

# Topological consistency via kernel estimation

OMER BOBROWSKI<sup>1</sup>, SAYAN MUKHERJEE<sup>2</sup> and JONATHAN E. TAYLOR<sup>3</sup>

<sup>1</sup>*Department of Mathematics, Duke University, Durham NC 27708, USA. E-mail: omer@math.duke.edu*

<sup>2</sup>*Departments of Statistical Science, Computer Science, and Mathematics, Duke University, Durham NC 27708, USA. E-mail: sayan@stat.duke.edu*

<sup>3</sup>*Department of Statistics, Stanford University, Stanford, CA 94305-4065, USA.  
E-mail: jonathan.taylor@stanford.edu*

We introduce a consistent estimator for the homology (an algebraic structure representing connected components and cycles) of level sets of both density and regression functions. Our method is based on kernel estimation. We apply this procedure to two problems: (1) inferring the homology structure of manifolds from noisy observations, (2) inferring the persistent homology (a multi-scale extension of homology) of either density or regression functions. We prove consistency for both of these problems. In addition to the theoretical results, we demonstrate these methods on simulated data for binary regression and clustering applications.

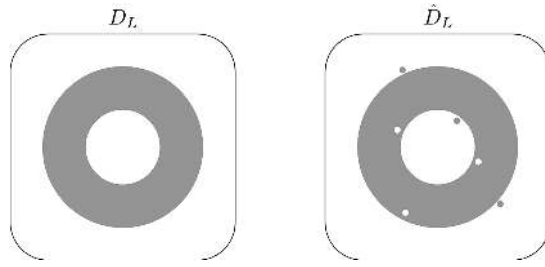
*Keywords:* clustering; homology; kernel density estimation; topological data analysis

## 1. Introduction

Level set estimation for probability density functions has been extensively studied in the past few decades. The basic formulation of the problem is as follows. Let  $p : \mathbb{R}^d \rightarrow \mathbb{R}$  be an unknown probability density function and define  $D_L := \{x \in \mathbb{R}^d : p(x) \geq L\}$  to be the  $L$ th super level set of  $p$  (from here on we will drop the word “super”). Given a sample  $\{X_1, \dots, X_n\}$  of i.i.d. observations drawn from  $p$ , we would like to estimate the set  $D_L$ . Recovering the level sets of density functions have shown to be useful in various applications such as clustering and cluster analysis [25,26,41,53,54,63], pattern recognition [24,32,40], anomaly detection [6,27], and econometrics [30,31,36,47] (where recovering the support of a distribution and its boundary is used for measuring efficiency).

Various solutions have been proposed to the level set estimation problem. Standard solutions include the plug-in estimator [4,5,27,51,52], the excess mass estimator [42,53,54,60,63,68], and the “naive” estimator [28,32,72]. The distance measure used to evaluate the performance of these estimators is usually either the Hausdorff distance or the Lebesgue distance (the volume of the difference between two sets). In this paper we wish to study level sets estimation from a topological perspective. Rather than trying to achieve an accurate recovery for the actual shape of the level sets, we wish to recover their qualitative topological properties (such as connected components and holes). Unfortunately, minimizing the Hausdorff or Lebesgue distance does not provide any guarantees for the quality of the topological recovery. Therefore, we have to consider a new type of an estimator. The sets in Figure 1 demonstrate the fact that minimizing the Hausdorff (or Lebesgue) distance can still result in very different topological spaces.

The motivation for studying the topology of level sets comes from the clustering problem. Given a set of observations generated by a probability density function  $p : \mathbb{R}^d \rightarrow \mathbb{R}$ , clustering



**Figure 1.** A schematic picture illustrating the difficulty in estimating the homology of level sets. Suppose that  $D_L$  is the annulus on the left and  $\hat{D}_L$  is its estimate on the right. While in both Hausdorff and Lebesgue distance the sets  $D_L$  and  $\hat{D}_L$  are close, the homology of these sets is completely different. In particular,  $D_L$  has a single connected component and a single hole, while  $\hat{D}_L$  has four of each. By taking the radius of the small circles to be as small as we wish, we can make both the Hausdorff and the Lebesgue distance to be arbitrarily small, while topologically we are looking at two different spaces.

can be loosely described as identifying and characterizing the connected components of either the support of  $p$  or one of its level sets (cf. [41,48,67,70]). From a topological perspective, clustering can be viewed as a question about the *homology* of the level sets. Briefly, the homology of a topological space  $X$  is a set of Abelian groups, denoted by  $\{H_0(X), H_1(X), \dots\}$ , where the elements of  $H_0(X)$  contain information about the connected components of  $X$ , and for  $k > 0$ , the group elements of  $H_k(X)$  contain information about “cycles,” or “holes” of different dimensions (see Section 2 for more details). From the perspective of algebraic topology, the clustering problem is thus equivalent to recovering  $H_0(X)$  where  $X$  is either the support of the distribution or a selected level set. A statistical perspective of the recent efforts in topological data analysis (TDA) [7,17,34,58,59] has been to extract topological invariants, and homology in particular, from random data. For example, recovering  $H_1$  provides information about holes or loops in the data, which is useful in various applications such as network coverage [29] or recovering periodic behavior [61]. The idea is that these topological summaries are useful for statistical inference and robust under various transformations. Our goal is therefore to examine level set estimation when the objective is not only to recover  $H_0(X)$  but rather the entire set of homology groups.

The idea of characterizing points or subsets of  $\mathbb{R}^d$  by their homology was developed in a series of papers in the late 1990s [64,65]. Asymptotic and non-asymptotic analysis of consistency and convergence of topological summaries as the number of observations increase has been examined for a variety of geometric objects using a variety of statistical and probabilistic tools [2,3,7–9, 11,12,15,19,44,45,58,59]. In the statistics and empirical process community, a version of the topology inference problem was presented as inference of the empirical geometry of data [46].

The main objective of this paper is to provide a consistent method for recovering the homology of the level sets  $D_L$  of functions  $f : \mathbb{R}^d \rightarrow \mathbb{R}$ , where  $f$  will be either a probability density function or a regression function. The standard plug-in idea would be to use a kernel-based estimator  $\hat{f}$  to construct an estimator  $\hat{D}_L$  to the level set. The problem with this approach is that due to the discrete nature of homology even a tiny error in the set estimate  $\hat{D}_L$  can introduce a significant error in homology. For example, an infinitesimally small region included by mistake can increase the number of components, while a small region excluded by mistake might intro-

duce a hole. Such errors in homology estimation may occur no matter how small the extraneous components and holes are. This problem is illustrated in Figure 1.

The main result in this paper presents a robust homology estimator for the level sets of both density and regression functions, that overcomes these difficulties. We show that instead of using  $\hat{D}_L$  as an estimate, one should consider the inclusion map between the nested pairs –  $\hat{D}_{L+\varepsilon} \subset \hat{D}_{L-\varepsilon}$  (for a properly chosen  $\varepsilon > 0$ ). The key object of interest is then the following induced map between the homology groups of the two level sets:

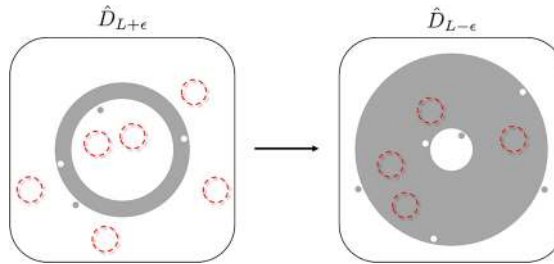
$$\iota_* : H_*(\hat{D}_{L+\varepsilon}) \rightarrow H_*(\hat{D}_{L-\varepsilon}),$$

where “\*” is a standard notation for an arbitrary degree. Inference of the homology at a single level is noisy, however the map  $\iota_*$  serves as a filter for the homological noise (see Figure 2). In particular, we will show that the image of this map –  $\text{Im}(\iota_*)$  – is isomorphic to the homology of  $D_L$  with a high probability. This statement is formalized by Theorem 3.3.

There are two direct implications for recovering the homology of level sets: recovering the homology of a manifold from a noisy sample and inference of the persistent homology of a function. For both applications, we make use of kernel density estimation to infer the image of the map  $\iota_*$  between the homology groups of different level sets. An interesting observation is that the conditions to recover the homology of the manifold or regression function do not require consistency of the kernel estimator.

The first application is inferring the homology of a manifold from a noisy sample. This problem was previously studied in [7,59]. In this paper, we show that for a wide class of noise models one can recover the homology of a manifold using fewer assumptions than previous methods and analysis. This result is stated in Theorem 3.6.

The second application is estimating the *persistent homology* of the function  $f$ . Persistent homology (described in Section 2) is a multi-scale topological summary. The main idea is instead of considering the homology of a single level  $D_L$ , the entire sequence of level sets is considered as  $L$  decreases from  $\infty$  to  $-\infty$ . One then tracks at what values of  $L$  changes in homology occur.



**Figure 2.** An illustration of the filtering mechanism underlying the homology estimator presented in this paper. Suppose that the set of interest  $D_L$  is the same as in Figure 1. Both estimates  $\hat{D}_{L-\varepsilon}$  and  $\hat{D}_{L+\varepsilon}$  have the wrong homology. The dashed circles in each figure mark the locations of the extraneous features (components and holes) in the other. We observe that none of the extraneous features exist in both sets. Since the image of the map  $\iota_*$  contains only the topological features that exist in both  $\hat{D}_{L+\varepsilon}$  and  $\hat{D}_{L-\varepsilon}$ , it will consist of a single component and a single hole – the correct homology of  $D_L$ .

The logic behind this computation is that homological features that persist across a wide range of levels are stable features while the other homological features are transient or noisy. This result is stated in Theorem 3.7.

The paper is structured as follows. In Section 2, we state the topological concepts and definitions we will use in this paper, namely homology and persistent homology. The main results of the paper are stated in Section 3 with the proofs in the Appendix. In Section 4, we provide a procedure to estimate the homology of level sets. Intuition about the estimator as well as results on simulated data are given in Section 5. We close with a discussion.

## 2. Topological preliminaries

In this section, we introduce the basic ideas of homology and persistent homology. To help fix ideas, we first present a particular example of persistent homology related to agglomerative hierarchical clustering.

