Topological Data Analysis

Larry Wasserman
Department of Statistics/Carnegie Mellon University
Pittsburgh, USA, 15217; email: larry@stat.cmu.edu

Abstract

Topological Data Analysis (TDA) can broadly be described as a collection of data analysis methods that find structure in data. This includes: clustering, manifold estimation, nonlinear dimension reduction, mode estimation, ridge estimation and persistent homology. This paper reviews some of these methods.

Contents

1 INTRODUCTION

Topological Data Analysis (TDA) refers to statistical methods that find structure in data. As the name suggests, these methods make use of topological ideas. Often, the term TDA is used narrowly to describe a particular method called *persistent homology* (discussed in Section 4). In this review, I take a broader perspective: I use the term TDA to refer to a large class of data analysis method that uses notions of shape and connectivity. The advantage of taking this broader definition of TDA is that it provides more context for recently developed methods. The disadvantage is that my review must necessarily be incomplete. In particular, I omit any reference to classical notions of shape such as shape manifolds (Kendall, 1984; Patrangenaru & Ellingson, 2015) and related ideas.

Clustering is the simplest example of TDA. Clustering is a huge topic and I will only discuss *density clustering* since this connects clustering to other methods in TDA. I will also selectively review aspects of manifold estimation (also called "manifold learning"), nonlinear dimension reduction, mode and ridge estimation and persistent homology.

In my view, the main purpose of TDA is to help the data analyst summarize and visualize complex datasets. Whether or not TDA can be used to make scientific discoveries is still unclear. There is another field that deals with the topological and geometric structure of data: computational geometry. The main difference is that in TDA we treat the data as random points whereas in computational geometry the data are usually seen as fixed.

Throughout this paper, we assume that we observe a sample

$$X_1, \ldots, X_n \sim P \tag{1}$$

where the distribution P is supported on some set $\mathcal{X} \subset \mathbb{R}^d$. Some of the technical results cited require either that P have sufficiently thin tails or that $\mathcal{X}$ be compact.

Software: many of the methods in this paper are implemented in the R package `TDA` available at `https://cran.r-project.org/web/packages/TDA/index.html`. A tutorial on the package can be found in Fasy et al. (2014a).

2 DENSITY CLUSTERS

Clustering is perhaps the oldest and simplest version of TDA. The connection between clustering and topology is clearest if we focus on density-based methods for clustering.

2.1 Level Set Clusters

Let $X_1, \ldots, X_n$ be a random sample from a distribution P with density p where $X_i \in \mathcal{X} \subset \mathbb{R}^d$. Density clusters are sets with high density. Hartigan (1975,

1981) formalized this as follows. For any $t \geq 0$ define the *upper level set*

$$L_t = \Big\{x : \ p(x) > t\Big\}. \tag{2}$$

The density clusters at level t, denoted by $\mathcal{C}_t$, are the connected components of L_t. The set of all density clusters is

$$\mathcal{C} = \bigcup_{t \geq 0} \mathcal{C}_t. \tag{3}$$

The leftmost plot in Figure 1 shows a density function. The middle plot shows the level set clusters corresponding to one particular value of t.

The estimated upper level set is

$$\widehat{L}_t = \Big\{x : \ \widehat{p}(x) > t\Big\} \tag{4}$$

where $\widehat{p}$ is any density estimator. A common choice is the kernel density estimator

$$\widehat{p}_h(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h^d} K\left(\frac{||x - X_i||}{h}\right) \tag{5}$$

where $h > 0$ is the bandwidth and K is the kernel. The theoretical properties of the estimator $\widehat{L}_t$ are discussed, for example, in Cadre (2006) and Rinaldo & Wasserman (2010). In particular, Cadre (2006) shows, under regularity conditions and appropriate h, that $\mu(\widehat{L}_t \Delta L_t) = O_P(1/\sqrt{nh^d})$ where μ is Lebesgue measure and $A\Delta B$ is the set difference between two sets A and B.

To find the clusters, we need to get the connected components of $\widehat{L}_t$. Let $I_t = \{i : \ \widehat{p}_h(X_i) > t\}$. Create a graph whose nodes correspond to $(X_i : \ i \in I_t)$. Put an edge between two nodes X_i and X_j if $||X_i - X_j|| \leq \epsilon$ where $\epsilon > 0$ is a tuning parameter. (In practice $\epsilon = 2h$ often seems to work well.) The connected conponenets $\widehat{C}_1, \widehat{C}_2, \ldots$ of the graph estimate the clusters at level t. The number of connected components is denoted by β_0 which is the zeroth-order Betti number. This is discussed in more detail in Section 4.1.

Related to level sets is the concept of excess mass. Given a class of sets $\mathcal{C}$, the *excess mass functional* is defined to be

$$E(t) = \sup\{P(C) - t\mu(C) : \ C \in \mathcal{C}\} \tag{6}$$

and any set $C \in \mathcal{C}$ such that $P(C) - t\mu(C) = E(t)$ is called a generalized t-cluster. If $\mathcal{C}$ is taken to be all measurable sets and the density is bounded and continous, then the upper level set L_t is the unique t-cluster. The excess mass functional is studied in Polonik (1995); Müller & Sawitzki (1991).

One question that arises in the use of level set clustering is: how do we choose t? One possibility is to choose t to cover some prescribed fraction $1 - \beta$ of the total mass; thus we choose t to satisfy $\int_{\widehat{L}_t} \widehat{p}(s)ds = 1 - \beta$. Another idea is to look at clusters at all levels t. This leads us to the idea of density trees.

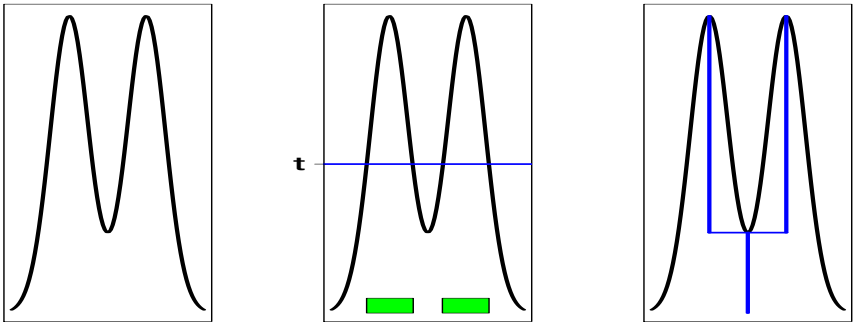

Figure 1: Left: a density function p. Middle: density clusters corresponding to $L_t = \{x : \ p(x) > t\}$. Right: the density tree corresponding to p is shown under the density. The leaves of the tree correspond to modes. The branches correspond to connected components of the level sets.

2.2 Density Trees

The set of all density clusters $\mathcal{C}$ has a tree structure: if $A, B \in \mathcal{C}$ then either $A \subset B$ or $B \subset A$ or $A \bigcap B = \emptyset$. For this reason, we can visually represent a density and its clusters as a tree which we denote by T_p or $T(p)$. Note that T_p is technically a collection of level sets, but it can be represented as a two-dimensional tree as in the right-most plot in Figure 1. The tree, shown under the density function, shows the number of level sets and shows when level sets merge. For example, if we cut across at some level t, then the number of braches of the tree corresponds to the number of connected components of the level set. The leaves of the tree correspond to the modes of the density.

The tree is called a *density tree* or *cluster tree*. This tree provides a convenient, two-dimensional visualization of a density regardless of the dimension d of the space in which the data lie.

Two density trees have the same "shape" if their tree structure is the same. Chen et al. (2016) make this precise as follows. For a given tree T_p define a distance on the tree by

$$d_{T_p}(x, y) = |p(x) + p(y) - 2m_p(x, y)|$$

where

$$m_p(x, y) = \sup\{t : \text{ there exists } C \in \mathcal{C}_t \text{ such that } x, y \in C\}$$

is called the merge height (Eldridge et al., 2015). For any two clusters $C_1, C_2 \in T_p$, we first define $\lambda_1 = \sup\{t : C_1 \in \mathcal{C}_t\}$, and λ_2 analogously. We then define the tree distance function on T_p by

$$d_{T_p}(C_1, C_2) = \lambda_1 + \lambda_2 - 2m_p(C_1, C_2) \tag{7}$$

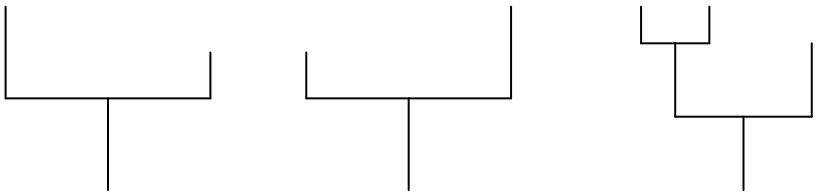

Figure 2: The first and second density trees are homeomorphic; there exists a bicontinous map from one tree to the other. The third tree is not homeomorphic to the other two. Thus the first two trees represent densities with the same shape.

where

$$m_p(C_1, C_2) = \sup\{\lambda \in \mathbb{R} : \text{ there exists } C \in T_p \text{ such that } C_1, C_2 \subset C\}.$$

Now d_{T_p} defines a distance on the tree and it induces a topology on T_p. Given two densities p and q, we say T_p is homeomorphic to T_q, written $T_p \cong T_q$, if there exists a bicontinuous map from T_p to T_q. This means that T_p and T_q have the same shape. In other words, they have the same tree structure. An example is shown in Figure 2.

The density tree can be estimated by plugging in any density estimator. The estimated tree is denoted by $\widehat{T}$ — usually based on a kernel density estimator $\widehat{p}_h$ which provides a nice visualization of the cluster structure of the data. Another choice of estimator is the k-nearest neighbor estimator as in Chaudhuri & Dasgupta (2010).

To estimate the shape of the density tree, it is not necessary to let the bandwidth h go to 0 as n increases. Let $p_h(x) = \mathbb{E}[\widehat{p}_h(x)]$ be the mean of the estimator. It can be shown that, under weak conditions, there exists $h_0 > 0$ such that, for all $0 < h < h_0$, $T(p_h) \cong T(p)$. This means that it suffices to estimate T_{p_h} for any small $h > 0$. It is not necessary to let $h \to 0$. This has important pratical implications since T_{p_h} can be estimated at the rate $O_P(n^{-1/2})$ independent of the dimensions d. Compare this to estimating p in the L_2 loss; the best rate under standard smoothness conditions is $O_P(n^{-2/(4+d)})$ which is slow for large dimensions d. The key point is: estimating the cluster structure is easier than estimating the density itself. In other words, you can estimate p poorly but still get the shape of the tree correct. See Chen et al. (2016) for more details.

The bootstrap can be used to get confidence sets for the density tree (Chen et al., 2016). Let P_n be the empirical measure that puts mass $1/n$ at each data point. Draw an iid sample $X_1^*, \ldots, X_n^* \sim P_n$ and compute the density estimator $\widehat{p}_h^*$. Repeat this process B times to get density estimates $\widehat{p}_h^{*(1)}, \ldots, \widehat{p}_h^{*(B)}$ and

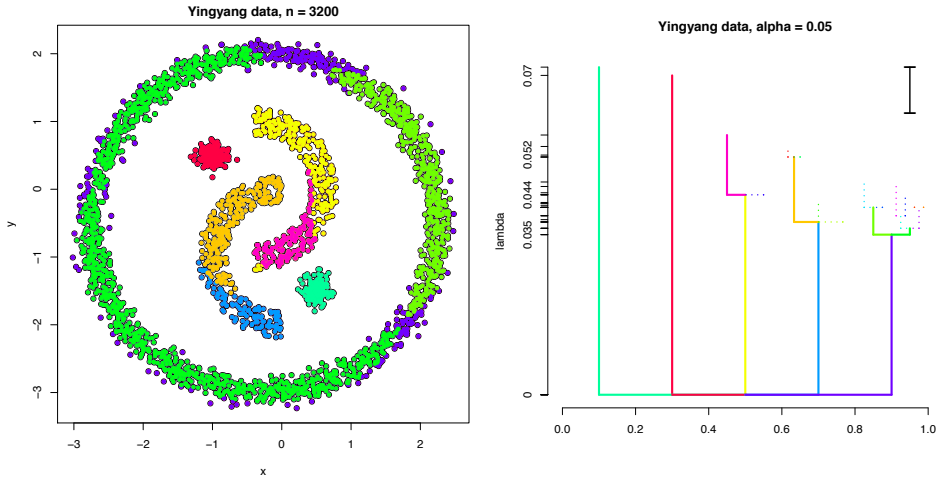

Figure 3: Example from Chen et al. (2016). Left: the data. Right: the tree. The solid lines are the pruned trees; The dashed lines are leaves and branches that have been pruned away because they are smaller than the bootstrap significance level $2\widehat{t}_\alpha$ (indicated in the top right corner).

define

$$\widehat{F}_n(t) = \frac{1}{B}\sum_{j=1}^{B} I(\sqrt{n}||\widehat{p}_h^{*(j)} - \widehat{p}_h||_\infty > t)$$

where I is the indicator function. For large B, $\widehat{F}_n$ approximates

$$F_n(t) = P(\sqrt{n}||\widehat{p}_h - p_h||_\infty > t).$$

Let $\widehat{t}_\alpha = \widehat{F}_n^{-1}(1-\alpha)$ which approximates $t_\alpha = F_n^{-1}(1-\alpha)$. Then

$$\lim_{n\to\infty} P(T(p_h) \in \mathcal{T}) = 1 - \alpha$$

where

$$\mathcal{T} = \left\{ T(p) : \ ||p - \widehat{p}_h||_\infty \leq \frac{\widehat{t}_\alpha}{\sqrt{n}} \right\}.$$

Thus, $\mathcal{T}$ is an asymptotic confidence set for the tree. The critical value $\widehat{t}_\alpha$ can be used to prune non-significant leaves and branches from $\widehat{T}$; see Figure 3.

