Statistical Inference for Topological Data Analysis

PhD Thesis Proposal

Fabrizio Lecci
Department of Statistics
Carnegie Mellon University

Thesis Committe

Jessi Cisewski
Frederic Chazal
Alessandro Rinaldo
Ryan Tibshirani
Larry Wasserman

January 31, 2014

Abstract

Topological Data Analysis (TDA) is an emerging area of research at the intersection of algebraic topology and computational geometry, aimed at describing, summarizing and analyzing possibly high-dimensional data using low-dimensional algebraic representations. Recent advances in computational topology have made it possible to actually compute topological invariants from data. These novel types of data summaries have been used successfully in a variety of applied problems, and their potential for high-dimensional statistical inference appears to be significant. Nonetheless, the statistical properties of the data summaries produced in TDA and, more generally, of the usually heuristic data-analytic methods they are part of, have remained largely unexplored by statisticians. Our analysis involves the tools of persistent homology, the main method of TDA for measuring the topological features of shapes and functions at different resolutions. A major part of our research also focuses on cluster trees and Reeb graphs, which provide a simple yet meaningful abstraction of the input domain of a function by means of the topological changes in its level sets.

The main goal of this thesis is to contribute to the development of a statistical theory for TDA and to further propose new and statistically principled methodologies to improve and extend the applicability of the algorithms of TDA. In particular, we will (1) study tests of significance and confidence intervals to separate topological signal from topological noise; (2) explore new methods for topological dimensional reduction; (3) determine how our methods contribute to reduce computational costs, which currently represent an obstacle in TDA.

Contents

1 Introduction

Topological Data Analysis (TDA) refers to a collection of methods for finding topological structure in data (Carlsson, 2009). Recent advances in computational topology have made it possible to actually compute topological invariants from data. The input of these procedures typically takes the form of a point cloud, regarded as possibly noisy observations from an unknown lower-dimensional set S whose interesting topological features were lost during sampling. The output is a collection of data summaries that are used to estimate the topological features of S.

These novel types of data summaries have been used successfully in a variety of applied problems, ranging from medical imaging and neuroscience (Chung et al., 2009; Pachauri et al., 2011) to cosmology (Sousbie, 2011; van de Weygaert et al., 2011; Cisewski et al., 2014), sensor networks (de Silva and Ghrist, 2007) landmark-based shape data analyses (Gamble and Heo, 2010), and cellular biology (Kasson et al., 2007). Nonetheless, the statistical properties of the data summaries produced in TDA and, more generally, of the usually heuristic data-analytic methods they are part of, have remained largely unexplored by statisticians.

One approach to TDA is persistent homology (Edelsbrunner and Harer, 2010), a method for studying the homology at multiple scales simultaneously. More precisely, it provides a framework and efficient algorithms to quantify the evolution of the topology of a family of nested topological spaces. Given a real-valued function f, persistent homology describes how the topology of the lower level sets $\{x : f(x) \leq t\}$ (or upper level sets $\{x : f(x) \geq t\}$) change as t increases from $-\infty$ to ∞ (or decreases from ∞ to $-\infty$). This information is encoded in the persistence diagram, a multiset of points in the plane, each corresponding to the birth and death of a homological feature that existed for some interval of t. Thanks to their stability properties (Cohen-Steiner et al., 2007; Chazal et al., 2012), persistence diagrams provide relevant multi-scale topological information about the data. One of the key challenges is persistent homology is to find a way to isolate the points of the persistence diagram representing the topological noise. In Fasy et al. (2013) and Chazal et al. (2013b) we 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 research objective of this proposal is to develop new theories and methods to improve and extend the applicability of the algorithms of persistent homology.

A second sect of research activities pertains to the related task of clustering in high dimensions. Density clustering 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 will consider the sublevel set tree as a topological descriptor and study its properties, including the notion of distance between trees that will lead to the definition of inferential procedures. We will also try to extend the results to a related descriptor, the Reeb graph, which encodes information on the level sets of a function and on the topology of the input domain (Biasotti et al., 2008).

2 Persistent Homology

We provide an informal description of the methods of homology and persistent homology. For rigorous expositions, the reader is referred to the textbooks Munkres (1984); Zomorodian (2005); Edelsbrunner and Harer (2010) and the introductory reviews Edelsbrunner and Harer (2008); Chazal et al. (2012).

2.1 Background

Homology is a mathematical formalism used to summarize the overall connectivity of a topological space. The homology of a space S is a collection of abelian groups of different dimensions, the pth dimensional group encoding the pth dimensional "holes" in S. The pth homology group $H_p(S)$ is the set of equivalence classes of loops enclosing the pth dimensional holes, and its rank β_p is called the pth *Betti number*. Roughly speaking, the pth Betti number β_p is the number of pth dimensional holes in S, so that β_0 is the number of connected components of S, β_1 is the number of loops, β_2 is the number of enclosed voids and so on. See Figure 1.

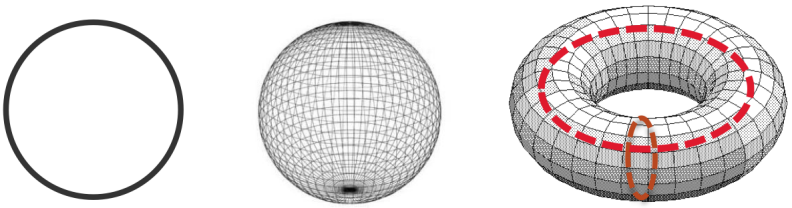

Figure 1: The circle has one connected component and one 1-dimensional hole: $\beta_0 = 1, \beta_1 = 1$. A sphere in $\mathbb{R}^3$ has one connected component and one 2-dimensional hole (void): $\beta_0 = 1, \beta_1 = 0, \beta_2 = 1$. The torus has one connected component, two 1-dimensional holes (the two non equivalent circles in red) and one enclosed void: $\beta_0 = 1, \beta_1 = 2, \beta_2 = 1$.

Persistent Homology is the main tool of TDA for measuring the scale or resolution of topological features. Given a function $f : \mathbb{X} \to \mathbb{R}$, defined for a triangulable subspace of $\mathbb{R}^D$, persistent homology describes the changes in the topology of the lower (or upper) level sets of f. For example, consider the lower level sets $L_t = \{x \in \mathbb{X} : f(x) \leq t\}$. The index t can be seen as a scale parameter leading to a filtration of subspaces, such that $L_t \subseteq L_s$ for all $t \leq s$. Such a filtration induces a family $\{H(L_t) : t \in \mathbb{R}\}$ of homology groups and the inclusions $L_t \hookrightarrow L_s$ induce a family of homomorphisms $H(L_t) \to H(L_s)$. Persistent homology describes f with the *persistence diagram*, a multiset of points in the plane, each corresponding to the birth and death of a homological feature that existed for some interval of t. The point (s, t) in the diagram represents a distinct topological feature that existed in $H(L_r)$ for $r \in [s, t)$.

In the following we focus on the persistent homology of distance functions and density functions, providing more details on the construction of the persistence diagrams and a few clarifying examples.

