

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/258818915>

# Confidence Sets for Persistence Diagrams

Article in *The Annals of Statistics* · March 2014

CITATIONS

245

READS

474

6 authors, including:



**Brittany Terese Fasy**

Montana State University

88 PUBLICATIONS 1,194 CITATIONS

SEE PROFILE



**Fabrizio Lecci**

Carnegie Mellon University

8 PUBLICATIONS 733 CITATIONS

SEE PROFILE



**Larry Wasserman**

Carnegie Mellon University

242 PUBLICATIONS 21,917 CITATIONS

SEE PROFILE



**Sivaraman Balakrishnan**

Carnegie Mellon University

22 PUBLICATIONS 969 CITATIONS

SEE PROFILE

## CONFIDENCE SETS FOR PERSISTENCE DIAGRAMS

BY BRITTANY TERESE FASY<sup>\*,1</sup>, FABRIZIO LECCI<sup>†,2</sup>,  
ALESSANDRO RINALDO<sup>†,2</sup>, LARRY WASSERMAN<sup>†,3</sup>,  
SIVARAMAN BALAKRISHNAN<sup>†</sup> AND AARTI SINGH<sup>†</sup>

*Tulane University\* and Carnegie Mellon University†*

Persistent homology is a method for probing topological properties of point clouds and functions. The method involves tracking the birth and death of topological features (2000) as one varies a tuning parameter. Features with short lifetimes are informally considered to be “topological noise,” and those with a long lifetime are considered to be “topological signal.” In this paper, we bring some statistical ideas to persistent homology. In particular, we derive confidence sets that allow us to separate topological signal from topological noise.

**1. Introduction.** Topological data analysis (TDA) refers to a collection of methods for finding topological structure in data [Carlsson (2009), Edelsbrunner and Harer (2010)]. TDA has been used in protein analysis, image processing, text analysis, astronomy, chemistry and computer vision, as well as in other fields.

One approach to TDA is *persistent homology*, a branch of computational topology that leads to a plot called a *persistence diagram*. This diagram can be thought of as a summary statistic, capturing multi-scale topological features. This paper studies the statistical properties of persistent homology. Homology detects the connected components, tunnels, voids, etc., of a topological space  $M$ . Persistent homology measures these features by assigning a birth and a death value to each feature. For example, suppose we sample  $\mathcal{S}_n = \{X_1, \dots, X_n\}$  from an unknown distribution  $P$ . We are interested in estimating the homology of the support of  $P$ . One method for doing so would be to take the union of the set of balls centered at the points in  $\mathcal{S}_n$  as we do in the supplementary material [Fasy et al. (2014)]; however, to do

---

Received March 2013; revised June 2014.

<sup>1</sup>Supported in part by NSF Grant CCF-1065106.

<sup>2</sup>Supported in part by NSF CAREER Grant DMS-11-49677.

<sup>3</sup>Supported in part by Air Force Grant FA95500910373, NSF Grant DMS-08-06009.

*AMS 2000 subject classifications.* Primary 62G05, 62G20; secondary 62H12.

*Key words and phrases.* Persistent homology, topology, density estimation.

This is an electronic reprint of the original article published by the  
Institute of Mathematical Statistics in *The Annals of Statistics*,  
2014, Vol. 42, No. 6, 2301–2339. This reprint differs from the original in  
pagination and typographic detail.

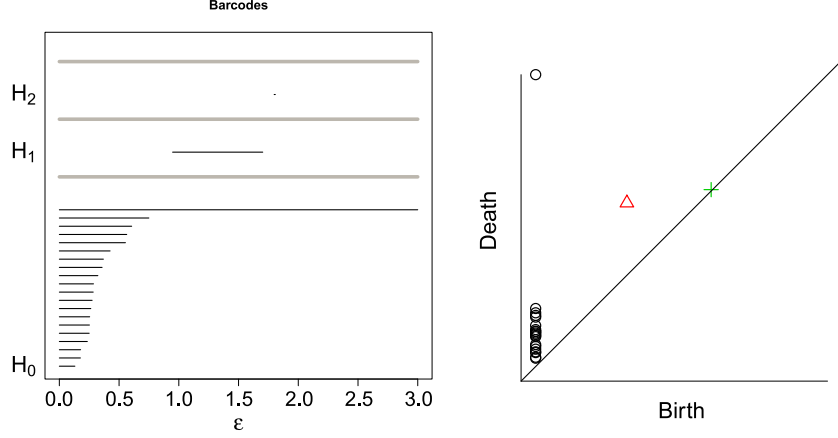


FIG. 1. (Left) barcode plot. Each horizontal bar shows the lifetime of a topological feature as a function of  $\epsilon$  for the homology groups  $H_0$ ,  $H_1$ , and  $H_2$ . Significant features have long horizontal bars. (Right) persistence diagram. The points in the persistence diagram are in one-to-one correspondence with the bars in the barcode plot. The birth and death times of the barcode become the  $x$ - and  $y$ -coordinates of the persistence diagram. Significant features are far from the diagonal.

so, we must choose a radius for these balls. Rather than choosing a single radius, we use persistent homology to summarize the homology for all radii, as shown in Figures 1 and 2. Persistence points far from the diagonal  $x = y$

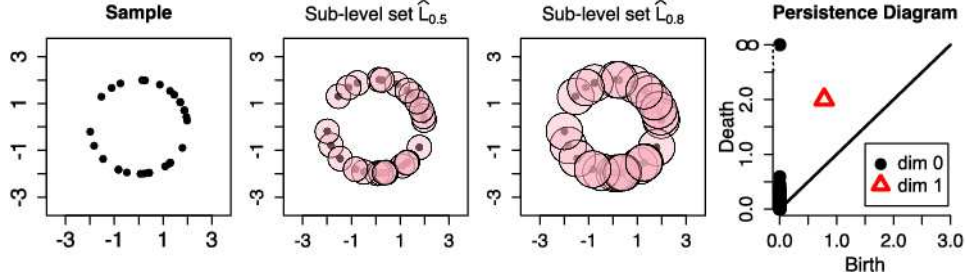


FIG. 2. (Left) 30 data points  $S_{30}$  sampled from the circle of radius 2. (Middle left) sub-levels set  $\hat{L}_{0.5} = \{x : d_{S_{30}} \leq 0.5\}$ ; at  $\epsilon = 0.5$ , the sub-level set consists of two connected components and zero loops. (Middle right) sub-levels set  $\hat{L}_{0.8} = \{x : d_{S_{30}} \leq 0.8\}$ ; as we keep increasing  $\epsilon$ , we assist at the birth and death of topological features; at  $\epsilon = 0.8$ , one of the connected components dies (is merged with the other one), and a one-dimensional hole appears; this loop will die at  $\epsilon = 2$ , when the union of the pink balls representing the distance function becomes simply connected. Right: the persistence diagram summarizes the topological features of the sampled points. The black dots represent the connected components: 30 connected components are present at  $\epsilon = 0$ , and they progressively die as  $\epsilon$  increases, leaving only one connected component. The red triangle represents the unique one-dimensional hole that appears at  $\epsilon = 0.8$  and dies at  $\epsilon = 2$ .

represent topological features common to a long interval of radii. We define persistent homology precisely in Section 2.

One of the key challenges in persistent homology is to find a way to separate the noise from the signal in the persistence diagram, and this paper suggests several statistical methods for doing so. In particular, we provide a confidence set for the persistence diagram  $\mathcal{P}$  corresponding to the distance function to a topological space  $\mathbb{M}$  using a sample from a distribution  $P$  supported on  $\mathbb{M}$ . The confidence set has the form  $\mathcal{C} = \{Q : W_\infty(\hat{\mathcal{P}}, Q) \leq c_n\}$ , where  $Q$  varies over the set of all persistence diagrams,  $\hat{\mathcal{P}}$  is an estimate of the persistence diagram constructed from the sample,  $W_\infty$  is a metric on the space of persistence diagrams, called the bottleneck distance, and  $c_n$  is an appropriate, data-dependent quantity. In addition, we study the upper level sets of density functions by computing confidence sets for their persistence diagrams.

*Goals.* There are two main goals in this paper. The first is to introduce persistent homology to statisticians. The second is to derive confidence sets for certain key quantities in persistent homology. In particular, we derive a simple method for separating topological noise from topological signal. The method has a simple visualization: we only need to add a band around the diagonal of the persistence diagram. Points in the band are consistent with being noise. We focus on simple, synthetic examples in this paper as proof of concept.

*Related work.* Some key references in computational topology are [Bubenik and Kim (2007), Carlsson (2009), Ghrist (2008), Carlsson and Zomorodian (2009), Edelsbrunner and Harer (2008), Chazal and Oudot (2008), Chazal et al. (2011)]. An [Introduction](#) to homology can be found in Hatcher (2002). The probabilistic basis for random persistence diagrams is studied in Mileyko, Mukherjee and Harer (2011) and Turner et al. (2014). Other relevant probabilistic results can be found in Kahle (2009, 2011), Penrose (2003), Kahle and Meckes (2013). Some statistical results for homology and persistent homology include Bubenik et al. (2010), Blumberg et al. (2012), Balakrishnan et al. (2011), Joshi et al. (2011), Bendich, Mukherjee and Wang (2010), Chazal et al. (2010), Bendich, Galkovskyi and Harer (2011), [Niyogi, Smale and Weinberger \(2008, 2011\)](#). The latter paper considers a challenging example which involves a set of the form in Figure 8 (top left), which we consider later in the paper. Heo, Gamble and Kim (2012) contains a detailed analysis of data on a medical procedure for upper jaw expansion that uses persistent homology as part of a nonlinear dimension reduction of the data. Chazal et al. (2013b) is closely related to this paper; they find convergence rates for persistence diagrams computed from data sampled from a

distribution on a metric space. As the authors point out, finding confidence intervals that follow from their methods is a challenging problem because their probability inequalities involve unknown constants. Restricting our attention to manifolds embedded in  $\mathbb{R}^D$ , we are able to provide several methods to compute confidence sets for persistence diagrams.

*Outline.* We define persistent homology formally in Section 2 and provide additional details in the supplementary material [Fasy et al. (2014)]. The statistical model is defined in Section 3. Several methods for constructing confidence intervals are presented in Section 4. Section 5 illustrates the ideas with a few numerical experiments. Proofs are contained in Section 6. Finally, Section 7 contains concluding remarks.

*Notation.* We write  $a_n \preceq b_n$  if there exists  $c > 0$  such that  $a_n \leq cb_n$  for all large  $n$ . We write  $a_n \asymp b_n$  if  $a_n \preceq b_n$  and  $b_n \preceq a_n$ . For any  $x \in \mathbb{R}^D$  and any  $r \geq 0$ ,  $B(x, r)$  denotes the  $D$ -dimensional ball of radius  $r > 0$  centered at  $x$ . For any closed set  $A \subset \mathbb{R}^D$ , we define the (Euclidean) distance function

$$(1) \quad d_A(x) = \inf_{y \in A} \|y - x\|_2.$$

In addition, for any  $\varepsilon \geq 0$ , the Minkowski sum is defined as

$$(2) \quad A \oplus \varepsilon = \bigcup_{x \in A} B(x, \varepsilon) = \{x : d_A(x) \leq \varepsilon\}.$$

The *reach* of  $A$ —denoted by  $\text{reach}(A)$ —is the largest  $\varepsilon \geq 0$  such that each point in  $A \oplus \varepsilon$  has a unique projection onto  $A$  [Federer (1959)]. If  $f$  is a real-valued function, we define the *upper level set*  $\{x : f(x) \geq t\}$ , the *lower level set*  $\{x : f(x) \leq t\}$  and the *level set*  $\{x : f(x) = t\}$ . If  $A$  is measurable, we write  $P(A)$  for the probability of  $A$ . For more general events  $A$ ,  $\mathbb{P}(A)$  denotes the probability of  $A$  on an appropriate probability space. In particular, if  $A$  is an event in the  $n$ -fold probability space under random sampling, then  $\mathbb{P}(A)$  means probability under the product measure  $P \times \cdots \times P$ . In some places, we use symbols like  $c, c_1, C, \dots$ , as generic positive constants. Finally, if two sets  $A$  and  $B$  are homotopic, we write  $A \cong B$ .

**2. Brief introduction to persistent homology.** In this section, we provide a brief overview of persistent homology. In the supplementary material [Fasy et al. (2014)], we provide a more details on relevant concepts from computational topology; however, for a more complete coverage of persistent homology, we refer the reader to Edelsbrunner and Harer (2010).

Given a real-valued function  $f$ , persistent homology describes how the topology of the lower level sets  $f^{-1}(-\infty, t]$  change as  $t$  increases from  $-\infty$  to  $\infty$ . In particular, persistent homology describes  $f$  with a multiset of points

in the plane, each corresponding to the birth and death of a homological feature that existed for some interval of  $t$ .

First, we consider the case where  $f$  is a distance function. Let  $K$  be a compact subset of  $\mathbb{R}^D$ , and let  $d_K: \mathbb{R}^D \rightarrow \mathbb{R}$  be the distance function to  $K$ . Consider the sub-level set  $L_t = \{x: d_K(x) \leq t\}$ ; note that  $K = L_0$ . As  $t$  varies from 0 to  $\infty$ , 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  $(b_1, d_1), \dots, (b_m, d_m)$ . These points can be plotted on the plane, resulting in a persistence diagram  $\mathcal{P}$ ; see Figures 1 and 2. Alternatively, we can represent the pair  $(b_i, d_i)$  as the interval  $[b_i, d_i]$ . The set of intervals is referred to as a *barcode plot*; see Figure 1. We view the persistence diagram and the barcode plot as topological summaries of the input function or data. Points near the diagonal in the persistence diagram (i.e., the short intervals in the barcode plot) have short lifetimes and are considered “topological noise.” Most applications are interested in features that we can distinguish from noise; that is, those features that persist for a large range of values  $t$ .

**2.1. Persistent homology.** We present persistent homology as a summary of the input function or data, as the goal of this paper is to define methods for computing confidence sets for that summary.

Given data points  $\mathcal{S}_n = \{X_1, \dots, X_n\}$ , we are interested in understanding the homology of the  $d$ -dimensional compact topological space  $\mathbb{M} \subset \mathbb{R}^D$  from which the data were sampled; see the supplementary material [Fasy et al. (2014)] for the definition of homology. If our sample is dense enough, and the topological space has a nice embedding in  $\mathbb{R}^D$ , then  $H_p(\mathbb{M})$  is a subgroup of the  $p$ th homology group of the sublevel set  $\hat{L}_\varepsilon = \{x: d_{\mathcal{S}_n}(x) \leq \varepsilon\}$  for an interval of values of  $\varepsilon$ . Choosing the right  $\varepsilon$  is a difficult task: small  $\varepsilon$  will have the homology of  $n$  points, and large  $\varepsilon$  will have the homology of a single point. Using persistent homology, we avoid choosing a single  $\varepsilon$  by assigning a persistence value to each nontrivial homology generator that is realized as  $\hat{L}_\varepsilon$  for some nonnegative  $\varepsilon$ . This persistence value is defined to be the length of the interval of  $\varepsilon$  for which that feature occurs. See Figure 2.

To consider  $\hat{L}_\varepsilon$  for every  $\varepsilon$  in  $(0, \infty)$  would be infeasible. Hence, we restrict our attention to equivalence classes of homology groups. Since  $H(\hat{L}_r) = H(\check{\text{Cech}}(\mathcal{S}_n, r))$ , we use the Čech complex to compute the homology of the

lower level sets; see the supplementary material [Fasy et al. (2014)] for the definition of a Čech complex. Let  $r_1, \dots, r_k$  be the set of radii such that the complexes  $\check{\text{Cech}}(\mathcal{S}_n, r_i)$  and  $\check{\text{Cech}}(\mathcal{S}_n, r_i - \varepsilon)$  are not identical for sufficiently small  $\varepsilon$ . Letting  $K_0 = \emptyset$ ,  $K_{k+1}$  be the maximal simplicial complex defined on  $\mathcal{S}_n$ , and  $K_i = \check{\text{Cech}}(\mathcal{S}_n, (r_i + r_{i-1})/2)$ , the sequence of complexes is the *Čech filtration* of  $d_{\mathcal{S}_n}$ . For all  $s < t$ , there is a natural inclusion  $i_{s,t}: K_s \hookrightarrow K_t$  that induces a group homomorphism  $i_{s,t}^*: H_p(K_s) \rightarrow H_p(K_t)$ . Thus we have the following sequence of homology groups:

$$(3) \quad H_p(|K_0|) \rightarrow H_p(|K_1|) \rightarrow \dots \rightarrow H_p(|K_n|).$$

We say that a homology class  $[\alpha]$  represented by a  $p$ -cycle  $\alpha$  is *born* at  $K_s$  if  $[\alpha]$  is not supported in  $K_r$  for any  $r < s$ , but is nontrivial in  $H_p(|K_s|)$ . The class  $[\alpha]$  born at  $K_s$  *dies* going into  $K_t$  if  $t$  is the smallest index such that the class  $[\alpha]$  is supported in the image of  $i_{s-1,t}^*$ . The birth at  $s$  and death at  $t$  of  $[\alpha]$  is recorded as the point  $(s, t)$  in the  $p$ th persistence diagram  $\mathcal{P}_p(d_{\mathcal{S}_n})$ , which we now formally define.