A density tree is *Hartigan consistent* if, with probability tending to 1, the correct cluster strucrure is recovered. Generally, density trees based on consistent density estimators will be Hartigan consistent. For more on Hartigan consistency, see Chaudhuri & Dasgupta (2010); Eldridge et al. (2015); Balakrishnan et al. (2013).

2.3 Mode Clustering and Morse Theory

Another density clustering method is *mode clustering* (Chacón et al., 2015, 2013; Chacón, 2012; Li et al., 2007; Comaniciu & Meer, 2002; Arias-Castro et al., 2015; Cheng, 1995). The idea is to find modes of the density and then define clusters as the basins of attraction of the modes. A point m is a (local) mode if there exists an open neighborhood N of x such that $p(x) > p(y)$ for every $y \in N$ such that $y \neq x$. Suppose that p has k local modes $\mathcal{M} = \{m_1, \ldots, m_k\}$. Assume that p has gradient g and Hessian H.

A point x is a *critical point* if $g(x) = (0, \ldots, 0)^T$. The function p is a *Morse function* if the Hessian is non-degenerate at each critical point (Milnor, 2016). We will assume that p is Morse. In this case, m is a local mode if and only if $g(m) = (0, \ldots, 0)^T$ and $\lambda_1(H(m)) < 0$ where $\lambda_1(A)$ denotes the largest eigenvalue of the matrix A.

Now let x be an arbitrary point. If we follow the steepest ascent path starting at x, we will eventually end up at one of the modes.[1] Thus, each point x in the sample space is assigned to a mode m_j. We say that m_j is the *destination* of x which is written

$$\text{dest}(x) = m_j.$$

The path $\pi_x : \mathbb{R} \to \mathbb{R}^d$ that leads from x to a mode is defined by the differential equation

$$\pi_x'(t) = \nabla p(\pi_x(t)), \quad \pi_x(0) = x.$$

The set of points assigned to mode m_j is called the *basin of attraction* of m_j and is denoted by C_j. The sets $C_1, \ldots, C_k$ are the population clusters. The left plot in Figure 4 shows a bivariate density with four modes. The right plot shows the partition induced by the modes.

To estimate the clusters, we find the modes $\widehat{\mathcal{M}} = \{\widehat{m}_1, \ldots, \widehat{m}_r\}$ of the density estimate. A simple algorithm called the *mean shift algorithm* (Cheng, 1995; Comaniciu & Meer, 2002) can be used to find the modes and to find the destiation of a any point x. For any given x, we define the iteration

$$x^{(j+1)} = \frac{\sum_i X_i K\left(\frac{||x^{(j)} - X_i||}{h}\right)}{\sum_i K\left(\frac{||x^{(j)} - X_i||}{h}\right)}.$$

See Figure 5. The convergence of this algorithm is studied in Arias-Castro et al. (2015).

It can be shown under suitable regularity conditions that the modes of the kernel density estimate are consistent estimates modes of the true density; see Genovese et al. (2016). Once again, however, it is not necessary to estimate the density well to estimate the mode clusters well. Specifically, define

$$c(x, y) = \begin{cases} 1 & \text{if } \text{dest}(x) = \text{dest}(y) \\ 0 & \text{if } \text{dest}(x) \neq \text{dest}(y). \end{cases}$$

[1] This is true for all x except on a set of Lebesgue measure 0.

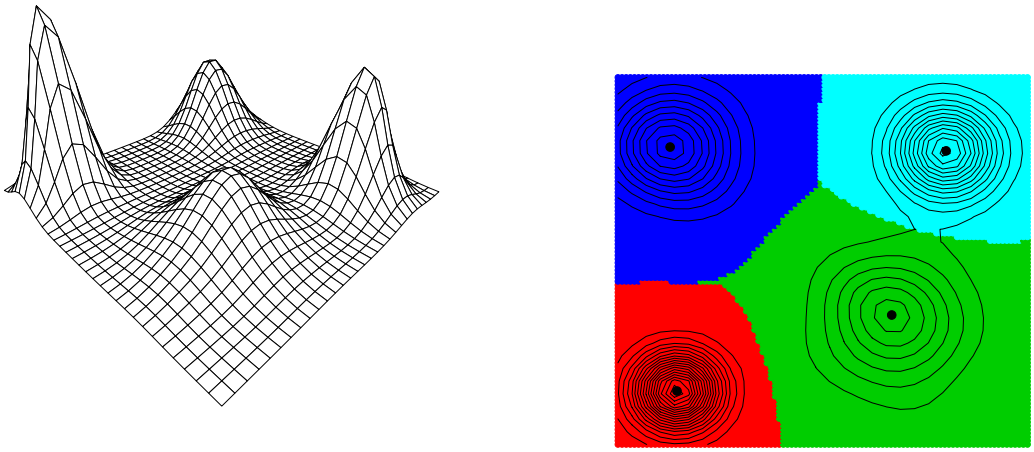

Figure 4: Left: a density with four modes. Right: the partition (basins of attraction) of the space induced by the modes. These are the population clusters.

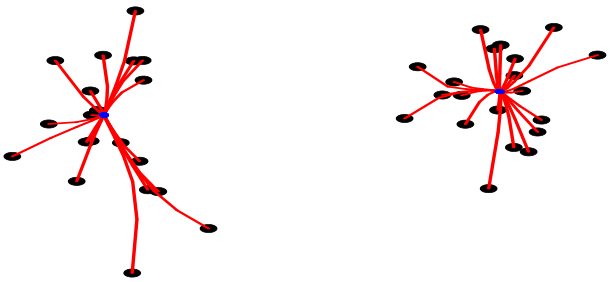

Figure 5: The mean shift algorithm. The data are represented by the black dots. The modes of the density estimate are the two blue dots. The red curves show the mean shift paths; each data point moves along its path towards a mode as we iterate the algorithm.

Thus, $c(x, y) = 1$ if x and y are in the same cluster. Similarly, the estimated clusters define a function $\widehat{c}$. Let $C_1, \ldots, C_k$ be the model clusters. Let $t_1, \ldots, t_k$ be constants and let $C_j(t_j) = \{x \in C_j : \ p(x) > t_j\}$. The sets $C_1(t_1), \ldots, C(t_k)$ are called *cluster cores*. These are the high density points within the clusters. Let Core $= \{X_i : \ X_i \in \bigcup_j C_j(t_j)\}$ be the data points in the cluster cores. Azizyan et al. (2015) show that, if $t_1, \ldots, t_k$ are sufficiently large, then

$$\mathbb{P}(\widehat{c}(X_j, X_k) \neq c(X_j, X_k) \text{ for any } X_j, X_k \in \text{Core}) \leq e^{-nb}$$

for some $b > 0$, independent of the dimension. This means that high density points can be accurately clustered using mode clustering.

3 LOW DIMENSIONAL SUBSETS

Sometimes the distribution P is supported on a set S of dimension r with $r < d$. (Recall that X_i has dimension d.) The set S might be of scientific interest and it is also useful for dimension reduction. Sometimes the support of P is d-dimensional but we are interested in finding a set S of dimension $r < d$ which has a high concentration of mass.

Figure 6 shows an example known as the Swiss-roll dataset. Here, the ambient dimension is $d = 3$ but the support of the distribution S is a manifold of instrinsic dimension $r = 2$. Figure 7 shows a more complex example. Here, $d = 2$ but clearly there is a $r = 1$ intrinsic dimensional subset S with a high concentration of data. (This dataset mimics what we often see in some datasets from astrophysics.) The set S is quite complex and is not a smooth manifold. The red lines show an estimate of S based on the techniques described in Section 3.3.

3.1 Manifolds

In the simplest case, the set S is a smooth, compact submanifold of dimension r. The term *manifold learning* can refer either to methods for estimating the set S or to dimension reduction methods that assume that the data are on (or near) a manifold. Principal component analysis can be thought of as a special case of manifold learning in which the data are assumed to lie near an affice subspace.

As a motivating example, consider images of a person's face as the person moves their head. Each image can be regareded as a high-dimensional vector. For example, a 16 by 16 image is a vector in $\mathbb{R}^d$ where $d = 16 \times 16 = 256$. However, the set of images will not fill up $\mathbb{R}^{256}$. As the person moves their head, these vectors are likely to trace out a surface of dimension $r = 3$, corresponding to the three degrees of freedom corresponding to the motion of the head.

Estimating S. An estimator of S is $\widehat{S} = \bigcup_{i=1}^n B(X_i, \epsilon_n)$ which was suggested (in a different context) by Devroye & Wise (1980). The estimator $\widehat{S}$ is

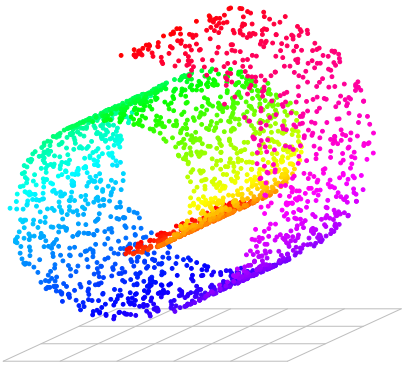

Figure 6: The swissroll dataset. The ambient dimension is $d = 3$ but the data are supported on a set S of dimension $r = 2$.

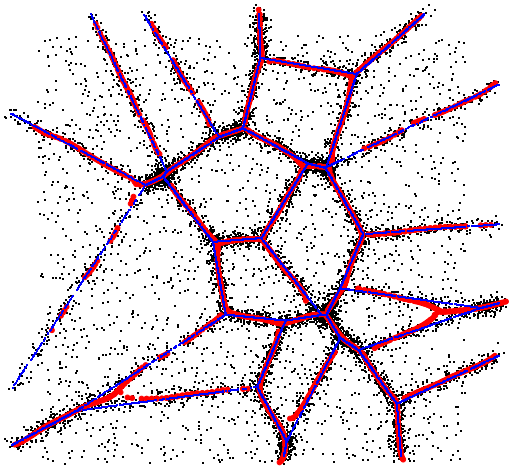

Figure 7: These data are two-dimensional but there is a set S of dimension $r = 1$ with a high concentration of data. The red lines show an estmate of S using the methods in Section 3.3.

d-dimensional but it does converge to S in the following sense (Cuevas, 2009; Fasy et al., 2014b; Niyogi et al., 2008; Cuevas et al., 2001; Chazal et al., 2014b). The *Hausdorff distance* $H(A, B)$ between two sets A and B is

$$H(A, B) = \inf\{\epsilon : \ A \subset B \oplus \epsilon \ \text{ and } \ B \subset A \oplus \epsilon\} \tag{8}$$

where

$$A \oplus \epsilon = \bigcup_{x \in A} B(x, \epsilon)$$

and $B(x, \epsilon)$ denotes a ball of radius ϵ centered at x. Suppose there exists $c > 0$ such that, for every $x \in S$ and every small ϵ, $P(B(x, \epsilon)) \geq c\epsilon^r$. Further, assume that the number of balls of size ϵ required to cover S is $C(1/\epsilon)^r$. These assumption mean that S is r-dimensional (and not too curved) and that P spreads its mass over all of S. Then

$$P(H(\widehat{S}, S) > \epsilon) \leq Cr^{-d} e^{-nc\epsilon^d}.$$

Hence, if we choose $\epsilon_n \asymp (\log n/n)^{1/r}$ then

$$H(\widehat{S}, S) = O_P\left(\frac{\log n}{n}\right)^{1/r}$$

where we recall that H is the Hausdorff distance defined in equation (8). However, better rates are possible under some conditions. The difficulty of estimating S as defined by minimax theory is given under various sets of assumptions, in Genovese et al. (2012b,a).

