frédéric Chazal (INRIA Saclay)
Jessi Cisewski (Yale University)

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy.
Abstract

Geometric structures can aid statistics in several ways. In high dimensional statistics, geometric structures can be used to reduce dimensionality. High dimensional data entails the curse of dimensionality, which can be avoided by if there are low dimensional geometric structures. On the other hand, geometric structures also provide useful information. Structures may carry scientific meaning about the data and can be used as features to enhance supervised or unsupervised learning. Lastly, geometry is also used in data visualization.

In this proposal, I will explore how statistical inference can be done on geometric structures, with regard to five areas. First, I will prove minimax rates for dimension estimation and reach estimation. Second, I will explore the minimax rates for persistent homology. Third, I will investigate inference on cluster trees, voids, persistent homology and multidimensional persistent homology. Fourth, I will extend and improve the R package TDA for computing topological data analysis. Fifth, I will study the embedding space for network data.

Contents

1 Introduction  2
   1.1 Differential Geometry ............................................. 3
   1.2 Reach .............................................................. 4
   1.3 Persistent Homology ............................................... 4

2 Ongoing work  5
   2.1 Minimax Rates for Estimating the Dimension of a Manifold ............ 5
      2.1.1 Upper Bound and Lower Bound .......................... 6
   2.2 The Origin of the Reach: Better Understanding Regularity Through Minimax Estimation Theory ................................................ 6
      2.2.1 Geometry of the Reach ...................................... 7
      2.2.2 Our Reach Estimator and Minimax Estimates .......... 7
   2.3 Statistical Inference for Cluster Trees ............................. 8
      2.3.1 Tree Metrics ................................................... 8
      2.3.2 Confidence Sets ............................................. 9
   2.4 Topological classification and sub-classification of voids using persistent homology .............................................. 9
      2.4.1 Methodology .................................................. 10
   2.5 R Package TDA: Statistical Tools for Topological Data Analysis .......... 10

3 Proposed work  11
   3.1 Statistical inference on persistent homology of density filtration on rips complex ................. 11
   3.2 Statistical inference on multi-dimensional persistent homology ............ 11
   3.3 Minimax rate of persistent homology .................................. 12
   3.4 Choosing between hyperbolic and euclidean as an embedding space for network data 12

1 Introduction

In high dimensional statistics, geometrical structures can be used to reduce dimensionality. High dimensional data suffers from the “curse of dimensionality” [Bellman, 1961; Lee and Verleysen, 2007a, Hastie et al., 2009], which refers to the fact that the number of data samples for an inference with
the desired accuracy grows exponentially with dimensions. The curse of dimensionality is mitigated if the data are to form geometrical structures. The assumed geometrical structures can both lower the dimensionality of the data and approximate complicated structure of the data.

On the other hand, geometrical structures of the data also provide information on data. First, geometrical structures carry scientific meaning about data in many scientific applications. For example, geometrical structures of galaxies, gas, and dark matter in the universe give clues on the initial state of the universe before the big bang. Also, geometrical structures of an enzyme determine its function. Second, the geometrical structures are used to enhance supervised or unsupervised learning. For this case, the interpretation of geometrical structures is unclear, but geometrical structures are extracted from data for higher performance in learning.

Lastly, geometry is also used in data visualization to provide insights on data through visual intuition. Some geometrical structures in data visualization such as size, orientation, shape are basic visual attributes that are perceived without conscious effort. Hence those geometrical structures are perceived in parallel and hence fast\cite{Few2004}. Nonquantitative information can be also conveyed by geometrical structures\cite{Few2013}. For example, a graph in 2d representing network data gives an immediate interpretation about which nodes are clustered or which nodes are influential.

In this proposal, I will explore how statistical inference can be done on geometrical structures, with regard to five areas. First, I will explore the minimax rates of dimension estimator (Section 2.1) and reach estimator (Section 2.2). Second, I will explore the minimax rate on persistent homology with regard to five areas. First, I will explore the minimax rates of dimension estimator (Section 2.1) and reach estimator (Section 2.2). Third, I will investigate inference on cluster trees (Section 2.3), voids (Section 2.4), and reach estimator (Section 2.2). Fourth, I will extend and improve R package TDA for computing topological persistent homology of density filtration on rips complex (Section 3.1), and multidimensional persistent homology (Section 3.3). Fourth, I will extend and improve R package TDA for computing topological persistent homology of density filtration on rips complex (Section 3.1), and multidimensional persistent homology (Section 3.3). Fifth, I will study embedding space for network data (Section 3.4).