**DEFINITION 1** (Persistence diagram). Given a function  $f: \mathbb{X} \rightarrow \mathbb{R}$ , defined for a triangulable subspace of  $\mathbb{R}^D$ , the  $p$ th *persistence diagram*  $\mathcal{P}_p(f)$  is the multiset of points in the extended plane  $\overline{\mathbb{R}}^2$ , where  $\overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty, +\infty\}$ , such that the each point  $(s, t)$  in the diagram represents a distinct topological feature that existed in  $H_p(f^{-1}((-\infty, r]))$  for  $r \in [s, t)$ . The *persistence barcode* is a multiset of intervals that encodes the same information as the persistence diagram by representing the point  $(s, t)$  as the interval  $[s, t]$ .

In sum, the zero-dimensional diagram  $\mathcal{P}_0(f)$  records the birth and death of components of the lower level sets; more generally, the  $p$ -dimensional diagram  $\mathcal{P}_p(f)$  records the  $p$ -dimensional holes of the lower level sets. We let  $\mathcal{P}(f)$  be the overlay of all persistence diagrams for  $f$ ; see Figures 2 and 3, for examples, of persistent homology of one-dimensional and two-dimensional distance functions.

In the supplementary material [Fasy et al. (2014)], we see that the homology of a lower level set is equivalent to a Čech complex and can be estimated by a Vietoris–Rips complex. Therefore, in Section 5, we use the Vietoris–Rips filtration to compute the confidence sets for persistence diagrams of distance functions.

**2.2. 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  $L_\infty$ -distance between functions and the bottleneck distance between

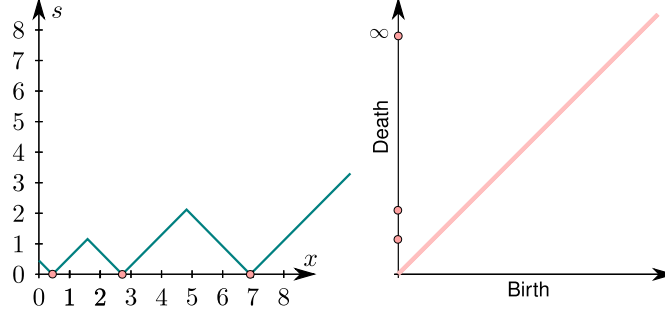


FIG. 3. (Left) distance function  $d_X$  for a set of three points. (Right) corresponding persistence diagram. If  $X \subset \mathbb{R}^1$ , then the persistence diagram records the death times of the initial components. Since all components are born at  $s=0$ , all points in the persistence diagram appear on the  $y$ -axis. The point labeled with  $\infty$  represents an essential class, that is, one that is inherited from the domain (in this case,  $\mathbb{R}^1$ ).

persistence diagrams; see the supplementary material [Fasy et al. (2014)] for the definition of bottleneck distance.

The  $L_\infty$ -distance  $\|f - g\|_\infty$  is the maximum difference between the function values

$$\|f - g\|_\infty = \sup_{x \in X} |f(x) - g(x)|.$$

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

**THEOREM 2 (Bottleneck stability).** *Let  $X$  be finitely triangulable, and let  $f, g: X \rightarrow \mathbb{R}$  be continuous. Then the bottleneck distance between the corresponding persistence diagrams is bounded from above by the  $L_\infty$ -distance between them,*

$$(4) \quad W_\infty(\mathcal{P}(f), \mathcal{P}(g)) \leq \|f - g\|_\infty.$$

The bottleneck stability theorem is one of the main requirements for our methods to work. We refer the reader to Cohen-Steiner, Edelsbrunner and Harer (2007) and to Chazal et al. (2012) for proofs of this theorem. Alternatively, one can consider different distance functions. The Wasserstein distance is defined by finding the perfect pairing that minimizes the sum (rather than the supremum) of the pairwise distances. For a restricted class of functions, there exists a stability theorem for the Wasserstein distance; see Cohen-Steiner et al. (2010). We note that the techniques to compute confidence sets presented in this paper can be extended to define a confidence set for the persistence diagram under the Wasserstein distance using a stronger set of assumptions on  $M$ .



**2.3. Hausdorff distance.** Let  $A, B$  be compact subsets of  $\mathbb{R}^D$ . One way to measure the distance between these sets is to take the *Hausdorff distance*, denoted by  $\mathbf{H}(A, B)$ , which is the maximum Euclidean distance from a point in one set to the closest point in the other set,

$$\begin{aligned}\mathbf{H}(A, B) &= \max\left\{\max_{x \in A} \min_{y \in B} \|x - y\|, \max_{x \in B} \min_{y \in A} \|x - y\|\right\} \\ &= \inf\{\epsilon : A \subset B \oplus \epsilon \text{ and } B \subset A \oplus \epsilon\}.\end{aligned}$$

The stability result Theorem 2 is key. Let  $\mathbb{M}$  be a  $d$ -manifold embedded in a compact subset  $\mathbb{X}$  of  $\mathbb{R}^D$ . If  $S$  is any subset of  $\mathbb{M}$ ,  $\mathcal{P}_S$  is the persistence diagram based on the lower level sets  $\{x : d_S(x) \leq t\}$ , and  $\mathcal{P}$  is the persistence diagram based on the lower level sets  $\{x : d_{\mathbb{M}}(x) \leq t\}$ , then, by Theorem 2,

$$(5) \quad W_\infty(\mathcal{P}_S, \mathcal{P}) \leq \|d_S - d_{\mathbb{M}}\|_\infty = \mathbf{H}(S, \mathbb{M}).$$

We bound  $\mathbf{H}(S, \mathbb{M})$  to obtain a bound on  $W_\infty(\mathcal{P}_S, \mathcal{P})$ . In particular, we obtain a confidence set for  $W_\infty(\mathcal{P}_S, \mathcal{P})$  by deriving a confidence set for  $\mathbf{H}(S, \mathbb{M})$ . Thus the stability theorem reduces the problem of inferring persistent homology to the problem of inferring Hausdorff distance. Indeed, much of this paper is devoted to the latter problem. We would like to point out that the Hausdorff distance plays an important role in many statistical problems. Examples include Cuevas (2009), Cuevas, Febrero and Fraiman (2001), Cuevas and Fraiman (1997, 1998), Cuevas, Fraiman and Pateiro-López (2012), Cuevas and Rodríguez-Casal (2004), Mammen and Tsybakov (1995). Our methods could potentially be useful for these problems as well.

**3. Statistical model.** As mentioned above, we want to estimate the homology of a set  $\mathbb{M}$ . We do not observe  $\mathbb{M}$  directly; rather, we observe a sample  $\mathcal{S}_n = \{X_1, \dots, X_n\}$  from a distribution  $P$  that is concentrated on or near  $\mathbb{M} \subset \mathbb{R}^D$ . For example, suppose  $\mathbb{M}$  is a circle. Then the homology of the data set  $\mathcal{S}_n$  is not equal to the homology of  $\mathbb{M}$ ; however, the set  $\widehat{L}_\varepsilon = \{x : d_{\mathcal{S}_n}(x) \leq \varepsilon\} = \bigcup_{i=1}^n B(X_i, \varepsilon)$ , where  $B(x, \varepsilon)$  denotes the Euclidean ball of radius  $\varepsilon$  centered at  $x$ , captures the homology of  $\mathbb{M}$  for an interval of values  $\varepsilon$ . Figure 2 shows  $\widehat{L}_\varepsilon$  for increasing values of  $\varepsilon$ .

Let  $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} P$  where  $X_i \in \mathbb{R}^D$ . Let  $\mathbb{M}$  denote the  $d$ -dimensional support of  $P$ . Let us define the following quantities:

$$(6) \quad \rho(x, t) = \frac{P(B(x, t/2))}{t^d}, \quad \rho(t) = \inf_{x \in \mathbb{M}} \rho(x, t).$$

We assume that  $\rho(x, t)$  is a continuous function of  $t$ , and we define

$$(7) \quad \rho(x, \downarrow 0) = \lim_{t \rightarrow 0} \rho(x, t), \quad \rho = \lim_{t \rightarrow 0} \rho(t).$$

Note that if  $P$  has continuous density  $p$  with respect to the uniform measure on  $\mathbb{M}$ , then  $\rho \propto \inf_{x \in \mathbb{M}} p(x)$ . Until Section 4.4, we make the following assumptions:

ASSUMPTION A1.  $\mathbb{M}$  is  $d$ -dimensional compact manifold (with no boundary), embedded in  $\mathbb{R}^D$  and  $\text{reach}(\mathbb{M}) > 0$ . (The definition of  $\text{reach}$  was given in Section 1.)

ASSUMPTION A2. For each  $x \in \mathbb{M}$ ,  $\rho(x, t)$  is a bounded continuous function of  $t$ , differentiable for  $t \in (0, t_0)$  and right differentiable at zero. Moreover,  $\partial\rho(x, t)/\partial t$  exists and is bounded away from zero and infinity for  $t$  in an open neighborhood of zero. Also, for some  $t_0 > 0$  and some  $C_1$  and  $C_2$ , we have

$$(8) \quad \sup_x \sup_{0 \leq t \leq t_0} \left| \frac{\partial\rho(x, t)}{\partial t} \right| \leq C_1 < \infty \quad \text{and} \quad \sup_{0 \leq t \leq t_0} |\rho'(t)| \leq C_2 < \infty.$$

REMARKS. The  $\text{reach}$  of  $\mathbb{M}$  does not appear explicitly in the results as the dependence is implicit and does not affect the rates in the asymptotics. Note that if  $P$  has a density  $p$  with respect to the Hausdorff measure on  $\mathbb{M}$ , then Assumption A2 is satisfied as long as  $p$  is smooth and bounded away from zero. Assumption A1 guarantees that as  $\varepsilon \rightarrow 0$ , the covering number  $N(\varepsilon)$  satisfies  $N(\varepsilon) \asymp (1/\varepsilon)^d$ . However, the conditions are likely stronger than needed. For example, it suffices that  $\mathbb{M}$  be compact and  $d$ -rectifiable. See, for example, Mattila (1995) and Ambrosio, Fusco and Pallara (2000).

Recall that the distance function is  $d_{\mathbb{M}}(x) = \inf_{y \in \mathbb{M}} \|x - y\|$ , and let  $\mathcal{P}$  be the persistence diagram defined by the lower level sets  $\{x : d_{\mathbb{M}}(x) \leq \varepsilon\}$ . Our target of inference is  $\mathcal{P}$ . Let  $\hat{\mathcal{P}}$  denote the persistence diagram of the  $\{x : d_{\mathcal{S}_n}(x) \leq \varepsilon\}$  where  $\mathcal{S}_n = \{X_1, \dots, X_n\}$ . We regard  $\hat{\mathcal{P}}$  as an estimate of  $\mathcal{P}$ . Our main goal is to find a confidence interval for  $W_{\infty}(\mathcal{P}, \hat{\mathcal{P}})$  as this implies a confidence set for the persistence diagram.

Until Section 4.4, we assume that the dimension of  $\mathbb{M}$  is known and that the support of the distribution is  $\mathbb{M}$  which is sometimes referred to as the noiseless case. These assumptions may seem unrealistic to statisticians but are, in fact, common in computational geometry. In Section 4.4, we weaken the assumptions. Specifically, we allow outliers, which means there may be points not on  $\mathbb{M}$ . Bendich, Galkovskyi and Harer (2011) show that methods based on the Čech complex perform poorly when there are outliers. Instead, we estimate the persistent homology of the upper level sets of the density function. We shall see that the methods in Section 4.4 are quite robust.

**4. Confidence sets.** Given  $\alpha \in (0, 1)$ , we will find  $c_n \equiv c_n(X_1, \dots, X_n)$  such that

$$(9) \quad \limsup_{n \rightarrow \infty} \mathbb{P}(W_{\infty}(\hat{\mathcal{P}}, \mathcal{P}) > c_n) \leq \alpha.$$

It then follows that  $C_n = [0, c_n]$  is an asymptotic  $1 - \alpha$  confidence set for the bottleneck distance  $W_\infty(\widehat{\mathcal{P}}, \mathcal{P})$ , that is,

$$(10) \quad \liminf_{n \rightarrow \infty} \mathbb{P}(W_\infty(\widehat{\mathcal{P}}, \mathcal{P}) \in [0, c_n]) \geq 1 - \alpha.$$

Recall that, from Theorem 2 and the fact that  $\|d_{\mathbb{M}} - d_{\mathcal{S}_n}\|_\infty = \mathbf{H}(\mathcal{S}_n, \mathbb{M})$ , we have

$$(11) \quad W_\infty(\widehat{\mathcal{P}}, \mathcal{P}) \leq \mathbf{H}(\mathcal{S}_n, \mathbb{M}),$$

where  $\mathcal{S}_n = \{X_1, \dots, X_n\}$  is the sample and  $\mathbf{H}$  is the Hausdorff distance. Hence it suffices to find  $c_n$  such that

$$(12) \quad \limsup_{n \rightarrow \infty} \mathbb{P}(\mathbf{H}(\mathcal{S}_n, \mathbb{M}) > c_n) \leq \alpha.$$

The confidence set  $\mathcal{C}_n$  is a subset of all persistence diagrams whose distance to  $\widehat{\mathcal{P}}$  is at most  $c_n$ ,

$$\mathcal{C}_n = \{\widetilde{\mathcal{P}} : W_\infty(\widehat{\mathcal{P}}, \widetilde{\mathcal{P}}) \leq c_n\}.$$

We can visualize  $\mathcal{C}_n$  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. Alternatively, we can visualize the confidence set by adding a band of width  $\sqrt{2}c_n$  around the diagonal of the persistence diagram  $\widehat{\mathcal{P}}$ . 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. That is, if the confidence set for a point on the diagram hits the diagonal, then we cannot rule out that the lifetime of that feature is 0, and we consider it to be noise. (This is like saying that if a confidence interval for a treatment effect includes 0, then the effect is not distinguishable from “no effect.”) This leads to the diagrams shown in Figure 4.

REMARK. This simple dichotomy of “signal” and “noise” is not the only way to quantify the uncertainty in the persistence diagram. Indeed, some points near the diagonal may represent interesting structure. One can imagine endowing each point in the diagram with a confidence set, possibly of different sizes and shapes. But for the purposes of this paper, we focus on the simple method described above.

The first three methods that we present are based on the persistence diagram constructed from the Čech complex. The fourth method takes a different approach completely and is based on density estimation. We define the methods in this section; we illustrate them in Section 5.

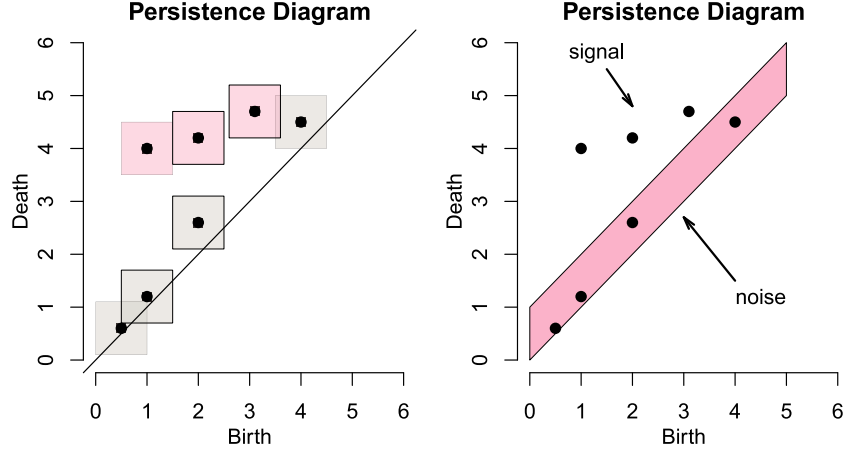


FIG. 4. First, we obtain the confidence interval  $[0, c_n]$  for  $W_\infty(\hat{\mathcal{P}}, \mathcal{P})$ . If a box of side length  $2c_n$  around a point in the diagram hits the diagonal, we consider that point to be noise. By putting a band of width  $\sqrt{2}c_n$  around the diagonal, we need 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  $c_n = 0.5$ .