It is unlikey that a sample will fall precisely on a submanifold S. A more realistic model is that we observe $Y_1, \ldots, Y_n$ where $Y_i = X_i + \epsilon_i$ where $X_1, \ldots, X_n \sim G$ is a sample from a distribution G supported on S and $\epsilon_1, \ldots, \epsilon_n$ are a sample from a noise distribution such as a Gaussian. In this case, Genovese et al. (2012a) showed that estimating S is hopeless; the minimax rate of convergence is logarithmic. However, it is possible to estimate an r-dimensional, high density region R that is close to S. The set R corresponds to a ridge in the density of Y; see Section 3.3.

Estimating the Topology of a Manifold. Another problem is to find an estimate $\widehat{S}$ of S that is topologically similar to S. If, for example, S is a three dimensional image, such as in Figure 23, then requring $\widehat{S}$ to be topologically similar ensures that $\widehat{S}$ "looks like" S in some sense. But what does "topologically similar" mean?

Two sets S and T (equipped with topologies) are *homeomorphic* if there exists a bi-continous map from S to T. Markov (1958) proved that, in general, the question of whether two spaces are homeomorphic is undecidable for dimension greater than 4.

Fortunately, it is possible to determine if two spaces are *homologically equivalent*. Homology is way of defining topological features algebraically using group

theory. The zero-th order homology of a set corresponds to its connected components. The first order homology corresponds to one-dimensional holes (like a donut). The second order homology corresponds to two-dimensional holes (like a soccer ball). And so on. If two sets are homeomorphic then they are homologically equivalent. However, the reverse is not true. This, homological equivalence is weaker than topological equivalence.

We'll discuss homology in more detail in Section 4.1. But here, we mention one of the first results about topology and statistics due to Niyogi et al. (2008). They showed that

$$\widehat{S} = \bigcup_{i=1}^{n} B(X_i, \epsilon)$$

has the same homology as S with high probability, as long as S has positive reach and ϵ is small relative to the reach. The *reach* of S is the largest real number r such that any point x that is a distance less than r from S, has a unique projection on S. The result assumes the data are sampled from a distribution supported on the submanifold S. Extensions that allow for noise are given in Niyogi et al. (2011). An unsolved problem is to find a data-driven method for choosing the tuning parameter ϵ. The assumption that S has positive reach can be weakened: Chazal et al. (2009) define a quantity called that μ-reach which is weaker than reach and they show that topological reconstructions are possible using this weaker regularity assumption.

Dimension Reduction. There are many methods that leverage the fact the the data are supported on a low dimensional set S without explicitly producing an estimate $\widehat{S}$ that is close to S in Hausdorff distance. Examples include: Isomap (Tenenbaum et al., 2000; De'ath, 1999), Local Linear Embedding (Roweis & Saul, 2000), diffusion maps (Coifman & Lafon, 2006), Laplacian eigenmaps (Belkin & Niyogi, 2001) and many others (Lee & Verleysen, 2007). Here, I will give a very brief description of Isomap.

The first step in Isomap is to form a graph from the data. For example, we connect two points X_i and X_j if $||X_i - X_j|| \leq \epsilon$ where ϵ is a tuning parameter. Next we define the distance between two points as the shortest path between the two points among all paths in the graph that connect them. We now have a distance matrix D where D_{ij} is the shortest path between X_i and X_j. The hope is that D_{ij} approximates the geodesic distance between X_i and X_j on the manifold. Finally, we use a standard dimension reduction method such as multidimensional scaling to embed the data in $\mathbb{R}^r$ while trying to preserve the distances D_{ij} as closely as possible. For example, we find a map ϕ to minimize the distorting $\sum_{i<j}[D_{i,j}^2 - ||\phi(X_i) - \phi(X_j)||^2]$. The transformed data $Z_i = \phi(X_i)$ now live in a lower dimensional space. Thus we have used the fact that the data live on a manifold, to perform a dimension reduction.

Figure 8 shows the result of applying isomap to the swissroll data using $\epsilon = 5$. In this case we perfectly recover the underlying structure. However, isomap is a fragile procedure. It is very sensitive to outliers and the choice of

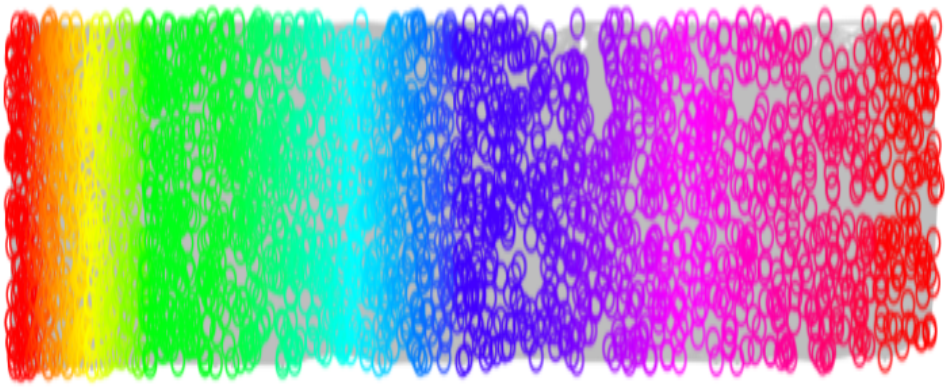

Figure 8: After applying isomap to the swissroll dataset with $\epsilon = 5$ we recover the undelrying two-dimensional structure.

tuning parameters. Other methods, such as diffusion maps and ridge estimation, are more robust.

3.2 Estimating Intrinsic Dimension

Many manifold estimation methods assume that the intrinsic dimension r of the manifold is known. In practice, we need to estimate the dimension. There is a large literature on this problem. Some examples include Little et al. (2011); Lombardi et al. (2011); Hein & Audibert (2005); Levina & Bickel (2004); Kégl (2002); Costa & Hero (2004). Minimax theory for dimension estimation is contained in Koltchinskii (2000) and Kim et al. (2016). Estimating the instrinsic dimension when the data are only approximately supported on a lower dimensional set is much harder than the case where ther support is precisely a lower dimensional set.

3.3 Ridges

Most manifold learning methods assume that the distribution P is supported on some manifold S. This is a very strong and unrealistic assumption. A weaker assumption is that there may exist some low dimensional sets where the density p has a relatively high local concentration. One way to make this more precise is through the idea of density ridges.

A *density ridge* is a low dimensional set with large density. But the distribution P may not even have a density. To deal with this issue, we define the smoothed distribution P_h obtained by convolving P with a Gaussian. Specifically, P_h is the distribution with density

$$p_h(x) = \int K_h(x-u)\, dP(u)$$

where $K_h(x) = h^{-d}(2\pi)^{-d/2} e^{-||x||^2/(2h^2)}$. Note that p_h is the mean of the kernel density estimator with bandwidth h. The smoothed distribution P_h always has a density, even if P does not. In topological inference, we imagine using a small but positive h. It is not necessary to let h tend to 0 as we usual do in density estimation. The salient topological features of P will be preserved by P_h.

Let g_h be the gradient of p_h and let H_h be the Hessian. Recall that a mode of p_h is a point x with $g_h(x) = (0,\ldots,0)^T$ and $\lambda_1(H_h(x)) < 0$. A mode is a 0-dimensonal ridge. More generally, an r-dimensional ridge is a set with sharp density in some directions, much like the ridge of a mountain. see Figure 9. In fact, there are many ways to define a ridge; see Eberly (1996). We use the following defintion. At a point x we will define a local tangent space of dimension r and local normal space of dimension $d-r$. Then x is a ridge point if it is a local mode in the direction of the normal. More precisely, let $\lambda_1(x) \geq \cdots \lambda_d(x)$ be the eigenvalues of the Hessian $H(x)$ and let $v_1(x),\ldots,v_d(x)$ be the correspdonding eigenvectors. Let $V(x) = [v_{r+1}(x)\cdots v_d(x)]$ and define the projected gradient of p by

$$G(x) = V(x)V(x)^T g(x).$$

The r-ridge is

$$R_r(p) = \{x:\ G(x) = 0,\ \lambda_{r+1}(x) < 0\}.$$

Under suitable regularity conditions, this is indeed an r-dimensional set.

The ridge can be estimated by the ridge of a kernel density estimate. Specifically, we take $\widehat{R} = R_r(\widehat{p}_h)$ to be the ridge of the kernel estimator. The properties of this estimator are studied in Genovese et al. (2014) and Chen et al. (2015b). An algorithm for finding the ridge set of $\widehat{p}_h$ was given by Ozertem & Erdogmus (2011) and is called the SCMS (subspace constrained mean shift algorithm). Examples are shown in Figure 7 and Figure 10. A further example is in Section 7.

Ridges can be related to manifolds as follows (Genovese et al. (2014)). Suppose we observe $Y_1,\ldots,Y_n$ where $Y_i = X_i + \sigma\epsilon_i$, $X_1,\ldots,X_n \sim G$ is a sample from a distribution G supported on a manifold S and $\epsilon_1,\ldots,\epsilon_n$ are a sample from a noise distribution such as a Gaussian. As mentioned earlier, S can only be estimated at a logarithmic rate. However, if σ is small enough and S has positive reach, then the density p of Y will have a well defined ridge R such that $H(R,S) = O(\sigma)$. Furthermore, R is "topologically similar" to S in a certain sense described in Genovese et al. (2014). In fact, p_h will have a ridge R_h such that $H(R_h,S) = O(\sigma + h)$ and R_h can be estimated at rate $O_P(\sqrt{\log n/n})$ independently of the dimension.

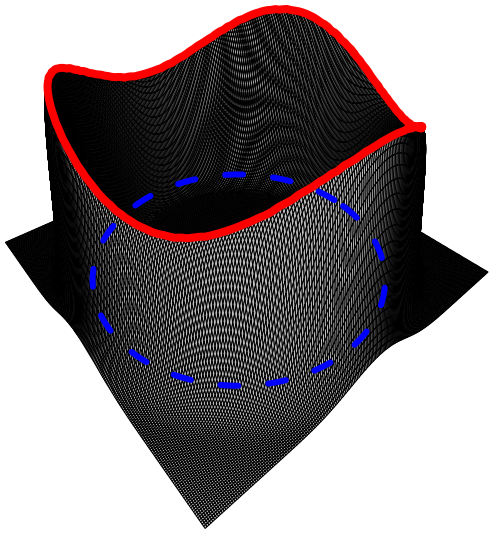

Figure 9: This is a plot of a two-dimensional density function with a clearly defined one-dimensional ridge. The ridge is the blue circle.

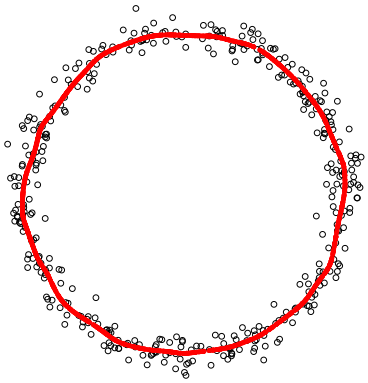

Figure 10: The data are generated as $Y_i = X_i + \epsilon_i$ where the X_i are sampled from a circle and ϵ_i are bivariate Gaussian. The ridge $\widehat{R}$ of the kernel density estimator is found using the SCMS algorithm and is shown in red.

An example is shown in Figure 10. The data are generated as follows. We sample $X_1, \ldots, X_n$ uniform form a circle. Then we set $Y_i = X_i + \epsilon_i$ where $\epsilon_1, \ldots, \epsilon_n$ are draws from a bivariate Normal with mean $(0, 0)$. Next we find the kernel density estimator based on $Y_1, \ldots, Y_n$ and we find the ridge $\widehat{R}$ of the kernel estimator using the SCMS algorithm. The data are the black points in the plot. The estimated ridge is shown in red. Notice that the data are full dimensional but the estimated ridge is one dimensional.