### 2.1. Homology

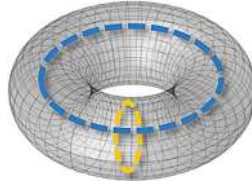
We develop the concept of homology intuitively, for a more rigorous and comprehensive treatment see [43,55]. Let  $X$  be a topological space. The *homology* of  $X$  is a set of Abelian groups  $\{H_k(X)\}_{k=0}^\infty$ , called homology groups. In this paper, we consider homology with coefficients in a field  $\mathbb{F}$ , in this case  $H_k(X)$  is actually a vector space. The zeroth homology group  $H_0(X)$  is generated by elements that represent connected components of  $X$ . For example, if  $X$  has three connected components, then  $H_0(X) \cong \mathbb{F} \oplus \mathbb{F} \oplus \mathbb{F}$  (here  $\cong$  denotes group isomorphism), and each of the three generators of this group corresponds to a different connected component of  $X$ . For  $k \geq 1$ , the  $k$ th homology group  $H_k(X)$  is generated by elements representing  $k$ -dimensional “holes” or “cycles” in  $X$ . An intuitive way to think about a  $k$ -dimensional hole is as the result of taking the boundary of a  $(k+1)$ -dimensional body. For example, if  $X$  is a circle then  $H_1(X) \cong \mathbb{F}$ , if  $X$  is a 2-dimensional sphere then  $H_2(X) \cong \mathbb{F}$ , and in general if  $X$  is a  $n$ -dimensional sphere, then

$$H_k(X) \cong \begin{cases} \mathbb{F}, & k = 0, n, \\ \{0\}, & \text{otherwise.} \end{cases}$$

Another interesting example is the 2-dimensional torus denoted by  $T$  (see Figure 3). The torus has a single connected component, therefore  $H_0(T) \cong \mathbb{F}$ , and a single 2-dimensional hole (the void inside the surface) implying that  $H_2(T) \cong \mathbb{F}$ . As for 1-cycles (or closed loops) the torus has two distinct features (see Figure 3) and therefore  $H_1(T) \cong \mathbb{F} \oplus \mathbb{F}$ .

The ranks of the homology groups (the number of generators) are called the Betti numbers, and are denoted by  $\beta_k(X) \triangleq \text{rank}(H_k(X))$ . When we refer to all the homology groups simultaneously, we use the notation  $H_*(X)$ .

In addition to providing a summary for a single space, homology can also characterize the topological behavior of functions. Let  $f : X \rightarrow Y$  be a map between two topological spaces, then homology theory provides a way to define the “induced map”  $f_* : H_*(X) \rightarrow H_*(Y)$  mapping between the homology groups of the two spaces.



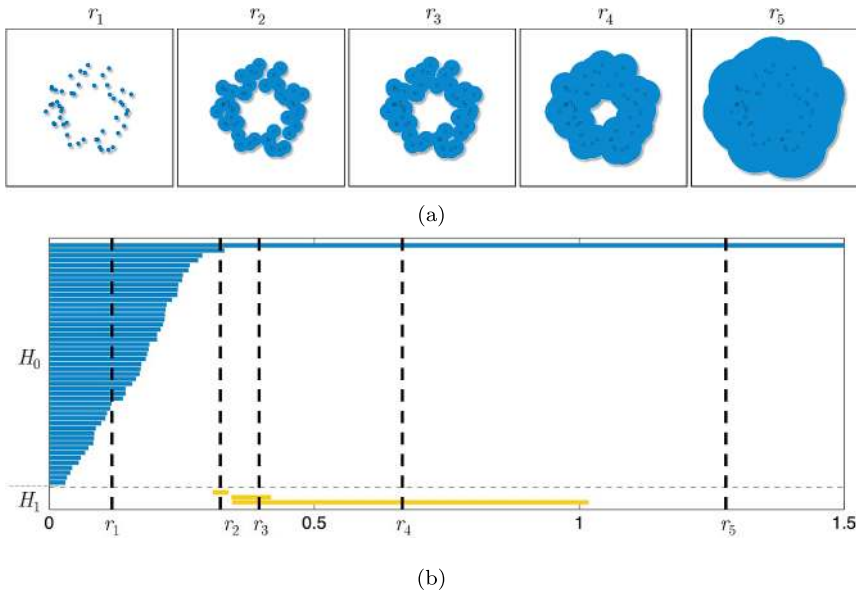
**Figure 3.** The 2-dimensional torus and its cycles. The torus has a single connected component and a single 2-cycle (the void locked inside the torus). In addition, it has two distinct 1-dimensional cycles (or closed loops) represented by the two curves in the figure. Consequently, the Betti numbers of the torus are  $\beta_0 = 1, \beta_1 = 2, \beta_2 = 1$ .

Another term we will use is *homotopy equivalence* (cf. [43,55]). Loosely speaking, two topological spaces  $X, Y$  are homotopy equivalent if we can continuously transform one into the other. We denote this property by  $X \simeq Y$ . If  $X \simeq Y$  then they have the same homology, that is,  $H_*(X) \cong H_*(Y)$ .

## 2.2. Persistent homology

Let  $\mathcal{X} = \{X_t\}_{t=a}^b$  be a filtration of topological spaces, such that  $X_{t_1} \subset X_{t_2}$  if  $t_1 < t_2$ . As the parameter  $t$  increases, the homology of the spaces  $X_t$  may change (e.g., components are added and merged, cycles are formed and filled up). The *persistent homology* of  $\mathcal{X}$ , denoted by  $\text{PH}_*(\mathcal{X})$ , keeps track of this process. Briefly,  $\text{PH}_*(\mathcal{X})$  contains the information about the homology of the individual spaces  $\{X_t\}$  as well as the mappings between the homology of  $X_{t_1}$  and  $X_{t_2}$  for every  $t_1 < t_2$ . The *birth time* of an element in  $\text{PH}_*(\mathcal{X})$  can be thought of as the value of  $t$  where this element appears for the first time. The *death time* is the value of  $t$  where an element vanishes, or merges with another existing element. We refer the reader to [34,35,39,74] for more details and formal definitions. Another perspective of persistence homology is as a summary statistic of point cloud data that is robust to certain invariances, this perspective has been developed in [10,14,49,69].

A useful way to describe persistent homology is via the notion of *barcodes*. A barcode for the persistent homology of a filtration  $\mathcal{X}$  is a collection of graphs, one for each order of homology group. A bar in the  $k$ th graph, starting at  $b$  and ending at  $d$  ( $b \leq d$ ) indicates the existence of a generator of  $H_k(X_t)$  (or a  $k$ -cycle) whose birth and death times are  $b, d$ , respectively. In Figure 4, we present an example for a barcode generated in the following way. We take a sample of  $n = 50$  points  $P_1, \dots, P_n \in \mathbb{R}^2$  sampled from a uniform distribution on an annulus. We then define  $X_r = \bigcup_i B_r(P_i)$  to be the union of closed balls around the sample points. Increasing  $r$  makes the space  $X_r$  grow. In this process connected components merge, and cycles are formed and then filled up. In Figure 4(a), we present a few snapshots of the space  $X_r$  for different values of  $r$  where different features show. The barcode in Figure 4(b) presents a summary of all the homology features in this process. We can see that there are two bars that are significantly longer than the others (one in  $H_0$  and one in  $H_1$ ) indicating that the underlying space has a single connected component, and a single cycle (as the annulus does).



**Figure 4.** (a)  $X_r$  is a union of balls of radius  $r$  around a random set of  $n = 50$  points, generated from a uniform distribution on an annulus in  $\mathbb{R}^2$ . We present five snapshots of this filtration. (b) The persistent homology of the filtration  $\{X_r\}_{r \geq 0}$ . The  $x$ -axis is the radius of the balls, and the bars represent the homology features that are born and died. For  $H_0$  we observe that at radius zero the number of components is exactly  $n$  and as the radius increases components merge (or die). Note that when two components merge, we terminate the bar for one of them, and the merged component is represented by the bar we keep. This is a standard representation that comes as the result of the algebraic structure underlying persistent homology (cf. [74]). The cycles show up later in this process. There are two bars that are significantly longer than the others (one in  $H_0$  and one in  $H_1$ ). These correspond to the true topological features of the annulus.

For a given space, there are many choices of filtrations (sequences of nested subspaces). In this paper the filtrations we work with are the (super) level sets of functions. Specifically, let  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  and let  $D_L$  be a level set of  $f$ . As the level  $L$  is decreased from  $\infty$  to  $-\infty$  the sets  $D_L$  grow, and in this process components and cycles are created and destroyed. We denote by  $\text{PH}_*(f)$  the persistent homology for this process.

To show later that we can recover the persistent homology structure, we will need a notion of distance between the persistent homology of two different filtrations. If  $\mathcal{X}$  is a filtration, the  $k$ th persistence diagram of  $\mathcal{X}$ , denoted by  $\text{Dgm}_k(\mathcal{X})$  is the set of all pairs  $(b, d)$  of birth–death times of features in  $\text{PH}_k(\mathcal{X})$ . The bottleneck distance between the persistent homology of the two filtrations  $\mathcal{X}$  and  $\mathcal{Y}$  is defined as

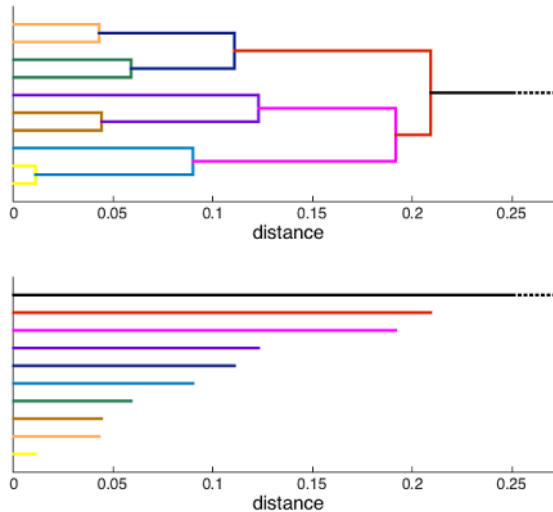
$$d_B(\text{PH}_k(\mathcal{X}), \text{PH}_k(\mathcal{Y})) = \inf_{\gamma \in \Gamma} \sup_{p \in \text{Dgm}_k(\mathcal{X})} \|p - \gamma(p)\|_\infty.$$

The set  $\Gamma$  consists of all the bijections  $\gamma : \text{Dgm}_k(\mathcal{X}) \cup \text{Diag} \rightarrow \text{Dgm}_k(\mathcal{Y}) \cup \text{Diag}$ , where  $\text{Diag} = \{(x, x) : x \in \mathbb{R}\} \subset \mathbb{R}^2$  is the diagonal line, and  $\|\cdot\|_\infty$  is the sup-norm in  $\mathbb{R}^2$ . In other words, we are looking for a matching between the points in  $\text{Dgm}_k(\mathcal{X})$  and  $\text{Dgm}_k(\mathcal{Y})$  that requires the minimal translations of birth and death times. We add the diagonal to each diagram for two reasons. First, we want to be able to consider diagrams with different numbers of features, and second, we want to allow deleting points from a diagram (by matching them to the diagonal) rather than forcing them to match.