1.1 Differential Geometry

We briefly review some notation from differential geometry. A topological manifold of dimension $d$ is a topological space $M$ and a family of homeomorphisms $\varphi_\alpha : U_\alpha \subset \mathbb{R}^d \to V_\alpha \subset M$ from an open subset of $\mathbb{R}^d$ to an open subset of $M$ such that $\bigcup_\alpha U_\alpha = M$. Such $d$ is unique and is called the dimension of a manifold. If, for any pair $\alpha, \beta$, with $\varphi_\alpha(U_\alpha) \cap \varphi_\beta(U_\beta) \neq \emptyset$, $\varphi_\beta^{-1} \circ \varphi_\alpha : U_\alpha \cap U_\beta \to U_\alpha \cap U_\beta$ is $C^k$, then $M$ is a $C^k$-manifold.

We assume that the topological manifold $M$ is embedded in $\mathbb{R}^m$, i.e. $M \subset \mathbb{R}^m$, and the metric is inherited from the metric of $\mathbb{R}^m$. For a topological manifold $M \subset \mathbb{R}^m$ and for any $q, r \in M$, a path joining $q_1$ to $q_2$ is a map $\gamma : [a, b] \to M$ for some $a, b \in \mathbb{R}$ such that $\gamma(a) = q_1, \gamma(b) = q_2$. The length of the curve $\gamma$ is defined as $\text{Length}(\gamma) = \int_a^b ||\gamma'(t)||_2 dt$. A topological manifold $M$ is equipped with the distance $\text{dist}_M : M \times M \to \mathbb{R}$ as $\text{dist}_M(q_1, q_2) = \inf_{\gamma: \text{path joining } q_1 \text{ and } q_2} \text{Length}(\gamma)$. A path $\gamma : [a, b] \to M$ is a geodesic if for all $t, t' \in [a, b], \text{dist}_M(\gamma(t), \gamma(t')) = |t - t'|$.

Let $T_q M$ denote the tangent space to $M$ at $q$. Given $q \in M$, there exist a set $0 \in \mathcal{E} \subset T_q(M)$ and a mapping $\exp_q : \mathcal{E} \subset T_q M \to M$ such that $t \to \exp_q(tv), t \in (-1, 1)$, is the unique geodesic of $M$ which, at $t = 0$, passes through $q$ with velocity $v$, for all $v \in \mathcal{E}$. The map $\exp_q : \mathcal{E} \subset T_q M \to M$ is called the exponential map on $q$. 

3
1.2 Reach

First introduced by Federer [Federer, 1959b], the reach is a regularity parameter defined as follows. Given a closed subset $A \subset \mathbb{R}^m$, the medial axis of $A$, denoted by $\text{Med}(A)$, is the subset of $\mathbb{R}^m$ composed of the points that have at least two nearest neighbors on $A$. Namely, denoting by $d(x, A) = \inf_{q \in A} ||q - x||$ the distance function to $A$,

$$\text{Med}(A) = \{ x \in \mathbb{R}^m | \exists q_1 \neq q_2 \in A, ||q_1 - x|| = ||q_2 - x|| = d(x, A) \} .$$

With the notation above, the reach of $A$ is defined as the minimal distance from $A$ to $\text{Med}(A)$.

**Definition 1.** The reach of a closed subset $A \subset \mathbb{R}^m$ is defined as

$$\tau_A = \inf_{q \in A} d(q, \text{Med}(A)) .$$

In the case of submanifolds, one can reformulate the definition of the reach in the following manner.

**Theorem 2.** [Federer, 1959a, Theorem 4.18] For all submanifold $M \subset \mathbb{R}^m$,

$$\tau_M = \inf_{q_1 \neq q_2 \in M} \frac{||q_1 - q_2||^2}{2d(q_2 - q_1, T_{q_1}M)} .$$

1.3 Persistent Homology

Suppose $X \subset \mathbb{R}^m$ be an observed data points. A filtration $\mathcal{F}$ is a collection of subsets in $\mathbb{R}^m$ that approximates the data points in different resolutions. Define a partial order on $\mathbb{R}^D$ by taking $(a_1, \cdots, a_D) \preceq (b_1, \cdots, b_D)$ if and only if $a_i \leq b_i$ for all $i$.

**Definition 3.** ($D$-dimensional) filtration $\mathcal{F} = \{ \mathcal{F}_a \subset \mathbb{R}^m : a \in \mathbb{R}^D \}$ is a collection of subsets in $\mathbb{R}^m$ satisfying that $a \preceq b$ implies $\mathcal{F}_a \subset \mathcal{F}_b$.

For filtration $\mathcal{F}$ and for each $k \in \mathbb{N}_0 = \mathbb{N} \cup \{0\}$, associated persistent homology $H_k \mathcal{F}$ is a collection of $k$-th dimensional homology of each subset in $\mathcal{F}$.

**Definition 4.** Let $\mathcal{F}$ be a $D$-dimensional filtration and let $k \in \mathbb{N}_0$. Associated ($D$-dimensional) $k$-th persistent homology $H_k \mathcal{F} = \{ H_k \mathcal{F}_a : a \in \mathbb{R}^D \}$ is a collection of vector spaces, where $H_k \mathcal{F}_a$ is a $k$-th dimensional homology of $\mathcal{F}_a$.