3.4 Stratified Spaces

Another generalization of manifold learning is to assume that the support of P is a *stratified space* which means that the space can be decomposed into several, intersecting submanifolds. Estimation of stratified spaces is much less developed than manifold estimation. Some examples include Bendich et al. (2007); Skraba & Wang (2014) and Bendich et al. (2007). Ridge based methods as discussed in Section 3.3 seem to work well in this case but, so far, this has not been established theoretically. A promising new approach due to Arias-Castro et al. (2011) is based on a version of local PCA.

4 PERSISTENT HOMOLOGY

Persistent homology is a multiscale approach to quantifying topological features in data (Edelsbrunner & Harer, 2010; Edelsbrunner et al., 2002; Edelsbrunner & Harer, 2008). This is the branch of TDA that gets the most attention and some researchers view TDA and persistent homology as synonymous.

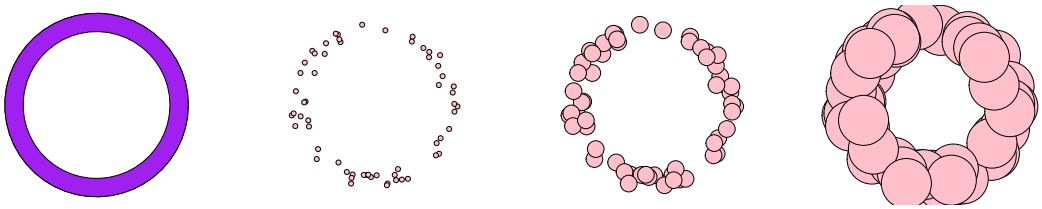

Figure 11: Plot 1: the support S of the distribution. Plots 2-4: Union of balls $\bigcup_{i=1}^n B(X_i, \epsilon)$ around 60 data points drawn from a uniform on S, with $\epsilon = 0.03, 0.10, 0.30$.

A quick, intuitive idea of persistent holomogy is given in Figures 11 and 12. Here, we see some data and we also see the set $\bigcup_{i=1}^n B(X_i, \epsilon)$ for various values of ϵ. The key observation is the topological features appear and disappear as ϵ increases. For example, when $\epsilon = 0$ there are n connected components. As ϵ increases some of the connected components die (that is, they merge) until only one connected component remains. Similarly, at a certain value of ϵ, a hole is born. The hole dies at a larger value of ϵ.

Thus, each feature has a birth time and a death time. The left plot in Figure 12 is a *barcode plot* which represents the birth time and death time of each feature as a bar. The right plot is a *persistence diagram* where each feature is a point on the diagram and the coordinates of the points are the birth time and death time. Features with a long lifetime correspond to points far from the diagonal. With this simple example in mind, we delve into more detail.

4.1 Homology

It is not possible to give a thorough review of homology given the present space constraints. But we can give a short, intuitive description which will suffice for what follows. More details are in the appendix and in Fasy et al. (2014b). Good introductions can be found in Hatcher (2000) and Edelsbrunner & Harer (2010).

Homology characterizes sets based on connected components and holes. Consider the set on the left in Figure 13. The set has one connected component and two holes. We write $\beta_0 = 1$ and $\beta_1 = 2$. The numbers $\beta_0, \beta_1, \ldots$ are called *Betti numbers*. Intuitively, β_0 is the number of connected components, β_1 is the number of one-dimensional holes, β_2 is the number of two-dimensional holes, etc. (More formally, β_j is the rank of the j^{th} homology group.) The set on the right in Figure 13 has two connected components and one hole, thus, $\beta_0 = 2$ and $\beta_1 = 1$. These holes are one-dimensional: they can be surrounded by a loop (like a piece of string). The inside of a soccer ball is a two dimensional hole. To surround it, we need a surface. For a soccer ball, $\beta_0 = 1, \beta_1 = 0, \beta_2 = 1$. For a torus (a hollowed out donut), $\beta_0 = 1, \beta_1 = 2, \beta_2 = 1$.

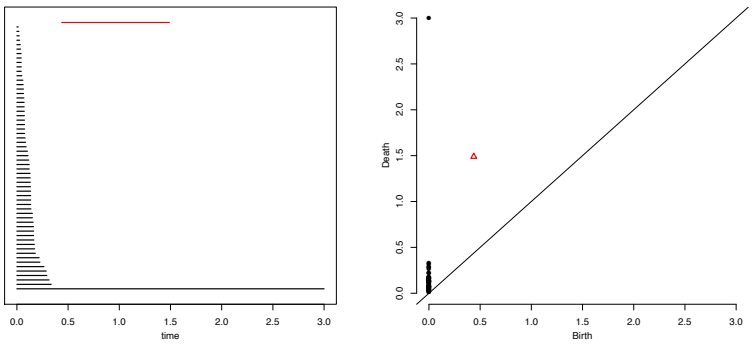

Figure 12: Left: the barcode plot corresponding to the data from Figure 11. The black lines show the birth and death of each connected component as ϵ increases. The red line shows the birth and death of the hole as ϵ increases. Right: the persistence diagram. In this case, the birth and death time of each feature is represented by a point on the diagram. The black points correspond to connected components. The red triangle corresponds to the hole. Points close to the diagonal have a short lifetime.

The formal definition of homology uses the language of group theory. (The equivalence class of loops surrounding a hole have a group structure.) The details are not needed to understand the rest of this paper. Persistent homology examines these homological features from a multiscale perspective.

4.2 Distance Functions and Persistent Homology

A good starting point for explaining persistent homology is the *distance function*. Given a set S, the distance function is defined to be

$$d_S(x) = \inf_{y \in S} ||x - y||.$$

The lower level sets of the distance function are

$$L_\epsilon = \{x : \ d_S(x) \leq \epsilon\}.$$

We also have that

$$L_\epsilon = \bigcup_{x \in S} B(x, \epsilon).$$

So L_ϵ can be thought of either as a union of balls, or as the lower level set of the distance function. As ϵ increases, the sets L_ϵ evolve. Topological features — connected components and holes — will appear and disappear. Consider the circle

$$S = \{(x, y) : \ x^2 + y^2 = 1\}.$$

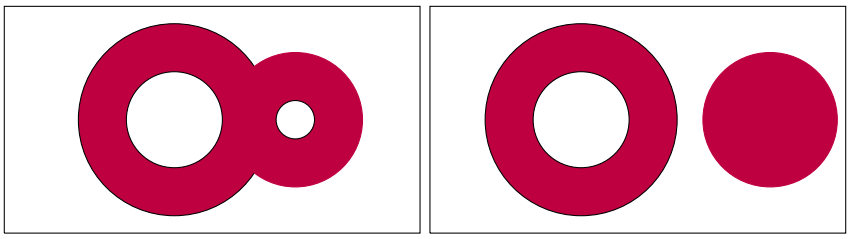

Figure 13: The set on the left has one connected component and two holes and hence $\beta_0 = 1$ and $\beta_1 = 2$. The set on the right has two connected components and one hole and hence $\beta_0 = 2$ and $\beta_1 = 1$.

The set L_ϵ is an annulus of radius ϵ. For all values of ϵ, L_ϵ has one connected component. For $0 \leq \epsilon < 1$, the set L_ϵ has one hole. The hole dies at $\epsilon = 1$. Thus, the hole has birthtime $\epsilon = 0$ and deathtime $\epsilon = 1$. In general, these features can be represented as a persistence diagram D as in Figure 12. The diagram D represents the persistent homology of S.

Technically, the persistence diagram D is a multiset consisting of all pairs of points on the plot as well as all points on the diagonal. Given two diagrams D_1 and D_2, the *bottleneck distance* defined by

$$\delta_\infty(D_1, D_2) = \inf_\gamma \sup_{z \in D_1} ||z - \gamma(z)||_\infty \tag{9}$$

where γ ranges over all bijections between D_1 and D_2. Intuitively, this is like overlaying the two diagrams and asking how much we have to shift the points on the diagrams to make them the same. See Figure 14.

Now suppose we observe a sample $X_1, \ldots, X_n$ drawn from a distribution P supported on S. The *empirical distance function* is

$$\widehat{d}(x) = \min_{1 \leq i \leq n} ||x - X_i||.$$

Note that the lower level sets of $\widehat{d}$ are precisely the union of balls described in the last section:

$$\widehat{L}_\epsilon = \{x : \ \widehat{d}(x) \leq \epsilon\} = \bigcup_{i=1}^n B(X_i, \epsilon).$$

The persistence diagram $\widehat{D}$ defined by these lower level sets is an estimate of the underlying diagram D.

The empirical distance function is the most commonly used method for defining the persistence diagram of a dataset in the field of computational topology. But from a statistical point of view, this is a very poor choice. It is clear that $\widehat{d}$ is highly non-robust. Even a few outliers will play havoc with the estimator.

Fortunately, more robust and statistically sound methods are available. The first, and perhaps most natural for statisticians, is to replace the lower level

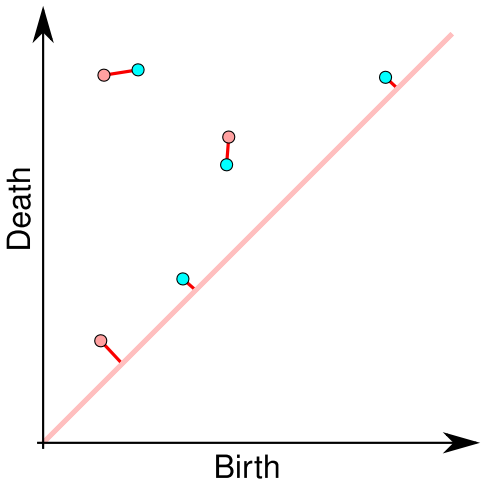

Figure 14: The bottleneck distance between two persistence diagrams is computed by finding the best matching between the two diagrams. This plot shows two diagrams that have been overlayed. The matching is indicated by the lines joining the points from the two diagrams. Note that some points — those with short lifetimes — are matched to the diagonal.

sets of the empirical distance function, with the upper level sets of a density estimator. This approach has been suggested by Phillips et al. (2015); Chazal et al. (2014a); Bobrowski et al. (2014); Chung et al. (2009); Bubenik (2015). The idea is to consider the upper level sets $\widehat{L}_t = \{x : \ \widehat{p}_h(x) > t\}$. As t varies from $\sup_x \widehat{p}_h(x)$ down to 0, the sets $\widehat{L}_t$ evolve and the birth and death times of features are again recorded on a persistence diagram. In this case, the birth times are actually after the death times. This is just an artifact from using upper level sets instead of lower level sets.

An alternative is to re-define the distance function to be intrinsically more robust. Specifically, Chazal et al. (2011) defined the *distance to a measure (DTM)* as follows. Let $0 \leq m \leq 1$ be a scale parameter and define

$$d_m^2(x) = \frac{1}{m}\int_0^m \delta_a^2(x)da$$

where

$$\delta_a(x) = \inf\{r > 0 : \ P(B(x,r)) > a\}.$$

We can think of d_m as a function $T(P)$ of the distribution P. The plug-in estimate of d_m obtained by inserting the empirical distribution in place of P is

$$\widehat{d}_m^2(x) = \frac{1}{k}\sum_{i=1}^{k} ||x - X_i(x)||^2$$

where $k = \lfloor mn \rfloor$ and $X_j(x)$ denote the data after re-ordering them so that $||X_1(x) - x|| \geq ||X_2(x) - m|| \geq \cdots$. In other words, $\widehat{d}_m^2(x)$ is just the average squared distance to the k-nearest neighbors.

The definition of d_m is not arbitrary. The function d_m preserves certain crucial properties that the distance function has, but it changes gracefully as we allow more and more noise. It is essentially a smooth, probabilistic version of the distance function. The properties of the DTM are discussed in Chazal et al. (2011, 2014a, 2015).

Whether we use the kernel density estimator or the DTM, we would like to have a way to decide when topological features are statistically significant. Fasy et al. (2014b); Chazal et al. (2014a) suggest the following method. Let

$$F(t) = P(\sqrt{n}\, \delta_\infty(\widehat{D}, D) \leq t)$$

where D is the true diagram and $\widehat{D}$ is the estimated diagram. Any point on the diagram that is farther than $t_\alpha = F^{-1}(1-\alpha)$ from the diagonal is considered significant at level α. Of course, F is not known but can be estimated by the bootstrap:

$$\widehat{F}(t) = \frac{1}{B}\sum_{j=1}^{B} I(\sqrt{n} d_\infty(\widehat{D}_j^*, \widehat{D}) \leq t)$$

where $\widehat{D}_1^*, \ldots, \widehat{D}_B^*$ are the diagrams based on B bootstrap samples. Then $\widehat{t}_\alpha = \widehat{F}^{-1}(1-\alpha)$ is an estimate of t_α.