To conclude this section, we note that the zeroth persistent homology,  $\text{PH}_0$ , is closely related to hierarchical clustering as the following example will illustrate. Let  $\mathcal{P} \subset \mathbb{R}^d$  be a finite set of points in Euclidean space. We define the distance function from the set  $d_{\mathcal{P}} : \mathbb{R}^d \rightarrow \mathbb{R}$  as

$$d_{\mathcal{P}}(x) = \min_{p \in \mathcal{P}} \|x - p\|.$$

In this case, computing the 0th persistent homology for the sub level set filtration of  $d_{\mathcal{P}}$  is very simple. We start at level 0 with just the finite set  $\mathcal{P}$ , and as we increase the level we merge connected components according to the distances between points in  $\mathcal{P}$ . The bottom of Figure 5 is the barcode generated by such a process, the top figure is the dendrogram generated by the same set of points. One can observe that the end points of the bars in the barcode are the nodes in the dendrogram.



**Figure 5.** Persistent homology and hierarchical clustering. The figure on top is the dendrogram generated by a set of 10 random points in the interval  $[0, 1]$ . The bottom figure is the barcode generated by the 0-persistent homology for the sub-level sets of the distance function from the same set of points. The  $x$ -axis represents function values (distance, in our case). In this example, all the connected components are created at distance zero, and only differ by their death point (when two components merge). Note, that one of the components (the top bar) lives forever. The death points in the barcode correspond to nodes in the dendrogram, we marked the bars with different colors matching the relevant part of the dendrogram.

### 3. Statistical model and main results

Given a function  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  the objects we analyze in this paper are the (super) level sets of  $f$

$$D_L \triangleq \{x \in \mathbb{R}^d : f(x) \geq L\}. \tag{3.1}$$

Note that for any  $L_1 < L_2$  we have  $D_{L_2} \subset D_{L_1}$ .

Previous results on level set estimation usually require some assumptions on either the function  $f$  (smooth, non-flat, etc.), or the shape of the level set (convex, star-shaped, elliptic, etc.). For the purpose of homology estimation, our main assumption on  $f$  is “tameness” as defined in [16].

**Definition 3.1.** Let  $f : \mathbb{R}^d \rightarrow \mathbb{R}$ , and  $D_L$  as defined in (3.1).

1. We say that  $L$  is a homological regular value if there exists  $\varepsilon > 0$  such that for every  $v_2 \leq v_1$  in  $(L - \varepsilon, L + \varepsilon)$  the map  $H_k(D_{v_1}) \rightarrow H_k(D_{v_2})$  induced by inclusion is an isomorphism for every  $k \geq 0$ .  
 Otherwise, we say that  $L$  is a homological critical value.
2. A function  $f$  is called tame if it has a finite number of homological critical values, and  $\text{rank}(H_k(D_L))$  is finite for all  $L$  and  $k$ .

Our main goal in this paper is to present a consistent method for recovering the homology of a given level set  $D_L$ . We will examine the level sets of two classical quantities of interest in statistics:

1. Density functions – Given Data =  $\{X_1, \dots, X_n\} \stackrel{\text{i.i.d.}}{\sim} p(x)$ , where  $p$  is a probability density function, our objective is to recover the level sets of  $f = p$ .
2. Regression functions – Given Data =  $\{(X_1, Y_1), \dots, (X_n, Y_n)\} \stackrel{\text{i.i.d.}}{\sim} p_{X,Y}(x, y)$ , where  $p_{X,Y}(x, y)$  is a joint probability density function and we state  $p : \mathbb{R}^d \rightarrow \mathbb{R}$  as the marginal density of  $X$ . Our objective is to recover the level sets of the regression function  $f(x) \triangleq \mathbb{E}\{Y \mid X = x\}$ .

A common procedure to recover the homology of an unknown space  $S$  from a random sample  $\mathcal{X} \subset S$  is to compute the homology of a union of closed balls around the sample points

$$U(\mathcal{X}, r) := \bigcup_{x \in \mathcal{X}} B_r(x), \tag{3.2}$$

for some choice of radius  $r$  (cf. [12,58]). In the level-set estimation literature, this procedure is known as the “naive” estimator [28,32,72]. We can use this idea to estimate the homology of the set  $D_L$  using the following procedure (P1):

1. Use the entire data set to construct an estimator  $\hat{f}$ .
2. Using the estimator  $\hat{f}$ , define

$$\mathcal{X}^L = \{X_i : \hat{f}(X_i) \geq L\},$$

as the set of data points lying in the  $L$ th level set of  $\hat{f}$ .



3. Consider  $U(\mathcal{X}^L, r)$  as an estimate of  $D_L$ , and the homology  $H_*(U(\mathcal{X}^L, r))$  as an estimate of  $H_*(D_L)$ .

We will use kernel estimators for  $\hat{f}$  in both the regression and density estimation case. A key difficulty in the above procedure is that the estimator  $\hat{f}$  may introduce errors in the filtering step 2 of the above procedure. In [28,32,72] it is shown that small errors in the estimate  $\hat{f}$  are translated to small errors in terms of the Hausdorff or Lebesgue distances. However, since homology is a discrete descriptor, even tiny errors in the filtering step can introduce large errors in the homology estimates. For example, even a single point incorrectly included in the level set assignment can form an extra connected component, and increase the zeroth Betti number by one (see Figure 1). One of the main challenges we will address in this paper is providing an estimator that is robust to this type of error.

Given a kernel function  $K : \mathbb{R}^d \rightarrow \mathbb{R}$  we construct our estimators as follows. In the density estimation case, we define

$$\hat{f}_n(x) = \hat{p}_n(x) \triangleq \frac{1}{n \times C_K r^d} \sum_{i=1}^n K_r(x - X_i),$$

where  $X_1, \dots, X_n$  are the observed data,  $K_r(x) = K(x/r)$ , and  $C_K$  is a normalizing constant defined below. In the regression setting, we use the Nadaraya–Watson estimator [56,66,73]

$$\hat{f}_n(x) \triangleq \frac{\sum_{i=1}^n Y_i K_r(x - X_i)}{\sum_{i=1}^n K_r(x - X_i)},$$

where  $\{(X_1, Y_1), \dots, (X_n, Y_n)\}$  are the observed data.

The kernel functions  $K(x)$  we consider satisfy the following conditions (C1):

1. The support of the kernel function is contained within the unit ball of radius 1, that is,  $\text{supp}(K) \subset B_1(0)$ .
2. The kernel function has a maximum at the origin, with  $K(0) = 1$ , and  $\forall x : K(x) \in [0, 1]$ .
3. The kernel function is smooth within the unit ball, and

$$\int_{\mathbb{R}^d} K(\xi) d\xi = C_K \quad \text{for } C_K \in (0, 1).$$

Note that the bounded support assumption is very common in level set estimation procedures (e.g., [5,27,72]). Weak regularity conditions on the density or regression function will be required to prove consistency of the estimates of the homology of level sets. For both density estimation and regression, we require the density function  $p$  to be tame and bounded, and we define

$$p_{\max} \triangleq \sup_{x \in \mathbb{R}^d} p(x).$$

For density estimation, we also require that for every  $L$  the set  $D_L \subset \mathbb{R}^d$  is bounded. For the regression case, we require in addition the following set of conditions (C2):

1. The marginal density of  $X$  has compact support, that is,  $\text{supp}(p)$  is compact.

2. The marginal density of  $X$  is bounded away from zero within its support, that is,  $p_{\min} \triangleq \inf_{x \in \text{supp}(p)} p(x) > 0$ .
3. The response variables are almost surely bounded, that is,  $|Y_i| \leq Y_{\max}$  almost surely for some non-random value  $Y_{\max} > 0$ .

Next, recall step 2 in the procedure (P1), and define

$$\mathcal{X}_n^L \triangleq \{X_i : \hat{f}_n(X_i) \geq L; 1 \leq i \leq n\}.$$

The subset  $\mathcal{X}_n^L$  can be used to construct an estimator to the level set  $D_L$ :

$$\hat{D}_L(n, r) \triangleq U(\mathcal{X}_n^L, r). \tag{3.3}$$

Note that the radius  $r$  is the same  $r$  as used for the bandwidth of the kernel function. This connection is crucial for the proofs.