For 1-dimensional persistent homology, its structure is completely represented as its decomposition. For $k$-th persistent homology $H_k \mathcal{F}$, the set of filtration values that a specific homology appears is always an interval $[b, d] \subset [-\infty, \infty]$, i.e. a specific homology is formed at some filtration value $b \in [-\infty, \infty]$ and dies when the inside hole is filled at some filtration value $d \in [-\infty, \infty]$.

**Definition 5.** Let $\mathcal{F}$ be a 1-dimensional filtration and let $k \in \mathbb{N}_0$. Associated $k$-th persistent diagram $Dgm_k(\mathcal{F})$ is a set of $(b,d)$ with multiplicity, where $[b,d]$ is the set of filtration values that a specific homology appears in $H_k \mathcal{F}$. $b$ is called a birth time and $d$ is called a death time.

Stability theorems and statistical inference have been developed for 1-dimensional filtrations, in particular when the filtration $\mathcal{F}$ is generated from sub-level sets or super-level sets of a function. Let $f : \mathbb{R}^m \to \mathbb{R}$ be a function that approximates the data points in different resolutions. The associated filtration $\mathcal{F}$ can be constructed from sub-level sets $\mathcal{F}_a = \{ x \in \mathbb{R}^m : f(x) \leq a \}$ or super-level sets $\mathcal{F}_a = \{ x \in \mathbb{R}^m : f(x) \geq a \}$. Common choices for the filtration function $f$ are as follows: (1) sub-level sets of distance function $f(x) = d(x, X) = \inf_{y \in X} d(x, y)$, (2) super-level set of density function $f(x) = \hat{p}_h(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h^m} K \left( \frac{|x-x_i|}{h} \right)$, with any kernel $K$ and a positive number $h$. Super-level sets of
function $f$ corresponds to sub-level sets of function $-f$, hence the same theory can be used. For each $k \in \mathbb{N}_0$, let $Dgm_k(f)$ be $k$-th persistent diagram from either sub-level sets or super-level sets of $f$.

Statistical inference have been developed for persistent homology in [Fasy et al., 2014]. When points of birth and death are close to the diagonal in the persistence diagram, corresponding homologies are not significant, since corresponding holes will be soon filled out right after when they are born. With detailed statistical analysis, a $1-\alpha$ confidence band $c_n$ for persistent homology can be calculated. Precisely, $c_n$ satisfies

$$
\lim \inf_{n \to \infty} \mathbb{P} \left( W_\infty(Dgm_k(f), Dgm_k(f)) \in [0, c_n] \right) \geq 1 - \alpha,
$$

where $Dgm_k(f)$ is persistence diagram for the true distribution of data, $Dgm_k(f)$ is persistence diagram computed on data, and $W_\infty(X, Y)$ is the bottleneck distance between two diagrams $X$ and $Y$ defined as $W_\infty(X, Y) = \inf_{\eta: X \to Y} \|x - \eta(x)\|_\infty$. Those holes above the confidence band are simultaneously statistically significant.

Sublevel sets of the distance to measure (DTM) [Caillerie et al., 2011] is considered to approximate holes in the data points in different resolutions. The DTM is a robustified version of the distance function. More precisely, the DTM $d_{\mu,m_0}$ for a probability distribution $\mu$ with parameter $m_0 \in [0, 1]$ is defined by

$$
d_{\mu,m_0} : \mathbb{R}^m \to \mathbb{R}^+, \ x \mapsto \sqrt{\frac{1}{m_0} \int_0^{m_0} (\delta_{\mu,m_0}(x))^2 \, dm},
$$

where $\delta_{\mu,m_0}(x) = \inf \{r > 0 : \mu(B(x, r)) > m_0 \}$. When $\mu$ is an empirical measure $P_n(x) = \frac{1}{n} \sum_{i=1}^{n} I_{X_i}(x)$, the empirical DTM is

$$
\hat{d}_{\mu,m_0}(x) = d_{P_n,m_0}(x) = \sqrt{\frac{1}{m_0n} \sum_{i \leq \lfloor m_0n \rfloor} \|X_{(i)} - x\|_2^2 + \left(1 - \frac{\lfloor m_0n \rfloor}{m_0n}\right) \|X_{\lfloor m_0n \rfloor} - x\|_2^2}, \quad (3)
$$

where for each $x$, $X_{(1)}, \ldots, X_{(n)}$ is ordered so that $\|X_{(1)} - x\|_2 \leq \cdots \leq \|X_{(n)} - x\|_2$. Hence the empirical DTM behaves similarly to the $k$-nearest distance with $k = \lfloor m_0n \rfloor$. The DTM is preferred choice for the filtration function, since the persistence diagram computed on the DTM is robust to noise.