Example. We sampled 1,000 observations from a circle in $\mathbb{R}^2$. Gaussian noise was then added to each observation. Then we added 100 outliers samples uniformly from the square. The data are shown in Figure 15. Figure 16 shows the kernel density estimator ($h = .02$) and the persistence diagram based on the upper level sets of the estimator. The points in the pink band are not significant at level $\alpha = 0.1$ (based on the bootstrap). The two points that are significant correspond to one connected component (black dot) and one hole (red triangle). Figure 17 shows a similar analysis of the same data using the DTM with $m = .1$. Generally, we find that the significant features are more prominent using the DTM rather than the kernel density estimator. Also, the DTM is less sensitive to the choice of tuning parameter although it is not known why this is true.

4.3 Simplicial Complexes

The persistence diagram is not computed directly from $\widehat{L}_\epsilon$. Instead, one forms an object called a *Čech complex*. The Čech complex C_ϵ is defined as follows. All singletons are included in C_ϵ; these are 0-dimensional simplices. All pairs of points X_i, X_j such that $||X_i - X_j|| \leq \epsilon$ are included in C_ϵ; these are 1-dimensional simplices. Each triple X_i, X_j, X_k such that $B(X_i, \epsilon/2) \cap B(X_j, \epsilon/2) \cap B(X_k, \epsilon/2)$ is non-empty, is included in C_ϵ; these are 2-dimensional simplices. And so on. The

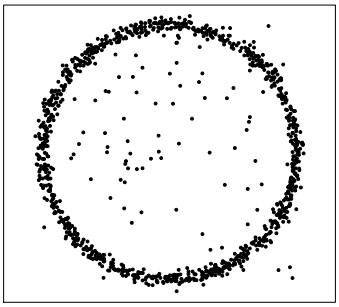

Figure 15: Data sampled from a circle, with Gaussian noise added. There are also 100 outliers sampled uniformly from the square.

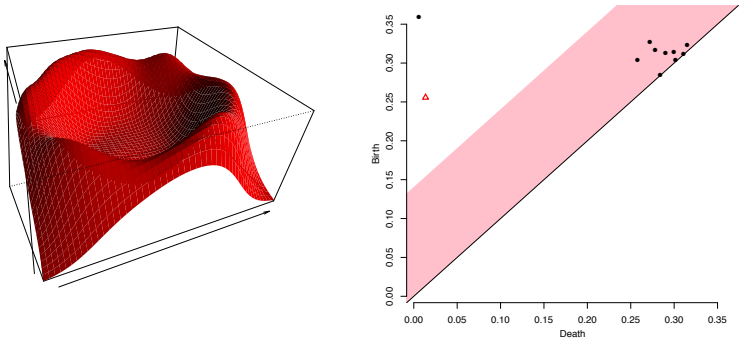

Figure 16: Left: the kernel density estimator. Right: the persistence diagram corresponding to the upper level sets of the estimator. The points above the pink band are significant compared to the bootstrap critical value. Note that one connected component (the black dot) and one hole (the red triangle) are significant.

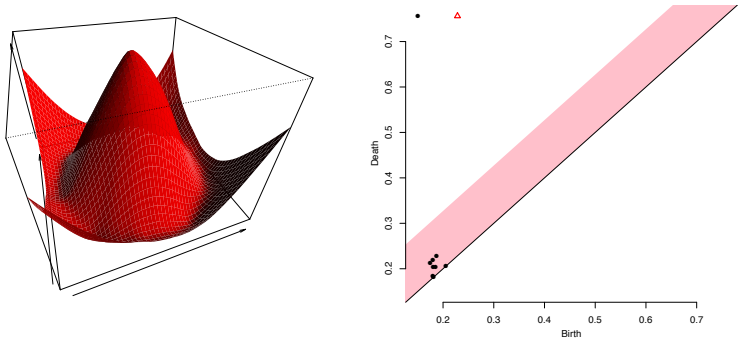

Figure 17: Left: the DTM. Right: the persistence diagram corresponding to the lower level sets of the DTM. The points above the pink band are significant compared to the bootstrap critical value. Note that one connected component (the black dot) and one hole (the red triangle) are significant.

Čech complex is an example of a *simplicial complex*. A collection of simplices is a simplicial complex if it satisfies the following condition: if F is a simplex in C_ϵ and E is a face of F, then E is also on C_ϵ. It can be shown that the homology of $\widehat{L}_\epsilon$ is the same as the homology of C_ϵ. But the homology of C_ϵ can be computed using basic matrix operations. This is how homology is computed in practice (Edelsbrunner & Harer, 2010). Persistent homology relates the complexes as ϵ varies. Again, all the relevant computations can be reduced to linear algebra. Working directly with the Čech complex is computationally prohibitive. In practice, one often uses the Vietoris-Rips complex V_ϵ which is defined as follows. A simplex is included in V_ϵ if each pair of vertices is no more than ϵ apart. It can be shown that the persistent homology defined by V_ϵ approximates the persistent homology defined by C_ϵ.

4.4 Back To Density Clustering

Chazal et al. (2013) have shown that persistent homology can be used as a tool for density clustering. This idea was futher examined in Genovese et al. (2016). Thus we have come full circle and returned to the topic of Section 2.

Recall the mode clustering method described in Section 2.3. We estimate the density, find the modes $\widehat{m}_1, \ldots, \widehat{m}_k$ and the basins of attraction $C_1, \ldots, C_k$ corresponding to the modes.

But we can use more information. In the language of persistent homology, each mode has a lifetime. See Figure 18. Suppose we start with $t = \sup_x p(x)$. We find the upper level set $L_t = \{x : \ p(x) \geq t\}$. Now we let t decrease. (We can think of t as "time" but, in this case, time runs backwards since it starts at a large number and tends to 0.) Everytime we get to a new mode, a

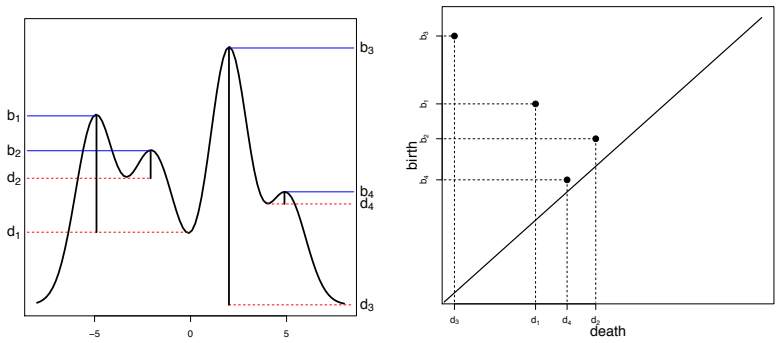

Figure 18: *Starting at the top of the density and moving down, each mode has a birth time b and a death time d. The persistence diagram (right) plots the points* $(d_1, b_1), \ldots, (d_4, b_4)$. *Modes with a long lifetime are far from the diagonal.*

new connected component of L_t is born. However, as t decreases, the connected components can merge. When they merge, the most recently created component is considered to be dead while the other component is still alive. This is called the "elder rule." Proceeding this way, small modes correspond to level sets with short lifetimes. Strong modes correspond to level sets with long lifetimes. We can plot the information as a persistence diagram as in the right plot of Figure 18.

We can use this representation of the modes to decide which modes of a density estimator are significant (Chazal et al., 2014a, 2013). Define $\widehat{t}_\alpha$ by

$$\mathbb{P}\big(\sqrt{n}||\widehat{p}_h^* - \widehat{p}_h|| > \widehat{t}_\alpha \;\big|\; X_1, \ldots, X_n\big) = \alpha,$$

where $\widehat{p}_h^*$ is based on a bootstrap sample $X_1^*, \ldots, X_n^*$ drawn from the empirical distribution P_n. The above probability can be estimated by

$$\frac{1}{B}\sum_{j=1}^{B} I(\sqrt{n}||\widehat{p}_h^* - \widehat{p}_h|| > t)$$

Any mode whose corresponding point on the persistence diagram is farther than $\widehat{t}_\alpha$ from the diagonal is considered a significant mode.

5 TUNING PARAMETERS AND LOSS FUNCTIONS

Virtually every method we have discussed in this paper requires the choice of a tuning parameter. For example, many of the methods involve a kernel density estimator which requires a bandwidth h. But the usual methods for choosing tuning parameters may not be appropriate for TDA. In fact, the problem of choosing tuning parameters is one of the biggest open challenges in TDA.

Let us consider the problem of estimating a density p with the kernel estimator $\widehat{p}_h$. The usual L_2 risk is $\mathbb{E}[\int(\widehat{p}_h(x)-p(x))^2dx]$. Under standard smoothness assumptions, the optimal bandwidth $h \asymp n^{-1/(4+d)}$ yielding a risk of order $n^{-4/(4+d)}$.

But in TDA we are interested in shape, not L_2 loss (or L_p loss for any p). And, as I have mentioned earlier, it may not even be necessary to let h tend to 0 to capture the relevant shape information. In Section 2.2 we saw that, in some cases, the density tree $T(p_h)$ has the same shape as the true tree $T(p)$ even for fixed $h > 0$. Here, $p_h(x) = \mathbb{E}[\widehat{p}_h(x)]$.

Similarly, consider estimating a ridge R of a density p. In general, the ridge can only be estimated at rate $O_P(n^{-2/(8+d)})$. Now suppose we use a small but fixed (non-decreasing) bandwidth h. Usually, the ridge R_h of p_h is a reasonably good but slightly biased approximation to R. But R can be estimated at rate $O_P(\sqrt{\log n/n})$. We are often better off living with the bias and estimating R_h instead of R.

In fact one could argue that any shape information that can only be recovered with small bandwidths is very subtle and cannot be reliably estimated. The salient structure can be recovered with a fixed bandwidth. To explain this in more detail, we consider two examples from Chen et al. (2015a).

The left plot in Figure 19 shows a density p. The blue points at the bottom show the level set $L = \{x : \ p > .05\}$. The right plot shows p_h for $h = .2$ and the blue points at the bottom show the level set $L_h = \{x : \ p_h > .05\}$. The smoothed out density p_h is biased and the level set L_h loses the small details of L. But L_h contains the main part of L and it may be more honest to say that $\widehat{L}_h$ is an estimate of L_h.

As a second example, let $P = (1/3)\phi(x; -5, 1) + (1/3)\delta_0 + (1/3)\phi(x; 5, 1)$ where ϕ is a Normal density and δ_0 is a point mass at 0. Of course, this distribution does not even have a density. The left plot in Figure (20) shows the density of the absolutely continuous part of P with a vertical line to who the point mass. The right plot shows p_h, which is a smooth, well-defined density. Again the blue points show the level sets. As before p_h is biased (as is L_h). But p_h is well-defined, as is L_h, and $\widehat{p}_h$ and $\widehat{L}_h$ are accurate estimators of p_h and L_h. Moreover, L_h contains the most important qualitative information about L, namely, that there are three connected components, one of which is small.

The idea of viewing p_h as the estimand is not new. The "scale space" approach to smoothing explictly argues that we should view $\widehat{p}_h$ as an estimate of p_h, and p_h is then regarded as a view of p at a particular resolution. This idea is discussed in detail in Chaudhuri & Marron (2000, 1999); Godtliebsen et al. (2002).

If we do decide to base TDA on tuning parameters that do not go to 0 as n increases then we need new methods for choosing tuning parameters. One possibility, suggested in Chazal et al. (2014a) and Guibas et al. (2013) is to choose the tuning parameter that maximizes the number of significant topological features. In particular, Chazal et al. (2014a) use the bootstrap to assess the significance of topological features and then they choose the smoothing pa-

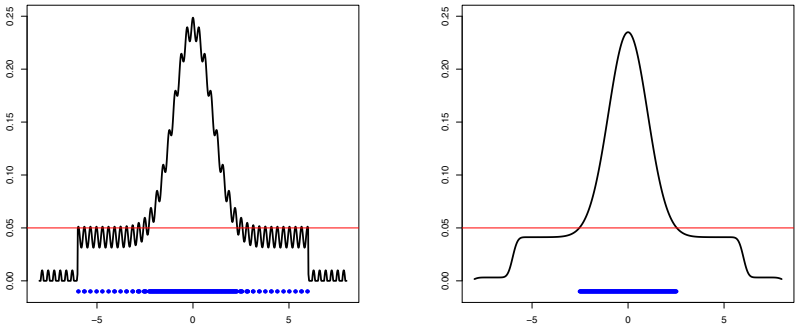

Figure 19: Left: a density p and a level set $\{p > t\}$. Right: the smoothed density p_h and the level set $\{p_h > t\}$.