2.1.1 Persistent Homology of the distance function

First, we consider the case where f is the distance function. Let $\mathbb{S}$ be a compact subset of $\mathbb{R}^D$ and let $d_{\mathbb{S}}\colon \mathbb{R}^D \to \mathbb{R}$ be the distance function to $\mathbb{S}$:

$$d_{\mathbb{S}}(x) = \inf_{y \in \mathbb{S}} \|y - x\|_2.$$

Consider the sub-level set $L_t = \{x : \ d_S(x) \leq t\}$; note that $\mathbb{S} = L_0$. As t varies from 0 to ∞, the set L_t changes. Persistent homology summarizes how the topological features of L_t change as a function of t. Key topological features of a set include the connected components (the zeroth order homology), the tunnels (the first order homology), voids (second order homology), etc. These features can appear (be born) and disappear (die) as t increases. For example, connected components of L_t die when they merge with other connected components. Each topological feature has a birth time b and a death time d. In general, there will be a set of features with birth and death times $(b_1, d_1), \ldots, (b_m, d_m)$. These points can be plotted on the plane, resulting in a persistence diagram $\mathcal{P}$. We view the persistence diagram as a topological summary of the input function or data.

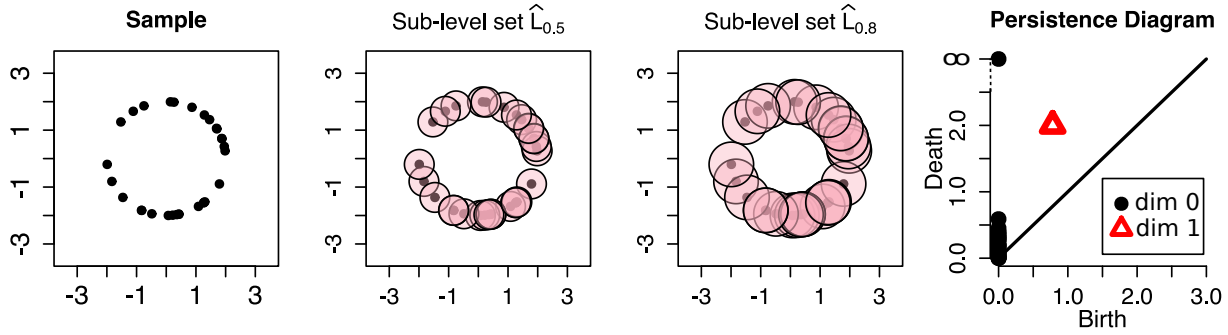

Figure 2: **Left**: 30 data points $\mathcal{S}_{30}$ sampled from the circle of radius 2. **Middle left**: sub-levels set $\hat{L}_{0.5} = \{x : d_{\mathcal{S}_{30}} \leq 0.5\}$; for $t = 0.5$ the sub-level set consists of two connected components and zero loops. **Middle right**: sub-levels set $\hat{L}_{0.8} = \{x : d_{\mathcal{S}_{30}} \leq 0.8\}$; as we keep increasing t we assist at the birth and death of topological features; at $t = 0.8$ one of the connected components dies (is merged with the other one) and a 1-dimensional hole is formed; this loop will die at $t = 2$, when the pink balls representing the distance function will touch each other in the center of the circle. **Right**: the empirical persistence diagram summarizes the topological features of the sampled points. The black dots represent the connected components: 30 connected components are present at $t = 0$ and they progressively die as t increases, leaving only one connected component that persists for large values of t. The red triangle represent the unique 1-dimensional hole that is formed at $t = 0.8$ and dies at $t = 2$.

Given data points $\mathcal{S}_n = \{X_1, \ldots, X_n\}$, we are interested in understanding the homology of the d-dimensional compact topological space $\mathbb{S} \subset \mathbb{R}^D$ from which the data were sampled. If our sample is dense enough then $\hat{L}_t = \{x : \ d_{\mathcal{S}_n}(x) \leq t\}$ has the same homology of $\mathbb{S}$ for an interval of values of t. Choosing the right t is a difficult task: small t will have the homology

of n points and large t will have the homology of a single point. Using persistent homology, we avoid choosing a single t by assigning a persistence value to each non-trivial topological feature that is realized for some non-negative t. As t varies, we summarize birth and death of topological features of $\mathcal{S}_n$ using the empirical persistence diagram $\hat{\mathcal{P}}$. We treat $\hat{\mathcal{P}}$ as an estimate of the unobserved persistence diagram $\mathcal{P}$ of the underlying space $\mathbb{S}$. Points near the diagonal in the persistence diagram have short lifetimes and are considered "topological noise". In most applications we are interested in features that we can distinguish from noise; that is, those features that persist for a large range of values of t. Figure 2 shows an example that clarifies the concepts described above.

Despite the seemingly geometric nature of homology invariants, they are in fact purely combinatorial quantities, which are computed by triangulating a topological space with simplicial complexes (Zomorodian and Carlsson, 2005). In some practical applications, the number of simplices can be so large that the exact computation of the persistent homology becomes prohibitive. Efficient approaches for approximating the persistent homology will be useful only if combined with statistical guarantees.

2.1.2 Persistent Homology of the density function.

Most of the literature on computational topology focuses on the distance function. Alternatively, one can use the data to construct a smooth density estimator and then find the persistence diagram defined by a filtration of the upper level sets of the density estimator. This strategy is discussed in detail in Fasy et al. (2013), where it is shown that the density-based method is very insensitive to outliers. A different approach to smoothing based on diffusion distances is discussed in Bendich et al. (2011).

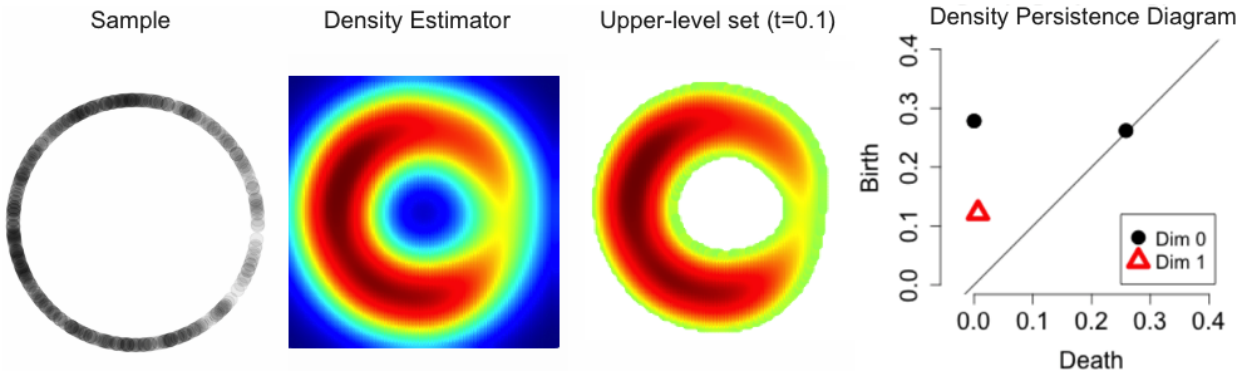

Figure 3: **Left**: 500 data points sampled from the circle of radius 1. **Middle left**: Gaussian kernel density estimator with bandwidth $h = 0.3$. **Middle right**: upper-levels set $\hat{U}_{0.1} = \{x : \hat{p}_h \leq 0.1\}$. **Right**: the empirical density persistence diagram summarizes the topological features of upper level sets of the kernel density estimator. The black dots represent the connected components: 2 connected components appear around $t = 0.27$, but one of them immediately dies (is merged to the other one). The red triangle represent the unique 1-dimensional hole that is formed at $t = 0.12$ and dies at $t = 0.01$.