2 Ongoing work

2.1 Minimax Rates for Estimating the Dimension of a Manifold

Suppose that $X_1, \ldots, X_n$ is an i.i.d. sample from a distribution $P$ whose support is an unknown, well behaved, manifold $M$ of dimension $d$ in $\mathbb{R}^m$, where $1 \leq d \leq m$. Manifold learning refers broadly to a suite of techniques from statistics and machine learning aimed at estimating $M$ or some of its features based on the data.

Most manifold learning techniques require, as input, the intrinsic dimension of the manifold. However, this quantity is almost never known in advance and therefore has to be estimated from the data.

Various intrinsic dimension estimators have been proposed and analyzed; [see, e.g., Lee and Verleysen, 2007b; Koltchinskii, 2000; Kegl, 2003; Levina et al., 2004; Hein and Audibert, 2005; Raginsky and Lazebnik, 2005; Little et al., 2009; 2011; Sricharan et al., 2010; Rozza et al., 2012; Camastra and
However, characterizing the intrinsic statistical hardness of estimating the dimension remains an open problem.

The traditional way of measuring the difficulty of a statistical problem is to bound its minimax risk, which in the present setting is loosely described as the worst possible statistical performance of an optimal dimension estimator. Formally, given a class of probability distribution $\mathcal{P}$, the minimax risk $R_n = R_n(\mathcal{P})$ is defined as

$$\begin{align*}
R_n = \inf_{\hat{d}} \sup_{P \in \mathcal{P}} \mathbb{E}_P \left[ 1(\hat{d} \neq d(P)) \right].
\end{align*}$$

(4)

In (4), $d(P)$ is the dimension of the support of $P$, $\mathbb{E}_P$ denotes the expectation with respect to the distribution $P$, $1(\cdot)$ is the indicator function, and the infimum is over all estimators (measurable functions of the data) $\hat{d} = \hat{d}(X_1, \ldots, X_n)$ of the dimension $d(P)$. The risk $\mathbb{E}_P[1(\hat{d} \neq d(P))]$ of a dimension estimator $\hat{d}$ is the probability that $\hat{d}$ differs from the true dimension $d(P)$ of the support of the data generating distribution $P$. The minimax risk $R_n(\mathcal{P})$, which is a function of both the sample size $n$ and the class $\mathcal{P}$, quantifies the intrinsic hardness of the dimension estimation problem, in the sense that any dimension estimator cannot have a risk smaller than $R_n$ uniformly over every $P \in \mathcal{P}$.

To define a meaningful minimax risk, it is essential to have some constraint on the set of probability distributions $\mathcal{P}$. If $\mathcal{P}$ is too large, then the minimax rate $R_n$ in (4) will not converge to 0 as $n$ goes to $\infty$, which means that the problem is statistically ill-posed. If $\mathcal{P}$ is too small, the data may not be well approximated by the distributions in $\mathcal{P}$ and hence the corresponding minimax estimator is not useful.

2.1.1 Upper Bound and Lower Bound

Determining the value of the minimax risk $R_n$ in (4) for a given problem requires two separate calculations: an upper bound on $R_n$ and a lower bound on $R_n$. In order to derive an upper bound, one analyzes the asymptotic risk of a specific estimator $\hat{d}_n$. Lower bounds are instead usually computed by measuring the difficulty of a multiple hypothesis testing problem that entails identifying finitely many distributions in $\mathcal{P}$ that are maximally difficult to discriminate [Tsybakov, 2008, Section 2.2].

We summarize our result in Theorem (6).

**Theorem 6.** Let $\mathcal{P}$ be a set of probability distributions supported on the submanifold embedded in $\mathbb{R}^m$ with mild conditions. Then the minimax risk $R_n$ in (4) is bounded as

$$\begin{align*}
C_{\mathcal{P}}^{(1)} n^{-2\alpha} \leq \inf_{\hat{d}_n} \sup_{P \in \mathcal{P}} \mathbb{E}_P \left[ 1(\hat{d}_n \neq d(P)) \right] \leq C_{\mathcal{P}}^{(2)} n^{-\frac{m}{m-1}},
\end{align*}$$

where $C_{\mathcal{P}}^{(1)}, C_{\mathcal{P}}^{(2)} \in (0, \infty)$ are constants depending on $\mathcal{P}$.

2.2 The Origin of the Reach: Better Understanding Regularity Through Minimax Estimation Theory

First introduced by Federer [Federer, 1959b] to unify volume formulas known in convex geometry on the one hand side and in differential geometry on the other, the reach $\tau_M$ of $M \subset \mathbb{R}^m$ is the largest length such that any point at distance less than $\tau_M$ of $M$ has a unique nearest neighbor on $M$. For a set, having positive reach greater than $\tau_{\min} > 0$ roughly means that one can roll freely a ball of radius $\tau_{\min}$ around it [Cuevas et al., 2012]. The reach informs on maximal directional curvature and on the width of possible narrow bottleneck structures on the shape. It corresponds to a minimal size of features $M$ contains. In statistical settings, such a scale corresponds to the least sampling density needed to recover geometric information.
Since sets with positive reach enjoy good geometric properties \cite{Federer1969, Thaele2008}, it has recently grown popular in the literature. In manifold reconstruction, reach regularity helps to formalize simple models on which minimax rates are well posed \cite{Genovese2012a, Kim2015}. The effective optimal estimators \cite{Boissonnat2014, Aamari2015} implicitly use it as a scale parameter in their construction. In homology inference \cite{Niyogi2008}, it drives the minimal sample size required to consistently estimate topological invariants. Reach has been explicitly used in geometric inference, such as volume estimation \cite{Niyogi2008} and manifold clustering \cite{Arias-Castro2016}. It is also a good regularity notion for dimension reduction techniques such as vector diffusions maps \cite{Singer2012}.