4.1. *Method I: Subsampling.* The first method uses subsampling. The usual approach to subsampling [see, e.g., Politis, Romano and Wolf (1999), Romano and Shaikh (2012)] is based on the assumption that we have an estimator  $\hat{\theta}$  of a parameter  $\theta$  such that  $n^\xi(\hat{\theta} - \theta)$  converges in distribution to some fixed distribution  $J$  for some  $\xi > 0$ . Unfortunately, our problem is not of this form. Nonetheless, we can still use subsampling as long as we are willing to have conservative confidence intervals.

Let  $b = b_n$  be such that  $b = o(\frac{n}{\log n})$  and  $b_n \rightarrow \infty$ . We draw all  $N$  subsamples  $\mathcal{S}_{b,n}^1, \dots, \mathcal{S}_{b,n}^N$ , each of size  $b$ , from the data where  $N = \binom{n}{b}$ . (In practice, as is always the case with subsampling, it suffices to draw a large number of subsamples randomly rather than use all  $N$  subsamples. But, for the theory, we assume all  $N$  subsamples are used.) Let  $T_j = \mathbf{H}(\mathcal{S}_{b,n}^j, \mathcal{S}_n)$ ,  $j = 1, \dots, N$ . Define

$$(13) \quad L_b(t) = \frac{1}{N} \sum_{j=1}^N I(T_j > t),$$

and let  $c_b = 2L_b^{-1}(\alpha)$ . Recalling the definition of  $\rho$  from (7), we can prove the following theorem:

THEOREM 3. Assume that  $\rho > 0$ . Then, for all large  $n$ ,

$$(14) \quad \mathbb{P}(W_\infty(\hat{\mathcal{P}}, \mathcal{P}) > c_b) \leq \mathbb{P}(\mathbf{H}(\mathcal{S}_n, \mathbb{M}) > c_b) \leq \alpha + O\left(\frac{b}{n}\right)^{1/4}.$$

4.2. *Method II: Concentration of measure.* The following lemma is similar to theorems in Devroye and Wise (1980), Cuevas, Febrero and Fraiman (2001) and Niyogi, Smale and Weinberger (2008).

LEMMA 4. *For all  $t > 0$ ,*

$$(15) \quad \mathbb{P}(W_\infty(\hat{\mathcal{P}}, \mathcal{P}) > t) \leq \mathbb{P}(\mathbf{H}(\mathcal{S}_n, \mathbb{M}) > t) \leq \frac{2^d}{\rho(t/2)t^d} \exp(-n\rho(t)t^d),$$

where  $\rho(t)$  is defined in (6). If, in addition,  $t < \min\{\rho/(2C_2), t_0\}$ , then

$$(16) \quad \mathbb{P}(\mathbf{H}(\mathcal{S}_n, \mathbb{M}) > t) \leq \frac{2^{d+1}}{t^d \rho} \exp\left(-n \frac{\rho t^d}{2}\right).$$

Hence if  $t_n(\alpha) < \min\{\rho/(2C_2), t_0\}$  is the solution to the equation

$$(17) \quad \frac{2^{d+1}}{t_n^d \rho} \exp\left(-n \frac{\rho t_n^d}{2}\right) = \alpha,$$

then

$$\mathbb{P}(W_\infty(\hat{\mathcal{P}}, \mathcal{P}) > t_n(\alpha)) \leq \mathbb{P}(\mathbf{H}(\mathcal{S}_n, \mathbb{M}) > t_n(\alpha)) \leq \alpha.$$

REMARKS. From the previous lemma, it follows that, setting  $t_n = \left(\frac{4}{\rho} \frac{\log n}{n}\right)^{1/d}$ ,

$$\mathbb{P}(\mathbf{H}(\mathcal{S}_n, \mathbb{M}) > t_n) \leq \frac{2^{d-1}}{n \log n},$$

for all  $n$  large enough. The right-hand side of (15) is known as the Lambert function [Lambert (1758)]. Equation (17) does not admit a closed form solution, but can be solved numerically.

To use the lemma, we need to estimate  $\rho$ . Let  $P_n$  be the empirical measure induced by the sample  $\mathcal{S}_n$ , given by

$$P_n(A) = \frac{1}{n} \sum_{i=1}^n I_A(X_i),$$

for any measurable Borel set  $A \subset \mathbb{R}^D$ . Let  $r_n$  be a positive small number and consider the plug-in estimator of  $\rho$ ,

$$(18) \quad \hat{\rho}_n = \min_i \frac{P_n(B(X_i, r_n/2))}{r_n^d}.$$

Our next result shows that, under our assumptions and provided that the sequence  $r_n$  vanishes at an appropriate rate as  $n \rightarrow \infty$ ,  $\hat{\rho}_n$  is a consistent estimator of  $\rho$ .

THEOREM 5. *Let  $r_n \asymp (\log n/n)^{1/(d+2)}$ . Then*

$$\hat{\rho}_n - \rho = O_P(r_n).$$

REMARK. We have assumed that  $d$  is known. It is also possible to estimate  $d$ , although we do not pursue that extension here.

We now need to use  $\hat{\rho}_n$  to estimate  $t_n(\alpha)$  as follows. Assume that  $n$  is even, and split the data randomly into two halves,  $\mathcal{S}_n = \mathcal{S}_{1,n} \sqcup \mathcal{S}_{2,n}$ . Let  $\hat{\rho}_{1,n}$  be the plug-in estimator of  $\rho$  computed from  $\mathcal{S}_{1,n}$ , and define  $\hat{t}_{1,n}$  to solve the equation

$$(19) \quad \frac{2^{d+1}}{t^d \hat{\rho}_{1,n}} \exp\left(-\frac{nt^d \hat{\rho}_{1,n}}{2}\right) = \alpha.$$

THEOREM 6. *Let  $\hat{\mathcal{P}}_2$  be the persistence diagram for the distance function to  $\mathcal{S}_{2,n}$ , then*

$$(20) \quad \begin{aligned} \mathbb{P}(W_\infty(\hat{\mathcal{P}}_2, \mathcal{P}) > \hat{t}_{1,n}) &\leq \mathbb{P}(\mathbf{H}(\mathcal{S}_{2,n}, \mathbb{M}) > \hat{t}_{1,n}) \\ &\leq \alpha + O\left(\frac{\log n}{n}\right)^{1/(2+d)}, \end{aligned}$$

where the probability  $\mathbb{P}$  is with respect to both the joint distribution of the entire sample and the randomness induced by the sample splitting.

In practice, we have found that solving (19) for  $\hat{t}_n$  without splitting the data also works well although we do not have a formal proof. Another way to define  $\hat{t}_n$  which is simpler but more conservative, is to define

$$(21) \quad \hat{t}_n = \left(\frac{2}{n\hat{\rho}_n} \log\left(\frac{n}{\alpha}\right)\right)^{1/d}.$$

Then  $\hat{t}_n = u_n(1 + O(\hat{\rho}_n - \rho))$  where  $u_n = (\frac{\rho}{n\hat{\rho}_n} \log(\frac{n}{\alpha}))^{1/d}$ , and so

$$\begin{aligned} \mathbb{P}(\mathbf{H}(\mathcal{S}_n, \mathbb{M}) > \hat{t}_n) &= \mathbb{P}(\mathbf{H}(\mathcal{S}_n, \mathbb{M}) > u_n) + O\left(\frac{\log n}{n}\right)^{1/(2+d)} \\ &\leq \alpha + O\left(\frac{\log n}{n}\right)^{1/(2+d)}. \end{aligned}$$

4.3. *Method III: The method of shells.* The dependence of the previous method on the parameter  $\rho$  makes the method very fragile. If the density is low in even a small region, then the method above is a disaster. Here

we develop a sharper bound based on shells of the form  $\{x : \gamma_j < \rho(x, \downarrow 0) < \gamma_{j+1}\}$  where we recall from (6) and (7) that

$$\rho(x, \downarrow 0) = \lim_{t \rightarrow 0} \rho(x, t) = \lim_{t \rightarrow 0} \frac{P(B(x, t/2))}{t^d}.$$

Let  $G(v) = P(\rho(X, \downarrow 0) \leq v)$ , and let  $g(v) = G'(v)$ .

**THEOREM 7.** *Suppose that  $g$  is bounded and has a uniformly bounded, continuous derivative. Then for any  $t \leq \rho/(2C_1)$ ,*

$$(22) \quad \begin{aligned} \mathbb{P}(W_\infty(\hat{\mathcal{P}}, \mathcal{P}) > t) &\leq P(\mathbf{H}(\mathcal{S}_n, \mathbb{M}) > t) \\ &\leq \frac{2^{d+1}}{t^d} \int_\rho^\infty \frac{g(v)}{v} e^{-nvt^d/2} dv. \end{aligned}$$

Let  $K$  be a smooth, symmetric kernel satisfying the conditions in Giné and Guillaou (2002) (which includes all the commonly used kernels), and let

$$(23) \quad \hat{g}(v) = \frac{1}{n} \sum_{i=1}^n \frac{1}{b} K\left(\frac{v - V_i}{b}\right),$$

where  $b > 0$ ,  $V_i = \hat{\rho}(X_i, r_n)$ , and

$$\hat{\rho}(x, r_n) = \frac{P_n(B(x, r_n/2))}{r_n^d}.$$

**THEOREM 8.** *Let  $r_n = (\frac{\log n}{n})^{1/(d+2)}$ .*

(1) *We have that*

$$\sup_v |\hat{g}(v) - g(v)| = O_P\left(b^2 + \sqrt{\frac{\log n}{nb}} + \frac{r_n}{b^2}\right).$$

Hence if we choose  $b \equiv b_n \asymp r_n^{1/4}$ , then

$$\sup_v |\hat{g}(v) - g(v)| = O_P\left(\frac{\log n}{n}\right)^{1/(2(d+2))}.$$

(2) *Suppose that  $n$  is even and that  $\rho > 0$ . Assume that  $g$  is strictly positive over its support  $[\rho, B]$ . Randomly split the data into two halves:  $\mathcal{S}_n = (\mathcal{S}_{1,n}, \mathcal{S}_{2,n})$ . Let  $\hat{g}_{1,n}$  and  $\hat{\rho}_{1,n}$  be estimators of  $g$  and  $\rho$ , respectively, computed from the first half of the data, and define  $\hat{t}_{1,n}$  to be the solution of the equation*

$$(24) \quad \frac{2^{d+1}}{\hat{t}_{1,n}^d} \int_{\hat{\rho}_n}^\infty \frac{\hat{g}(v)}{v} e^{-nvt^d/2} dv = \alpha.$$

Then

$$(25) \quad \mathbb{P}(W_\infty(\widehat{\mathcal{P}}_2, \mathcal{P}) > \hat{t}_{1,n}) \leq \mathbb{P}(\mathbf{H}(\mathcal{S}_{2,n}, \mathbb{M}) > \hat{t}_{1,n}) \leq \alpha + O(r_n),$$

where  $\widehat{\mathcal{P}}_2$  is the persistence diagram associated to  $\mathcal{S}_{2,n}$  and the probability  $\mathbb{P}$  is with respect to both the joint distribution of the entire sample and the randomness induced by the sample splitting.

4.4. *Method IV: Density estimation.* In this section, we take a completely different approach. We use the data to construct a smooth density estimator, and then we find the persistence diagram defined by a filtration of the upper level sets of the density estimator; see Figure 5. A different approach to smoothing based on diffusion distances is discussed in Bendich, Galkovskyi and Harer (2011).

Again, let  $X_1, \dots, X_n$  be a sample from  $P$ . Define

$$(26) \quad p_h(x) = \int_{\mathbb{M}} \frac{1}{h^D} K\left(\frac{\|x - u\|_2}{h}\right) dP(u).$$

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  $P$ . Our target of inference in this section is  $\mathcal{P}_h$ , the persistence diagram of the upper level sets of  $p_h$ . The standard estimator for  $p_h$  is the kernel density estimator

$$(27) \quad \hat{p}_h(x) = \frac{1}{n} \sum_{i=1}^n \frac{1}{h^D} K\left(\frac{\|x - X_i\|_2}{h}\right).$$

It is easy to see that  $\mathbb{E}(\hat{p}_h(x)) = p_h(x)$ . Let us now explain why  $\mathcal{P}_h$  is of interest.

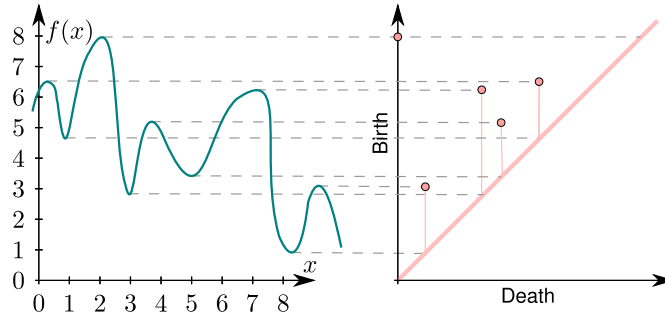


FIG. 5. We plot the persistence diagram corresponding to the upper level set filtration of the function  $f(x)$ . For consistency, we swap the birth and death axes so all persistence points appear above the line  $y = x$ . The point born first does not die, but for convenience we mark its death at  $f(x) = 0$ . This is analogous to the point marked with  $\infty$  in our previous diagrams.



First, the upper level sets of a density are of intrinsic interest in statistics in 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, first suppose that  $\mathbb{M}$  is a smooth, compact  $d$ -manifold, and suppose that  $P$  is supported on  $\mathbb{M}$ . Let  $p$  be the density of  $P$  with respect to Hausdorff measure on  $\mathbb{M}$ . In the special case where  $P$  is the uniform distribution, every upper level set  $\{p > t\}$  of  $p$  is identical to  $\mathbb{M}$  for  $t > 0$  and small enough.

Thus if  $\mathcal{P}$  is the persistence diagram defined by the upper level sets  $\{x: p(x) > t\}$  of  $p$ , and  $\mathcal{Q}$  is the persistence diagram of the distance function  $d_{\mathbb{M}}$ , then the points of  $\mathcal{Q}$  are in one-to-one correspondence with the generators of  $H(\mathbb{M})$ , and the points with higher persistence in  $\mathcal{P}$  are also in 1–1 correspondence with the generators of  $H(\mathbb{M})$ . For example, suppose that  $\mathbb{M}$  is a circle in the plane with radius  $\tau$ . Then  $\mathcal{Q}$  has two points: one at  $(0, \infty)$  representing a single connected component, and one at  $(0, \tau)$  representing the single cycle.  $\mathcal{P}$  also has two points: both at  $(0, 1/2\pi\tau)$  where  $1/2\pi\tau$  is simply the maximum of the density over the circle. In sum, these two persistence diagrams contain the same information; furthermore,  $\{x: p(x) > t\} \cong \mathbb{M}$  for all  $0 < t < 1/2\pi\tau$ .

If  $P$  is not uniform but has a smooth density  $p$ , bounded away from 0, then there is an interval  $[a, A]$  such that  $\{x: p(x) > t\} \cong \mathbb{M}$  (i.e., is homotopic) for  $a \leq t \leq A$ . Of course, one can create examples where no level sets are equal to  $\mathbb{M}$ , but it seems unlikely that any method can recover the homology of  $\mathbb{M}$  in those cases.

Next, suppose there is noise; that is, we observe  $Y_1, \dots, Y_n$ , where  $Y_i = X_i + \sigma\varepsilon_i$  and  $\varepsilon_1, \dots, \varepsilon_n \sim \Phi$ . We assume that  $X_1, \dots, X_n \sim Q$  where  $Q$  is supported on  $\mathbb{M}$ . Note that  $X_1, \dots, X_n$  are unobserved. Here,  $\Phi$  is the noise distribution and  $\sigma$  is the noise level. The distribution  $P$  of  $Y_i$  has density

$$(28) \quad p(y) = \int_{\mathbb{M}} \phi_{\sigma}(y - u) dQ(u),$$

where  $\phi$  is the density of  $\varepsilon_i$  and  $\phi_{\sigma}(z) = \sigma^{-D} \phi(y/\sigma)$ . In this case, no level set  $L_t = \{y: p(y) > t\}$  will equal  $\mathbb{M}$ . But as long as  $\phi$  is smooth and  $\sigma$  is small, there will be a range of values  $a \leq t \leq A$  such that  $L_t \cong \mathbb{M}$ .

The estimator  $\hat{p}_h(x)$  is consistent for  $p$  if  $p$  is continuous, as long as we let the bandwidth  $h = h_n$  change with  $n$  in such a way that  $h_n \rightarrow 0$  and  $nh_n \rightarrow \infty$  as  $n \rightarrow \infty$ . However, for exploring topology, we need not let  $h$  tend to 0. Indeed, more precise topological inference is obtained by using a bandwidth  $h > 0$ . Keeping  $h$  positive smooths the density, but the level sets can still retain the correct topological information.