Let $X_1, \ldots, X_n \overset{i.i.d.}{\sim} P$, where $X_i \in \mathbb{R}^D$. We assume that the support of P is d-dimensional compact manifold $\mathbb{M}$. Define

$$p_h(x) = \int_{\mathbb{M}} \frac{1}{h^D} K\left(\frac{||x-u||}{h}\right) dP(u). \tag{1}$$

Then p_h is the density of the probability measure P_h which is the convolution $P_h = P \star \mathbb{K}_h$ where $\mathbb{K}_h(A) = h^{-D}\mathbb{K}(h^{-1}A)$ and $\mathbb{K}(A) = \int_A K(t)dt$. That is, P_h is a smoothed version of P. The standard estimator for p_h is the kernel density estimator

$$\hat{p}_h(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h^D} K\left(\frac{||x-X_i||}{h}\right). \tag{2}$$

It is easy to see that $\mathbb{E}(\hat{p}_h(x)) = p_h(x)$.

Our target of inference is $\mathcal{P}_h$, the persistence diagram of the upper level sets $\{x : p_h(x) \geq t\}$. We estimate $\mathcal{P}_h$ using the empirical diagram $\hat{\mathcal{P}}_h$ of the upper level sets $\{x : \hat{p}_h(x) \geq t\}$. See Figure 3 for an example.

$\mathcal{P}_h$ is of interest for several reasons. First, the upper level sets of a density are of intrinsic interest in statistics and machine learning. The connected components of the upper level sets are often used for clustering. The homology of these upper level sets provides further structural information about the density. Second, under appropriate conditions, the upper level sets of p_h may carry topological information about a set of interest $\mathbb{M}$. To see this, suppose that p is the density of P with respect to Hausdorff measure on $\mathbb{M}$. If p is smooth and bounded away from 0, then there is an interval $[a, A]$ such that $\{x : \ p(x) \geq t\} \cong \mathbb{M}$ for some $a \leq t \leq A$.

In the language of computational topology, $\mathcal{P}_h$ can be considered a topological simplification of $\mathcal{P}$, the persistence diagram of the upper level sets $\{x : \ p(x) \geq t\}$. $\mathcal{P}_h$ may omit subtle details that are present in $\mathcal{P}$ but is much more stable.

2.1.3 Persistence Diagrams and Stability

We say that the persistence diagram is stable if a small change in the input function produces a small change in the persistence diagram. There are many variants of the stability result for persistence diagrams, as we may define different ways of measuring distance between functions or distance between persistence diagrams. We are interested in using the L_∞-distance between functions and the bottleneck distance between persistence diagrams, that we define through the notion of matching.

A *matching* between $A \subset \mathbb{R}^2$ and $B \subset \mathbb{R}^2$ is a set of edges (a, b), with $a \in A$ and $b \in B$, such that no vertex is incident to two edges. A matching is *perfect* if every vertex is incident on exactly one edge. We want to find a matching between the points of two persistence diagrams $\mathcal{P}_1$ and $\mathcal{P}_2$ that minimizes the cost associated with the matching. Let the L_∞ distance between two points $a, b \in \mathbb{R}$ be

$$d_\infty(a, b) = \max\{|a_x - b_x|, |a_y - b_y|\},$$

where (a_x, a_y) and (b_x, b_y) are the coordinates of a and b. To resolve the issue where the number of off-diagonal points in both diagrams is not equal, we allow an off-diagonal point to be matched to a point on the diagonal $y = x$. Given a matching M, the cost of a matching is

$$C(M) = \max_{(a,b)\in M} d_\infty(a, b)$$

The bottleneck distance between persistence diagrams $\mathcal{P}_1$ and $\mathcal{P}_2$ is

$$W_\infty(\mathcal{P}_1, \mathcal{P}_2) = \min_M C(M)$$

where the minimum is over all the perfect matching between $\mathcal{P}_1$ and $\mathcal{P}_2$. The set $\mathcal{D}$ of persistence diagrams is equipped with the the bottleneck metric W_∞ and its completion $\overline{\mathcal{D}}$ is Polish, i.e. complete and separable, which makes it amenable to probability theory (Blumberg et al., 2012; Mileyko et al., 2011).

We can upper bound the bottleneck distance between two persistence diagrams by the L_∞-distance between the corresponding functions:

Theorem 1 (Bottleneck Stability)**.** *Let $\mathbb{X}$ be finitely triangulable, and let $f, g\colon \mathbb{X} \to \mathbb{R}$ be continuous. Then, the bottleneck distance between the corresponding persistence diagrams is bounded from above by the L_∞-distance between them:*

$$W_\infty(Dgm_f, Dgm_g) \leq ||f - g||_\infty. \tag{3}$$

The bottleneck stability theorem is one of the main requirements for our methods to work, as we will see in Section 2.2. We refer the reader to Cohen-Steiner et al. (2007) and to Chazal et al. (2012) for proofs of this theorem.

2.1.4 Persistence Landscapes

Bubenik (2012) introduced another representation called the persistence landscape, which is in one-to-one correspondence with persistence diagrams. A persistence landscape is a continuous, piecewise linear function $\lambda\colon \mathbb{Z}^+ \times \mathbb{R} \to \mathbb{R}$. The advantage of landscapes and, more generally, of any function-valued summaries of persistent homology is that we can analyze them using existing techniques and theories from nonparametric statistics.

To define the persistence landscape function, we first consider the persistence diagram with a different set of coordinates. Each point, representing a feature born at b and dead at d, is plotted with coordinates $(x, y) = \left(\frac{b+d}{2}, \frac{b-d}{2}\right)$. Then we replace each persistence point $p = (x, y)$ with the triangle function

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

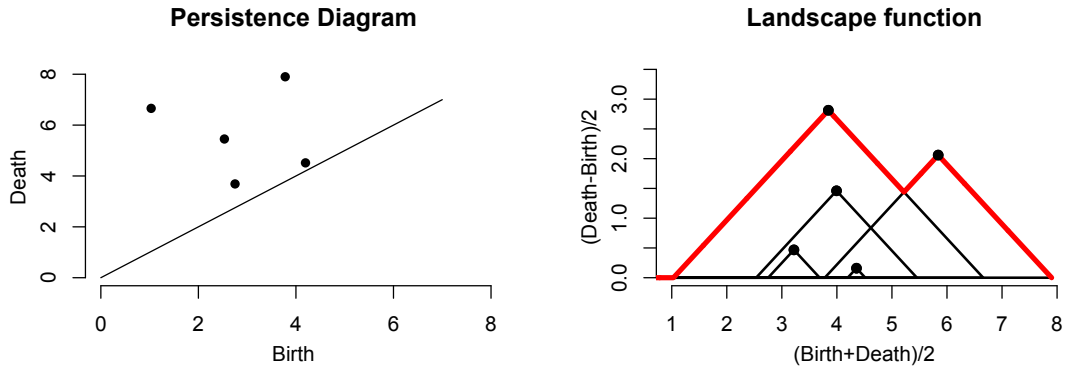

Figure 4: **Left**: a persistence diagrams with coordinates (birth, death). **Right**: the same persistence diagram with coordinates ((birth+death)/2, (death-birth)/2). The red curve is the landscape $\lambda(1, \cdot)$.