This work gives new geometric results on what the reach relates to, and tackles the question of its estimation in minimax frameworks. Formally, given a class of probability distribution \( \mathcal{P} \), the minimax risk \( R_n = R_n(\mathcal{P}) \) is defined as

\[
R_n = \inf_{\tau_n} \sup_{P \in \mathcal{P}} \mathbb{E}_P \left[ \left| \frac{1}{\tau(P)} - \frac{1}{\tau_n} \right| \right].
\]

In (5), \( \tau(P) \) is the reach of the support of \( P \), \( \mathbb{E}_P \) denotes the expectation with respect to the distribution \( P \), and the infimum is over all estimators (measurable functions of the data) \( \tau_n(X_1, \ldots, X_n) \) of the reach \( \tau(P) \). The minimax risk \( R_n(\mathcal{P}) \) has an interpretation that any reach estimator cannot have a risk smaller than \( R_n \) uniformly over every \( P \in \mathcal{P} \).

In our model, we assumed that tangent spaces are observed at all the sample points. In other words, we assume that when \( X_1, \ldots, X_n \) are observed, \( T_{X_1}M, \ldots, T_{X_n}M \) are observed as well.

### 2.2.1 Geometry of the Reach

We first define the concept of reach attaining pair and bottleneck related to how reach is attained on the manifold.

**Definition 7.** Let \( M \subset \mathbb{R}^m \) be a submanifold with reach \( \tau_M \).

- A pair of points \( (q_1, q_2) \in M^2 \) is called reach attaining if there exists a point \( z_0 \in \text{Med}(M) \) such that \( q_1, q_2 \in B(z_0, \tau_M) \). We call \( z_0 \) the axis point of \( (q_1, q_2) \), and \( ||q_1 - q_2|| \in (0, 2\tau_M) \) its size.
- A reach attaining pair \( (q_1, q_2) \in M^2 \) is said to be a bottleneck of \( M \) if its size equals \( 2\tau_M \), i.e. \( ||q_1 - q_2|| = 2\tau_M \).

Then, we show that either the manifold has a bottleneck structure (global case), or the reach is coming from the local structure of the manifold (local case), which is stated in Theorem 8.

**Theorem 8.** Let \( M \subset \mathbb{R}^m \) be a compact submanifold with reach \( \tau_M \). At least one of the following two assertions holds.

- (Global case) \( M \) has a bottleneck \( (q_1, q_2) \in M^2 \), i.e. there exists \( z_0 \in \text{Med}(M) \) such that \( q_1, q_2 \in \partial B(z_0, \tau_M) \) and \( ||q_1 - q_2|| = 2\tau_M \).
- (Local case) There exists \( q_0 \in M \) and an arc-length parametrized geodesic \( \gamma_0 \) such that \( \gamma_0(0) = q_0 \) and \( ||\gamma''_0(0)|| = \frac{1}{\tau_M} \).

### 2.2.2 Our Reach Estimator and Minimax Estimates

Given a finite point cloud \( \mathcal{X} = \{x_1, \ldots, x_n\} \subset M \), we define our reach estimator \( \hat{\tau} : \mathbb{R}^m \rightarrow \mathbb{R} \) as

\[
\hat{\tau}(\mathcal{X}) = \inf_{1 \leq i < j \leq n} \frac{||x_j - x_i||^2}{2d(x_j - x_i, T_{x_i}M)}. \tag{6}
\]
We analyze the asymptotic risk of this reach estimator to get the upper bound of the minimax rate. Lower bound of the minimax rate is computed by measuring the difficulty of identifying distributions in $\mathcal{P}$ that are similar to each other.

We summarize our result in Theorem 9.

**Theorem 9.** Let $\mathcal{P}$ be a set of probability distributions supported on the $d$-dimensional submanifold embedded in $\mathbb{R}^m$ with mild conditions. The minimax risk $R_n$ in $\mathbb{E}$ is bounded as

$$C_{\mathcal{P}}^{(1)} n^{-\frac{d}{2}} \leq \inf_{\hat{\tau}_n} \sup_{P \in \mathcal{P}} \mathbb{E}_{P(n)} \left[ \frac{1}{\hat{\tau}_n} - \frac{1}{\tau_n} \right] \leq C_{\mathcal{P}}^{(2)} n^{-\frac{2d}{2d+1}},$$

where $C_{\mathcal{P}}^{(1)}, C_{\mathcal{P}}^{(2)} \in (0, \infty)$ are constants depending on $\mathcal{P}$.