We would also like to point out that the quantities  $\mathcal{P}_h$  and  $\widehat{\mathcal{P}}_h$  are more robust and much better behaved statistically than the Čech complex of the raw data. In the language of computational topology,  $\mathcal{P}_h$  can be considered a topological simplification of  $\mathcal{P}$ .  $\mathcal{P}_h$  may omit subtle details that are present in  $\mathcal{P}$  but is much more stable. For these reasons, we now focus on estimating  $\mathcal{P}_h$ .

Recall that, from the stability theorem,

$$(29) \quad W_\infty(\widehat{\mathcal{P}}_h, \mathcal{P}_h) \leq \|\hat{p}_h - p_h\|_\infty.$$

Hence it suffices to find  $c_n$  such that

$$(30) \quad \limsup_{n \rightarrow \infty} \mathbb{P}(\|\hat{p}_h - p_h\|_\infty > c_n) \leq \alpha.$$

*Finite sample band.* Suppose that the support of  $P$  is contained in  $\mathcal{X} = [-C, C]^D$ . Let  $p$  be the density of  $P$ . Let  $K$  be a kernel with the same assumptions as above, and choose a bandwidth  $h$ . Let

$$\hat{p}_h(x) = \frac{1}{n} \sum_{i=1}^n \frac{1}{h^D} K\left(\frac{\|x - X_i\|_2}{h}\right)$$

be the kernel density estimator, and let

$$p_h(x) = \frac{1}{h^D} \int_{\mathcal{X}} K\left(\frac{\|x - u\|_2}{h}\right) dP(u)$$

be the mean of  $\hat{p}_h$ .

LEMMA 9. Assume that  $\sup_x K(x) = K(0)$  and that  $K$  is  $L$ -Lipschitz, that is,  $|K(x) - K(y)| \leq L\|x - y\|_2$ . Then

$$(31) \quad \mathbb{P}(\|\hat{p}_h - p_h\|_\infty > \delta) \leq 2 \left( \frac{4CL\sqrt{D}}{\delta h^{D+1}} \right)^D \exp\left(-\frac{n\delta^2 h^{2D}}{2K^2(0)}\right).$$

REMARK. The proof of the above lemma uses Hoeffding's inequality. A sharper result can be obtained by using Bernstein's inequality; however, this introduces extra constants that need to be estimated.

We can use the above lemma to approximate the persistence diagram for  $p_h$ , denoted by  $\mathcal{P}_h$ , with the diagram for  $\hat{p}_h$ , denoted by  $\widehat{\mathcal{P}}_h$ :

COROLLARY 10. Let  $\delta_n$  solve

$$(32) \quad 2 \left( \frac{4CL\sqrt{D}}{\delta_n h^{D+1}} \right)^D \exp\left(-\frac{n\delta_n^2 h^{2D}}{2K^2(0)}\right) = \alpha.$$

Then

$$(33) \quad \sup_{P \in \mathcal{Q}} \mathbb{P}(W_\infty(\widehat{\mathcal{P}}_h, \mathcal{P}_h) > \delta_n) \leq \sup_{P \in \mathcal{Q}} \mathbb{P}(\|\hat{p}_h - p_h\|_\infty > \delta_n) \leq \alpha,$$

where  $\mathcal{Q}$  is the set of all probability measures supported on  $\mathcal{X}$ .

Now we consider a different finite sample band. Computationally, the persistent homology of the upper level sets of  $\hat{p}_h$  is actually based on a piecewise linear approximation to  $\hat{p}_h$ . We choose a finite grid  $G \subset \mathbb{R}^D$  and form a triangulation over the grid. Define  $\hat{p}_h^\dagger$  as follows. For  $x \in G$ , let  $\hat{p}_h^\dagger(x) = \hat{p}_h(x)$ . For  $x \notin G$ , define  $\hat{p}_h^\dagger(x)$  by linear interpolation over the triangulation. Let  $p_h^\dagger(x) = \mathbb{E}(\hat{p}_h^\dagger(x))$ . The real object of interest is the persistence diagram  $\mathcal{P}_h^\dagger$  of the upper level set filtration of  $p_h^\dagger(x)$ . We approximate this diagram with the persistence diagram  $\widehat{\mathcal{P}}_h^\dagger$  of the upper level set filtration of  $\hat{p}_h^\dagger(x)$ . As before,

$$W_\infty(\widehat{\mathcal{P}}_h^\dagger, \mathcal{P}_h^\dagger) \leq \|\hat{p}_h^\dagger - p_h^\dagger\|_\infty.$$

But due to the piecewise linear nature of these functions, we have that

$$\|\hat{p}_h^\dagger - p_h^\dagger\|_\infty \leq \max_{x \in G} |\hat{p}_h^\dagger(x) - p_h^\dagger(x)|.$$

LEMMA 11. *Let  $N = |G|$  be the size of the grid. Then*

$$(34) \quad \mathbb{P}(\|\hat{p}_h^\dagger - p_h^\dagger\|_\infty > \delta) \leq 2N \exp\left(-\frac{2n\delta^2 h^{2D}}{K^2(0)}\right).$$

Hence, if

$$(35) \quad \delta_n = \left(\frac{K(0)}{h}\right)^D \sqrt{\frac{1}{2n} \log\left(\frac{2N}{\alpha}\right)},$$

then

$$(36) \quad \mathbb{P}(\|\hat{p}_h^\dagger - p_h^\dagger\|_\infty > \delta_n) \leq \alpha.$$

This band can be substantially tighter as long we do not use a grid that is too fine. In a sense, we are rewarded for acknowledging that our topological inferences take place at some finite resolution.

*Asymptotic confidence band.* A tighter—albeit only asymptotic—bound can be obtained using large sample theory. The simplest approach is the bootstrap.

Let  $X_1^*, \dots, X_n^*$  be a sample from the empirical distribution  $P_n$ , and let  $\hat{p}_h^*$  denote the density estimator constructed from  $X_1^*, \dots, X_n^*$ . Define the random measure

$$(37) \quad J_n(t) = \mathbb{P}(\sqrt{nh^D} \|\hat{p}_h^* - \hat{p}_h\|_\infty > t | X_1, \dots, X_n)$$

and the bootstrap quantile  $Z_\alpha = \inf\{t : J_n(t) \leq \alpha\}$ .

THEOREM 12. As  $n \rightarrow \infty$ ,

$$\mathbb{P}\left(W_\infty(\hat{\mathcal{P}}_h, \mathcal{P}_h) > \frac{Z_\alpha}{\sqrt{nh^D}}\right) \leq \mathbb{P}(\sqrt{nh^D}\|\hat{p}_h - p_h\|_\infty > Z_\alpha) = \alpha + O\left(\sqrt{\frac{1}{n}}\right).$$

The proof follows from standard results; see, for example, Chazal et al. (2013a). As usual, we approximate  $Z_\alpha$  by Monte Carlo. Let  $T = \sqrt{nh^D}\|\hat{p}_h - \hat{p}_h^*\|_\infty$  be from a bootstrap sample. Repeat bootstrap  $B$  times yielding values  $T_1, \dots, T_B$ . Let

$$Z_\alpha = \inf\left\{z : \frac{1}{B} \sum_{j=1}^B I(T_j > z) \leq \alpha\right\}.$$

We can ignore the error due to the fact that  $B$  is finite since this error can be made as small as we like.

REMARK. We have emphasized fixed  $h$  asymptotics since, for topological inference, it is not necessary to let  $h \rightarrow 0$  as  $n \rightarrow \infty$ . However, it is possible to let  $h \rightarrow 0$  if one wants. Suppose  $h \equiv h_n$  and  $h \rightarrow 0$  as  $n \rightarrow \infty$ . We require that  $nh^D/\log n \rightarrow \infty$  as  $n \rightarrow \infty$ . As before, let  $Z_\alpha$  be the bootstrap quantile. It follows from Theorem 3.4 of Neumann (1998), that

$$\begin{aligned} \mathbb{P}\left(W_\infty(\hat{\mathcal{P}}_h, \mathcal{P}_h) > \frac{Z_\alpha}{\sqrt{nh^D}}\right) &\leq \mathbb{P}(\sqrt{nh^D}\|\hat{p}_h - p_h\|_\infty > Z_\alpha) \\ (38) \qquad \qquad \qquad &= \alpha + \left(\frac{\log n}{nh^D}\right)^{(4+D)/(2(2+D))}. \end{aligned}$$

OUTLIERS. Now we explain why the density-based method is very insensitive to outliers. Let  $P = \pi U + (1 - \pi)Q$  where  $Q$  is supported on  $\mathbb{M}$ ,  $\pi > 0$  is a small positive constant and  $U$  is a smooth distribution supported on  $\mathbb{R}^D$ . Apart from a rescaling, the bottleneck distance between  $\mathcal{P}_P$  and  $\mathcal{P}_Q$  is at most  $\pi$ . The kernel estimator is still a consistent estimator of  $p$ , and hence the persistence diagram is barely affected by outliers. In fact, in the examples in Bendich, Galkovskiy and Harer (2011), there are only a few outliers which formally corresponds to letting  $\pi = \pi_n \rightarrow 0$  as  $n \rightarrow \infty$ . In this case, the density method is very robust. We show this in more detail in the experiments section.

**5. Experiments.** As is common in the literature on computational topology, we focus on a few simple, synthetic examples. For each of them we compute the Rips persistence diagram and the density persistence diagram introduced in Section 4.4. We use a Gaussian kernel with bandwidth  $h = 0.3$ . This will serve to illustrate the different methods for the construction of confidence bands for the persistence diagrams.

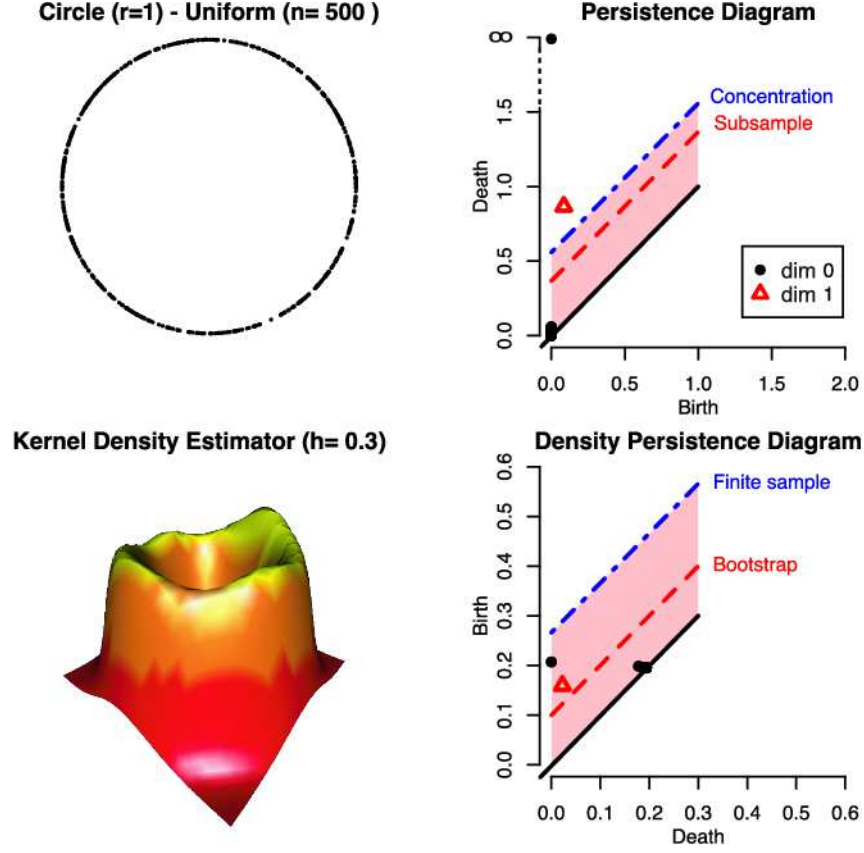


FIG. 6. Uniform distribution over the unit Circle. (Top left) sample  $S_n$ . (Top right) corresponding persistence diagram. The black circles indicate the life span of connected components, and the red triangles indicate the life span of 1-dimensional holes. (Bottom left) kernel density estimator. (Bottom right) density persistence diagram. For more details see Example 13.

EXAMPLE 13. Figure 6 shows the methods described in the previous sections applied to a sample from the uniform distribution over the unit circle ( $n = 500$ ). In the top right plot the different 95% confidence bands for the persistence diagram are computed using methods I (subsampling) and II (concentration of measure). Note that the uniform distribution does not satisfy the assumptions for the method of shells. The subsampling method and the concentration method both correctly show one significant connected component and one significant loop. In the bottom right plot the finite sample density estimation method and the bootstrap method are applied to the density persistence diagram. The first method does not have sufficient power to detect the topological features. However, the bootstrap density estimation method does find that one connected component and one loop are significant.

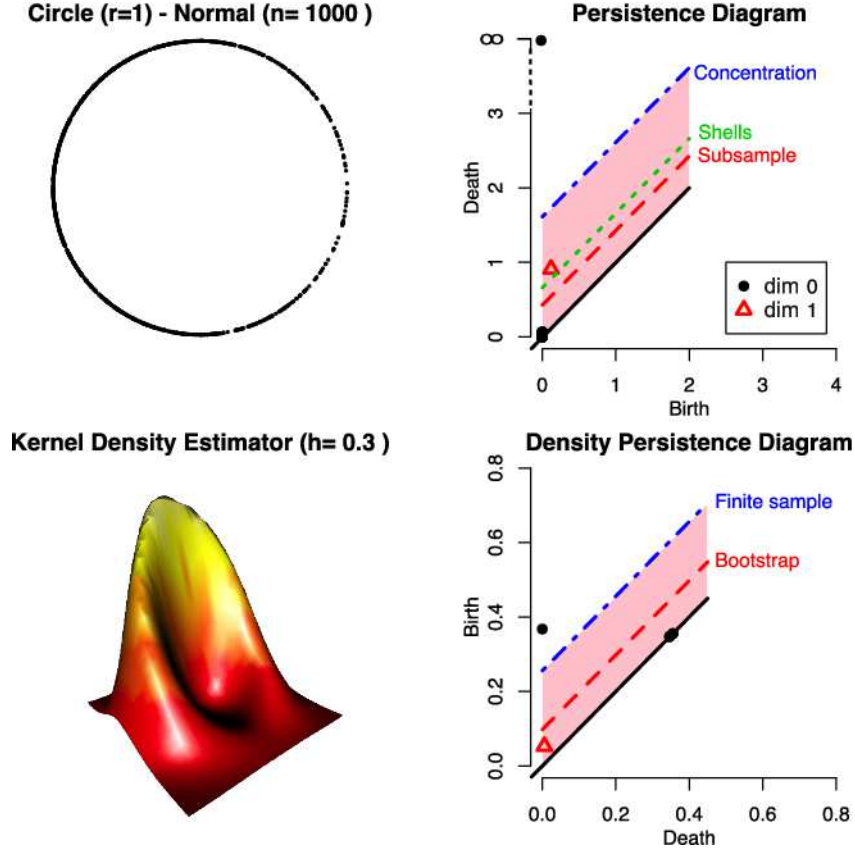


FIG. 7. Truncated Normal distribution over the unit Circle. (Top left) sample  $S_n$ . (Top right) corresponding persistence diagram. The black circles indicate the life span of connected components, and the red triangles indicate the life span of 1-dimensional holes. (Bottom left) kernel density estimator. (Bottom right) density persistence diagram. For more details see Example 14.

EXAMPLE 14. Figure 7 shows the methods described in the previous sections applied to a sample from the truncated Normal distribution over the unit circle ( $n = 1000$ ). The top left plot shows the sample, and the bottom left plot shows the kernel density estimator constructed using a Gaussian kernel with bandwidth  $h = 0.3$ . The plots on the left show the different methods for the construction of 95% confidence bands around the diagonal of the persistence diagrams. This case is challenging because there is a portion of the circle that is sparsely sampled. The concentration method fails to detect the loop, as shown in the top right plot. However, the method of shells and the subsampling method both declare the loop to be significant. The bottom right plot shows that both the finite sample method and the bootstrap method fail to detect the loop.