Notice that p is itself on the graph of $t_p(z)$. We obtain an arrangement of curves by overlaying the graphs of the functions $\{t_p(z)\}_{p\in\mathcal{P}}$; see Figure 4.

The persistence landscape is defined formally as a walk through this arrangement:

$$\lambda_{\mathcal{P}}(k, z) = \underset{p\in\mathcal{P}}{\text{kmax}}\ t_p(z), \tag{4}$$

where kmax is the kth maximum value in the set; in particular, 1max is the usual maximum function. Observe that $\lambda_{\mathcal{P}}(k, z)$ is 1-Lipschitz.

For a fixed $k \geq 1$ we define $\lambda(\cdot) := \lambda(k, \cdot)$. Let P be a probability distribution on the space of persistence landscapes upper bounded by $T/2$, for some $T > 0$. Let $\lambda_1, \ldots, \lambda_n \sim P$. We define the mean landscape as

$$\mu(t) = \mathbb{E}[\lambda_i(t)], \quad t \in [0, T].$$

The mean landscape is an unknown function that we would like to estimate. We estimate μ with the sample average

$$\overline{\lambda}_n(t) = \frac{1}{n}\sum_{i=1}^{n} \lambda_i(t), \quad t \in [0, T].$$

Note that since $\mathbb{E}(\overline{\lambda}_n(t)) = \mu(t)$, we have that $\overline{\lambda}_n$ is a point-wise unbiased estimator of the unknown function μ. Bubenik (2012) showed that $\overline{\lambda}_n$ converges pointwise to μ and that the pointwise Central Limit Theorem holds.

2.2 Research Plan

Several recent attempts have been made, with different approaches, to study persistence diagrams from a statistical point of view. See for example Turner et al. (2012); Robinson

and Turner (2013); Munch et al. (2013); Chazal et al. (2013c). In the following we describe the first results that we obtained in the study of persistence diagrams and other summary functions, as well as the open questions that we propose to address.

2.2.1 Preliminary Work

In Fasy et al. (2013), Chazal et al. (2013a) and Chazal et al. (2013b), we have taken the first steps towards a rigorous statistical analysis of persistent homology.

In particular, we have derived confidence sets for persistence diagrams that allow us to separate topological signal from topological noise. An asymptotic $1-\alpha$ confidence set for the bottleneck distance $W_\infty(\hat{\mathcal{P}}, \mathcal{P})$ is an interval $[0, c_n]$ such that

$$\liminf_{n\to\infty} \mathbb{P}\Big(W_\infty(\hat{\mathcal{P}}, \mathcal{P}) \in [0, c_n]\Big) \geq 1-\alpha. \tag{5}$$

We can visualize the confidence interval by centering a box of side length $2c_n$ at each point p on the persistence diagram. The point p is considered indistinguishable from noise if the corresponding box, formally defined as $\{q \in \mathbb{R}^2 : \ d_\infty(p,q) \leq c_n\}$, intersects the diagonal. The union of boxes forms the confidence set for the unobserved persistence diagram $\mathcal{P}$. Alternatively, we can visualize the confidence set by adding a band of width $\sqrt{2}c_n$ around the diagonal of the persistence diagram $\hat{\mathcal{P}}$. The interpretation is this: points in the band are not significantly different from noise. Points above the band can be interpreted as representing a significant topological feature. This leads to the diagrams shown in Figure 5.

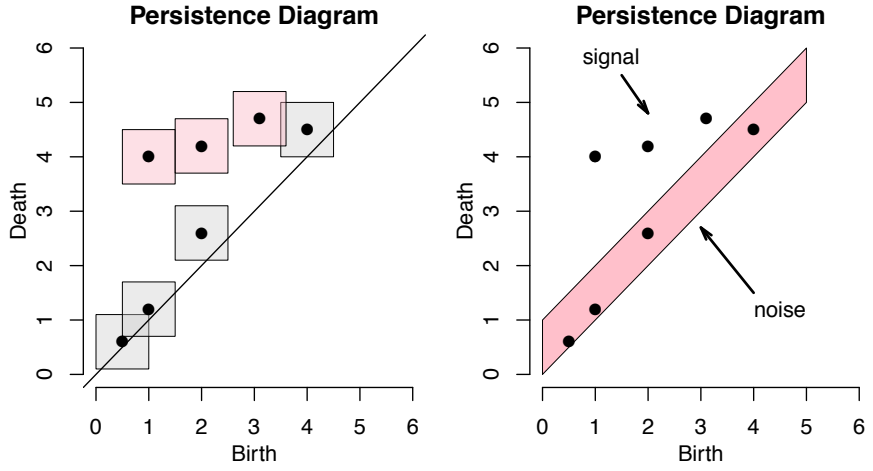

Figure 5: First, we obtain the confidence interval $[0, c_n]$ for $W_\infty(\hat{\mathcal{P}}, \mathcal{P})$. If a box of side length $2c_n$ around a point in the diagram hits the diagonal, we consider that point to be noise. By putting a band of width $\sqrt{2}c_n$ around the diagonal, we need only check which points fall inside the band and outside the band. The plots show the two different ways to represent the confidence interval $[0, c_n]$. For this particular example $c_n = 0.5$.

In Fasy et al. (2013) we proposed several methods for the construction of asymptotic confidence intervals for $W_\infty(\hat{\mathcal{P}}, \mathcal{P})$. The general strategy is as follows.

When f is the distance function (see Section 2.1.1), from the stability theorem (Theorem 1), we know that $W_\infty(\mathcal{P}, \hat{\mathcal{P}}) \leq \|d_{\mathbb{M}} - d_{\mathcal{S}_n}\|_\infty$. Hence, it suffices to find c_n such that

$$\liminf_{n\to\infty} \mathbb{P}\Big(\|d_{\mathbb{M}} - d_{\mathcal{S}_n}\|_\infty \in [0, c_n]\Big) \geq 1 - \alpha \tag{6}$$

to conclude that

$$\liminf_{n\to\infty} \mathbb{P}\Big(W_\infty(\mathcal{P}, \hat{\mathcal{P}}) \in [0, c_n]\Big) \geq 1 - \alpha. \tag{7}$$

Similarly when f is the density function p_h (see Section 2.1.2), it suffices to find c_n such that

$$\liminf_{n\to\infty} \mathbb{P}\Big(\|p_h - \hat{p}_h\|_\infty \in [0, c_n]\Big) \geq 1 - \alpha \tag{8}$$

to conclude that

$$\liminf_{n\to\infty} \mathbb{P}\Big(W_\infty(\mathcal{P}_h, \hat{\mathcal{P}}_h) \in [0, c_n]\Big) \geq 1 - \alpha. \tag{9}$$

In the example of Figure 6 we use the bootstrap technique to construct an asymptotic 95% confidence set for the persistence diagram of the uniform density over the torus. We described this method in details in Chazal et al. (2013a).