### 2.3 Statistical Inference for Cluster Trees

In this work, we study statistical inference for the cluster tree of an unknown density. We assume that we observe an i.i.d. sample $\{X_1, \ldots, X_n\}$ from a distribution $P_0$ with unknown density $p_0$. Here, $X_i \in \mathcal{X} \subset \mathbb{R}^m$. The connected components $C(\lambda)$, of the upper level set $\{x : p_0(x) \geq \lambda\}$, are called high-density clusters. The set of high-density clusters forms a nested hierarchy which is referred to as the cluster tree of $p_0$, which we denote as $T_{p_0}$.

Methods for density clustering fall broadly in the space of hierarchical clustering algorithms, and inherit several of their advantages: they allow for extremely general cluster shapes and sizes and do not, in general, require the pre-specification of the number of clusters. Furthermore, unlike flat clustering methods, hierarchical methods are able to provide a multi-resolution summary of the underlying density. The cluster tree, irrespective of the dimensionality of the input random variable, is displayed as a two-dimensional object and this makes it an ideal tool to visualize data. In the context of statistical inference, density clustering has another important advantage over other clustering methods: the object of inference, the cluster tree of the unknown density $p_0$, is clearly specified.

Some topological structure of the estimated tree will often arise due to random fluctuations in the data, and in a scientific application, we are often most interested in reliably distinguishing topological features genuinely present in the cluster tree of the unknown $T_{p_0}$, from topological features that arise due to finite sample artifacts. In this paper, we focus our inference on the cluster tree of the kernel density estimator, $T_{\hat{p}_h}$, where $\hat{p}_h$ is the kernel density estimator,

$$\hat{p}_h(x) = \frac{1}{nh^m} \sum_{i=1}^n K \left( \frac{||x - X_i||}{h} \right),$$

where $K$ is a kernel and $h$ is an appropriately chosen bandwidth.

#### 2.3.1 Tree Metrics

In this work, we use standard $\ell_\infty$ metric $d_\infty(T_p, T_q) = \sup_{x \in \mathcal{X}} |p(x) - q(x)|$. This metric has several advantages in statistical inference. The large sample behavior of $d_\infty$ is well understood (see [Chen et al., 2015, Chernozhukov et al., 2016]). In particular, $d_\infty(T_{\hat{p}_h}, T_{p_0})$ converges to the supremum of an appropriate Gaussian process, on the basis of which we can construct confidence intervals for the $d_\infty$ metric.
2.3.2 Confidence Sets

We consider the construction of valid confidence intervals centered around the kernel density estimator (KDE) in Equation (7). As a conceptual first step, suppose that for a specified value $\alpha$ we could compute the $1 - \alpha$ quantile of the distribution of $d_\infty(T_{\hat{p}_h}, T_{p_h})$, and denote this value $t_\alpha$. Then a valid confidence set for the unknown $T_{p_h}$ is $C_\alpha = \{T : d_\infty(T, \hat{T}_{p_h}) \leq t_\alpha\}$. To estimate $t_\alpha$, we use the bootstrap. Specifically, we generate $B$ bootstrap samples, $\tilde{X}_1^n, \ldots, \tilde{X}_B^n$, by sampling with replacement from the original sample. On each bootstrap sample, we compute the KDE, and the associated cluster tree. We denote the cluster trees $\tilde{T}_1^n, \ldots, \tilde{T}_B^n$. Finally, we estimate $t_\alpha$ by

$$\hat{t}_\alpha = \hat{F}^{-1}(1 - \alpha), \text{ where } \hat{F}(s) = \frac{1}{B} \sum_{i=1}^{n} \mathbb{I}(d_\infty(\tilde{T}_{\hat{p}_h}, \tilde{T}_{p_h}) < s).$$

Then the data-driven confidence set is $\hat{C}_\alpha = \{T : d_\infty(T, \hat{T}_h) \leq \hat{t}_\alpha\}$. The following can be shown:

**Theorem 10.** Under mild regularity conditions on the kernel, we have that the constructed confidence set is asymptotically valid and satisfies,

$$\mathbb{P}(T_h \in \hat{C}_\alpha) = 1 - \alpha + O\left(\left(\frac{\log^7 n}{nh^m}\right)^{1/6}\right).$$

While we establish the theorem in full generality, when we are interested in inference for the cluster tree, we take $h$ to be a fixed small constant, and the above theorem guarantees that the non-parametric bootstrap is consistent at a dimension independent $O\left((\log n)^7 / n h^m\right)^{1/6}$ rate. This rate is believed to be optimal.

The confidence set $\hat{C}_\alpha$ is an infinite set with a complex structure. One way to understand the structure of the confidence set is to focus on simple trees in the confidence set. Intuitively, these trees only contain topological features (splits and branches) that are sufficiently strongly supported by the data.