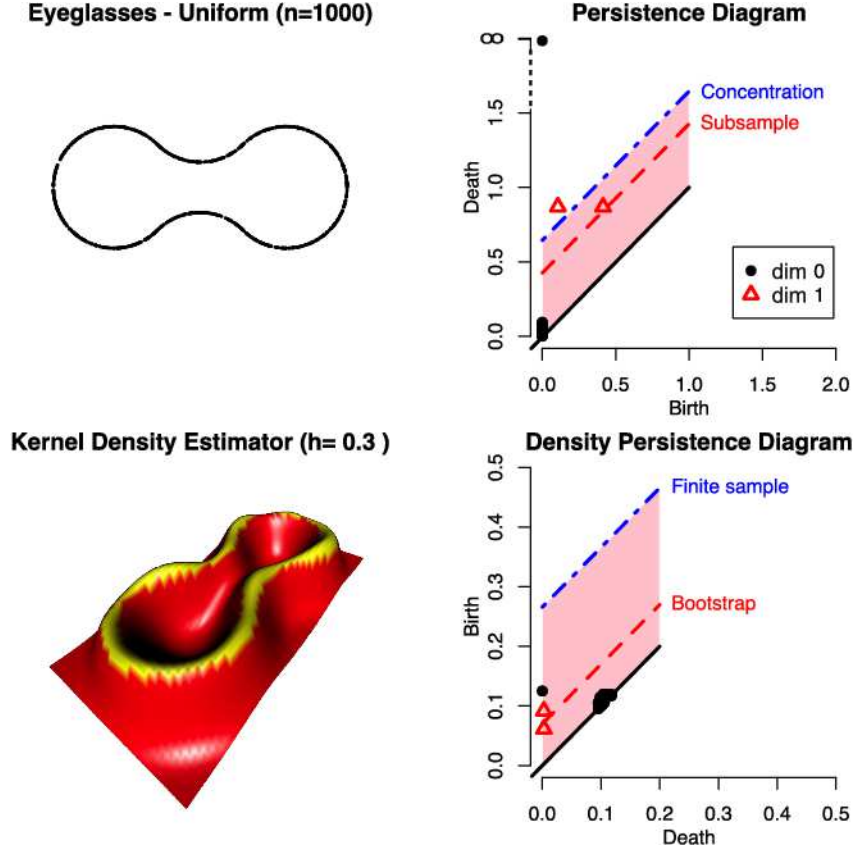


FIG. 8. Uniform distribution over the eyeglasses curve. (Top left) sample  $S_n$ . (Top right) corresponding persistence diagram. The black circles indicate the life span of connected components and the red triangles indicate the life span of 1-dimensional holes. Bottom left: kernel density estimator. (Bottom right) density persistence diagram. For more details see Example 15.

EXAMPLE 15. Figure 8 shows the methods described in the previous sections applied to a sample of size  $n = 1000$  from the uniform distribution over the eyeglasses, a figure similar to the Cassini Curve obtained by gluing together two unit circles. Note that the uniform distribution does not satisfy the assumptions for the method of shells. Each method provides a 95% confidence band around the diagonal for the persistence diagrams. The top right plot shows that the subsample method declares both the loops to be significant, while the concentration method detects only one of them, as the bootstrap method for the density persistence diagram, shown in the bottom right plot.

EXAMPLE 16. In this example, we replicate Examples 13 and 15, adding some outliers to the uniform distributions over the unit circle and the eye-

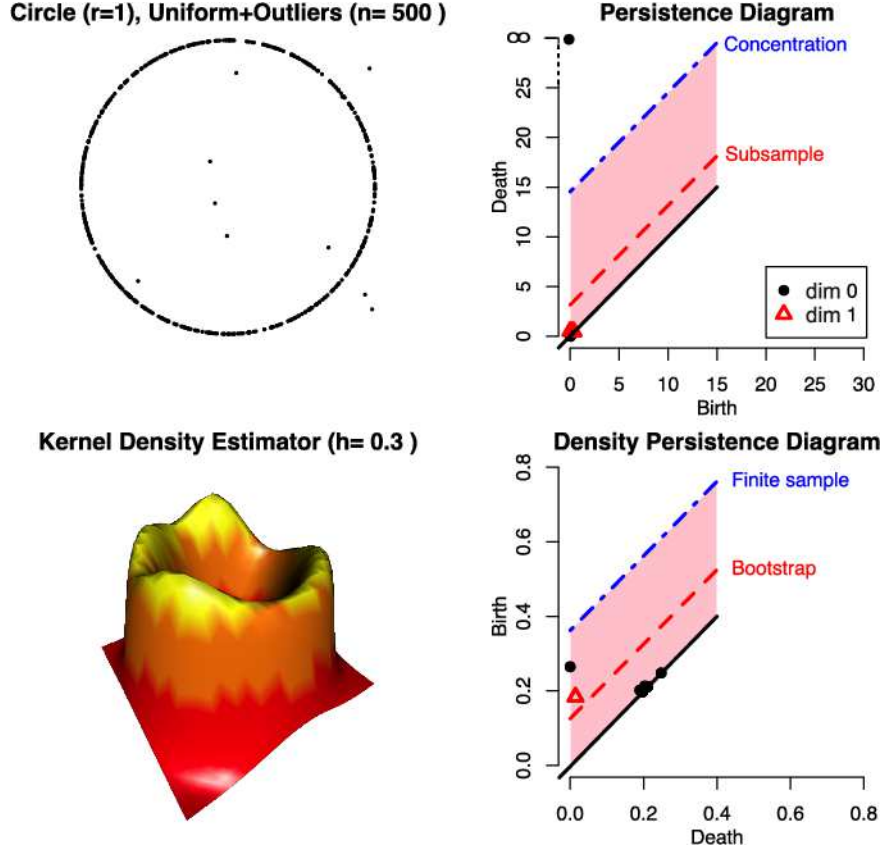


FIG. 9. Uniform distribution over the unit Circle with some outliers. See Example 16.

glasses. Figures 9 and 10 show the persistence diagrams with different methods for the construction of 95% confidence bands. Much of the literature on computational topology focuses on methods that use the distance function to the data. As we see here, and as discussed in Bendich, Galkovskyi and Harer (2011), such methods are quite fragile. A few outliers are sufficient to drastically change the persistence diagram and force the concentration method and the subsample method to declare the topological features to be not significant. On the other hand the density-based methods are very insensitive to the presence of outliers, as shown in the bottom right plots of the two figures.

**6. Proofs.** In this section, we provide proofs of the theorems and lemmas found in Section 3.

Recall that the  $\delta$ -covering number  $N$  of a manifold  $\mathbb{M}$  is the smallest number of Euclidean balls of radius  $\delta$  required to cover the set. The  $\delta$ -



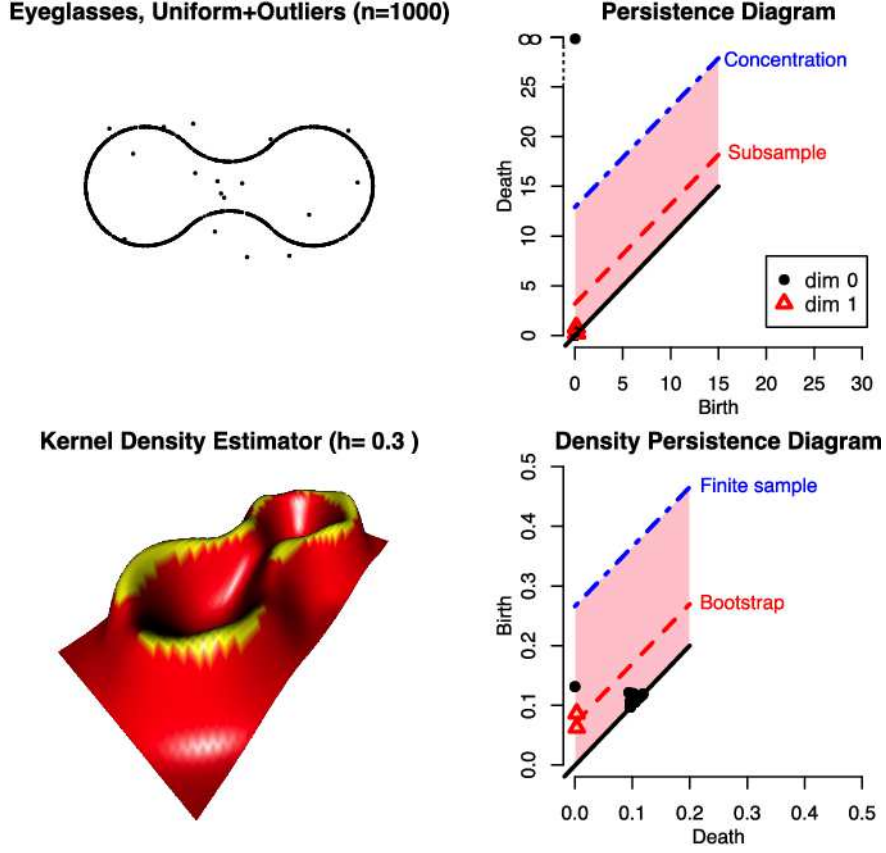


FIG. 10. Uniform distribution over the eyeglasses curve with some outliers. See Example 16.

packing number  $N'$  is the maximum number of sets of the form  $B(x, \delta) \cap \mathbb{M}$ , where  $x \in \mathbb{M}$ , that may be packed into  $\mathbb{M}$  without overlap. First, we prove the following lemma.

LEMMA 17. Let  $r_n = o(1)$  and let  $\mathcal{D}_n = \{d_1, \dots, d_{N'}\}$  be the set of centers of  $\{B(d_i, r_n) \cap \mathbb{M} : i = 1, \dots, N'\}$ , a  $r_n$ -packing set for  $\mathbb{M}$  (this set is nonempty and of cardinality  $N'$  increasing in  $n$ ). For large  $n$ , the size of the  $r_n$ -packing satisfies

$$(39) \quad \frac{\text{vol}(\mathbb{M})}{A_2(d)r_n^d} \leq N' \leq \frac{A_1(d)}{r_n^d},$$

for some constants  $A_1(d)$  and  $A_2(d)$  depending on the reach and dimension  $d$  of  $\mathbb{M}$ .

PROOF. From the proof of Lemma 4 we know that

$$(40) \quad N' \leq \frac{A_1(d)}{r_n^d},$$

where  $A_1(d)$  is a constant depending on the dimension of the manifold. A similar lower bound for  $N'$  can be obtained. Let  $N$  be the size of a  $2r_n$ -covering set for  $\mathbb{M}$ , formed by Euclidean balls  $B(c_i, 2r_n)$  with centers  $\mathcal{C} = \{c_1, \dots, c_N\}$ . By Lemma 5.2 in Niyogi, Smale and Weinberger (2008) and a simple volume argument we have

$$(41) \quad N' \geq N \geq \frac{\text{vol}(\mathbb{M})}{\max_{i=1, \dots, N} \text{vol}(B(c_i, 2r_n) \cap \mathbb{M})}.$$

For large  $n$ , by Corollary 1.3 in Chazal (2013),

$$(42) \quad \max_{i=1, \dots, N} \text{vol}(B(c_i, 2r_n) \cap \mathbb{M}) \leq A_2(d)r_n^d,$$

where  $A_2(d)$  is a constant depending on the dimension of the manifold. Combining (40), (41) and (42) we obtain (39).  $\square$

PROOF OF THEOREM 3. We begin the proof by showing that there exists an event of probability approaching one (as  $n \rightarrow \infty$ ) such that, over this event, for any subsample  $\mathcal{S}_{b,n}$  of the data  $\mathcal{S}_n$  of size  $b$ ,

$$(43) \quad \mathbf{H}(\mathcal{S}_n, \mathbb{M}) \leq \mathbf{H}(\mathcal{S}_{b,n}, \mathcal{S}_n).$$

Toward this end, let  $t_n = (\frac{4}{\rho} \frac{\log n}{n})^{1/d}$ , and define the event  $\mathcal{A}_n = \{\mathbf{H}(\mathcal{S}_n, \mathbb{M}) < t_n\}$ . Then, by the remark following Lemma 4,  $\mathbb{P}(\mathcal{A}_n^c) \leq \frac{2^{d-1}}{n \log n}$ , for all  $n$  large enough. Next, let  $\mathcal{D}_n = \{d_1, \dots, d_{N'}\}$  be the set of centers of  $\{B(d_i, 2t_n) \cap \mathbb{M} : i = 1, \dots, N'\}$ , a  $2t_n$ -packing set for  $\mathbb{M}$ . By Lemma 17, the size of  $\mathcal{D}_n$  is of order  $t_n^{-d} = \Theta(\frac{n}{\log n})$ .

We can now show (43). Suppose  $\mathcal{A}_n$  holds and, arguing by contradiction, also that  $\mathbf{H}(\mathcal{S}_{b,n}, \mathcal{S}_n) < \mathbf{H}(\mathcal{S}_n, \mathbb{M})$ . Then

$$(44) \quad \begin{aligned} \mathbf{H}(\mathcal{S}_{b,n}, \mathcal{D}_n) &\leq \mathbf{H}(\mathcal{S}_{b,n}, \mathcal{S}_n) + \mathbf{H}(\mathcal{S}_n, \mathcal{D}_n) < \mathbf{H}(\mathcal{S}_n, \mathbb{M}) + \mathbf{H}(\mathcal{S}_n, \mathcal{D}) \\ &\leq 2\mathbf{H}(\mathcal{S}_n, \mathbb{M}) \\ &\leq 2t_n. \end{aligned}$$

Because of our assumption on  $b$  and of (39),  $\frac{b}{N'} \rightarrow 0$  as  $n \rightarrow \infty$ , which implies that a  $(1 - o(1))$  fraction of the balls  $\{B(d_j, 2t_n), j = 1, \dots, N'\}$  contains no points from  $\mathcal{S}_{b,n}$ . So  $\mathbf{H}(\mathcal{S}_{b,n}, \mathcal{D}) > 2t_n$ , which, in light of (44), yields a contradiction. Thus, on  $\mathcal{A}_n$ , (43) holds, for any subsample  $\mathcal{S}_{b,n}$  of size  $b$ , as claimed.

Next, let  $\{\mathcal{S}_{b,n}^j, j = 1, \dots, N\}$  be an enumeration of all possible subsamples of  $\mathcal{S}_n$  of size  $b$ , where  $N = \binom{n}{b}$ , and define

$$\tilde{L}_b(t) = \frac{1}{N} \sum_{j=1}^N I(\mathbf{H}(\mathcal{S}_{b,n}^j, \mathbb{M}) > t).$$

Using (43) we obtain that, on the event  $\mathcal{A}_n$ ,

$$\mathbf{H}(\mathcal{S}_b^j, \mathbb{M}) \leq \mathbf{H}(\mathcal{S}_b^j, \mathcal{S}_n) + \mathbf{H}(\mathcal{S}_n, \mathbb{M}) \leq 2\mathbf{H}(\mathcal{S}_b^j, \mathcal{S}_n),$$

and therefore,  $\mathbf{H}(\mathcal{S}_b^j, \mathcal{S}_n) \geq \mathbf{H}(\mathcal{S}_b^j, \mathbb{M})/2$ , simultaneously over all  $j = 1, \dots, N$ . Thus, on that event

$$(45) \quad \tilde{L}_b(t) \leq L_b(t/2) \quad \text{for all } t > 0,$$

where  $L_b$  is defined in (13). Thus, letting  $\mathbf{1}_{\mathcal{A}_n} = \mathbf{1}_{\mathcal{A}_n}(\mathcal{S}_n)$  be the indicator function of the event  $\mathcal{A}_n$ , we obtain the bound

$$\tilde{L}_b(c_b) = \tilde{L}_b(c_b)(\mathbf{1}_{\mathcal{A}_n} + \mathbf{1}_{\mathcal{A}_n^c}) \leq L_b(c_b/2) + \mathbf{1}_{\mathcal{A}_n^c} \leq \alpha + \mathbf{1}_{\mathcal{A}_n^c},$$

where the first inequality is due to (45) and the second inequality to the fact that  $L_b(c_b/2) = \alpha$ , by definition of  $c_b$ . Taking expectations, we obtain that

$$(46) \quad \mathbb{E}(\tilde{L}_b(c_b)) \leq \alpha + \mathbb{P}(\mathcal{A}_n^c) = \alpha + O\left(\frac{1}{n \log n}\right).$$

Next, define

$$J_b(t) = \mathbb{P}(\mathbf{H}(\mathcal{S}_b, \mathbb{M}) > t), \quad t > 0,$$

where we recall that  $\mathcal{S}_b$  is an i.i.d. sample of size  $b$ . Then Lemma A.2 in Romano and Shaikh (2012) yields that, for any  $\epsilon > 0$ ,

$$(47) \quad \mathbb{P}\left(\sup_{t>0} |\tilde{L}_b(t) - J_b(t)| > \epsilon\right) \leq \frac{1}{\epsilon} \sqrt{\frac{2\pi}{k_n}},$$

where  $k_n = \lfloor \frac{n}{b} \rfloor$ . Let  $\mathcal{B}_n$  be the event that

$$\sup_{t>0} |\tilde{L}_b(t) - J_b(t)| \leq \frac{\sqrt{2\pi}}{k_n^{1/4}}$$

and  $\mathbf{1}_{\mathcal{B}_n} = \mathbf{1}_{\mathcal{B}_n}(\mathcal{S}_n)$  be its indicator function. Then

$$\mathbb{E}\left(\sup_{t>0} |\tilde{L}_b(t) - J_b(t)| \mathbf{1}_{\mathcal{B}_n}\right) \leq \frac{\sqrt{2\pi}}{k_n^{1/4}},$$

and, using (47) and the fact that  $\sup_t |\tilde{L}_b(t) - J_b(t)| \leq 1$  almost everywhere,

$$\mathbb{E}\left(\sup_{t>0} |\tilde{L}_b(t) - J_b(t)| \mathbf{1}_{\mathcal{B}_n^c}\right) \leq \mathbb{P}(\mathcal{B}_n^c) \leq \frac{1}{k_n^{1/4}}.$$