To overcome the noisiness of the estimator  $\hat{D}_L(n, r)$  discussed above, we present the following procedure. First, note that for any  $\varepsilon \in (0, L)$ , we have that  $\hat{D}_{L+\varepsilon}(n, r) \subset \hat{D}_{L-\varepsilon}(n, r)$ . The inclusion map

$$\iota : \hat{D}_{L+\varepsilon}(n, r) \hookrightarrow \hat{D}_{L-\varepsilon}(n, r)$$

induces a map in homology

$$\iota_* : H_*(\hat{D}_{L+\varepsilon}(n, r)) \rightarrow H_*(\hat{D}_{L-\varepsilon}(n, r)). \tag{3.4}$$

We use this map to define

$$\hat{H}_*(L, \varepsilon; n) \triangleq \text{Im}(\iota_*). \tag{3.5}$$

We will use  $\hat{H}_*(L, \varepsilon; n)$  as an estimator for  $H_*(D_L)$ . The intuition behind using this inclusion map is as follows. Using Lemma A.2, we can show that with a high probability we have

$$\begin{array}{ccccc} D_{L+2\varepsilon} & & D_L & & D_{L-2\varepsilon} \\ & \searrow & \nearrow & \searrow & \nearrow \\ & \hat{D}_{L+\varepsilon}(n, r) & \xrightarrow{\iota} & \hat{D}_{L-\varepsilon}(n, r) & \end{array}, \tag{3.6}$$

where  $\hookrightarrow$  represents inclusion. Assuming that  $H_*(D_{L+2\varepsilon}) \cong H_*(D_L) \cong H_*(D_{L-2\varepsilon})$ , then all the cycles in  $H_*(D_L)$  must persist throughout this entire sequence of inclusions and in particular they should be present in  $\hat{H}_*(L, \varepsilon; n)$ . In contrast, any cycles in  $\hat{D}_{L\pm\varepsilon}(n, r)$  that do not belong to  $D_L$  must be terminated as we move from  $\hat{D}_{L+\varepsilon}(n, r)$  to  $\hat{D}_{L-\varepsilon}(n, r)$  via  $D_L$ , and therefore should not be in  $\hat{H}_*(L, \varepsilon; n)$ . To prove that the inclusion sequence in (3.6) holds, we require the following regularity condition on  $L$ .

**Definition 3.2.** Given a level  $L > 0$  and  $\varepsilon \in (0, L/2)$ , we say that  $L$  is  $\varepsilon$ -regular if

$$\begin{aligned} \partial D_{L+2\varepsilon} \cap \partial D_{L+(3/2)\varepsilon} &= \partial D_{L+(1/2)\varepsilon} \cap \partial D_L = \partial D_L \cap \partial D_{L-(1/2)\varepsilon} \\ &= \partial D_{L-(3/2)\varepsilon} \cap \partial D_{L-2\varepsilon} = \emptyset, \end{aligned}$$

where “ $\partial$ ” is the set boundary.

This regularity condition basically guarantees sufficient “separation” between the level sets involved in the estimation process (its importance will become clearer in the proofs). In particular, if  $f$  is continuous in  $f^{-1}([L - 2\varepsilon, L + 2\varepsilon])$ , then  $L$  is  $\varepsilon$ -regular. We will assume that the levels we are studying are always  $\varepsilon$ -regular.

We now state the main result in this paper which holds for both the density estimation as well as regression setting.

**Theorem 3.3.** *Let  $L > 0$  and  $\varepsilon \in (0, L/2)$  be such that the function  $f(x)$  has no critical values in the range  $[L - 2\varepsilon, L + 2\varepsilon]$ . If  $r \rightarrow 0$ , and  $nr^d \rightarrow \infty$ , then for  $n$  large enough we have*

$$\mathbb{P}(\hat{H}_*(L, \varepsilon; n) \cong H_*(D_L)) \geq 1 - 6ne^{-C_\varepsilon^* nr^d}.$$

In particular, if  $nr^d \geq D \log n$  with  $D > (C_{\varepsilon/2}^*)^{-1}$ , then

$$\lim_{n \rightarrow \infty} \mathbb{P}(\hat{H}_*(L, \varepsilon; n) \cong H_*(D_L)) = 1.$$

The constant value  $C_\varepsilon^*$  in the theorem above is

$$C_\varepsilon^* = \frac{\varepsilon^2 C_K}{3p_{\max} + \varepsilon}, \tag{3.7}$$

for density estimation, and

$$C_\varepsilon^* = \frac{\varepsilon^2 p_{\min}^2 C_K}{3(Y_{\max}^2 + \varepsilon^2)p_{\max} + 2\varepsilon p_{\min}(Y_{\max} + \varepsilon)}, \tag{3.8}$$

for regression (see the [Appendix](#) for more details).

Theorem 3.3 states that if we want to recover the homology of the level set  $D_L$  we can compute the image of the homology map as we move from  $\hat{D}_{L+\varepsilon}(n, r)$  to the slightly larger complex  $\hat{D}_{L-\varepsilon}(n, r)$ . We note that another possible solution to this estimation problem is to dilate the estimated set  $\hat{D}_L$  directly (e.g., by covering the points with a slightly larger balls), as suggested by the results in [19]. However, such a method will require further knowledge about the level sets (such as their feature size), and the gradient of the function  $f$ , which is not required by the method we propose here.

**Remark.** In order to choose  $D$ , we need to know the values of  $p_{\min}$ ,  $p_{\max}$  and  $Y_{\max}$ , which might not be directly available. There are a few possible ways to address this problem:

1. Since all we need are bounds and not the precise values, one option is to make the broad assumption that  $p$  belongs to a class of density functions bounded by some fixed values, and use a similar assumption for  $Y$ .
2. Another option is to estimate these values from the data, taking values as high as we want for the upper bounds  $p_{\max}$ ,  $Y_{\max}$  (and as low as we want for the lower bound  $p_{\min}$ ), to guarantee that the estimated values are indeed valid bounds with high probability. Using estimated values instead of the true ones affects the theoretical validity of Theorem 3.3, but we believe it should have a negligible effect in practice.

3. Finally, another option is to take  $nr^d \gg \log n$  (e.g.,  $nr^d = (\log n)^2$ ). Then it is guaranteed that the probability converges to one, and we do not need to know the value of  $C_\varepsilon$ .

In the following sections we describe two applications for the estimator we proposed, addressing problems that are of significant interest in the fields of topological data analysis and machine learning.

### 3.1. An application to manifold learning

Let  $\mathcal{M}$  be a smooth  $m$ -dimensional, closed manifold (compact and without a boundary), embedded in  $\mathbb{R}^d$ . Given a random sample  $\mathcal{X}_n = \{X_1, \dots, X_n\} \subset \mathbb{R}^d$  we wish to recover the homology of  $\mathcal{M}$ . The case where the observations are drawn directly from the manifold (i.e.,  $\mathcal{X}_n \subset \mathcal{M}$ ), has been extensively studied (see [12,58]). In [12], the following asymptotic result was presented.

**Theorem 3.4 (Theorem 4.9 in [12]).** *If  $nr^d \geq C \log n$ , and  $C > (\omega_d p_{\min})^{-1}$ , then:*

$$\lim_{n \rightarrow \infty} \mathbb{P}(H_*(U(\mathcal{X}_n, r)) \cong H_*(\mathcal{M})) = 1,$$

where  $\omega_d$  is the volume of a  $d$ -dimensional unit ball, and  $p_{\min} = \inf_{x \in \mathcal{M}} p(x) > 0$ .

In this section, we extend this result to the case where noise is present. The term “noise” in this context refers to the fact that the observations do not necessarily lie on  $\mathcal{M}$ , but rather in its vicinity. As an example, consider the observations  $X_1, \dots, X_n$  defined as

$$X_i = Y_i + Z_i \quad \text{where } Y_i \stackrel{\text{i.i.d.}}{\sim} \rho(\mathcal{M}) \quad \text{and} \quad Z_i \stackrel{\text{i.i.d.}}{\sim} \mathbf{N}(\mathbf{0}, \sigma^2 \mathbf{I}_d), \tag{3.9}$$

where  $Y_i$  is drawn from a distribution  $\rho$  that is supported on a manifold  $\mathcal{M}$ , and  $Z_i$  is drawn from the normal distribution in the ambient space  $\mathbb{R}^d$ . For this model, the methods used to prove consistency of the estimator in [12,58] no longer apply since the outliers produced by the noise create their own topology, and interfere with our ability to recover  $H_*(\mathcal{M})$ .

The seminal work in [59] studies the following special case. Let  $Y_i \stackrel{\text{i.i.d.}}{\sim} \rho(\mathcal{M})$ , for each  $i \in \{1, \dots, n\}$  let  $\mathcal{N}_i$  be the normal space to  $\mathcal{M}$  at  $Y_i$ , and let  $Z_i \sim \mathbf{N}(0, \sigma^2 \mathbf{I}_{d-m})$  be a multivariate normal variable in the normal space  $\mathcal{N}_i$ . Our observations are then taken to be  $X_i = Y_i + Z_i$ . Under explicit assumptions on  $\sigma$  and  $\mathcal{M}$ , they show that the homology of  $\mathcal{M}$  can be recovered from  $\mathcal{X}_n$  with a high probability. The work in [7] extends this idea to a few other noise models. The results and proofs in [7,59] are tied to specific noise models and rely on the parameters of the noise model and the geometry of  $\mathcal{M}$ . We wish to use the result in Theorem 3.3 to study the same homology inference problem for a large class of distributions, and with as few assumptions as possible.

We start by defining a general class of density functions on  $\mathbb{R}^d$ , from which it would be possible to extract the homology of  $\mathcal{M}$ .

**Definition 3.5.** *Let  $p : \mathbb{R}^d \rightarrow \mathbb{R}_+$  be a probability density function. We say that  $p$  represents a noisy version of  $\mathcal{M}$ , if there exist  $0 < A < B < \infty$  such that:*

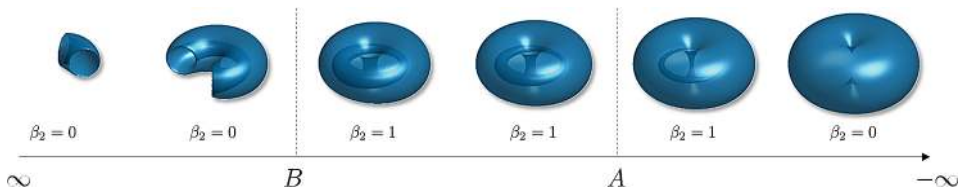
1. For every  $L \in [A, B]$  we have  $D_L \simeq \mathcal{M}$ .
2. For every  $L > B$ , we have  $D_L \simeq \mathcal{M}'$ , where  $\mathcal{M}' \subset \mathcal{M}$  is a compact locally contractible proper subset of  $\mathcal{M}$ ,

where “ $\simeq$ ” stands for homotopy equivalence (see Section 2).