We propose two pruning schemes to find trees, that are simpler than the empirical tree $\hat{T}_{p_h}$ that are in the confidence set. Pruning the empirical tree aids visualization as well as de-noises the empirical tree by eliminating some topological features that arise solely due to the stochastic variability of the finite-sample. The algorithms are:

1. **Pruning only leaves:** Remove all leaves of length less than $2\hat{t}_\alpha$.
2. **Pruning leaves and internal branches:** In this case we first prune the leaves as above. This yields a new tree. Now we again prune (using cumulative length) any leaf of length less than $2\hat{t}_\alpha$. We continue iteratively until all remaining leaves are of cumulative length larger than $2\hat{t}_\alpha$.

The remaining tree $\tilde{T}$ is a valid tree with a statistical guarantee that is simpler than the original estimate $\hat{T}_{p_h}$.

2.4 Topological classification and sub-classification of voids using persistent homology

Though there is not a generally accepted definition of voids, void-finding methods often use the geometry at the apparent boundary of higher densities of matter. Several existing methods for finding voids follow this approach, such as Watershed Void Finder (WVF, \cite{Platen2007}), ZOnes Bordering On Voidsness (ZOBOV, \cite{Neyrinck2008}), and Void IDentication and Examination (VIDE, \cite{Sutter2014}). These methods are designed to find voids, but are not generally capable of determining if those
voids are statistically significant. The proposed method, “TopVoids”, finds statistically significant voids by considering a multiscale topological signature of the data, called persistent homology. In addition to locating statistically significant cosmological voids under the usual interpretation of voids, the proposed method defines two types, or classes, of voids with respect to certain topological characteristics. In particular, we can locate voids that are low-density regions completely surrounded by high-density regions (“wall voids”), and voids that are loops in the cosmic web formed by filaments (“filament voids”).

2.4.1 Methodology

For finding voids, TopVoids computes persistent homology on sublevel sets of DTM functions. Suppose we observe iid data $X_1, \cdots, X_n \sim \mu$, and let the parameter $m_0 \in (0, 1)$ be fixed. For the empirical DTM function $\hat{d}_{\mu,m_0}$ in (3), we consider the filtration

$$F = \left\{ \{x \in \mathbb{R}^m : \hat{d}_{\mu,m_0}(x) \leq t\} : t \in (0, \infty) \right\},$$

where for each $x \in \mathbb{R}^m$, $X(1), \cdots, X(n)$ is ordered so that $\|X(1) - x\|_2 \leq \cdots \leq \|X(n) - x\|_2$. From this filtration, TopVoids computes its persistent homology.

TopVoids then computes $1 - \alpha$ confidence band $c_n$ using bootstrap, where $c_n$ satisfies

$$\lim \inf_{n \to \infty} \mathbb{P} \left( W_{\infty}(Dgm_k(\hat{d}_{\mu,m_0}), Dgm_k(d_{\mu,m_0})) \in [0, c_n] \right) \geq 1 - \alpha$$

for $k = 1, 2$, where $Dgm_k(d_{\mu,m_0})$ is the $k$-th persistent diagram for true distribution $\mu$ and $Dgm_k(\hat{d}_{\mu,m_0})$ is the empirical $k$-th persistent diagram computed on data. Then, any homological feature $e \in Dgm_k(\hat{d}_{\mu,m_0})$ whose lifetime is longer than $2c_n$ is considered to be a significant void, for either $k = 1$ or $k = 2$.

TopVoids further classify voids by their dimension. Any significant 1-dim void $e \in Dgm_1(\hat{d}_{\mu,m_0})$ is called a “filament void”, and any significant 2-dim void $e \in Dgm_2(\hat{d}_{\mu,m_0})$ is called a “wall void”. The wall void is surrounded by matter, and the filament void is surrounded by a loop of matter, but the vacant space is open to the normal direction of the loop; it is essentially void inside the loop.

2.5 R Package TDA: Statistical Tools for Topological Data Analysis

This work is devoted to the presentation of the R package TDA, which provides a user-friendly interface for the efficient algorithms of the C++ libraries GUDHI, Dionysus, and PHAT.

First, we describe how to compute some widely studied functions that, starting from a point cloud, provide some topological information about the underlying space: the distance function (distFct), the distance to a measure function (dtm), the k Nearest Neighbor density estimator (knnDE), the kernel density estimator (kde), and the kernel distance (kernelDist). Next, we compute persistence diagrams: the function gridDiag can be used to compute persistent homology of sublevel sets (or superlevel sets) of functions evaluated over a grid of points; the function ripsDiag returns the persistence diagram of the Rips filtration built on top of a point cloud.