Thus

$$(48) \quad \mathbb{E} \left( \sup_{t>0} |\tilde{L}_b(t) - J_b(t)| \right) = O \left( \frac{1}{k_n^{1/4}} \right) = O \left( \frac{b}{n} \right)^{1/4}.$$

We can now prove the claim of the theorem. First notice that the following bounds hold:

$$\begin{aligned} \mathbb{P}(\mathbf{H}(\mathcal{S}_n, \mathbb{M}) > c_b) &\leq \mathbb{P}(\mathbf{H}(\mathcal{S}_b, \mathbb{M}) > c_b) \\ &= J_b(c_b) \\ &\leq \tilde{L}_b(c_b) + \sup_{t>0} |\tilde{L}_b(t) - J_b(t)|. \end{aligned}$$

Now take expectations on both sides, and use (46) and (48) to obtain that

$$\mathbb{P}(\mathbf{H}(\mathcal{S}_n, \mathbb{M}) > c_b) \leq O \left( \frac{b}{n} \right)^{1/4} + O \left( \frac{1}{n \log n} \right) = O \left( \frac{b}{n} \right)^{1/4},$$

as claimed.  $\square$

**PROOF OF LEMMA 4.** Let  $\mathcal{C} = \{c_1, \dots, c_N\}$  be the set of centers of Euclidean balls  $\{B_1, \dots, B_N\}$ , forming a minimal  $t/2$ -covering set for  $\mathbb{M}$ , for  $t/2 < \text{diam}(\mathbb{M})$ . Then  $\mathbf{H}(\mathcal{C}, \mathbb{M}) \leq t/2$  and

$$\begin{aligned} \mathbb{P}(\mathbf{H}(\mathcal{S}_n, \mathbb{M}) > t) &\leq \mathbb{P}(\mathbf{H}(\mathcal{S}_n, \mathcal{C}) + \mathbf{H}(\mathcal{C}, \mathbb{M}) > t) \\ &\leq \mathbb{P}(\mathbf{H}(\mathcal{S}_n, \mathcal{C}) > t/2) \\ &= \mathbb{P}(B_j \cap \mathcal{S}_n = \emptyset \text{ for some } j) \\ &\leq \sum_j P(B_j \cap \mathcal{S}_n = \emptyset) \\ &= \sum_j [1 - P(B_j)]^n \\ &\leq N[1 - \rho(t)t^d]^n \\ &\leq N \exp(-n\rho(t)t^d), \end{aligned}$$

where the second-to-last inequality follows from the fact that  $\min_j P(B_j) \geq \rho(t)t^d$  by definition of  $\rho(t)$  and the last inequality from the fact that  $\rho(t)t^d \leq 1$ . Next, let  $\mathcal{D} = \{d_1, \dots, d_{N'}\}$  be the set of centers of  $\{B'_1 \cap \mathbb{M}, \dots, B'_{N'} \cap \mathbb{M}\}$ , a maximal  $t/4$ -packing set for  $\mathbb{M}$ . Then  $N \leq N'$  [see, e.g., Lemma 5.2, Niyogi, Smale and Weinberger (2008), for a proof of this standard fact], and by definition, the balls  $\{B'_j \cap \mathbb{M}, j = 1, \dots, N'\}$  are disjoint. Therefore,

$$1 = P(\mathbb{M}) \geq \sum_{j=1}^{N'} P(B'_j \cap \mathbb{M}) = \sum_{j=1}^{N'} P(B'_j) \geq N' \rho(t/2) \frac{t^d}{2^d},$$

where we have used again the fact that  $\min_j P(B'_j) \geq \rho(t/2) \frac{t^d}{2^d}$ . We conclude that  $N \leq N' \leq (2^d)/(\rho(t/2)t^d)$ . Hence

$$\mathbb{P}(\mathbf{H}(\mathcal{S}_n, \mathbb{M}) > t) \leq \frac{2^d}{\rho(t/2)t^d} \exp(-n\rho(t)t^d).$$

Now suppose that  $t < \min\{\rho/(2C_2), t_0\}$  ( $C_2$  and  $t_0$  are defined in Assumption A2). Then, since we assume that  $\rho(t)$  is differentiable on  $[0, t_0]$  with derivative bounded in absolute value by  $C_2$ , there exists a  $0 \leq \tilde{t} \leq t$  such that

$$\rho(t) = \rho + t\rho'(\tilde{t}) \geq \rho - C_2 t \geq \frac{\rho}{2}.$$

Similarly, under the same conditions on  $t$ , we have that  $\rho(t/2) \geq \frac{\rho}{2}$ . The result follows.  $\square$

PROOF OF THEOREM 5. Let

$$\hat{\rho}(x, t) = \frac{P_n(B(x, t/2))}{t^d}.$$

Note that

$$(49) \quad \sup_{x \in \mathbb{M}} P(B(x, r_n/2)) \leq C r_n^d$$

for some  $C > 0$ , since  $\rho(x, t)$  is bounded by Assumption A2. Let  $r_n = (\frac{\log n}{n})^{1/(d+2)}$ , and consider all  $n$  large enough so that

$$(50) \quad \frac{\rho}{2}(\log n)^{d/(d+2)} > 1 \quad \text{and} \quad n^{2/(d+2)} - \frac{2}{d+2} \log n > 0.$$

Let  $\mathcal{E}_{1,n}$  be the event that the sample  $\mathcal{S}_n$  forms an  $r_n$ -cover for  $\mathbb{M}$ . Then, by Lemma 4 and since  $\mathbb{P}(\mathcal{E}_{1,n}^c) = \mathbb{P}(\mathbf{H}(\mathcal{S}_n, \mathbb{M}) \leq r_n)$ , we have

$$\begin{aligned} \mathbb{P}(\mathcal{E}_{1,n}^c) &\leq \frac{2^{d+1}}{\rho} \left( \frac{n}{\log n} \right)^{d/(d+2)} \exp \left\{ -\frac{\rho}{2} n \left( \frac{\log n}{n} \right)^{d/(d+2)} \right\} \\ &\leq \frac{2^{d+1}}{\rho} n^{d/(d+2)} \exp \left\{ -\left( \frac{\rho}{2} (\log n)^{d/(d+2)} \right) n^{2/(d+2)} \right\} \\ &\leq \frac{2^{d+1}}{\rho} n^{d/(d+2)} \exp \{ -n^{2/(d+2)} \} \\ &\leq \frac{2^{d+1}}{\rho} \exp \left\{ -n^{2/(d+2)} + \frac{d}{d+2} \log n \right\} \\ &\leq \frac{2^{d+1}}{\rho} \frac{1}{n}, \end{aligned}$$

where the third and last inequalities hold since  $n$  is assumed large enough to satisfy (50).

Let  $C_3 = \max\{C_1, C_2\}$  where  $C_1$  and  $C_2$  are defined in Assumption A2. Let

$$\epsilon_n = \sqrt{\bar{C} \frac{2 \log n}{(n-1)r_n^d}}$$

with  $\bar{C} = \max\{4C_3, 2(C+1/3)\}$ , for some  $C$  satisfying (49). Assume further that  $n$  is large enough so that, in addition to (50),  $\epsilon_n < 1$ . With this choice of  $\epsilon_n$ , define the event

$$\mathcal{E}_{2,n} = \left\{ \max_{i=1, \dots, n} \left| \frac{P_{i,n-1}(B(X_i, r_n/2))}{r_n^d} - \frac{P(B(X_i, r_n/2))}{r_n^d} \right| \leq c^* \epsilon_n \right\},$$

where  $P_{i,n-1}$  is the empirical measure corresponding to the data points  $\mathcal{S}_n \setminus \{X_i\}$ , and  $c^*$  is a positive number satisfying

$$c^* \epsilon_n r_n^d - \frac{1}{n} \geq \epsilon_n r_n^d$$

for all  $n$  large enough (which exists by our choice of  $\epsilon_n$  and  $r_n$ ). We will show that  $\mathbb{P}(\mathcal{E}_{2,n}) \geq 1 - \frac{2}{n}$ . To this end, let  $\mathbb{P}^i$  denote the probability induced by  $\mathcal{S}_n \setminus \{X_i\}$  (which, by independence is also the conditional probability of  $\mathcal{S}_n$  given  $X_i$ ) and  $\mathbb{P}_i$  be the marginal probability induced by  $X_i$ , for  $i = 1, \dots, n$ . Then

$$\begin{aligned} \mathbb{P}(\mathcal{E}_{2,n}^c) &\leq \sum_{i=1}^n \mathbb{P} \left( \left| \frac{P_n(B(X_i, r_n/2))}{r_n^d} - \frac{P(B(X_i, r_n/2))}{r_n^d} \right| > c^* \epsilon_n \right) \\ &\leq \sum_{i=1}^n \mathbb{P} \left( |P_{i,n-1}(B(X_i, r_n/2)) - P(B(X_i, r_n/2))| > c^* \epsilon_n r_n^d - \frac{1}{n} \right) \\ (51) \quad &\leq \sum_{i=1}^n \mathbb{P}(|P_{i,n-1}(B(X_i, r_n/2)) - P(B(X_i, r_n/2))| > \epsilon_n r_n^d) \\ &= \sum_{i=1}^n \mathbb{E}_i[\mathbb{P}^i(|P_{i,n-1}(B(X_i, r_n/2)) - P(B(X_i, r_n/2))| > \epsilon_n r_n^d)], \end{aligned}$$

where the first inequality follows from the union bound, the second from the fact that  $|P_n(B(X_i, r_n/2)) - P_{i,n-1}(B(X_i, r_n/2))| \leq \frac{1}{n}$  for all  $i$  (almost everywhere with respect to the join distribution of the sample) and the third inequality from the definition of  $c^*$ .

By Bernstein's inequality, for any  $i = 1, \dots, n$ ,

$$\mathbb{P}^i(|P_{i,n-1}(B(X_i, r_n/2)) - P(B(X_i, r_n/2))| > \epsilon_n r_n^d)$$

$$\begin{aligned}
&\leq 2 \exp \left\{ -\frac{1}{2} \frac{(n-1)\epsilon_n^2 r_n^{2d}}{C r_n^d + 1/3 r_n^d \epsilon_n} \right\} \\
&\leq 2 \exp \left\{ -\frac{1}{2} \frac{(n-1)\epsilon_n^2 r_n^d}{(C + 1/3)} \right\} \\
&\leq \frac{2}{n^2},
\end{aligned}$$

where in the first inequality we have used the fact that  $P(B(x, r_n/2))(1 - P(B(x, r_n/2))) \leq C r_n^d$ . Therefore, from (51),

$$\mathbb{P}(\mathcal{E}_{2,n}^c) \leq \frac{2}{n}.$$

Let  $j = \arg \min_i \hat{\rho}(X_i, r_n)$  and  $k = \arg \min_i \rho(X_i, r_n)$ . Suppose  $\mathcal{E}_{2,n}$  holds and, arguing by contradiction, that

$$(52) \quad |\hat{\rho}(X_j, r_n) - \rho(X_k, r_n)| = |\hat{\rho}_n - \rho(X_k, r_n)| > c^* \epsilon_n.$$

Since  $\mathcal{E}_{2,n}$  holds, we have  $|\hat{\rho}(X_j, r_n) - \rho(X_j, r_n)| \leq c^* \epsilon_n$  and  $|\hat{\rho}(X_k, r_n) - \rho(X_k, r_n)| \leq c^* \epsilon_n$ . This implies that if  $\hat{\rho}(X_j, r_n) < \rho(X_k, r_n)$ , then  $\rho(X_j, r_n) < \rho(X_k, r_n)$ , while if  $\hat{\rho}(X_j, r_n) > \rho(X_k, r_n)$ , then  $\hat{\rho}(X_k, r_n) < \hat{\rho}(X_j, r_n)$ , which is a contradiction.

Therefore, with probability  $\mathbb{P}(\mathcal{E}_{1,n} \cap \mathcal{E}_{2,n}) \geq 1 - \frac{(2^{d+1}/\rho)+2}{n}$ , the sample points  $\mathcal{S}_n$  forms an  $r_n$ -covering of  $\mathbb{M}$  and

$$\left| \hat{\rho}_n - \min_i \rho(X_i, r_n) \right| \leq c^* \epsilon_n.$$

Since the sample  $\mathcal{S}_n$  is a covering of  $\mathbb{M}$ ,

$$(53) \quad \left| \min_i \rho(X_i, r_n) - \inf_{x \in \mathbb{M}} \rho(x, r_n) \right| \leq \max_i \sup_{x \in B(X_i, r_n)} |\rho(x, r_n) - \rho(X_i, r_n)|.$$

Because  $\rho(x, t)$  has a bounded continuous derivative in  $t$  uniformly over  $x$ , we have, if  $r_n < t_0$ ,

$$\sup_{x \in B(X_i, r_n)} |\rho(x, r_n) - \rho(X_i, r_n)| \leq C_3 r_n,$$

almost surely. Furthermore, since  $\rho(t)$  is right-differentiable at zero,

$$|\rho(r_n) - \rho| \leq C_3 r_n,$$

for all  $r_n < t_0$ . Combining the last two observations with (53) and using the triangle inequality, we conclude that

$$|\hat{\rho}_n - \rho| \leq c^* \epsilon_n + 2C_3 r_n,$$

with probability at least  $1 - \frac{(2^{d+1}/\rho)+2}{n}$ , for all  $n$  large enough. Because our choice of  $r_n$  satisfies the equation  $\epsilon_n = \Theta(\sqrt{\frac{\log n}{nr_n^d}})$ , the terms on the right-hand side of the last display are balanced, so that

$$|\hat{\rho}_n - \rho| \leq C_4 \left( \frac{\log n}{n} \right)^{1/(d+2)},$$

for some  $C_4 > 0$ .  $\square$

**PROOF OF THEOREM 6.** Let  $\mathbb{P}_1$  denote the unconditional probability measure induced by the first random half  $\mathcal{S}_{1,n}$  of the sample,  $\mathbb{P}_2$  the conditional probability measure induced by the second half of the sample  $\mathcal{S}_{2,n}$  given the outcome of the data splitting and the values of the first sample, and  $\mathbb{P}_{1,2}$  the probability measure induced by the whole sample and the random splitting. Then

$$\mathbb{P}(\mathbf{H}(\mathcal{S}_{2,n}, \mathbb{M}) > \hat{t}_{1,n}) = \mathbb{P}_{1,2}(\mathbf{H}(\mathcal{S}_{2,n}, \mathbb{M}) > \hat{t}_{1,n}) = \mathbb{E}_1(\mathbb{P}_2(\mathbf{H}(\mathcal{S}_{2,n}, \mathbb{M}) > \hat{t}_{1,n})),$$

where  $\mathbb{E}_1$  denotes the expectation corresponding to  $\mathbb{P}_1$ . By Theorem 5, there exist constants  $C$  and  $C'$  such that the event  $\mathcal{A}_n$  that  $|\hat{\rho}_{1,n} - \rho| \leq C(\frac{\log n}{n})^{1/(d+2)}$  has  $\mathbb{P}_1$ -probability no smaller than  $1 - \frac{C'}{n}$ , for  $n$  large enough. Then

$$\begin{aligned} (54) \quad & \mathbb{E}_1(\mathbb{P}_2(\mathbf{H}(\mathcal{S}_{2,n}, \mathbb{M}) > \hat{t}_{1,n})) \\ & \leq \mathbb{E}_1(\mathbb{P}_2(\mathbf{H}(\mathcal{S}_{2,n}, \mathbb{M}) > \hat{t}_{1,n}); \mathcal{A}_n) + \mathbb{P}_1(\mathcal{A}_n^c), \end{aligned}$$

where, for a random variable  $X$  with expectation  $\mathbb{E}_X$  and an event  $\mathcal{E}$  measurable with respect to  $X$ , we write  $\mathbb{E}[X; \mathcal{E}]$  for the expectation of  $X$  restricted to the sample points in  $\mathcal{E}$ .