In other words, we consider density functions  $p$  for which there is a range where the level sets are “similar” to  $\mathcal{M}$ . For levels higher than this range, the level sets are “similar” to nice subsets of  $\mathcal{M}$ . For example, the distribution in (3.9) satisfies this conditions for small enough  $\sigma$ . By “locally contractible” we refer to the property that every point  $x$  has a neighborhood  $\mathcal{N}_x$  that is homotopy equivalent to a single point. For example, if  $\mathcal{M}'$  is a compact manifold with boundary, then it is locally contractible. We need this requirement to rule out the appearance of highly twisted topological spaces. In Figure 6, we present a sequence of level sets for a density function that represents a noisy version of the torus. This density was generated by taking a uniform distribution on the latitude angle, a wrapped normal distribution on the longitude angle, and adding independent Gaussian noise. Note that the level sets are 3-dimensional whereas the torus is 2-dimensional. Nevertheless, we can see that there is a whole range of levels where they are topologically equivalent.

The model described in Definition 3.5 generalizes the additive Gaussian noise model discussed in [7,59] but is essentially different than the other noise models in [7]. This model is very broad in the sense that it is not tied to any specific assumptions on the distribution (e.g., uniform in the “clutter” and “tubular” noise models, or having Fourier transform bounded away from zero in the “additive” model [7]). In addition, we believe that this model is more “natural” for topological estimation since it emphasizes the topological behavior of the density rather than making analytic assumptions on its functional structure.

If we know a priori the values of  $A$  and  $B$ , then the recovery method would be simple. Given a sample  $\mathcal{X}_n = \{X_1, \dots, X_n\} \stackrel{\text{i.i.d.}}{\sim} p$ , and setting  $f = p$ , we choose  $L$  and  $\varepsilon$  such that  $[L - 2\varepsilon, L + 2\varepsilon] \subset (A, B)$ , and compute  $\hat{H}_*(L, \varepsilon; n)$ . Theorem 3.3 guarantees that with high probability  $\hat{H}_*(L, \varepsilon; n) \cong H_*(D_L) \cong H_*(\mathcal{M})$ .



**Figure 6.** In this figure we demonstrate a sequence of level sets for a density function  $p$  that is a noisy version of the 2-dimensional torus. The horizontal axis represents the function levels in a decreasing order. For very high values ( $L > B$ ) we see that the level sets look like a subset of the torus. Note that they are not real subsets, since these are 3-dimensional shapes, whereas the torus is 2-dimensional. Inside the range  $(A, B)$  the level sets look like the torus (where  $\beta_0 = \beta_2 = 1$ , and  $\beta_1 = 2$ ). For low levels the topology changes again, be we no longer require any assumptions.

However, in real problems we are not given  $A, B$  so the real challenge is to recover  $\mathcal{M}$  without knowing the stable range. To show that the procedure described below is consistent, we require the following assumptions to hold.

- (i)  $\mathcal{M}$  is connected and orientable;
- (ii)  $B - A > 8\varepsilon$ .

The following procedure (P2) will be used to estimate the homology of  $\mathcal{M}$  from the a noisy sample  $\mathcal{X}_n$ . In this procedure, we will use the estimated Betti numbers defined as  $\hat{\beta}_k(L, \varepsilon; n) \triangleq \text{rank}(\hat{H}_*(L, \varepsilon; n))$ . Define

$$N_\varepsilon := \sup_{x \in \mathbb{R}^d} \lceil f(x)/2\varepsilon \rceil, \quad L_{\max} = 2\varepsilon N_\varepsilon \quad \text{and} \quad L_i = L_{\max} - 2i\varepsilon. \quad (3.10)$$

The procedure (P2) is as follows.

1. Compute  $\hat{H}_*(L_i, \varepsilon; n)$  for all  $i = 1, \dots, N_\varepsilon$ .
2. Define

$$i^* \triangleq 1 + \min\{i \in \{1, \dots, N_\varepsilon\} : \hat{\beta}_m(L_i, \varepsilon; n) = 1\}.$$

This index will be shown to be the first point where we are guaranteed to observe the homology of  $\mathcal{M}$ .

3. Our estimator for the homology of  $\mathcal{M}$  will then be  $\hat{H}_*(L_{i^*}, \varepsilon; n)$ .

Note that in this procedure a choice has to be made for the parameter  $r$  (the radius of the balls and the bandwidth of the kernel). The following theorem states that if  $r$  is chosen appropriately we can estimate the homology of a manifold from noisy observations with high probability.

**Theorem 3.6.** *Let  $\mathcal{M}$  be a  $m$ -dimensional closed, connected, orientable manifold embedded in  $\mathbb{R}^d$ . Let  $X_1, \dots, X_n$  be data points sampled from a density function  $p$  satisfying the conditions in Definition 3.5. Choose  $r \rightarrow 0$  that satisfies  $nr^d \geq D \log n$  with  $D > (C_{\varepsilon/2}^*)^{-1}$ , where  $C_\varepsilon$  is defined in (3.7). Applying procedure (P2), we then have*

$$\lim_{n \rightarrow \infty} \mathbb{P}(\hat{H}_*(L_{i^*}, \varepsilon; n) \cong H_*(\mathcal{M})) = 1.$$

We state here the main ideas used in proving the above, while the detailed proof is given in the [Appendix](#). We use Poincaré duality, a fundamental idea in algebraic topology. Poincaré duality relates homology groups to co-homology groups of closed orientable  $m$ -dimensional manifolds, stating that  $H_k(\mathcal{M}) \cong H^{m-k}(\mathcal{M})$ , where  $H^{m-k}(\mathcal{M})$  is the co-homology of  $\mathcal{M}$  (cf. [43,55]). An important consequence of Poincaré duality is that  $\beta_k(\mathcal{M}) = \beta_{m-k}(\mathcal{M})$  for every  $k = 0, \dots, m$ , and in particular  $\beta_0(\mathcal{M}) = \beta_m(\mathcal{M})$ . Our assumption that  $\mathcal{M}$  is connected implies that  $\beta_0(\mathcal{M}) = 1$ , and from Poincaré duality we conclude that  $\beta_m(\mathcal{M}) = 1$  as well. In contrast, if  $\mathcal{M}' \subset \mathcal{M}$  is a proper compact locally contractible subset of  $\mathcal{M}$  then using a different type of duality one can show that  $\beta_m(\mathcal{M}') = 0$  (see Proposition 3.46 in [43]). Our assumptions on  $A, B$  then implies that if  $L_i > B$  we have  $\beta_m(D_{L_i}) = 0$ , while if  $L_i \in (A, B)$  then  $\beta_m(D_{L_i}) = 1$ . Therefore, the first  $L_i$  for which the  $m$ -th Betti number switches from 0 to 1 necessarily lies in  $(A, B)$ , and we can use this  $L_i$  to recover the homology of  $\mathcal{M}$ . In practice, we defined  $i^*$  to be

the second level at which we have  $\hat{\beta}_m(L_i, \varepsilon; n) = 1$ . This is a precautionary measure which we discuss in the proof.

**Remark.**

1. To use the result in Theorem 3.6 one needs to know the values of  $m$  and  $\varepsilon$ . We consider these values to be crucial information required to “extract” the topology of the manifold. Their knowledge replaces other assumptions about the geometry of the manifold which we want to avoid. Note that for  $\varepsilon$  we do not require a precise value but any lower bound would suffice.
2. Also required is the knowledge  $L_{\max}$  (or equivalently  $N_\varepsilon$ ). Note, that when we have a finite sample  $\{X_1, \dots, X_n\}$  we can estimate  $L_{\max}$  using  $\hat{L}_{\max} := \max_i \lceil f_n(X_i)/2\varepsilon \rceil$ . For every  $L > \hat{L}_{\max}$  we have  $\hat{D}_L(n, r) = \emptyset$ . Therefore, in practice, even if the true  $L_{\max}$  is higher than  $\hat{L}_{\max}$ , it does not affect the procedure, since the higher levels are empty anyway.
3. It is possible that small perturbations in the density function will generate  $m$ -dimensional cycles at level sets with  $L > B$ . To be able to ignore these cycles when they appear, additional information about the geometry of the underlying manifold should be provided (e.g., its feature size), otherwise it will be impossible to determine which of the  $m$ -dimensional cycles belongs to the manifold (even if the function  $f$  is known completely), and the homology inference problem is ill-posed. If we want to limit ourselves to use only the fact that the data is “concentrated” around a  $m$ -dimensional manifold, then we need to assume the density function allows us to identify it properly, and that is the essence of Definition 3.5.

### 3.2. Persistent homology and application to clustering

A common topological summary used in TDA is persistent homology (see Section 2). Given a function  $f$  the persistent homology of  $f$ ,  $\text{PH}_*(f)$ , tracks when the homology of (super) level-sets of  $f$  changes and serves as a summary of the function. This summary contains information about the creation and destruction of connected components and cycles of the level sets. In the case where  $f = p$  is a density function, the zeroth persistent homology  $\text{PH}_0(f)$  can be viewed as a summary of the evolution of clusters in the data, and can be useful for clustering algorithms as discussed in Section 2.2. By definition,  $\text{PH}_*(f)$  is computed from the continuous filtration  $\mathcal{D} = \{D_L\}_{L \in \mathbb{R}}$  as  $L$  decreases from  $\infty$  to  $-\infty$ . Note that the persistent homology  $\text{PH}_*(f)$  contains much more information than just the homology at each level  $D_L$ . It also contains information about mappings between different levels, and hence enables us to track the evolution of cycles.

In this section, we wish to address the estimation of  $\text{PH}_*(f)$  where  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  is either a density function (tame and bounded) or a regression function (satisfying the conditions (C2) as well). In both cases, we have shown that the estimator  $\hat{H}_*(L, \varepsilon, n)$  defined in (3.5), can recover the homology of  $D_L$  for every  $L$ . In order to recover the persistent homology we also need to make sure that the mappings between different levels are recovered as well. The error measure we use is the commonly used “bottleneck distance” (see Section 2). To estimate  $\text{PH}_*(f)$ , recall the definitions of  $N_\varepsilon$ ,  $L_{\max}$ , and  $L_i$  in (3.10) and consider the following discrete filtration

$$\hat{\mathcal{D}}^\varepsilon \triangleq \{\hat{D}_{L_i}(n, r)\}_{i \in \mathbb{Z}},$$

where  $\hat{D}_{L_i}(n, r)$  is defined by (3.3). Denoting the persistent homology of  $\hat{D}^\varepsilon$  by  $\widehat{\text{PH}}_*^\varepsilon(f)$ , and using the methods presented in this section we prove the following.