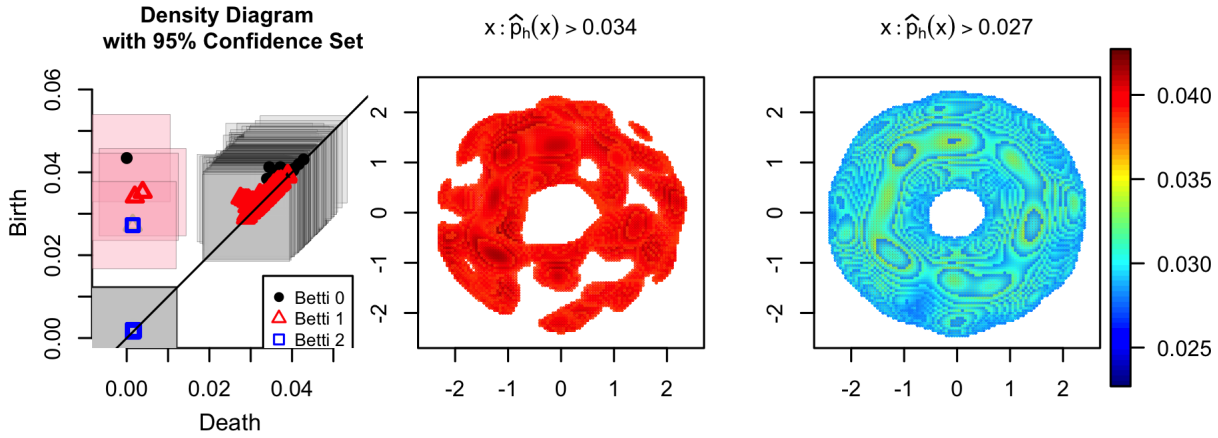

Figure 6: We embed the torus $\mathbb{S}^1 \times \mathbb{S}^1$ in $\mathbb{R}^3$ and we use the rejection sampling algorithm of Diaconis et al. (2012) ($R = 1.5, r = 0.8$) to sample 10,000 points uniformly from the torus. Then, we compute the persistence diagram $\widehat{\mathcal{P}}_h$ using the Gaussian kernel with bandwidth $h = 0.25$ and use the bootstrap to construct the 0.95% confidence interval $[0\,,\,0.01]$ for $W_\infty(\widehat{\mathcal{P}}_h, \mathcal{P}_h)$. Note that the confidence set correctly captures the topology of the torus. That is, only the points representing real features of the torus are significantly far from the diagonal.

In Chazal et al. (2013b) we derived similar results for the landscape function, described in Section 2.1.4). We showed that the average persistence landscape converges weakly to

a Gaussian process and we constructed 95% confidence bands for the average landscape using the multiplier bootstrap. A pair of functions $\ell_n, u_n \colon \mathbb{R} \to \mathbb{R}$ is an asymptotic $(1-\alpha)$ confidence band for μ if, as $n \to \infty$,

$$\mathbb{P}\Big(\ell_n(t) \leq \mu(t) \leq u_n(t) \text{ for all } t\Big) \geq 1-\alpha, \tag{10}$$

Confidence bands are valuable tools for statistical inference, as they allow to quantify and visualize the uncertainty about the mean persistence landscape function μ and to screen out topological noise. See Figure 7 for an example.

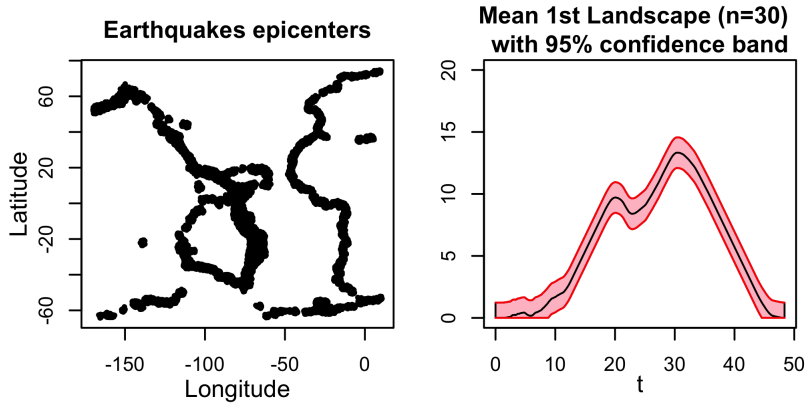

Figure 7: The plot on the left shows 8000 epicenters of earthquakes in the latitude/longitude rectangle $[-75, 75] \times [-170, 10]$ of magnitude greater than 5.0 recorded between 1970 and 2009 (USGS data). We randomly sampled $m = 400$ epicenters and computed the approximated persistence diagram of the distance function (Betti 1). We repeated this procedure $n = 30$ times and computed the empirical average landscape $\overline{\lambda}_n$. Using the multiplier bootstrap described in Chazal et al. (2013b), we obtained a uniform 95% confidence band for the average landscape $\mu(t)$ (right).

2.2.2 Research Aim

We will continue to study the objects and tools of persistent homology from a statistical point of view. Our work will result in various non-parametric inferential procedures based on persistence diagrams and summary functions such as persistence landscapes. These methods will partially solve the important practical issue of approximating the persistent homology in cases where exact computations are prohibitive.

An immediate application of the confidence sets described above will be the formalization of **hypothesis tests** that will be able to discriminate sampling artifacts (the topological noise) from the true topological features. More generally, we will treat diagrams as non-parametric test statistics and, given two separate samples, we will study the power of tests that reject the null hypothesis of population homogeneity solely based on homological features. Preliminary results on hypothesis tests for persistent homology are presented in Bubenik (2012) and Robinson and Turner (2013), although they are mainly based on permutation tests and they do not provide a rigorous statistical analysis of the power of these procedures. We

will consider two non-parametric tests that have shown good performance in preliminary experiments: the Rosenbaum test (Rosenbaum, 2005) and Kernel Tests (Gretton et al., 2012). Tests of this kind will be extremely useful for instance in the medical imaging (Chung et al., 2009; Pachauri et al., 2011) and cosmology (Sousbie, 2011; van de Weygaert et al., 2011; Cisewski et al., 2014).

Much of the literature on computational topology focuses on using the distance function to the data. As discussed in Bendich et al. (2011), such methods are quite sensitive to the presence of outliers. In Fasy et al. (2013) we showed that density-based methods are more robust: we can use the data to construct a smooth density estimator and then find the persistence diagram defined by a filtration of the upper level sets of the density estimator. Another promising idea in this direction is the concept of **distance of a measure to a set** (Chazal et al., 2010). These functions share many properties with classical distance functions, which makes them suitable for inference purposes.

Another summary function for persistence diagrams can be obtained by modelling the diagram as a **point process** on the plane (see e.g. Daley and Vere-Jones, 2002). As described in Edelsbrunner et al. (2012) one can construct the empirical function on the plane $\phi : \mathbb{R}^2 \to \mathbb{R}$, whose integral over every region $A \subset \mathbb{R}^2$ is the expected number of points in A. Alternatively, one can construct an **intensity function**, that is a smooth 2-dimensional density estimation of the process on the plane. See Figure 8. We will derive a rigorous statistical analysis to prove the convergence in distribution of the average intensity function and to measure the significance of topological properties encoded in the corresponding persistence diagram.