Define  $F(t, \rho) = \frac{2^{d+1}}{t^d \rho} e^{-((\rho n)/2)t^d}$ . Then, by Lemma 4,

$$\mathbb{P}_2(\mathbf{H}(\mathcal{S}_{2,n}, \mathbb{M}) > \hat{t}_{1,n}) \leq F(\hat{t}_{1,n}, \rho).$$

The rest of the proof is devoted to showing that, on  $\mathcal{A}_n$ ,  $F(\hat{t}_{1,n}, \rho) \leq \alpha + O((\log n/2)^{1/(d+2)})$ . To simplify the derivation, we will write  $O(R_n)$  to indicate a term that is in absolute value of order  $O((\log n/n)^{1/(d+2)})$ , where the exact value of the constant may change from line to line. Accordingly, on the event  $\mathcal{A}_n$  and for  $n$  large enough,  $\hat{\rho}_{1,n} - \rho = O(R_n)$ . Since  $\rho > 0$  by assumption, this implies that, on the same event and for  $n$  large,

$$\frac{\hat{\rho}_{1,n}}{\rho} = 1 + O(R_n) \quad \text{and} \quad \frac{\rho}{\hat{\rho}_{1,n}} = 1 + O(R_n).$$



Now, on  $\mathcal{A}_n$  and for all  $n$  large enough,

$$\begin{aligned}
 F(\hat{t}_{1,n}, \rho) &= \frac{2^{d+1}}{\hat{t}_{1,n}^d \rho} \exp\left(-\frac{n\hat{t}_{1,n}^d \rho}{2}\right) \\
 &= \left(\frac{\hat{\rho}_{1,n}}{\rho}\right) \frac{2^{d+1}}{\hat{t}_{1,n}^d \hat{\rho}_{1,n}} \exp\left(-\frac{n\hat{t}_{1,n}^d \hat{\rho}_{1,n}(\rho/\hat{\rho}_{1,n})}{2}\right) \\
 (55) \quad &= (1 + O(R_n)) F(\hat{t}_{1,n}, \hat{\rho}_{1,n}) \exp\left(-\frac{n\hat{t}_{1,n}^d \hat{\rho}_{1,n} O(R_n)}{2}\right) \\
 &= \alpha(1 + O(R_n)) \left[\exp\left(-\frac{n\hat{t}_{1,n}^d \hat{\rho}_{1,n}}{2}\right)\right]^{O(R_n)} \\
 &= \alpha(1 + O(R_n)) \left[\frac{\alpha}{2} \hat{t}_{1,n}^d \hat{\rho}_{1,n}\right]^{O(R_n)},
 \end{aligned}$$

where the last two identities follow from the fact that

$$(56) \quad F(\hat{t}_{1,n}, \hat{\rho}_{1,n}) = \alpha$$

for all  $n$ .

Next, let  $t_n^* = (\frac{2}{\alpha\rho} \frac{\log n}{n})^{1/d}$ . We then claim that, for all  $n$  large enough and on the event  $\mathcal{A}_n$ ,  $\hat{t}_{1,n} \leq t_n^*$ . In fact, using similar arguments,

$$F(t_n^*, \hat{\rho}_{1,n}) = F(t_n^*, \rho)(1 + O(R_n)) \exp(-nt_n^* O(R_n)) \rightarrow 0 \quad \text{as } n \rightarrow \infty,$$

since  $O(R_n) = o(1)$  and, by Lemma 4,  $F(t_n^*, \rho) \rightarrow 0$  and  $n \rightarrow \infty$ .

By (56), it then follows that, for all  $n$  large enough,  $F(\hat{t}_{1,n}, \hat{\rho}_{1,n}) > F(t_n^*, \hat{\rho}_{1,n})$ . Because  $F(t, \rho)$  is decreasing in  $t$  for each  $\rho$ , the claim is proved.

Thus, substituting  $\hat{t}_{1,n}$  with  $t_n^*$  in equation (55) yields

$$\begin{aligned}
 F(\hat{t}_{1,n}, \rho) &\leq F(t_n^*, \rho) \\
 &\leq \alpha(1 + O(R_n)) \left[\frac{\alpha}{2} (t_n^*)^d \hat{\rho}_{1,n}\right]^{O(R_n)} \\
 &= \alpha(1 + O(R_n)) \left[\frac{\alpha}{2} (t_n^*)^d \rho(1 + O(R_n))\right]^{O(R_n)} \\
 &= \alpha(1 + O(R_n)) \left[\frac{\log n}{n} + o(1)\right]^{O(R_n)} \\
 &= \alpha(1 + O(R_n))(1 + o(1)) \\
 &= \alpha + O(R_n),
 \end{aligned}$$

as  $n \rightarrow \infty$ , where we have written as  $o(1)$  all terms that are lower order than  $O(R_n)$ . The second-to-last step follows from the limit

$$\begin{aligned} \lim_{n \rightarrow \infty} \left( \frac{\log n}{n} \right)^{O(R_n)} &= \lim_{n \rightarrow \infty} \exp \left\{ \log \left( \frac{\log n}{n} \right) C \left( \frac{\log n}{n} \right)^{1/(d+2)} \right\} \\ &= \exp \left\{ \lim_{n \rightarrow \infty} \log \left( \frac{\log n}{n} \right) C \left( \frac{\log n}{n} \right)^{1/(d+2)} \right\} = 1, \end{aligned}$$

for some constant  $C$ .

Therefore, on the event  $\mathcal{A}_n$  and for all  $n$  large enough,

$$\mathbb{P}_2(\mathbf{H}(\mathcal{S}_{2,n}, \mathbb{M}) > \hat{t}_{1,n}) \leq \alpha + O(R_n).$$

Since  $\mathbb{P}_1(\mathcal{A}_n) = O(R_n)$  [in fact, it is of lower order than  $O(R_n)$ ], the result follows from (54).  $\square$

**PROOF OF THEOREM 7.** Let  $B = \sup_{x \in \mathbb{M}} \rho(x, \downarrow 0)$ . Fix some  $\delta \in (0, B - \rho)$ . Choose equally spaced values  $\rho \equiv \gamma_1 < \gamma_2 < \dots < \gamma_m < \gamma_{m+1} \equiv B$  such that  $\delta \geq \gamma_{j+1} - \gamma_j$ . Let  $\Omega_j = \{x : \gamma_j \leq \rho(x, \downarrow 0) < \gamma_{j+1}\}$ , and define  $h(\mathcal{S}_n, \Omega_j) = \sup_{y \in \Omega_j} \min_{x \in \mathcal{S}_n} \|x - y\|_2$  for  $j = 1, \dots, m$ . Now  $\mathbf{H}(\mathcal{S}_n, \mathbb{M}) = h(\mathcal{S}_n, \mathbb{M}) \leq \max_j h(\mathcal{S}_n, \Omega_j)$ , and so

$$\mathbb{P}(\mathbf{H}(\mathcal{S}_n, \mathbb{M}) > t) \leq \mathbb{P}\left(\max_j h(\mathcal{S}_n, \Omega_j) > t\right) \leq \sum_{j=1}^m \mathbb{P}(h(\mathcal{S}_n, \Omega_j) > t).$$

Let  $C_j = \{c_{j1}, \dots, c_{jN_j}\}$  be the set of centers of  $\{B(c_{j1}, t/2), \dots, B(c_{jN_j}, t/2)\}$ , a  $t/2$ -covering set for  $\Omega_j$ , for  $j = 1, \dots, m$ . Let  $B_{jk} = B(c_{jk}, t/2)$ , for  $k = 1, \dots, N_j$ . Then, for all  $j$  and for  $t < \min\{\gamma_j/(2C_1), t_0\}$  ( $C_1$  and  $t_0$  are defined in Assumption A2), there is  $0 \leq \tilde{t} \leq t$  such that

$$\frac{P(B_{jk})}{t^d} = \rho(c_{jk}, t) \geq \rho(c_{jk}, \downarrow 0) + t\rho'(c_{jk}(\tilde{t})) \geq \rho(c_{jk}, \downarrow 0) - C_1 t \geq \frac{\gamma_j}{2}$$

and

$$\begin{aligned} \mathbb{P}(h(\mathcal{S}_n, \Omega_j) > t) &\leq \mathbb{P}(h(\mathcal{S}_n, C_j) + h(C_j, \Omega_j) > t) \\ &\leq \mathbb{P}(h(\mathcal{S}_n, C_j) > t/2) \\ &\leq \mathbb{P}(B_{jk} \cap \mathcal{S}_n = \emptyset \text{ for some } k) \\ &\leq \sum_{k=1}^{N_j} \mathbb{P}(B_{jk} \cap \mathcal{S}_n = \emptyset) \\ &\leq \sum_{k=1}^{N_j} [1 - P(B_{jk})]^n \leq \sum_{k=1}^{N_j} \left[1 - t^d \frac{\gamma_j}{2}\right]^n \\ &= N_j \left[1 - t^d \frac{\gamma_j}{2}\right]^n \leq N_j \exp\left(-\frac{n\gamma_j t^d}{2}\right). \end{aligned}$$

Following the strategy in the proof of Lemma 4,

$$P(\Omega_j) \geq \sum_k P(B_{jk}) \geq N_j t^d \frac{\gamma_j}{2^{d+1}},$$

so that  $N_j \leq 2^{d+1} P(\Omega_j) / (\gamma_j t^d)$ . Therefore, for  $t < \min\{\rho/(2C_1), t_0\}$ ,

$$\begin{aligned} \mathbb{P}(\mathbf{H}(\mathcal{S}_n, \mathbb{M}) > t) &\leq \frac{2^{d+1}}{t^d} \sum_{j=1}^m \frac{P(\Omega_j)}{\gamma_j} \exp\left(-\frac{n\gamma_j t^d}{2}\right) \\ &= \frac{2^{d+1}}{t^d} \sum_{j=1}^m \frac{G(\gamma_j + \delta) - G(\gamma_j)}{\delta} \frac{\delta}{\gamma_j} \exp\left(-\frac{n\gamma_j t^d}{2}\right) \\ &\rightarrow \frac{2^{d+1}}{t^d} \int_{\rho}^B \frac{g(v)}{v} \exp\left(-\frac{nv t^d}{2}\right) dv \end{aligned}$$

as  $\delta \rightarrow 0$ .  $\square$

PROOF OF THEOREM 8. Let

$$(57) \quad g^*(v) = \frac{1}{n} \sum_{i=1}^n \frac{1}{b} K\left(\frac{v - W_i}{b}\right),$$

where  $W_i = \rho(X_i, \downarrow 0)$ . Then

$$\sup_v |\hat{g}(v) - g(v)| \leq \sup_v |g(v) - g^*(v)| + \sup_v |g^*(v) - \hat{g}(v)|.$$

By standard asymptotics for kernel estimators,

$$\sup_v |g(v) - g^*(v)| \leq C_1 b^2 + O_P\left(\sqrt{\frac{\log n}{nb}}\right).$$

Next, note that

$$(58) \quad \left| K\left(\frac{v - W_i}{b}\right) - K\left(\frac{v - V_i}{b}\right) \right| \leq \frac{C|W_i - V_i|}{b} \preceq \frac{r_n}{b}$$

from Theorem 5. Hence,

$$|g^*(v) - \hat{g}(v)| \preceq \frac{r_n}{b^2}.$$

Statement (1) follows.

To show the second statement, we will use the same notation and a similar strategy as in the proof of Theorem 6. Thus

$$\begin{aligned} \mathbb{P}(\mathbf{H}(\mathcal{S}_{2,n}, \mathbb{M}) > \hat{t}_{1,n}) &= \mathbb{P}_{1,2}(\mathbf{H}(\mathcal{S}_{2,n}, \mathbb{M}) > \hat{t}_{1,n}) \\ &= \mathbb{E}_1(\mathbb{P}_2(\mathbf{H}(\mathcal{S}_{2,n}, \mathbb{M}) > \hat{t}_{1,n})). \end{aligned}$$

Let  $\mathcal{A}_n$  be the event defined in the proof of Theorem 6 and  $\mathcal{B}_n$  the event that  $\sup_v |\hat{g}(v) - g(v)| \leq r_n$ . Then  $\mathbb{P}_1(\mathcal{A}_n \cap \mathcal{B}_n) \geq 1 - O(1/n)$ . Now,

$$\begin{aligned} \mathbb{E}_1(\mathbb{P}_2(\mathbf{H}(\mathcal{S}_{2,n}, \mathbb{M}) > \hat{t}_{1,n})) \\ \leq \mathbb{E}_1(\mathbb{P}_2(\mathbf{H}(\mathcal{S}_{2,n}, \mathbb{M}) > \hat{t}_{1,n}); \mathcal{A}_n \cap \mathcal{B}_n) + \mathbb{P}_1((\mathcal{A}_n \cap \mathcal{B}_n)^c). \end{aligned}$$

By Theorem 7, conditionally on  $\mathcal{S}_{1,n}$  and the randomness of the data splitting,

$$(59) \quad \mathbb{P}_2(\mathbf{H}(\mathcal{S}_{2,n}, \mathbb{M}) > \hat{t}_{1,n}) \leq \frac{2^{d+1}}{\hat{t}_{1,n}^d} \int_{\rho}^{\infty} \frac{g(v)}{v} e^{-nv(\hat{t}_{1,n}^d/2)} dv.$$

We will show next that, on the event  $\mathcal{A}_n \cap \mathcal{B}_n$ , the right-hand side of the previous equation is bounded by  $\alpha + O(r_n)$  as  $n \rightarrow \infty$ . The second claim of the theorem will then follow. Throughout the proof, we will assume that the event  $\mathcal{A}_n \cap \mathcal{B}_n$  holds.

Recall that  $\hat{t}_{1,n}$  solves the equation

$$(60) \quad \frac{2^{d+1}}{\hat{t}_{1,n}^d} \int_{\rho}^{\infty} \frac{\hat{g}(v)}{v} e^{-nv(\hat{t}_{1,n}^d/2)} dv = \alpha$$

(and this solution exists for all large  $n$ ). By assumption,  $g(v)$  is uniformly bounded away from 0. Hence  $g(v)/\hat{g}(v) = 1 + O(r_n)$  and so

$$\begin{aligned} & \frac{2^{d+1}}{\hat{t}_{1,n}^d} \int_{\rho}^{\infty} \frac{g(v)}{v} e^{-nv(\hat{t}_{1,n}^d/2)} dv \\ &= \frac{2^{d+1}}{\hat{t}_{1,n}^d} \int_{\hat{\rho}_n}^B \frac{g(v)}{v} e^{-nv(\hat{t}_{1,n}^d/2)} dv + z_n \\ &= (1 + O(r_n)) \frac{2^{d+1}}{\hat{t}_{1,n}^d} \int_{\hat{\rho}_n}^B \frac{\hat{g}(v)}{v} e^{-nv(\hat{t}_{1,n}^d/2)} dv + z_n \\ &= \alpha(1 + O(r_n)) + O(\hat{\rho}_n - \rho) + z_n \\ &= \alpha + O(r_n) + z_n, \end{aligned}$$

where

$$z_n = \frac{2^{d+1}}{\hat{t}_{1,n}^d} \int_{\rho}^{\hat{\rho}_n} \frac{g(v)}{v} e^{-nv(\hat{t}_{1,n}^d/2)} dv.$$

Now  $\hat{t}_{1,n} \geq c_1(\log n/n)^{1/d} \equiv u_n$  for some  $c > 0$ , for all large  $n$  since otherwise the left-hand side of (60) would go to zero, and (60) would not be satisfied. So, for some positive  $c_2, c_3$ ,

$$z_n \leq \frac{2^{d+1}}{u_n^d} \int_{\rho}^{\hat{\rho}_n} \frac{g(v)}{\rho} e^{-n\rho(u_n^d/2)} dv \leq |\hat{\rho}_n - \rho| \frac{c_2}{n^{c_3}} = O(|\hat{\rho}_n - \rho|). \quad \square$$

PROOF OF LEMMA 9. First we show that

$$|p_h(x) - p_h(y)| \leq \frac{L}{h^{D+1}} \|x - y\|_2.$$

This follows from the definition of  $p_h$ ,

$$\begin{aligned} & |p_h(x) - p_h(y)| \\ &= \left| \frac{1}{h^D} \int_{\mathcal{X}} K\left(\frac{\|x - u\|}{h}\right) dP(u) - \frac{1}{h^D} \int_{\mathcal{X}} K\left(\frac{\|y - u\|}{h}\right) dP(u) \right| \\ &\leq \frac{1}{h^D} \int_{\mathcal{X}} \left| K\left(\frac{\|x - u\|}{h}\right) - K\left(\frac{\|y - u\|}{h}\right) \right| dP(u) \\ &\leq \frac{1}{h^D} \int_{\mathcal{X}} L \left| \frac{\|x - u\|}{h} - \frac{\|y - u\|}{h} \right| dP(u) \\ &\leq \frac{L}{h^D} \frac{\|x - y\|}{h} = \frac{L}{h^{D+1}} \|x - y\|. \end{aligned}$$