**Theorem 3.7.** *If  $r \rightarrow 0$  and  $nr^d \rightarrow \infty$ , then*

$$\mathbb{P}(d_B(\widehat{\text{PH}}_*^\varepsilon(f), \text{PH}_*(f)) \leq 5\varepsilon) \geq 1 - 3N_\varepsilon n e^{-C_\varepsilon^* nr^d},$$

where  $C_\varepsilon^*$  is defined in (3.7) (density) and (3.8) (regression). In particular, if  $nr^d \geq D \log n$  with  $D > (C_{\varepsilon/2}^*)^{-1}$ , we have

$$\lim_{n \rightarrow \infty} \mathbb{P}(d_B(\widehat{\text{PH}}_*^\varepsilon(f), \text{PH}_*(f)) \leq 5\varepsilon) = 1.$$

In other words, we state that the estimator  $\widehat{\text{PH}}_*^\varepsilon(f)$  is “consistent” up to a given precision of  $5\varepsilon$ . Note that we will always have some discretization error since our estimator is discrete (having an inherent step size  $\varepsilon$ ) while the filtration we wish to study is continuous. However, one can make  $\varepsilon$  arbitrarily small to achieve higher precision. The smaller value of  $\varepsilon$  we choose the smaller  $C_{\varepsilon/2}^*$  will be and the convergence of  $\widehat{\text{PH}}_*^\varepsilon(f)$  to  $\text{PH}_*(f)$  will be slower.

To prove this theorem (see Appendix), we invoke Lemma A.2  $M$  times in order to form a sequence of inclusions alternating between level sets  $D_L$  and their estimates  $\hat{D}_L(n, r)$ . This alternating sequence is called “interleaving” and the work in [18] provides means to bound the distance between the persistent homology computed for these two types of filtrations. In Section 5, we provide several examples for the estimation of persistent homology using  $\widehat{\text{PH}}_*^\varepsilon(f)$ .

As we discuss in Section 4, Theorem 3.7 can be adjusted to use the filtration of Rips complexes  $\{R_{L_i}(n, r)\}_{i \in \mathbb{Z}}$  instead of  $\{\hat{D}_{L_i}(n, r)\}_{i \in \mathbb{Z}}$ . The work in [20,21] studies a different method to recover the persistent homology of  $f$  using Rips complexes. In order to recover  $\text{PH}_*(f)$ , [20] considers the maps  $\iota_*^L : H_*(R_L(n, r)) \hookrightarrow H_*(R_L(n, 2r))$  induced by inclusion for all values of  $L$  and for a fixed  $r$ . The persistence module for the family of images  $\{\text{Im}(\iota_*^L)\}_L$  is then used as an approximation for  $\text{PH}_*(f)$ . In a way, one can think of the transition  $R_L(n, r) \hookrightarrow R_L(n, 2r)$  as playing the same role as the transition  $R_{L+\varepsilon}(n, r) \hookrightarrow R_{L-\varepsilon}(n, r)$  we study in this paper, “filtering” the noisy homology. Changing the radius rather than the level, allows one to avoid the level discretization that our method relies on, which leads to a more accurate approximation. On the other hand, this method requires further assumptions on the model parameters, and computing the estimator is more complicated. It remains future work to study whether these two methods could be combined into a more powerful and robust one.

In a different line of work [15,22,37] persistent homology is recovered by constructing a kernel-based estimator  $\hat{f}$  for the function at hand and then computing the persistent homology of the estimator  $\text{PH}(\hat{f})$ . The work in [62] presents a different approach by recovering the sublevel sets of distance-like functions called “kernel distance” functions. The validity of these methods is established by using the stability theorem [23] stating that  $d_B(\text{PH}_*(f), \text{PH}_*(\hat{f})) \leq \|f - \hat{f}\|_\infty$ . There are two significant advantages to the estimator we propose in this paper. First, we do not require assumptions about the global sup-norm convergence of the estimator. Second, computing the estimator  $\text{PH}(\hat{f})$  in practice involves discretizing the space, and this may have a significant effect on the ability to recover small features in the data (see, e.g., the clustering examples in Section 5). The estimator we propose does not require such a discretization.



## 4. Computing the homology estimator

The estimator we propose in Section 3 requires the computation of the image between the homology groups of  $\hat{D}_{L+\varepsilon}(n, r)$  and  $\hat{D}_{L-\varepsilon}(n, r)$  (defined in (3.3)). As a review for a more statistical audience, we state the fundamental tools required to compute this estimator. In general, algorithms for computing homology of unions of balls require two steps. The first step is to obtain a combinatorial representation of the geometric object that is either equivalent in homology or approximately equivalent in homology to the original geometric object. This step is outlined in Section 4.1. The combinatorial representation reduces homology computation to a linear algebra problem. The second step is to apply a set of linear transformations to this combinatorial representation to compute the image of the homology groups under the inclusion map between two complexes. This step is outlined in Section 4.2.

### 4.1. The Čech and Vietoris–Rips complex

Let  $S$  be a set, and  $\Sigma \subset 2^S$  be a collection of finite subsets of  $S$ . We say that  $\Sigma$  is an *abstract simplicial complex* if for every  $A \in \Sigma$  and  $B \subset A$  we also have  $B \in \Sigma$ . In this section we introduce two special types of abstract simplicial complex that can be useful for computing the estimators presented in this paper.

Let  $\mathcal{X} = \{x_1, \dots, x_n\}$  be a set of points in  $\mathbb{R}^d$ , and suppose that we wish to compute the homology of the union of balls  $U(\mathcal{X}, r)$  (see (3.2)) for some  $r > 0$ . The Čech complex is an abstract simplicial complex that allows us to convert the homology computation problem into linear algebra. The Vietoris–Rips (or just Rips) complex can be thought of as an approximation to the Čech complex. This approximation offers computational advantages over the Čech complex but suffers from not sharing the same direct relation to the homology of  $U(\mathcal{X}, r)$  as the Čech complex. We first provide the definitions for these complexes.

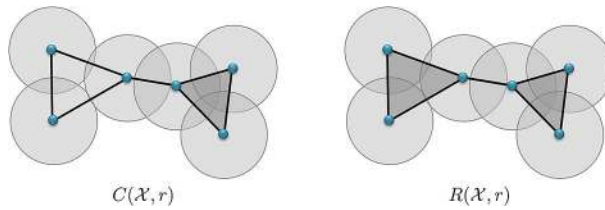
**Definition 4.1 (Čech complex).** Let  $\mathcal{X} = \{x_1, x_2, \dots, x_n\}$  be a collection of points in  $\mathbb{R}^d$ , and let  $r > 0$ . The Čech complex  $C(\mathcal{X}, r)$  is constructed as follows:

1. The 0-simplices (vertices) are the points in  $\mathcal{X}$ .
2. A  $k$ -simplex  $[x_{i_0}, \dots, x_{i_k}]$  is in  $C(\mathcal{X}, r)$  if  $\bigcap_{j=0}^k B_r(x_{i_j}) \neq \emptyset$ .

**Definition 4.2 (Rips complex).** Let  $\mathcal{X} = \{x_1, x_2, \dots, x_n\}$  be a collection of points in  $\mathbb{R}^d$ , and let  $r > 0$ . The Rips complex  $R(\mathcal{X}, r)$  is constructed as follows:

1. The 0-simplices (vertices) are the points in  $\mathcal{X}$ .
2. A  $k$ -simplex  $[x_{i_0}, \dots, x_{i_k}]$  is in  $R(\mathcal{X}, r)$  if  $\|x_{i_j} - x_{i_l}\| \leq 2r$  for all  $0 \leq j, l \leq k$ .

Figure 7 depicts a simple example of a Čech and Rips complex in  $\mathbb{R}^2$ . The figure also highlights the contrast between the two complexes. The main difference is that the Rips complex is constructed simply from pairwise intersection information while the Čech complex requires high-order information. This difference is realized in Figure 7 in the far left triangle in either complex. In the Rips complex, the left triangle is filled in to be a face, since all three pairwise



**Figure 7.** On the left – the Čech complex  $C(\mathcal{X}, r)$ , on the right – the Rips complex  $R(\mathcal{X}, r)$  with the same set of vertices and the same radius. We see that the three left-most balls do not have a common intersection and therefore do not generate a 2-dimensional face in the Čech complex. However, since all the pairwise intersections occur, the Rips complex does include the corresponding face.

intersections occur. In the Čech complex higher-order interactions are also computed, in this case one observes that the three pairwise intersections do not overlap resulting in three edges rather than a filled in face. The main advantage of the Rips complex is computational – all we need in order to construct the Rips complex is to compute the pairwise distances between all the points, rather than to check for all possible orders of intersections of balls as we would have to for the Čech complex.

The Rips complex can be considered as an approximation to the Čech complex. It is clear from the definitions that  $C(\mathcal{X}, r) \subset R(\mathcal{X}, r)$ . In addition, it is shown in [29] that  $R(\mathcal{X}, r) \subset C(\mathcal{X}, \sqrt{2}r)$ . Combining these two statements we have that

$$R(\mathcal{X}, r) \subset C(\mathcal{X}, \sqrt{2}r) \subset R(\mathcal{X}, \sqrt{2}r).$$

An important result in algebraic topology called the “Nerve lemma” (cf. [13]) states that the Čech complex  $C(\mathcal{X}, r)$  is homotopy equivalent to the neighborhood set  $U(\mathcal{X}, r)$ . In particular it follows  $H_*(C(\mathcal{X}, r)) \cong H_*(U(\mathcal{X}, r))$ . As a consequence, any statement made about the homology of  $U(\mathcal{X}, r)$  applies to  $C(\mathcal{X}, r)$  and vice versa.