One of the key challenges in persistent homology is to find a way to isolate the points of the persistence diagram representing the topological noise. Statistical methods for persistent homology provide an alternative to its exact computation. Knowing with high confidence that an approximated persistence diagram is close to the true—computationally infeasible—diagram is often enough for practical purposes. [Fasy et al., 2014], [Chazal et al., 2014b], and [Chazal et al., 2014a] propose several statistical methods to construct confidence sets for persistence diagrams and other summary functions that allow us to separate topological signal from topological noise. The methods are implemented in the TDA package.
Finally, the TDA package provides the implementation of an algorithm for density clustering. This method allows us to identify and visualize the spatial organization of the data, without specific knowledge about the data generating mechanism and in particular without any a priori information about the number of clusters. We describe the function clusterTree, that, given a density estimator, encodes the hierarchy of the connected components of its superlevel sets into a dendrogram, the cluster tree.

3 Proposed work

3.1 Statistical inference on persistent homology of density filtration on rips complex

The statistical inference on 1-dimensional persistent homology has been developed in [Fasy et al., 2014]. However, this approach is computationally infeasible in high dimensional space: since this approach is bootstrapping $l_\infty$ distance of the density estimator and its mean, density estimator functions are to be computed on a grid of points, which is computationally infeasible if the dimension is high.

One approach, suggested in [Bobrowski et al., 2014], is to compute the kernel density estimator on Rips filtration, and compute persistent homology on that. While the computational complexity of a density estimator on grid of points grows super-exponentially with the dimension, the computational complexity of a Rips filtration is bounded by only in terms of the sample size. And such persistent homology is also consistent, as shown in [Bobrowski et al., 2014].

Based on this approach, the goal of this work is to build a confidence set for the persistent homology of the density estimator on Rips filtration. We will show the stability theorem between density estimator on Rips filtration and true density function, and use that to build a confidence set as in [Fasy et al., 2014].

3.2 Statistical inference on multi-dimensional persistent homology

For 1-dimensional persistent homology, the statistical inference has been developed in [Fasy et al., 2014]. Extension of this result to multi-dimensional persistent homology is not trivial. For 1-dimensional persistent homology, statistical inference is based on the stability theorem, which measures the difference of two persistent homologies by their corresponding representations in persistent diagrams. For the multi-dimensional persistent homology, its structure is much more complicated and cannot be simply represented in the persistent diagram.

One way to avoid this difficulty is to consider projected 1-dimensional persistent homology from multi-dimensional persistent homology. The structure of the multi-dimensional persistent homology can be understood by considering its projection onto all possible 1-dimensional filtration, as in [Lesnick and Wright, 2015]. A distance between two multi-dimensional persistent homologies can be also measured by taking supremum over distances between two projected persistent homologies onto all possible 1-dimensional filtrations. The stability theorem for multi-dimensional persistent homology is built on this distance in [Landi, 2014].

Using this approach, the goal of this work is to build up statistical inference on multi-dimensional persistent homology. As statistical inference on 1-dimensional persistent homology is derived from stability theorem on 1-dimensional persistent homology in [Fasy et al., 2014], we will build statistical inference based on stability theorem in [Landi, 2014].
3.3 Mimimax rate of persistent homology

Usually, the persistent homology of the underlying manifold is approximated by the persistent homology of the sample points. As mentioned before, the statistical inference on the sampled persistent homology has been developed in [Fasy et al., 2014]. Though, the optimal guarantee of the estimated persistent homology is still suspicious: its convergence to the persistent homology of the underlying manifold is not compared with the convergence of other possible estimators of the persistent homology.

One traditional way to argue theoretical optimality of the estimator is to argue that the estimator achieves the minimax rate. Since the minimax rate is the worst possible statistical performance of an optimal estimator, the estimator achieving the minimax rate implies that the estimator is the best in terms of the worst case performance. Hence computing the minimax rate is theoretically interesting.

Based on these motivations, the goal of this work is to compute minimax rate of the persistent homology estimator. Computing the minimax rate requires computing both the upper bound and the lower bound. We will compute the upper bound by computing the asymptotic risk of the sampled persistent homology. Then we will compute the lower bound by evaluating the difficulty of identifying similar distributions that are maximally different in persistent homology.

3.4 Choosing between hyperbolic and euclidean as an embedding space for network data

For visualizing network data in 2d, data should be embedded to ambient space. Traditionally, data is embedded to euclidean space, such as multidimensional scaling. However, embedding network data to euclidean space is sometimes inappropriate. Considering the volume of a ball with increasing radius, then the volume grows polynomially in euclidean space, while the volume grows exponentially in tree network data, and hence embedding tree network data into euclidean space distort the distance structure in the data. An alternative is to embed data to hyperbolic space. First introduced by [Lamping et al., 1995], embedding network data in hyperbolic space has been vastly tried in visualizing internet network, where the degree of a network is believed to follow power law distribution.

Hence a natural question arises: when we should embed network data to euclidean space and to hyperbolic space. The goal of this work is to build some criteria for determining whether embedding space should be euclidean space or hyperbolic space.

References