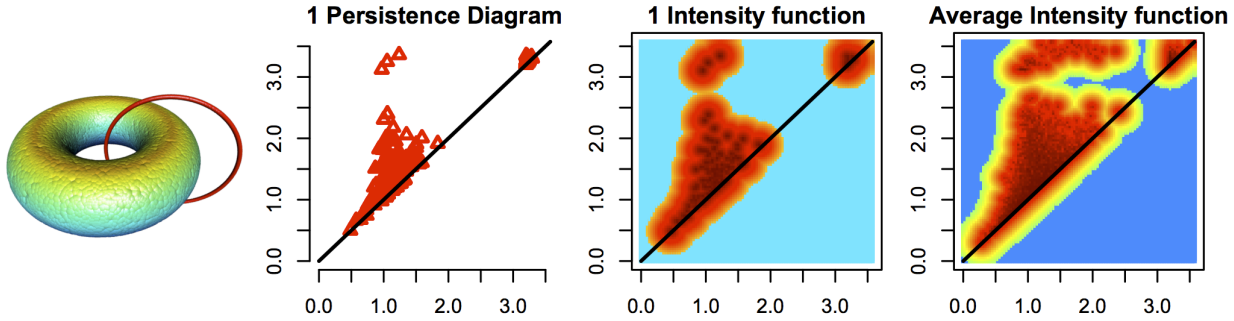

Figure 8: We embed the torus $\mathbb{S}^1 \times \mathbb{S}^1$ in $\mathbb{R}^3$ and we use the rejection sampling algorithm of Diaconis et al. (2012) (R = 5, r = 1.8) to sample 10,000 points uniformly from the torus. Then we link it with a circle of radius 5, from which we sample 1,800 points. These $N = 11,800$ points constitute the sample space (left). We randomly sample $m = 600$ of these points, estimate the corresponding persistence diagram (Betti 1) (middle left) and the corresponding intensity function, using a Gaussian kernel with bandwidth $h = 0.1$ (middle right). We repeat this procedure $n = 30$ times to construct the average intensity function (right).

Finally we will implement all the methods described above in a publicly available R package.

3 Density Clustering

The other set of research objectives of this proposal pertains to the classic data-analytic task of clustering in high dimensions. Suppose we observe a collection of points $\mathcal{S}_n = \{X_1, \ldots, X_n\}$ in $\mathbb{R}^D$. Density clustering allows us to identify and visualize the spatial organization of $\mathcal{S}_n$, without specific knowledge about the data generating mechanism and in particular without any a priori information about the number of clusters.

3.1 Background

3.1.1 Level Set Trees

Let f be the density of the probability distribution P generating the observed sample $\mathcal{S}_n$. For a threshold value $\lambda > 0$, the corresponding (super) level set of f is $L_f(\lambda) := \text{cl}(\{x \in \mathbb{R}^D : f(x) > \lambda\})$, and its D-dimensional subsets are called high-density regions. The λ high-density clusters of P are the maximal connected subsets of $L_f(\lambda)$ (see Figure 9). The statistical literature addressing the problem of estimating the high density regions corresponding to a fixed λ under a variety of metrics is extensive. See, e.g., Cuevas et al. (2001); Azzalini and Torelli (2007); Rigollet and Vert (2009); Rinaldo and Wasserman (2010). In these works, the level set is estimated non-parametrically either by selecting an element from a carefully chosen class of sets of manageable complexity, or, more frequently, as the level set of a non-parametric estimate of f itself. These methods often come with minimax guarantees that hold only under strong regularity conditions on λ, f or P that are typically not verifiable in practice.

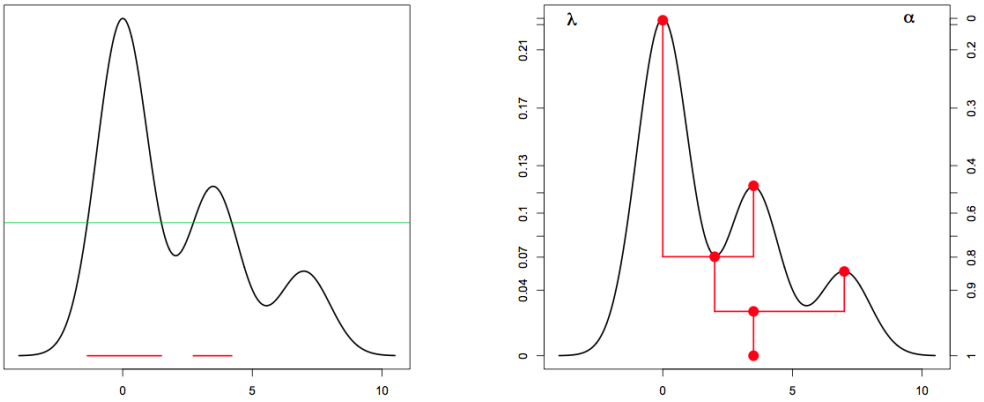

Figure 9: **Left**: a density function and two high-density clusters (red intervals) corresponding to the density level $\lambda = 0.09$ (green line). **Right**: the density tree, indexed by both the density level λ (left vertical axis) and by the probability content α (right vertical axis).

A more fundamental approach to clustering and data visualization is to consider not just one level set of P at a time, but all the level sets simultaneously. This naturally leads to the notion of the cluster density tree of P (see, e.g., Hartigan (1981)), defined as the collection of sets $T := \{L_f(\lambda), \lambda \geq 0\}$, which satisfies the tree property: $A, B \in T$ implies that $A \subset B$

or $B \subset A$ or $A \cap B = \emptyset$. We will refer to this construction as the λ-tree. Alternatively, in Rinaldo et al. (2012) the authors re-parametrize the depth of the tree by a probability content parameter $\alpha \in (0, 1)$, and re-define the density α-tree as $\{L(\alpha), \alpha \in (0, 1)\}$ where $L(\alpha) := L(\lambda_\alpha)$ with $\lambda_\alpha := \sup\{\lambda, P(L_f(\lambda)) \geq \alpha\}$. More recently, Kent et al. (2013) introduced the cluster κ-*tree*, which facilitates the interpretation of the tree by precisely encoding the probability content of each tree branch rather than density level. This new descriptor further improves the interpretability and generality of level set trees.
While the cluster density tree has long been known to be the most informative representation of P for the purposes of clustering, only very recently have statisticians and mathematicians begun to analyze it thoroughly (see, e.g., Stuetzle and Nugent (2010); Carlsson and Mémoli (2010); Chaudhuri and Dasgupta (2010); Kpotufe and von Luxburg (2011); Rinaldo et al. (2012)), and much work remains to be done in order to understand the statistics of density trees.