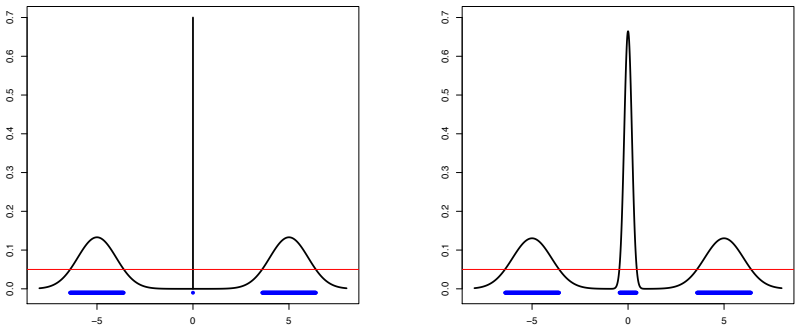

Figure 20: Left: a distribution with a continuous component and a point mass at 0. Right: the smoothed density p_h. The level set L_h is biased but is estimable and it approximates the main features of L.

rameter to maximize the number of such features. This maximal significance approach is promising but so far there is no theory to support the idea.

The problem of choosing tuning parameters thus remains one of the greatest challenges in TDA. In fact, the same problem permeates the clustering literature. To date, there is no agreement on how to choose k in k-means clustering, for example.

6 DATA VISUALIZATION AND EMBEDDINGS

Topological ideas play a role in data visualization either explicitly or implicitly. In fact, many TDA methods may be regarded as visualization methods. For example, density trees, persistence diagrams and manifold learning all provide low dimensional representations of the data that are easy to visualize.

Some data visualization methods work by embedding the data in $\mathbb{R}^2$ and then simply plotting the data. Consider a point cloud $X_1, \ldots, X_n$ where $X_i \in \mathbb{R}^d$. Let $\psi : \mathbb{R}^d \to \mathbb{R}^2$ and let $Z_i = \psi(X_i)$. Because the points $Z_1, \ldots, Z_n$ are in $\mathbb{R}^2$, we can easily plot the Z_i's. Perhaps the most familiar version is multidimensional scaling (MDS) where ψ is chosen to be a linear function minimizing some measure of distance between the original pairwise distances $||X_i - X_j||^2$ and the embedded distances $||Z_i - Z_j||^2$. In particular, if we minimize $\sum_{i \neq j}(||X_i - X_j||^2 - ||Z_i - Z_j||^2)$ then the solution is to project the data onto the first two principal components.

But traditional MDS does a poor job of preserving local structure such as clusters. Local, nonlinear versions of MDS do a better job of preserving local structure. An example is *Laplacian Eigenmaps* which was proposed by Belkin & Niyogi (2003). Here, we choose ψ to minimize $\sum_{i,j} W_{ij}||Z_i - Z_j||^2$ (subject to some consraints) where the W_{ij} are localization weights such as $W_{ij} = e^{-||X_i - X_j||^2/(2h^2)}$. The resulting embedding does a good job of preserving local structure. However, Maaten & Hinton (2008) noted that local methods of this type can cause the data to be too crowded together. They proposed a new method called t-SNE which seems to work better but they provided no justification for the method. Carreira-Perpinán (2010) provided an explanation of why t-SNE works. He showed that t-SNE optimizes a criterion that essentially contains two terms, one promoting localization and the other which causes points to repel each other. Based on this insight, he proposed a new method called *elastic embedding* that explicitly has a term encouraging clusters to stay together and a term that repels points from each other. What is notable about t-SNE and elastic embedding is that they preserve clusters and loops. The loops are preserved apparently due to the repelling term. It appears, in other words that these methods preserve topological features of the data.

This leads to the following question: is it possible to derive low-dimensional embedding methods that explicitly preserve topological features of the data? This is an interesting open question.

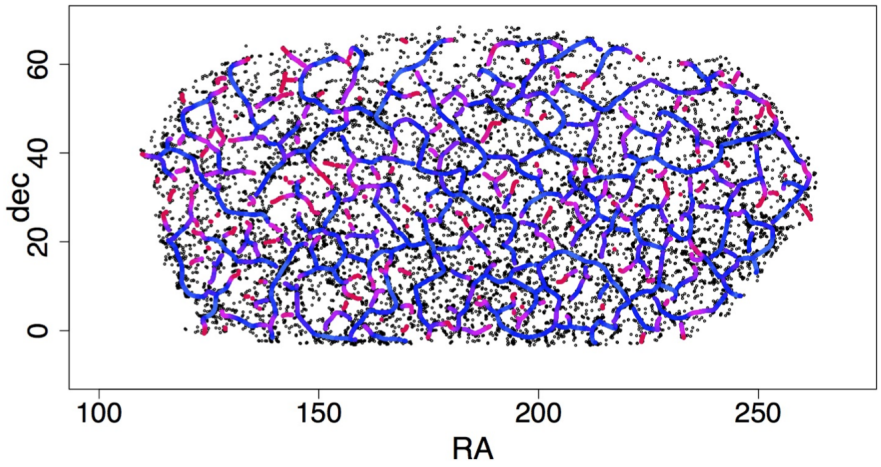

Figure 21: A filament map from Chen et al. (2015c). The data are galaxies from the Sloan Digital Sky Survey. The blue lines are detected filaments. The red dots are clusters.

7 APPLICATIONS

7.1 The Cosmic Web

The matter in the Universe is distributed in a complex, spiderweb-like pattern known as the Cosmic web. Understanding and quantifying this structure is one of the challenges of modern cosmology. Figure 21 shows a two-dimensional slice of data consisting of some galaxies from the Sloan Digital Sky Survey (`www.sdss.org`) as analyzed in Chen et al. (2015c). (RA refers to "right ascension" and DEC refers to "declination." These measure position in the sky using essentially longitude and latitude). The blue lines are filaments that were found using the ridge methods discussed in Section 3.3. Also shown are clusters (red dots) that were found by previous researchers. Filament maps like this permit researchers to investigate questions about how structure formed in our Universe. For example, Chen et al. (2015d) investigated how the properties of galaxies differ depending on the distance from filaments.

Several papers, such as Van de Weygaert et al. (2011); van de Weygaert et al. (2011, 2010) have used homology and persistent homology to study the structure of the cosmic web. These papers use TDA to quantify the clusters, holes and voids in astronomical data. Sousbie et al. (2011); Sousbie (2011) uses Morse theory to model the filamentary structures of the cosmic web.

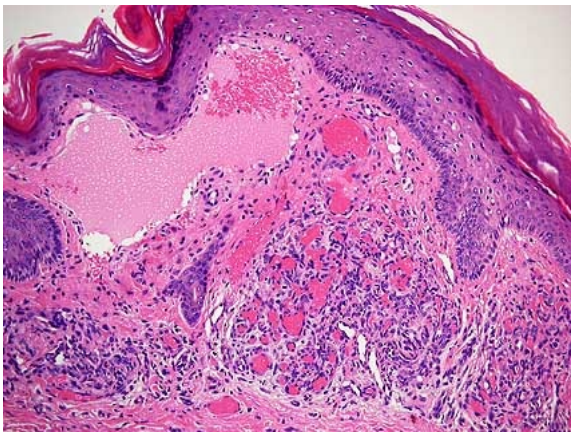

Figure 22: An example of a histology image from http://medicalpicturesinfo.com/histology/.

7.2 Images

Many researchers have used some form of TDA for image analysis. Consider Figure 23 which shows a 3d image of a rabbit. Given a large collection of such images, possibly corrupted by noise, we would like to define features that can be used for classifying such images. It is critical that the features be invariant to shifts, rotations and small deformations. Topological are thus a promising source of relevant features. A number of papers have used TDA to define such features, for example, Bonis et al. (2016); Li et al. (2014); Carrière et al. (2015).

TDA has also been used in the classification of 2d images. For example Singh et al. (2014) considered breast cancer histology images. These images show the arrangement of cells of tissue samples. An example of a histology image is given in Figure 22.

A typical image has many clumps and voids so TDA may be an appropriate method for summarizing the images. Singh et al. (2014) used the Betti numbers as a function of scale, as features for a classifier. The goal was to discriminate different sub-types of cancer. They achieved a classification accuracy of 69.86 percent.

7.3 Proteins

Kovacev-Nikolic et al. (2016) used TDA to study the maltose binding protein which is a protein found in Escherichia coli. An example of such a protein is given in Figure 24; the figure is from http://lilith.nec.aps.anl.gov/Structures/Publications.htm. The protein is a dynamic structure and the changes in structure are of biological relevance. Quoting from Kovacev-Nikolic et al. (2016):

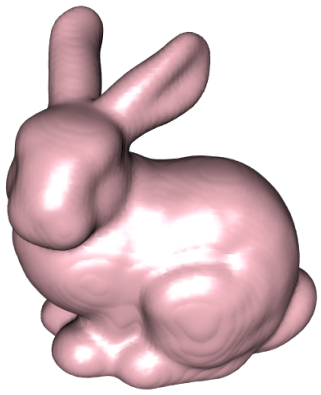

Figure 23: A three-dimensional image. Classifying such images requires features that are invariant to small deformations of the image. TDA can potentially provide such features.

> A major conformational change in the protein occurs when a smaller molecule called a ligand attaches to the protein molecule ... Ligand-induced conformational changes are important because the biological function of the protein occurs through a transition from a ligand-free (apo) to a ligand-bound (holo) structure ...

The protein can be in an open or closed conformation, and the closed conformation is due to having a captured ligand. The goal of the authors is to classify the state of the protein.

Each protein is represented by 370 points (corresponding to amino acids) in three dimension space. The authors construct a dynamic model of the protein structure (since the structure changes over time) from which they define dynamical distances between the 370 points. Thus a protein is represented by a 370 by 370 distance matrix. From the distance matrix they construct a persistence diagram. Next, they convert the persistence diagrams into a set of functions called *landscapes* as defined in Bubenik (2015). Turning the diagram into a set of one-dimensional functions makes it easier to use standard statistical tools. In particular, they do a two-sample permutation test using the integrated distances between the landscape functions as a test statistic. The p-value is 5.83×10^{-4} suggesting a difference between the open and closed conformations. This suggests that landscapes can be used to classify proteins as open or closed. They also show that certain sites on the protein, known as *active sites*, are associated

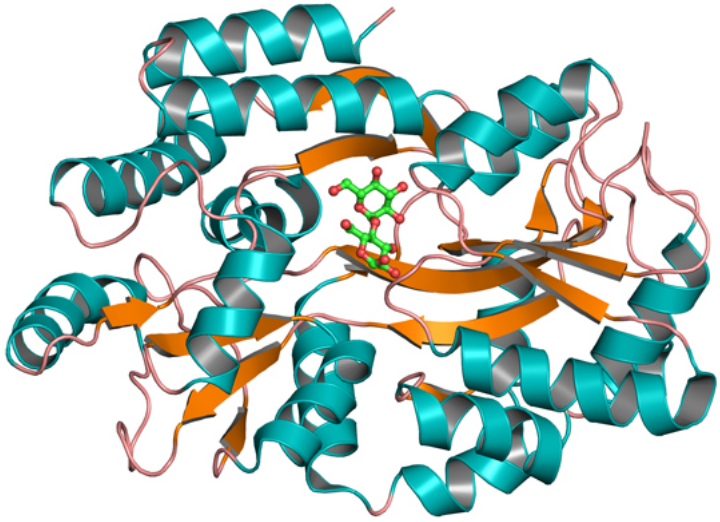

Figure 24: A maltose binding protein. The image of the protein is from `http://lilith.nec.aps.anl.gov/Structures/Publications.htm`.

with loops in the protein.

7.4 Other Applications

Here I briefly mention a few other examples of TDA.