Denote the Čech complex generated by the filtered point set  $\mathcal{X}_n^L$  as  $C_L(n, r) \triangleq C(\mathcal{X}_n^L, r)$ . We can then define

$$i_* : H_*(C_{L+\varepsilon}(n, r)) \rightarrow H_*(C_{L-\varepsilon}(n, r))$$

to be the map induced by the inclusion map between the simplicial complexes. Defining

$$\hat{H}_*^C(L, \varepsilon; n) \triangleq \text{Im}(i_*),$$

then by the Nerve lemma, since  $\hat{D}_{L\pm\varepsilon}(n, r)$  and  $C_{L\pm\varepsilon}(n, r)$  are completely equivalent structures (in terms of homology), Theorem 3.3 holds without changes for  $\hat{H}_*^C(L, \varepsilon; n)$ .

Next, we denote the Rips complex constructed from the filtered sample as  $R_L(n, r) \triangleq R(\mathcal{X}_n^L, r)$  and define the following inclusion map for any  $\varepsilon \in (0, L/2)$

$$\iota : R_{L+\varepsilon}(n, r) \hookrightarrow R_{L-\varepsilon}(n, r).$$

This inclusion induces a map in homology

$$\iota_* : H_*(R_{L+\varepsilon}(n, r)) \rightarrow H_*(R_{L-\varepsilon}(n, r)),$$

and we denote

$$\hat{H}_*^R(L, \varepsilon; n) \triangleq \text{Im}(t_*).$$

Note that the Nerve lemma applies only to the Čech complex and not the Rips. Nevertheless, the following theorem states that we can compute the homology of  $D_L$  using the Rips complex as well. The importance of providing a consistent estimator for  $H_*(D_L)$  that uses the Rips complex is due to its computational efficiency.

**Theorem 4.3.** *Let  $L > 0$  and  $\varepsilon \in (0, L/2)$  be such that the function  $f(x)$  has no critical values in the range  $[L - 2\varepsilon, L + 2\varepsilon]$ . If  $r \rightarrow 0$  and  $nr^d \rightarrow \infty$ , then for  $n$  large enough we have*

$$\mathbb{P}(\hat{H}_*^R(L, \varepsilon; n) \cong H_*(D_L)) \geq 1 - 6ne^{-C_\varepsilon^* nr^d}.$$

In particular, if  $nr^d \geq D \log n$  with  $D > (C_\varepsilon^*)^{-1}$ , then

$$\lim_{n \rightarrow \infty} \mathbb{P}(\hat{H}_*^R(L, \varepsilon; n) \cong H_*(D_L)) = 1.$$

In the next subsection, we provide an algorithm for computing the image of the inclusion map using either the Čech or Rips complex.

## 4.2. Computing the homology of the image

Our estimator for  $H_k(D_L)$  requires the computation of the image of the map between the homology of two nested simplicial complexes  $\Delta^{(1)} \subset \Delta^{(2)}$  (either Čech or Rips). This map is denoted by  $t_k : H_k(\Delta^{(1)}) \rightarrow H_k(\Delta^{(2)})$ . In this section, we present an algebraic algorithm to compute the rank of this image, namely the estimated Betti number  $\beta_k$ . Note that there are several efficient algorithms to compute persistent homology that can also be used here (see [1,35,50]). We present a relatively simple algorithm, in the interest of clarity for a statistical audience, for the case where  $\mathbb{F}$  is a field of characteristic zero (e.g.,  $\mathbb{R}, \mathbb{Q}$ ). For a fixed homology degree  $0 \leq k \leq d$  the algorithm will consist of two steps:

- (1) Finding a basis for the kernel of a square matrix defined later as  $L_k^{(1)}$ .
- (2) Computing the rank of two matrices, defined later as  $\partial_{k+1}^{(2)}$  and  $\hat{\partial}_{k+1}^{(2)}$ , and then we will have that

$$\text{rank}(\text{Im}(t_k)) = \text{rank}(\hat{\partial}_{k+1}^{(2)}) - \text{rank}(\partial_{k+1}^{(2)}).$$

In the following, we provide more details about homology computation for simplicial complexes, and in particular the definitions of the matrices  $L_k^{(1)}$ ,  $\partial_{k+1}^{(2)}$ , and  $\hat{\partial}_{k+1}^{(2)}$  mentioned above.

### 4.2.1. Computing the homology of a simplicial complex

Let  $\Delta$  be a simplicial complex, let  $\Delta_k$  be the set of  $k$ -simplexes in  $\Delta$ , and let  $n_k = |\Delta_k|$ , so we can write

$$\Delta_k = \{\sigma_1, \sigma_2, \dots, \sigma_{n_k}\}.$$

We assume that every  $k$ -simplex  $\sigma_i \in \Delta_k$  is attached with a unique orientation (an ordering on its set of vertices), denoted by  $\sigma_i = [x_0^i, \dots, x_k^i]$ . Defining  $C_k \triangleq \mathbb{F}^{n_k}$ , we wish to map the simplexes of  $\Delta_k$  into a basis of  $C_k$  in a way that preserves orientation information. To do that we first define  $\Delta_k^\pi$  to be the set containing all the simplexes in  $\Delta_k$  in all possible orientations. We then define the map  $T_k : \Delta_k^\pi \rightarrow C_k$  in the following way. For every simplex  $\sigma_i \in \Delta_k$  we define  $T_k(\sigma_i) = \mathbf{e}_i$ , where  $\mathbf{e}_i$  consists of one at the  $i$ th entry, and zero elsewhere. For every permutation  $\pi$  on  $0, \dots, k$ , we then define

$$T_k([x_{\pi(0)}^i, \dots, x_{\pi(k)}^i]) = \text{sign}(\pi)\mathbf{e}_i,$$

where  $\text{sign}(\pi) = (-1)^{P(\pi)}$ , and  $P(\pi)$  is the parity of the permutation  $\pi$ . The vector space  $C_k$  is usually referred to as the “space of  $k$ -chains” of  $\Delta$ .

Next, using the map  $T_k$ , we define the matrix  $\partial_k$  to be a  $n_{k-1} \times n_k$  matrix where the  $i$ th column is given by

$$(\partial_k)_i = \sum_{\substack{\sigma \in \Delta_{k-1} \\ \text{is a face of } \sigma_i}} T_{k-1}(\sigma).$$

We note that the orientation of  $\sigma$  used in the sum is the one inherited from the orientation of  $\sigma_i$ . In other words, the nonzero entries in the  $i$ th column correspond to the  $(k-1)$ -dimensional faces of  $\sigma_i \in \Delta_k$  (with the proper sign representing their orientation). The matrix  $\partial_k$  can be thought of as a linear transformation from  $C_k$  to  $C_{k-1}$  and is referred to as “the boundary operator.” The  $k$ th homology of  $\Delta$  is then defined to be the quotient space given by

$$H_k(\Delta) \triangleq \ker(\partial_k) / \text{Im}(\partial_{k+1}). \tag{4.1}$$

One way to find a basis for  $H_k(\Delta)$  is via the combinatorial Laplacian, defined as the following  $n_k \times n_k$  matrix

$$L_k \triangleq \partial_{k+1} \partial_{k+1}^T + \partial_k^T \partial_k.$$

Note that  $L_0$  is the well-known graph Laplacian. If  $\mathbb{F}$  is a field with characteristic zero (e.g.  $\mathbb{R}, \mathbb{Q}$ ) then it is shown in [38] that the kernel of  $L_k$  is isomorphic to  $H_k(\Delta)$  and in particular, the Betti numbers of  $\Delta$  are given by  $\beta_k(\Delta) = \dim(\ker(L_k))$ .

#### 4.2.2. The homology of the map

Our goal is not only to compute the homology of  $\Delta^{(1)}$  and  $\Delta^{(2)}$  separately, but rather to compute the image of the map  $\iota_k : H_k(\Delta^{(1)}) \rightarrow H_k(\Delta^{(2)})$ . For  $j = 1, 2$  let  $\Delta_k^{(j)}$  be the set of  $k$ -simplexes in  $\Delta^{(j)}$ , and let  $n_k^{(j)} = |\Delta_k^{(j)}|$ . Since  $\Delta^{(1)} \subset \Delta^{(2)}$  we can list the simplexes in the following way:

$$\begin{aligned} \Delta_k^{(1)} &= \{\sigma_1, \sigma_2, \dots, \sigma_{n_k^{(1)}}\}, \\ \Delta_k^{(2)} &= \{\sigma_1, \sigma_2, \dots, \sigma_{n_k^{(1)}}, \sigma_{n_k^{(1)}+1}, \dots, \sigma_{n_k^{(2)}}\}. \end{aligned}$$

Using this ordering on the simplexes, we define the boundary operators  $\partial_k^{(j)}$  and the combinatorial Laplacians  $L_k^{(j)}$  for each of the complexes. It is then easy to see that

$$\partial_k^{(2)} = \begin{pmatrix} \partial_k^{(1)} & \cdots \\ 0 & \ddots \end{pmatrix}. \quad (4.2)$$

Now, if  $\{v_1, \dots, v_m\} \subset C_k^{(1)}$  is a basis for  $\ker(L_k^{(1)})$  then it represents a basis for  $H_k(\Delta^{(1)})$ , such that  $\beta_k(\Delta^{(1)}) = m$ . Let  $\hat{v}_i \in C_k^{(2)}$  be a zero padded version of  $v_i \in C_k^{(1)}$ . From (4.1), we know that  $v_i \in \ker(\partial_k^{(1)})$ , and thus from (4.2) it is clear that  $\hat{v}_i \in \ker(\partial_k^{(2)})$  as well. This implies that the vectors in  $\{\hat{v}_1, \dots, \hat{v}_m\}$  are candidates to form a basis for  $\text{Im}(t_k)$ . Note, however, that while  $\hat{v}_i \in \ker(\partial_k^{(2)})$ , it is possible that some linear combinations of  $\hat{v}_1, \dots, \hat{v}_m$  are in  $\text{Im}(\partial_{k+1}^{(2)})$ , which means that they are considered as trivial in  $H_k(\Delta^{(2)})$ . This means that  $\{\hat{v}_1, \dots, \hat{v}_m\}$  might be larger than a basis for  $\text{Im}(t_k)$ , and we need to reduce this set. This can be done by solving several sets of linear equations, which we avoid describing here. However, the rank of  $\text{Im}(t_k)$  can be computed easily by

$$\text{rank}(\text{Im}(t_k)) = \text{rank}(\hat{\partial}_{k+1}^{(2)}) - \text{rank}(\partial_{k+1}^{(2)}),$$

where

$$\hat{\partial}_{k+1}^{(2)} = (\partial_{k+1}^{(2)}, \hat{v}_1, \dots, \hat{v}_m)$$

is a  $n_k^{(2)} \times (n_{k+1}^{(2)} + m)$  matrix we get by concatenating the boundary matrix  $\partial_{k+1}^{(2)}$  with the column vectors  $\hat{v}_i$ . In other words, we measure how many vectors from the set  $\{\hat{v}_1, \dots, \hat{v}_m\}$  can be added to the set of column vectors of  $\partial_{k+1}^{(2)}$  without generating linear dependency.