3.2 Research Plan

The notion of density tree offers a principled way for visualizing a distribution in arbitrary dimensions and is clearly important for clustering. Many results have been published about clustering at a fixed density levels, but the first strong results about the accuracy of estimators for the entire level set tree appeared only recently (see Chaudhuri and Dasgupta (2010); Kpotufe and von Luxburg (2011); Rinaldo et al. (2012)). However a great deal of work remains to quantify the statistical properties of these methods. For example, the consistency results in Chaudhuri and Dasgupta (2010) and Kpotufe and von Luxburg (2011) are based on the λ-tree. Since there is a one-to-one map between the levels of the λ-tree and those of the α-tree, it is likely that the α-tree is also **consistent**. We will also analyze the theoretical properties of the κ-tree, whose statistical consistency has never been studied before.

The cluster density tree can be seen as a topological descriptor. In order to study the stability of this object and to use it for statistical inference it is important to define a **distance between two trees**. In Morozov et al. (2013) the authors define the *interleaving distance*, based on continuous maps between the trees, which leads to a stability results, but whose cost of computation is prohibitive. Several definitions of distance are also presented in Kent (2013). One of the most promising is the *paint mover distance*, which is based on the Wasserstein metric and can be easily computed by solving a linear program. We will analyze the statistical properties of the paint mover distance and we will use it to construct an **average tree** that will lead to a rigorous definition of inferential procedures. A first step in this direction will consist in borrowing the concept of landscape function from persistent homology. The idea is to summarize the information contained in the level set tree using a continuous real valued function and then take advantage of its simplicity for statistical inference. This procedure automatically defines a map from a d-dimensional density to an unnormalized 1-dimensional density that preserves the critical value of the original function. As done for the persistence landscape, we will construct an **average tree landscape** and a confidence band, using appropriate bootstrapping procedure, which

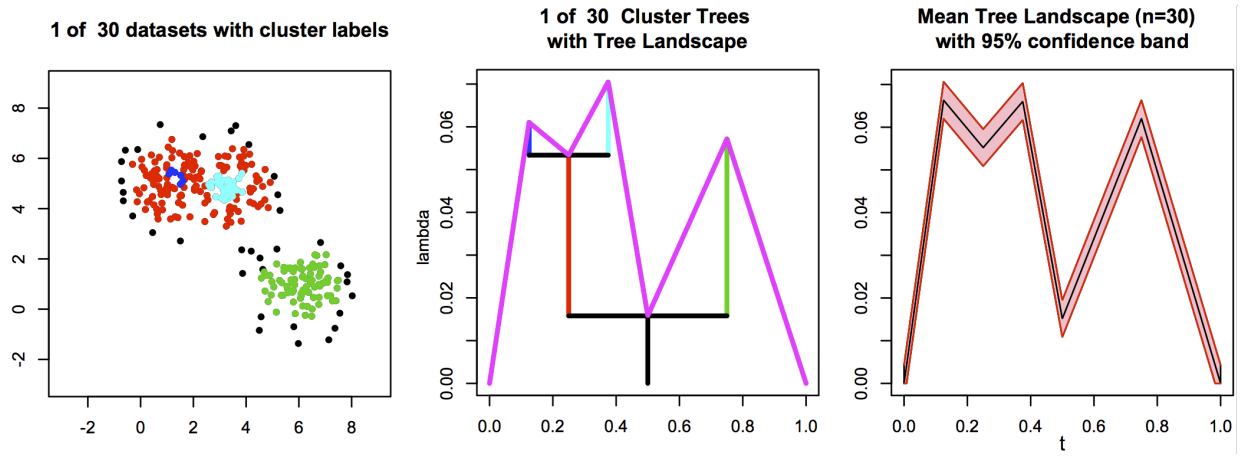

Figure 10: We simulate 30 datasets, each of them formed by 300 points sampled from three 2d-Gaussian distributions. The plot on the left shows one of these dataset. The plot in the middle shows the cluster tree and corresponding tree landscape function (piecewise linear curve). The plot on the right shows the empirical average landscape (black curve) obtained as a pointwise average of the 30 Landscape functions. The pink band is a 95% confidence band for $\mu(t)$, the mean Landscape associated to the sampling scheme. It is obtained using the multiplier bootstrap.

will lead to the formalization of **hypothesis tests**. See Figure 10 for a toy example with preliminary results.

We will also try to extend these statistical techniques to a similar topological descriptor, the **Reeb graph**. Given a continuous function $f : \mathbb{X} \to \mathbb{R}$ defined on a triangulable topological space $\mathbb{X}$ we consider the level set $f^{-1}(t) = \{x \in \mathbb{X} : f(x) = t\}$, for $t \in \mathbb{R}$. Each level set may contain several connected components. We say that two points $x, y \in f^{-1}(t)$ are equivalent, denoted by $x \sim y$, if they are in the same connected component. The Reeb graph of the function f is the quotient space $\mathbb{X}/\sim$, which is the set of equivalence classes equipped with the quotient topology induced by this equivalence relation. Beside the level sets of the function, the Reeb graph provides information on the topological space on which the function is defined. Even though the Reeb graph loses aspects of the original topological structure, it can reflect the 1-dimensional connectivity of the space in some cases (Biasotti et al., 2008; Edelsbrunner and Harer, 2010; Bauer et al., 2013). See Figure 11 for an example.

Finally, a critical factor in the usefulness of a data analysis method is computational speed and memory efficiency. As in persistent homology, a rigorous statistical analysis of level set trees and Reeb graphs will allow us to approximate their construction in cases where exact computations are prohibitive.

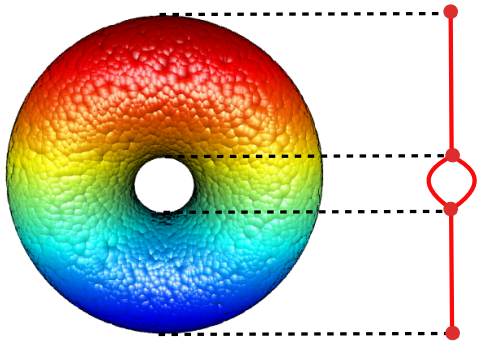

Figure 11: Reeb graph of the height function of the torus.

References

Adelchi Azzalini and Nicola Torelli. Clustering via nonparametric density estimation. *Statistics and Computing*, 17(1):71–80, 2007.

Ulrich Bauer, Xiaoyin Ge, and Yusu Wang. Measuring distance between reeb graphs. *arXiv preprint arXiv:1307.2839*, 2013.

Paul Bendich, Taras Galkovskyi, and John Harer. Improving homology estimates with random walks. *Inverse Problems*, 27(12):124002, 2011.

Silvia Biasotti, Daniela Giorgi, Michela Spagnuolo, and Bianca Falcidieno. Reeb graphs for shape analysis and applications. *Theoretical Computer Science*, 392(1):5–22, 2008.

Andrew J Blumberg, Itamar Gal, Michael A Mandell, and Matthew Pancia. Persistent homology for metric measure spaces, and robust statistics for hypothesis testing and confidence intervals. *arXiv preprint arXiv:1206.4581*, 2012.

Peter Bubenik. Statistical topology using persistence landscapes, 2012. arXiv preprint 1207.6437.

Gunnar Carlsson. Topology and data. *Bulletin of the American Mathematical Society*, 46 (2):255–308, 2009.

Gunnar Carlsson and Facundo Mémoli. Multiparameter hierarchical clustering methods. In *Classification as a Tool for Research*, pages 63–70. Springer, 2010.