The Euler characteristic is a topological quantity which I did not mention in this paper. It has played an important role in various aspects of probability as well as to applications in astrophysics and neuroscience (Worsley, 1995; Taylor & Worsley, 2007; Adler & Taylor, 2009; Worsley, 1994, 1996; Taylor & Worsley, 2007). The Euler characteristic has also been used for classification of shapes (Richardson & Werman, 2014). See also Turner et al. (2014). Bendich et al. (2010) use topological methods to study the interactions between root systems of plants. Carstens & Horadam (2013) use persistent homology to describe the structure of collaboration networks. Xia et al. (2015) use TDA in the analysis of biomolecules. Adcock et al. (2014) use TDA to classify images of lesions of the liver. Chung et al. (2009) use persistence diagrams constructed from data on cortical thickness to distinguish control subjects and austistic subjects. Offroy & Duponchel (2016) reviews the role of TDA in chemometrics. Bendich et al. (2016) use persistent homology to study the structure of brain arteries. There is now a substantial literature on TDA in neuroscience including Arai et al. (2014); Babichev & Dabaghian (2016); Basso et al. (2016); Bendich et al. (2014); Brown & Gedeon (2012); Cassidy et al. (2015); Chen et al. (2014); Choi et al. (2014); Chung et al. (2009); Curto & Itskov (2008); Curto et al. (2013, 2015); Curto & Youngs (2015); Curto (2016); Dabaghian et al. (2011, 2012,

2014); Dabaghian (2015); Dlotko et al. (2016); Ellis & Klein (2014); Giusti & Itskov (2013); Giusti et al. (2015, 2016); Hoffman et al. (2016); Jeffs et al. (2015); Kanari et al. (2016); Khalid et al. (2014); Kim et al. (2014); Lee et al. (2011); Lienkaemper et al. (2015); Manin (2015); Masulli & Villa (2015); Petri et al. (2014); Pirino et al. (2014); Singh et al. (2008); Sizemore et al. (2016a,b); Spreemann et al. (2015); Stolz (2014); Yoo et al. (2016); Zeeman (1965). The website `http://www.chadgiusti.com/algtop-neuro-bibliography.html` maintain a bibliography of references in this area.

8 CONCLUSION: THE FUTURE OF TDA

TDA is an exciting area and is full of interesting ideas. But so far, it has had little impact on data analysis. Is this because the techniques are new? Is it because the techniques are too complicated? Or is it because the methods are simply not that useful in practice?

Right now, it is hard to know the answer. My personal opinion is that TDA is a very specialized tool that is useful in a small set of problems. For example, it seems to be an excellent tool for summarizing data relating to the cosmic web. But, I doubt that TDA will ever become a general purpose tool like regression. The exception is clustering, which of course is used routinely, although some might argue that it is a stretch to consider clustering part of TDA. I have seen a number of examples where complicated TDA methods were used to analyze data but no effort was made to compare these methods to simpler, more traditional statistical methods. It is my hope that, in the next few years, researchers will do thorough comparisons of standard statistical methods with TDA in a number of scientific areas so that we can truly assess the value of these new methods.

DISCLOSURE STATEMENT

The author is not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

Thanks to Steve Fienberg for suggesting that I write this review. And thanks to Robert Adler, Eddie Amari, Omer Bobrowski, Bertrand Michel, Vic Patrangenaru, JaeHyeok Shin, Isabella Verdinelli, for providing comments on an earlier draft.

References

Adcock A, Rubin D, Carlsson G. 2014. Classification of hepatic lesions using the matching metric. *Computer vision and image understanding* 121:36–42

Adler RJ, Taylor JE. 2009. Random fields and geometry. Springer Science & Business Media

Arai M, Brandt V, Dabaghian Y. 2014. The effects of theta precession on spatial learning and simplicial complex dynamics in a topological model of the hippocampal spatial map. *PLoS Comp. Bio.* 10

Arias-Castro E, Chen G, Lerman G, et al. 2011. Spectral clustering based on local linear approximations. *Electronic Journal of Statistics* 5:1537–1587

Arias-Castro E, Mason D, Pelletier B. 2015. On the estimation of the gradient lines of a density and the consistency of the mean-shift algorithm. *Journal of Machine Learning Research*

Azizyan M, Chen YC, Singh A, Wasserman L. 2015. Risk bounds for mode clustering. *arXiv preprint arXiv:1505.00482*

Babichev A, Dabaghian Y. 2016. Persistent memories in transient networks. *arXiv:1602.00681 [qbio.NC]*

Balakrishnan S, Narayanan S, Rinaldo A, Singh A, Wasserman L. 2013. Cluster trees on manifolds, In *Advances in Neural Information Processing Systems*

Basso E, Arai M, Dabaghian Y. 2016. Gamma synchronization of the hippocampal spatial map—topological model. *arXiv:1603.06248 [qbio.NC]*

Belkin M, Niyogi P. 2001. Laplacian eigenmaps and spectral techniques for embedding and clustering., In *NIPS*, vol. 14

Belkin M, Niyogi P. 2003. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation* 15:1373–1396

Bendich P, Cohen-Steiner D, Edelsbrunner H, Harer J, Morozov D. 2007. Inferring local homology from sampled stratified spaces, In *Foundations of Computer Science, 2007. FOCS'07. 48th Annual IEEE Symposium on*. IEEE

Bendich P, Edelsbrunner H, Kerber M. 2010. Computing robustness and persistence for images. *IEEE transactions on visualization and computer graphics* 16:1251–1260

Bendich P, Marron J, Miller E, Pieloch A, Skwerer S. 2014. Persistent homology analysis of brain artery trees. *Ann. Appl. Stat.* to appear

Bendich P, Marron JS, Miller E, Pieloch A, Skwerer S, et al. 2016. Persistent homology analysis of brain artery trees. *The Annals of Applied Statistics* 10:198–218

Bobrowski O, Mukherjee S, Taylor JE. 2014. Topological consistency via kernel estimation. *arXiv preprint arXiv:1407.5272*

Bonis T, Ovsjanikov M, Oudot S, Chazal F. 2016. Persistence-based pooling for shape pose recognition, In *International Workshop on Computational Topology in Image Context*. Springer

Brown J, Gedeon T. 2012. Structure of the afferent terminals in terminal ganglion of a cricket and persistent homology. *PLoS ONE* 7

Bubenik P. 2015. Statistical topological data analysis using persistence landscapes. *Journal of Machine Learning Research* 16:77–102

Cadre B. 2006. Kernel estimation of density level sets. *Journal of multivariate analysis* 97:999–1023

Carreira-Perpinán MA. 2010. The elastic embedding algorithm for dimensionality reduction., In *ICML*, vol. 10

Carrière M, Oudot SY, Ovsjanikov M. 2015. Stable topological signatures for points on 3d shapes, In *Computer Graphics Forum*, vol. 34. Wiley Online Library

Carstens C, Horadam K. 2013. Persistent homology of collaboration networks. *Mathematical problems in engineering* 2013

Cassidy B, Rae C, Solo V. 2015. Brain activity: Conditional dissimilarity and persistent homology, In *Biomedical Imaging (ISBI), 2015 IEEE 12th International Symposium on*. IEEE

Chacón. 2012. Clusters and water flows: a novel approach to modal clustering through morse theory. *arXiv preprint arXiv:1212.1384*

Chacón JE, Duong T, et al. 2013. Data-driven density derivative estimation, with applications to nonparametric clustering and bump hunting. *Electronic Journal of Statistics* 7:499–532

Chacón JE, et al. 2015. A population background for nonparametric density-based clustering. *Statistical Science* 30:518–532

Chaudhuri K, Dasgupta S. 2010. Rates of convergence for the cluster tree, In *Advances in Neural Information Processing Systems*

Chaudhuri P, Marron JS. 1999. Sizer for exploration of structures in curves. *Journal of the American Statistical Association* 94:807–823

Chaudhuri P, Marron JS. 2000. Scale space view of curve estimation. *Annals of Statistics* :408–428

Chazal F, Cohen-Steiner D, Lieutier A. 2009. A sampling theory for compact sets in euclidean space. *Discrete & Computational Geometry* 41:461–479

Chazal F, Cohen-Steiner D, Mérigot Q. 2011. Geometric inference for probability measures. *Foundations of Computational Mathematics* 11:733–751

Chazal F, Fasy BT, Lecci F, Michel B, Rinaldo A, Wasserman L. 2014a. Robust topological inference: Distance to a measure and kernel distance. *arXiv preprint arXiv:1412.7197*

Chazal F, Glisse M, Labruère C, Michel B. 2014b. Convergence rates for persistence diagram estimation in topological data analysis, In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*

Chazal F, Guibas LJ, Oudot SY, Skraba P. 2013. Persistence-based clustering in riemannian manifolds. *Journal of the ACM (JACM)* 60:41

Chazal F, Massart P, Michel B. 2015. Rates of convergence for robust geometric inference. *arXiv preprint arXiv:1505.07602*

Chen YC, Genovese CR, Wasserman L. 2015a. Density level sets: Asymptotics, inference, and visualization. *arXiv preprint arXiv:1504.05438*

Chen YC, Genovese CR, Wasserman L, et al. 2015b. Asymptotic theory for density ridges. *The Annals of Statistics* 43:1896–1928

Chen YC, Ho S, Freeman PE, Genovese CR, Wasserman L. 2015c. Cosmic web reconstruction through density ridges: method and algorithm. *Monthly Notices of the Royal Astronomical Society* 454:1140–1156

Chen YC, Ho S, Tenneti A, Mandelbaum R, Croft R, et al. 2015d. Investigating galaxy-filament alignments in hydrodynamic simulations using density ridges. *Monthly Notices of the Royal Astronomical Society* 454:3341–3350

Chen YC, Kim J, Balakrishnan S, Rinaldo A, Wasserman L. 2016. Statistical inference for cluster trees. *arXiv preprint arXiv:1605.06416*

Chen Z, Gomperts SN, Yamamoto J, Wilson MA. 2014. Neural representation of spatial topology in the rodent hippocampus. *Neural Comput.* 26:1–39

Cheng Y. 1995. Mean shift, mode seeking, and clustering. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 17:790–799

Choi H, Kim YK, Kang H, Lee H, Im HJ, et al. 2014. Abnormal metabolic connectivity in the pilocarpine-induced epilepsy rat model: a multiscale network analysis based on persistent homology. *NeuroImage* 99:226–236

Chung MK, Bubenik P, Kim PT. 2009. Persistence diagrams of cortical surface data, In *International Conference on Information Processing in Medical Imaging*. Springer

Coifman RR, Lafon S. 2006. Diffusion maps. *Applied and computational harmonic analysis* 21:5–30

Comaniciu D, Meer P. 2002. Mean shift: A robust approach toward feature space analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 24:603–619

Costa JA, Hero AO. 2004. Geodesic entropic graphs for dimension and entropy estimation in manifold learning. *IEEE Transactions on Signal Processing* 52:2210–2221

Cuevas A. 2009. Set estimation: Another bridge between statistics and geometry. *Boletín de Estadística e Investigación Operativa* 25:71–85

Cuevas A, Febrero M, Fraiman R. 2001. Cluster analysis: a further approach based on density estimation. *Computational Statistics & Data Analysis* 36:441–459

Curto C. 2016. What can topology tell us about the neural code? *forthcoming*

Curto C, Gross E, Jeffries J, Morrison K, Omar M, et al. 2015. What makes a neural code convex? *arXiv:1508.00150 [q-bio.NC]*

Curto C, Itskov V. 2008. Cell groups reveal structure of stimulus space. *PLoS Comp. Bio.* 4:e1000205

Curto C, Itskov V, Veliz-Cuba A, Youngs N. 2013. The neural ring: an algebraic tool for analyzing the intrinsic structure of neural codes. *Bull. Math. Biol.* 75:1571–1611

Curto C, Youngs N. 2015. Neural ring homomorphisms and maps between neural codes. *arXiv:1511.00255 [math.NC]*

Dabaghian Y. 2015. Geometry of spatial memory replay. *arXiv:1508.06579 [q-bio.NC]*

Dabaghian Y, Brandt VL, Frank LM. 2014. Reconceiving the hippocampal map as a topological template. *Elife* 3:e03476

Dabaghian Y, Cohn AG, Frank L. 2011. Topological coding in the hippocampus. In *Computational modeling and simulation of intellect: Current state and future prospectives*. IGI Global Hershey, PA, 293–320

Dabaghian Y, Mémoli F, Frank L, Carlsson G. 2012. A topological paradigm for hippocampal spatial map formation using persistent homology. *PLoS Comp. Bio.* 8:e1002581

De'ath G. 1999. Extended dissimilarity: a method of robust estimation of ecological distances from high beta diversity data. *Plant Ecology* 144:191–199

Devroye L, Wise GL. 1980. Detection of abnormal behavior via nonparametric estimation of the support. *SIAM Journal on Applied Mathematics* 38:480–488

Dlotko P, Hess K, Levi R, Nolte M, Reimann M, et al. 2016. Topological analysis of the connectome of digital reconstructions of neural microcircuits. *arXiv:1601.01580 [q-bio.NC]*

Eberly D. 1996. Ridges in image and data analysis, vol. 7. Springer Science & Business Media

Edelsbrunner H, Harer J. 2008. Persistent homology-a survey. *Contemporary mathematics* 453:257–282

Edelsbrunner H, Harer J. 2010. Computational topology: an introduction. American Mathematical Soc.