By a similar argument,

$$|\hat{p}_h(x) - \hat{p}_h(y)| \leq \frac{L}{h^{D+1}} \|x - y\|.$$

Divide  $\mathcal{X}$  into a grid based on cubes  $A_1, \dots, A_N$  with length of size  $\varepsilon$ . Note that  $N \asymp (C/\varepsilon)^D$ . Each cube  $A_j$  has diameter  $\sqrt{D}\varepsilon$ . Let  $c_j$  be the center of  $A_j$ . Now

$$\begin{aligned} \|\hat{p}_h - p_h\|_{\infty} &= \sup_x |\hat{p}_h(x) - p_h(x)| = \max_j \sup_{x \in A_j} |\hat{p}_h(x) - p_h(x)| \\ &\leq \left( \max_j |\hat{p}_h(c_j) - p_h(c_j)| \right) + 2v, \end{aligned}$$

where  $v = \frac{L\varepsilon\sqrt{D}}{2h^{D+1}}$ . We have that

$$\begin{aligned} \mathbb{P}(\|\hat{p}_h - p_h\|_{\infty} > \delta) &\leq \mathbb{P}\left(\max_j |\hat{p}_h(c_j) - p_h(c_j)| > \delta - 2v\right) \\ &\leq \sum_j \mathbb{P}(|\hat{p}_h(c_j) - p_h(c_j)| > \delta - 2v). \end{aligned}$$

Let  $\varepsilon = \frac{\delta h^{D+1}}{2L\sqrt{D}}$ . Then  $2v = \delta/2$ , and so

$$\mathbb{P}(\|\hat{p}_h - p_h\|_{\infty} > \delta) \leq \sum_j \mathbb{P}\left(|\hat{p}_h(c_j) - p_h(c_j)| > \frac{\delta}{2}\right).$$

Note that  $\hat{p}_h(x)$  is an average of quantities bounded between zero and  $K(0)/(nh^D)$ . Hence, by Hoeffding's inequality,

$$\mathbb{P}\left(|\hat{p}_h(c_j) - p_h(c_j)| > \frac{\delta}{2}\right) \leq 2 \exp\left(-\frac{n\delta^2 h^{2D}}{2K^2(0)}\right).$$

Therefore, summing over  $j$ , we conclude

$$\begin{aligned}
\mathbb{P}(\|\hat{p}_h - p_h\|_\infty > \delta) &\leq 2N \exp\left(-\frac{n\delta^2 h^{2D}}{2K^2(0)}\right) \\
&= 2\left(\frac{C}{\varepsilon}\right)^D \exp\left(-\frac{n\delta^2 h^{2D}}{2K^2(0)}\right) \\
&= 2\left(\frac{4CL\sqrt{D}}{\delta h^{D+1}}\right)^D \exp\left(-\frac{n\delta^2 h^{2D}}{2K^2(0)}\right). \quad \square
\end{aligned}$$

**7. Conclusion.** We have presented several methods for separating noise from signal in persistent homology. The first three methods are based on the distance function to the data. The last uses density estimation. There is a useful analogy here: methods that use the distance function to the data are like statistical methods that use the empirical distribution function. Methods that use density estimation use smoothing. The advantage of the former is that it is more directly connected to the raw data. The advantage of the latter is that it is less fragile; that is, it is more robust to noise and outliers.

We conclude by mentioning some open questions that we plan to address in future work:

(1) We focus on assessing the uncertainty of persistence diagrams. Similar ideas can be applied to assess uncertainty in barcode plots. This requires assessing uncertainty at different scales  $\varepsilon$ . This suggests examining the variability of  $\mathbf{H}(\hat{S}_\varepsilon, S_\varepsilon)$  at different values of  $\varepsilon$ . From Molchanov (1998), we have that

$$(61) \quad \sqrt{n\varepsilon^D} \mathbf{H}(\hat{S}_\varepsilon, S_\varepsilon) \rightsquigarrow \inf_{x \in \partial S_\varepsilon} \left| \frac{\mathbb{G}(x)}{L(x)} \right|,$$

where  $\mathbb{G}$  is a Gaussian process,

$$(62) \quad L(x) = \frac{d}{dt} \inf\{p_\varepsilon(y) - p_\varepsilon(x) : \|x - y\| \leq t\} \Big|_{t=0}$$

and  $p_\varepsilon(x)$  is the mean of the kernel estimator using a spherical kernel with bandwidth  $\varepsilon$ . The limiting distribution it is not helpful for inference because it would be very difficult to estimate  $L(x)$ . We are investigating practical methods for constructing confidence intervals on  $\mathbf{H}(\hat{S}_\varepsilon, S_\varepsilon)$  and using this to assess uncertainty of the barcodes.

(2) Confidence intervals provide protection against type I errors, that is, false detections. It is also important to investigate the power of the methods to detect real topological features. Similarly, we would like to quantify the minimax bounds for persistent homology.

(3) In the density estimation method, we use a fixed bandwidth. Spatially adaptive bandwidths might be useful for more refined inferences.

(4) It is also of interest to construct confidence intervals for other topological parameters such as *the degree  $p$  total persistence* defined by

$$\theta = 2 \sum d(x, \text{Diag})^p,$$

where the sum is over the points in  $\mathcal{P}$  whose distance from the diagonal is greater than some threshold, and  $\text{Diag}$  denotes the diagonal.

(5) Our experiments are meant to be a proof of concept. Detailed simulation studies are needed to see under which conditions the various methods work well.

(6) The subsampling method is very conservative due to the fact that  $b = o(n)$ . Essentially, there is a bias of order  $H(\mathcal{S}_b, \mathbb{M}) - H(\mathcal{S}_n, \mathbb{M})$ . We conjecture that it is possible to adjust for this bias.

(7) The optimal bandwidth for the density estimation method is an open question. The usual theory is based on  $L_2$  loss which is not necessarily appropriate for topological estimation.

**Acknowledgments.** The authors would like to thank the anonymous reviewers for their suggestions and feedback, as well as Frédéric Chazal and Peter Landweber for their comments.

## SUPPLEMENTARY MATERIAL

### Supplement to “Confidence sets for persistence diagrams”

(DOI: [10.1214/14-AOS1252SUPP](https://doi.org/10.1214/14-AOS1252SUPP); .pdf). In the supplementary material we give a brief introduction to persistence homology and provide additional details about homology, simplicial complexes and stability of persistence diagrams.

## REFERENCES

- AMBROSIO, L., FUSCO, N. and PALLARA, D. (2000). *Functions of Bounded Variation and Free Discontinuity Problems. Oxford Mathematical Monographs.* Clarendon Press, New York. [MR1857292](#)
- BALAKRISHNAN, S., RINALDO, A., SHEEHY, D., SINGH, A. and WASSERMAN, L. (2011). Minimax rates for homology inference. Preprint. Available at [arXiv:1112.5627](#).
- BENDICH, P., GALKOVSKIY, T. and HARER, J. (2011). Improving homology estimates with random walks. *Inverse Problems* **27** 124002. [MR2854318](#)
- BENDICH, P., MUKHERJEE, S. and WANG, B. (2010). Towards stratification learning through homology inference. Preprint. Available at [arXiv:1008.3572](#).
- BLUMBERG, A. J., GAL, I., MANDELL, M. A. and PANCIA, M. (2012). Persistent homology for metric measure spaces, and robust statistics for hypothesis testing and confidence intervals. Preprint. Available at [arXiv:1206.4581](#).

- BUBENIK, P. and KIM, P. T. (2007). A statistical approach to persistent homology. *Homology, Homotopy Appl.* **9** 337–362. [MR2366953](#)
- BUBENIK, P., CARLSSON, G., KIM, P. T. and LUO, Z.-M. (2010). Statistical topology via Morse theory persistence and nonparametric estimation. In *Algebraic Methods in Statistics and Probability II. Contemp. Math.* **516** 75–92. Amer. Math. Soc., Providence, RI. [MR2730741](#)
- CARLSSON, G. (2009). Topology and data. *Bull. Amer. Math. Soc. (N.S.)* **46** 255–308. [MR2476414](#)
- CARLSSON, G. and ZOMORODIAN, A. (2009). The theory of multidimensional persistence. *Discrete Comput. Geom.* **42** 71–93. [MR2506738](#)
- CHAZAL, F. (2013). An upper bound for the volume of geodesic balls in submanifolds of Euclidean spaces. Technical report, INRIA.
- CHAZAL, F. and OUDOT, S. Y. (2008). Towards persistence-based reconstruction in Euclidean spaces. In *Computational Geometry (SCG'08)* 232–241. ACM, New York. [MR2504289](#)
- CHAZAL, F., COHEN-STEINER, D., MÉRIGOT, Q. et al. (2010). Geometric inference for measures based on distance functions. INRIA RR-6930.
- CHAZAL, F., OUDOT, S., SKRABA, P. and GUIBAS, L. J. (2011). Persistence-based clustering in Riemannian manifolds. In *Computational Geometry (SCG'11)* 97–106. ACM, New York. [MR2919600](#)
- CHAZAL, F., DE SILVA, V., GLISSE, M. and OUDOT, S. (2012). The structure and stability of persistence modules. Preprint. Available at [arXiv:1207.3674](#).
- CHAZAL, F., FASY, B. T., LECCI, F., RINALDO, A., SINGH, A. and WASSERMAN, L. (2013a). On the bootstrap for persistence diagrams and landscapes. Preprint. Available at [arXiv:1311.0376](#).
- CHAZAL, F., GLISSE, M., LABRUÈRE, C. and MICHEL, B. (2013b). Optimal rates of convergence for persistence diagrams in topological data analysis. Preprint. Available at [arXiv:1305.6239](#).
- COHEN-STEINER, D., EDELSBRUNNER, H. and HARER, J. (2007). Stability of persistence diagrams. *Discrete Comput. Geom.* **37** 103–120. [MR2279866](#)
- COHEN-STEINER, D., EDELSBRUNNER, H., HARER, J. and MILEYKO, Y. (2010). Lipschitz functions have  $L_p$ -stable persistence. *Found. Comput. Math.* **10** 127–139. [MR2594441](#)
- CUEVAS, A. (2009). Set estimation: Another bridge between statistics and geometry. *Bol. Estad. Investig. Oper.* **25** 71–85. [MR2750781](#)
- CUEVAS, A., FEBRERO, M. and FRAIMAN, R. (2001). Cluster analysis: A further approach based on density estimation. *Comput. Statist. Data Anal.* **36** 441–459. [MR1855727](#)
- CUEVAS, A. and FRAIMAN, R. (1997). A plug-in approach to support estimation. *Ann. Statist.* **25** 2300–2312. [MR1604449](#)
- CUEVAS, A. and FRAIMAN, R. (1998). On visual distances in density estimation: The Hausdorff choice. *Statist. Probab. Lett.* **40** 333–341. [MR1664548](#)
- CUEVAS, A., FRAIMAN, R. and PATEIRO-LÓPEZ, B. (2012). On statistical properties of sets fulfilling rolling-type conditions. *Adv. in Appl. Probab.* **44** 311–329. [MR2977397](#)
- CUEVAS, A. and RODRÍGUEZ-CASAL, A. (2004). On boundary estimation. *Adv. in Appl. Probab.* **36** 340–354. [MR2058139](#)
- DEVROYE, L. and WISE, G. L. (1980). Detection of abnormal behavior via nonparametric estimation of the support. *SIAM J. Appl. Math.* **38** 480–488. [MR0579432](#)
- EDELSBRUNNER, H. and HARER, J. (2008). Persistent homology—a survey. In *Surveys on Discrete and Computational Geometry. Contemp. Math.* **453** 257–282. Amer. Math. Soc., Providence, RI. [MR2405684](#)



- EDELSBRUNNER, H. and HARER, J. L. (2010). *Computational Topology: An Introduction*. Amer. Math. Soc., Providence, RI. [MR2572029](#)
- FASY, B. T., LECCI, F., RINALDO, A., WASSERMAN, L., BALAKRISHNAN, S. and SINGH, A. (2014). Supplement to “Confidence sets for persistence diagrams.” DOI:[10.1214/14-AOS1252SUPP](#).
- FEDERER, H. (1959). Curvature measures. *Trans. Amer. Math. Soc.* **93** 418–491. [MR0110078](#)
- GHRIST, R. (2008). Barcodes: The persistent topology of data. *Bull. Amer. Math. Soc. (N.S.)* **45** 61–75. [MR2358377](#)
- GINÉ, E. and GUILLOU, A. (2002). Rates of strong uniform consistency for multivariate kernel density estimators. *Ann. Inst. Henri Poincaré Probab. Stat.* **38** 907–921. [MR1955344](#)
- HATCHER, A. (2002). *Algebraic Topology*. Cambridge Univ. Press, Cambridge. [MR1867354](#)
- HEO, G., GAMBLE, J. and KIM, P. T. (2012). Topological analysis of variance and the maxillary complex. *J. Amer. Statist. Assoc.* **107** 477–492. [MR2980059](#)
- JOSHI, S., KOMMARAJU, R. V., PHILLIPS, J. M. and VENKATASUBRAMANIAN, S. (2011). Comparing distributions and shapes using the kernel distance. In *Computational Geometry (SCG’11)* 47–56. ACM, New York. [MR2919594](#)
- KAHLE, M. (2009). Topology of random clique complexes. *Discrete Math.* **309** 1658–1671. [MR2510573](#)
- KAHLE, M. (2011). Random geometric complexes. *Discrete Comput. Geom.* **45** 553–573. [MR2770552](#)
- KAHLE, M. and MECKES, E. (2013). Limit theorems for Betti numbers of random simplicial complexes. *Homology, Homotopy Appl.* **15** 343–374. [MR3079211](#)
- LAMBERT, J. H. (1758). Observationes variae in mathesin puram. *Acta Helveticae Physico-mathematico-anatomico-botanico-medica* **III** 128–168.
- MAMMEN, E. and TSYBAKOV, A. B. (1995). Asymptotical minimax recovery of sets with smooth boundaries. *Ann. Statist.* **23** 502–524. [MR1332579](#)
- MATTLA, P. (1995). *Geometry of Sets and Measures in Euclidean Spaces: Fractals and Rectifiability*. Cambridge Studies in Advanced Mathematics **44**. Cambridge Univ. Press, Cambridge. [MR1333890](#)
- MILEYKO, Y., MUKHERJEE, S. and HARER, J. (2011). Probability measures on the space of persistence diagrams. *Inverse Problems* **27** 124007, 22. [MR2854323](#)
- MOLCHANOV, I. S. (1998). A limit theorem for solutions of inequalities. *Scand. J. Stat.* **25** 235–242. [MR1614288](#)
- NEUMANN, M. H. (1998). Strong approximation of density estimators from weakly dependent observations by density estimators from independent observations. *Ann. Statist.* **26** 2014–2048. [MR1673288](#)
- NIYOGI, P., SMALE, S. and WEINBERGER, S. (2008). Finding the homology of submanifolds with high confidence from random samples. *Discrete Comput. Geom.* **39** 419–441. [MR2383768](#)
- NIYOGI, P., SMALE, S. and WEINBERGER, S. (2011). A topological view of unsupervised learning from noisy data. *SIAM J. Comput.* **40** 646–663. [MR2810909](#)
- PENROSE, M. (2003). *Random Geometric Graphs*. Oxford Studies in Probability **5**. Oxford Univ. Press, Oxford. [MR1986198](#)
- POLITIS, D. N., ROMANO, J. P. and WOLF, M. (1999). *Subsampling*. Springer, New York. [MR1707286](#)
- ROMANO, J. P. and SHAIKH, A. M. (2012). On the uniform asymptotic validity of subsampling and the bootstrap. *Ann. Statist.* **40** 2798–2822. [MR3097960](#)

TURNER, K., MILEYKO, Y., MUKHERJEE, S. and HARER, J. (2014). Fréchet Means for Distributions of Persistence Diagrams. *Discrete Comput. Geom.* **52** 44–70. [MR3231030](#)

B. T. FASY  
COMPUTER SCIENCE DEPARTMENT  
TULANE UNIVERSITY  
NEW ORLEANS, LOUISIANA 70118  
USA  
E-MAIL: [brittany.fasy@alumni.duke.edu](mailto:brittany.fasy@alumni.duke.edu)

F. LECCI  
A. RINALDO  
L. WASSERMAN  
DEPARTMENT OF STATISTICS  
CARNEGIE MELLON UNIVERSITY  
PITTSBURGH, PENNSYLVANIA 15213  
USA  
E-MAIL: [lecci@cmu.edu](mailto:lecci@cmu.edu)  
[arinaldo@cmu.edu](mailto:arinaldo@cmu.edu)  
[larry@cmu.edu](mailto:larry@cmu.edu)

S. BALAKRISHNAN  
A. SINGH  
COMPUTER SCIENCE DEPARTMENT  
CARNEGIE MELLON UNIVERSITY  
PITTSBURGH, PENNSYLVANIA 15213  
USA  
E-MAIL: [sbalakri@cs.cmu.edu](mailto:sbalakri@cs.cmu.edu)  
[aarti@cs.cmu.edu](mailto:aarti@cs.cmu.edu)