Kamalika Chaudhuri and Sanjoy Dasgupta. Rates of convergence for the cluster tree. In *Advances in Neural Information Processing Systems*, pages 343–351, 2010.

Frédéric Chazal, David Cohen-Steiner, Quentin Mérigot, et al. Geometric inference for measures based on distance functions. 2010.

Frédéric Chazal, Vin de Silva, Marc Glisse, and Steve Oudot. The structure and stability of persistence modules. *arXiv preprint arXiv:1207.3674*, 2012.

Frédéric Chazal, Brittany Terese Fasy, Fabrizio Lecci, Alessandro Rinaldo, Aarti Singh, and Larry Wasserman. On the bootstrap for persistence diagrams and landscapes, 2013a. arXiv preprint 1311.0376.

Frédéric Chazal, Brittany Terese Fasy, Fabrizio Lecci, Alessandro Rinaldo, and Larry Wasserman. Stochastic convergence of persistence landscapes and silhouettes, 2013b. arXiv preprint 1312.0308.

Frédéric Chazal, Marc Glisse, Catherine Labruère, and Bertrand Michel. Optimal rates of convergence for persistence diagrams in topological data analysis. *arXiv preprint arXiv:1305.6239*, 2013c.

Moo K Chung, Peter Bubenik, and Peter T Kim. Persistence diagrams of cortical surface data. In *Information Processing in Medical Imaging*, pages 386–397. Springer, 2009.

Jessi Cisewski, Rupert AC Croft, Peter E Freeman, Christopher R Genovese, Nishikanta Khandai, Melih Ozbek, and Larry Wasserman. Nonparametric 3d map of the igm using the lyman-alpha forest. *arXiv preprint arXiv:1401.1867*, 2014.

David Cohen-Steiner, Herbert Edelsbrunner, and John Harer. Stability of persistence diagrams. *Discrete & Computational Geometry*, 37(1):103–120, 2007.

Antonio Cuevas, Manuel Febrero, and Ricardo Fraiman. Cluster analysis: a further approach based on density estimation. *Computational Statistics & Data Analysis*, 36(4):441–459, 2001.

Vin de Silva and Robert Ghrist. Coverage in sensor networks via persistent homology. *Algebraic & Geometric Topology*, 7(339-358):24, 2007.

Persi Diaconis, Susan Holmes, and Mehrdad Shahshahani. Sampling from a manifold. *arXiv preprint arXiv:1206.6913*, 2012.

Herbert Edelsbrunner and John Harer. Persistent homology-a survey. *Contemporary mathematics*, 453:257–282, 2008.

Herbert Edelsbrunner and John Harer. *Computational Topology: An Introduction*. Amer Mathematical Society, 2010.

Herbert Edelsbrunner, A Ivanov, and R Karasev. Current open problems in discrete and computational geometry. 2012.

Brittany Fasy, Fabrizio Lecci, Alessandro Rinaldo, Larry Wasserman, Sivaraman Balakrishnan, and Aarti Singh. Statistical inference for persistent homology, 2013. arXiv preprint 1303.7117.

Jennifer Gamble and Giseon Heo. Exploring uses of persistent homology for statistical analysis of landmark-based shape data. *Journal of Multivariate Analysis*, 101(9):2184–2199, 2010.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13:723–773, 2012.

John A Hartigan. Consistency of single linkage for high-density clusters. *Journal of the American Statistical Association*, 76(374):388–394, 1981.

Peter M Kasson, Afra Zomorodian, Sanghyun Park, Nina Singhal, Leonidas J Guibas, and Vijay S Pande. Persistent voids: a new structural metric for membrane fusion. *Bioinformatics*, 23(14):1753–1759, 2007.

Brian Kent. *Level Set Trees for Applied Statistics*. PhD thesis, Department of Statistics, Carnegie Mellon University, 2013.

Brian P Kent, Alessandro Rinaldo, and Timothy Verstynen. Debacl: A python package for interactive density-based clustering. *arXiv preprint arXiv:1307.8136*, 2013.

S. Kpotufe and U. von Luxburg. Pruning nearest neighbor cluster trees. In *International Conference on Machine Learning (ICML)*, 2011.

Yuriy Mileyko, Sayan Mukherjee, and John Harer. Probability measures on the space of persistence diagrams. *Inverse Problems*, 27(12):124007, 2011.

Dmitriy Morozov, Kenes Beketayev, and Gunther Weber. Interleaving distance between merge trees. In *Workshop on Topological Methods in Data Analysis and Visualization: Theory, Algorithms and Applications (TopoInVis 13)*, 2013.

Elizabeth Munch, Paul Bendich, Katharine Turner, Sayan Mukherjee, Jonathan Mattingly, and John Harer. Probabilistic Fréchet means and statistics on vineyards, 2013. arXiv preprint 1307.6530.

James R Munkres. *Elements of algebraic topology*, volume 2. Addison-Wesley Reading, 1984.

Deepti Pachauri, Chris Hinrichs, Moo K Chung, Sterling C Johnson, and Vikas Singh. Topology-based kernels with application to inference problems in alzheimer's disease. *Medical Imaging, IEEE Transactions on*, 30(10):1760–1770, 2011.

Philippe Rigollet and Régis Vert. Optimal rates for plug-in estimators of density level sets. *Bernoulli*, 15(4):1154–1178, 2009.

Alessandro Rinaldo and Larry Wasserman. Generalized density clustering. *The Annals of Statistics*, 38(5):2678–2722, 2010.

Alessandro Rinaldo, Aarti Singh, Rebecca Nugent, and Larry Wasserman. Stability of density-based clustering. *Journal of Machine Learning Research*, 13:905–948, 2012.

Andrew Robinson and Katharine Turner. Hypothesis testing for topological data analysis. *arXiv preprint arXiv:1310.7467*, 2013.

Paul R Rosenbaum. An exact distribution-free test comparing two multivariate distributions based on adjacency. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(4):515–530, 2005.

Thierry Sousbie. The persistent cosmic web and its filamentary structure–i. theory and implementation. *Monthly Notices of the Royal Astronomical Society*, 414(1):350–383, 2011.

Werner Stuetzle and Rebecca Nugent. A generalized single linkage method for estimating the cluster tree of a density. *Journal of Computational and Graphical Statistics*, 19(2), 2010.

Katharine Turner, Yuriy Mileyko, Sayan Mukherjee, and John Harer. Fréchet means for distributions of persistence diagrams, 2012. arXiv preprint 1206.2790.

Rien van de Weygaert, Gert Vegter, Herbert Edelsbrunner, Bernard JT Jones, Pratyush Pranav, Changbom Park, Wojciech A Hellwing, Bob Eldering, Nico Kruithof, EGP Patrick Bos, et al. Alpha, betti and the megaparsec universe: on the topology of the cosmic web. In *Transactions on Computational Science XIV*, pages 60–101. Springer, 2011.

Afra Zomorodian and Gunnar Carlsson. Computing persistent homology. *Discrete & Computational Geometry*, 33(2):249–274, 2005.

Afra J Zomorodian. *Topology for computing*. Number 16. Cambridge University Press, 2005.