Edelsbrunner H, Letscher D, Zomorodian A. 2002. Topological persistence and simplification. *Discrete and Computational Geometry* 28:511–533

Eldridge J, Wang Y, Belkin M. 2015. Beyond hartigan consistency: Merge distortion metric for hierarchical clustering. *arXiv preprint arXiv:1506.06422*

Ellis SP, Klein A. 2014. Describing high-order statistical dependence using concurrence topology, with application to functional mri brain data. *H. H. A.* 16

Fasy BT, Kim J, Lecci F, Maria C. 2014a. Introduction to the r package tda. *arXiv preprint arXiv:1411.1830*

Fasy BT, Lecci F, Rinaldo A, Wasserman L, Balakrishnan S, et al. 2014b. Confidence sets for persistence diagrams. *The Annals of Statistics* 42:2301–2339

Genovese CR, Perone-Pacifico M, Verdinelli I, Wasserman L. 2012a. Manifold estimation and singular deconvolution under hausdorff loss. *The Annals of Statistics* 40:941–963

Genovese CR, Perone-Pacifico M, Verdinelli I, Wasserman L. 2012b. Minimax manifold estimation. *Journal of Machine Learning Research* :1263–1291

Genovese CR, Perone-Pacifico M, Verdinelli I, Wasserman L. 2016. Nonparametric inference for density modes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 78:99–126

Genovese CR, Perone-Pacifico M, Verdinelli I, Wasserman L, et al. 2014. Nonparametric ridge estimation. *The Annals of Statistics* 42:1511–1545

Giusti C, Ghrist R, Bassett DS. 2016. Two's company, three (or more) is a simplex: Algebraic-topological tools for understanding higher-order structure in neural data. *J. Comput. Neurosci.* 41

Giusti C, Itskov V. 2013. A no-go theorem for one-layer feedforward networks. *Neural Comput.* 26:2527–2540

Giusti C, Pastalkova E, Curto C, Itskov V. 2015. Clique topology reveals intrinsic geometric structure in neural correlations. *Proc. Nat. Acad. Sci. USA* 112:13455–13460

Godtliebsen F, Marron J, Chaudhuri P. 2002. Significance in scale space for bivariate density estimation. *Journal of Computational and Graphical Statistics* 11:1–21

Guibas L, Morozov D, Mérigot Q. 2013. Witnessed k-distance. *Discrete & Computational Geometry* 49:22–45

Hartigan JA. 1975. Clustering algorithms. Wiley

Hartigan JA. 1981. Consistency of single linkage for high-density clusters. *Journal of the American Statistical Association* 76:388–394

Hatcher A. 2000. Algebraic topology. Cambridge Univ. Press

Hein M, Audibert JY. 2005. Intrinsic dimensionality estimation of submanifolds in r d, In *Proceedings of the 22nd international conference on Machine learning*. ACM

Hoffman K, Babichev A, Dabaghian Y. 2016. Topological mapping of space in bat hippocampus. *arXiv:1601.04253 [q-bio.NC]*

Jeffs RA, Omar M, Suaysom N, Wachtel A, Youngs N. 2015. Sparse neural codes and convexity. *arXiv:1511.00283 [math.CO]*

Kanari L, Dłotko P, Scolamiero M, Levi R, Shillcock J, et al. 2016. Quantifying topological invariants of neuronal morphologies. *arXiv:1603.08432 [q-bio.NC]*

Kégl B. 2002. Intrinsic dimension estimation using packing numbers, In *Advances in neural information processing systems*

Kendall DG. 1984. Shape manifolds, procrustean metrics, and complex projective spaces. *Bulletin of the London Mathematical Society* 16:81–121

Khalid A, Kim BS, Chung MK, Ye JC, Jeon D. 2014. Tracing the evolution of multi-scale functional networks in a mouse model of depression using persistent brain network homology. *Neuroimage* 101:351–363

Kim E, Kang H, Lee H, Lee HJ, Suh MW, et al. 2014. Morphological brain network assessed using graph theory and network filtration in deaf adults. *Hear. Res.* 315:88–98

Kim J, Rinaldo A, Wasserman L. 2016. Minimax rates for estimating the dimension of a manifold. *arXiv preprint arXiv:1605.01011*

Koltchinskii VI. 2000. Empirical geometry of multivariate data: A deconvolution approach. *The Annals of Statistics* 28:591–629

Kovacev-Nikolic V, Bubenik P, Nikolić D, Heo G. 2016. Using persistent homology and dynamical distances to analyze protein binding. *Statistical applications in genetics and molecular biology* 15:19–38

Lee H, Chung MK, Kang H, Kim BN, Lee DS. 2011. Discriminative persistent homology of brain networks, In *Biomedical Imaging: From Nano to Macro, 2011 IEEE International Symposium on*. IEEE

Lee JA, Verleysen M. 2007. Nonlinear dimensionality reduction. Springer Science & Business Media

Levina E, Bickel PJ. 2004. Maximum likelihood estimation of intrinsic dimension, In *Advances in neural information processing systems*

Li C, Ovsjanikov M, Chazal F. 2014. Persistence-based structural recognition, In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*

Li J, Ray S, Lindsay B. 2007. A nonparametric statistical approach to clustering via mode identification. *Journal of Machine Learning Research* 8:1687–1723

Lienkaemper C, Shiu A, Woodstock Z. 2015. Obstructions to convexity in neural codes. *arXiv:1509.03328 [q-bio.NC]*

Little AV, Maggioni M, Rosasco L. 2011. Multiscale geometric methods for estimating intrinsic dimension. *Proc. SampTA* 4:2

Lombardi G, Rozza A, Ceruti C, Casiraghi E, Campadelli P. 2011. Minimum neighbor distance estimators of intrinsic dimension, In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer

Maaten Lvd, Hinton G. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research* 9:2579–2605

Manin YI. 2015. Neural codes and homotopy types: mathematical models of place field recognition. *arXiv:1501.00897 [math.HO]*

Markov A. 1958. Insolubility of the problem of homeomorphy. *Proc. Intern. Congress of Mathematicians* :300–306

Masulli P, Villa AE. 2015. The topology of the directed clique complex as a network invariant. *arXiv:1510.00660 [q-bio.NC]*

Milnor J. 2016. Morse theory.(am-51), vol. 51. Princeton university press

Müller DW, Sawitzki G. 1991. Excess mass estimates and tests for multimodality. *Journal of the American Statistical Association* 86:738–746

Niyogi P, Smale S, Weinberger S. 2008. Finding the homology of submanifolds with high confidence from random samples. *Discrete & Computational Geometry* 39:419–441

Niyogi P, Smale S, Weinberger S. 2011. A topological view of unsupervised learning from noisy data. *SIAM Journal on Computing* 40:646–663

Offroy M, Duponchel L. 2016. Topological data analysis: A promising big data exploration tool in biology, analytical chemistry and physical chemistry. *Analytica chimica acta* 910:1–11

Ozertem U, Erdogmus D. 2011. Locally defined principal curves and surfaces. *The Journal of Machine Learning Research* 12:1249–1286

Patrangenaru V, Ellingson L. 2015. Nonparametric statistics on manifolds and their applications to object data analysis. CRC Press

Petri G, Expert P, Turkheimer F, Carhart-Harris R, Nutt D, et al. 2014. Homological scaffolds of brain functional networks. *J. Roy. Soc. Int.* 11:20140873

Phillips JM, Wang B, Zheng Y. 2015. Geometric inference on kernel density estimates, In *31st International Symposium on Computational Geometry (SoCG 2015)*, vol. 34. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik

Pirino V, Riccomagno E, Martinoia S, Massobrio P. 2014. A topological study of repetitive co-activation networks in in vitro cortical assemblies. *Phys. Bio.* 12:016007–016007

Polonik W. 1995. Measuring mass concentrations and estimating density contour clusters-an excess mass approach. *The Annals of Statistics* :855–881

Richardson E, Werman M. 2014. Efficient classification using the euler characteristic. *Pattern Recognition Letters* 49:99–106

Rinaldo A, Wasserman L. 2010. Generalized density clustering. *The Annals of Statistics* :2678–2722

Roweis ST, Saul LK. 2000. Nonlinear dimensionality reduction by locally linear embedding. *Science* 290:2323–2326

Singh G, Memoli F, Ishkhanov T, Sapiro G, Carlsson G, Ringach DL. 2008. Topological analysis of population activity in visual cortex. *J. Vis.* 8:11

Singh N, Couture HD, Marron JS, Perou C, Niethammer M. 2014. Topological descriptors of histology images, In *International Workshop on Machine Learning in Medical Imaging*. Springer

Sizemore A, Giusti C, Bassett DS. 2016a. Classification of weighted networks through mesoscale homological features. *Journal of Complex Networks*

Sizemore A, Giusti C, Betzel RF, Bassett DS. 2016b. Closures and cavities in the human connectome. *arxiv:1608.03520 [q-bio.NC]*

Skraba P, Wang B. 2014. Approximating local homology from samples, In *Proceedings of the Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms*. Society for Industrial and Applied Mathematics

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

Sousbie T, Pichon C, Kawahara H. 2011. The persistent cosmic web and its filamentary structure–ii. illustrations. *Monthly Notices of the Royal Astronomical Society* 414:384–403

Spreemann G, Dunn B, Botnan MB, Baas NA. 2015. Using persistent homology to reveal hidden information in neural data. *arXiv:1510.06629 [q-bio.NC]*

Stolz B. 2014. Computational topology in neuroscience. Master's thesis, University of Oxford

Taylor JE, Worsley KJ. 2007. Detecting sparse signals in random fields, with an application to brain mapping. *Journal of the American Statistical Association* 102:913–928

Tenenbaum JB, De Silva V, Langford JC. 2000. A global geometric framework for nonlinear dimensionality reduction. *science* 290:2319–2323

Turner K, Mukherjee S, Boyer DM. 2014. Persistent homology transform for modeling shapes and surfaces. *Information and Inference* :iau011

van de Weygaert R, Platen E, Vegter G, Eldering B, Kruithof N. 2010. Alpha shape topology of the cosmic web, In *Voronoi Diagrams in Science and Engineering (ISVD), 2010 International Symposium on*. IEEE

van de Weygaert R, Pranav P, Jones BJ, Bos E, Vegter G, et al. 2011. Probing dark energy with alpha shapes and betti numbers. *arXiv preprint arXiv:1110.5528*

Van de Weygaert R, Vegter G, Edelsbrunner H, Jones BJ, Pranav P, et al. 2011. Alpha, betti and the megaparsec universe: on the topology of the cosmic web, In *Transactions on Computational Science XIV*. Springer-Verlag

Worsley KJ. 1994. Local maxima and the expected euler characteristic of excursion sets of χ 2, f and t fields. *Advances in Applied Probability* :13–42

Worsley KJ. 1995. Boundary corrections for the expected euler characteristic of excursion sets of random fields, with an application to astrophysics. *Advances in Applied Probability* :943–959

Worsley KJ. 1996. The geometry of random images. *Chance* 9:27–40

Xia K, Zhao Z, Wei GW. 2015. Multiresolution topological simplification. *Journal of Computational Biology* 22:887–891

Yoo J, Kim EY, Ahn YM, Ye JC. 2016. Topological persistence vineyard for dynamic functional brain connectivity during resting and gaming stages. *Journal of Neuroscience Methods* 267:1–13

Zeeman EC. 1965. The topology of the brain and visual perception. In *Mathematics and computer science in biology and medicine*. London: H.M. Stationary Office, 240–256