## 5. Results on simulated data

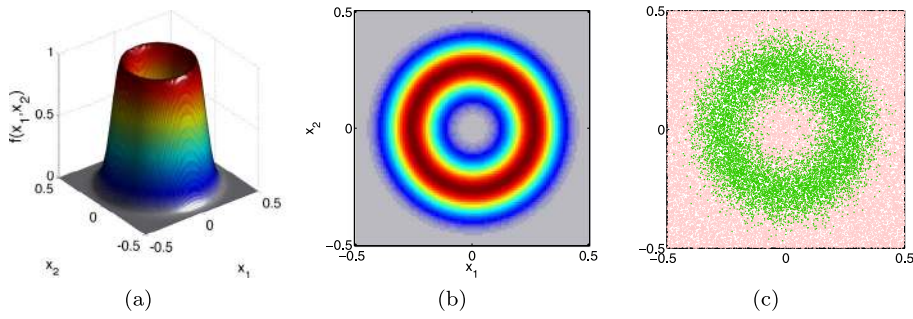
In this section, we illustrate how we can use the methods in Section 3 for data analysis using some simulated examples. The examples we chose relate to classical problems in statistics: classification, non-parametric regression, and clustering. We use these examples to demonstrate the novelty and strength of the methods proposed in this paper.

### 5.1. Binary regression

We illustrate how we can recover the homology of a classification function. The marginal density of the explanatory variables is uniform in the unit square  $X \sim U([-1/2, 1/2]^2)$ . We then set the conditional probability of the binary response  $Y$  as

$$\mathbb{P}(Y = 1 \mid X = x) = f(x) \triangleq C(1 + \sin(4\pi \|x\|^2))e^{-100(\|x\| - 1/4)^2}, \quad (5.1)$$

where  $C$  is a normalization factor guaranteeing that  $f(x)$  is indeed a conditional probability. The graph of this conditional probability is given in Figure 8.



**Figure 8.** (a) The graph of the conditional probability on the unit square. (b) The level sets of the image of the conditional probability. (c) For a set of points drawn from the marginal distribution on the unit square we label them red or green based on the conditional probability given by (5.1). The green points are those assigned to a response of *one* and the red points are those assigned zeros.

We generate i.i.d. observations  $\{(X_1, Y_1), \dots, (X_n, Y_n)\}$  from the joint distribution and our objective is to recover the topology of the level set  $D_L$  for  $L = 0.5$  which is used as the binary classifier in this case, and has the shape of an annulus. We use the Rips construction presented in Theorem 4.3, with  $n = 50,000$ ,  $r = 0.01$ , and  $\varepsilon = 0.2$ . This gives us two complexes:  $S_1 = R_{0.3}(n, r)$  and  $S_2 = R_{0.7}(n, r)$ . Figure 9 shows the sets of disks used to create the two Rips complexes. The light blue disks are the ones corresponding to  $S_1$  and the orange ones corresponds to  $S_2$ . Computing the Betti numbers yields:

	$S_1$	$S_2$	$S_1 \hookrightarrow S_2$
$\beta_0$	34	53	1
$\beta_1$	23	49	1

Indeed, while the homology of each of the complexes  $S_1, S_2$  is extremely noisy, the image of the map between them looks exactly like an annulus.

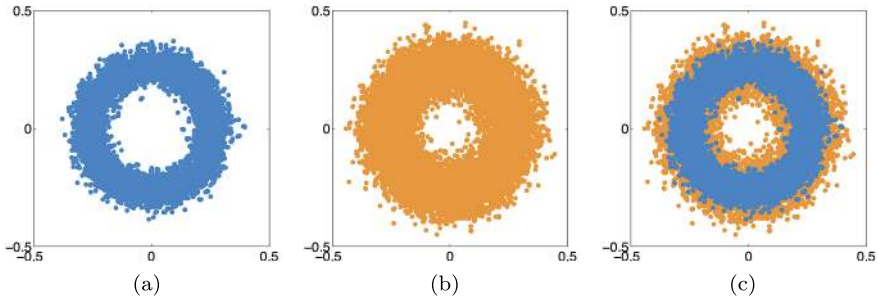
### 5.2. Kernel regression

In this example, we consider a regression function on the unit square  $f : [-1, 1]^2 \rightarrow \mathbb{R}$  with additive noise

$$Y_i = f(X_i) + \xi_i. \tag{5.2}$$

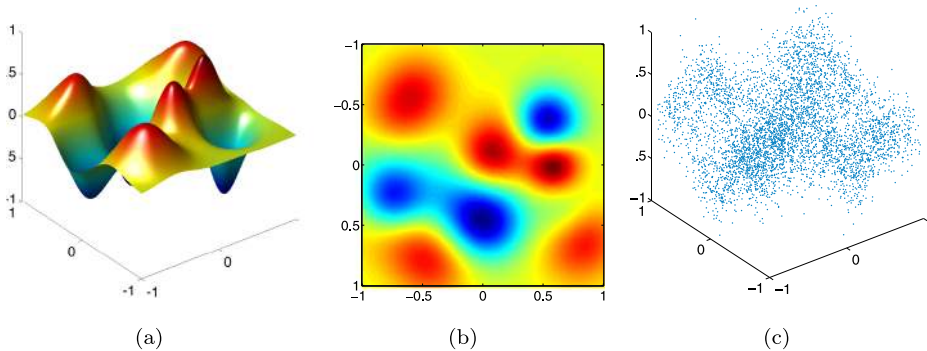
Our objective will be to recover the barcode or persistent homology of the above function from noisy observations.

The regression function  $f$  was generated from a random mixture of Gaussians, and its graph is presented in Figure 10(a). The “true” barcode of the function  $f$  is presented in Figure 11(a).

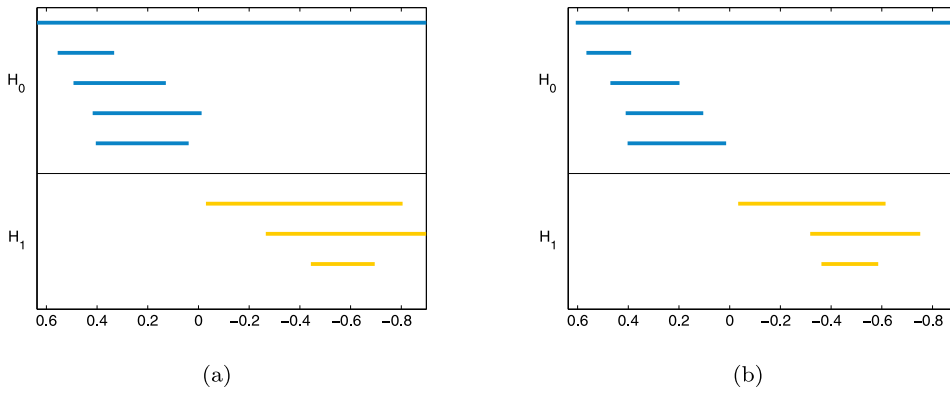


**Figure 9.** Computing the homology of a level set for a regression function. We generated  $\{(X_i, Y_i)\}_{i=1}^{50,000}$  i.i.d. observations from the marginal and conditional distributions given in equation (5.1). For  $L = 0.5$  and  $\varepsilon = 0.2$  we present the following: (a) the set  $\hat{D}_{L+\varepsilon}(n, r)$ , (b) the set  $\hat{D}_{L-\varepsilon}(n, r)$ , (c) the two sets combined. Note that both individual sets in (a) and (b) contain many connected components and cycles. However, in (c) we observe that most of these homological features do not survive the transition. All the extra connected components in (a) are merged into the large component in (b). Similarly, all the extra cycles in (a) are filled up in (b).

This barcode was computed by evaluating  $f$  directly on a dense grid and computing the persistent homology of this discretized version. The independent variables  $X_i$  are generated from a uniform distribution in the box  $[-1, 1]^2$ . The noise  $\xi_i$  is independent of  $X_i$ , and generated by a normal distribution with  $\sigma = 0.2$  truncated at  $5\sigma$  (we require in (C2) for the response variables to be bounded). To estimate this barcode, we used  $\widehat{\text{PH}}_*^\varepsilon(f)$  (defined in Section 3.2) with  $n = 5000$ ,  $r = 0.1$ ,  $\varepsilon = 0.001$ . The result is presented in Figure 11(b).



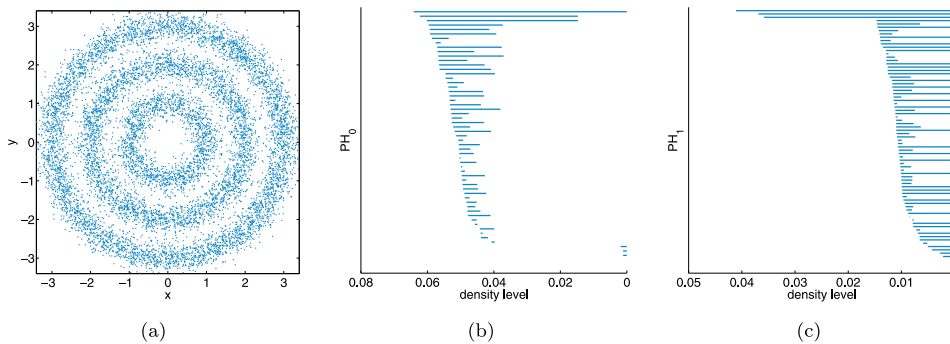
**Figure 10.** A regression function in  $\mathbb{R}^2$ . (a) The graph of the function in the box  $[-1, 1]^2$ . (b) The level sets of the function. It is easy to spot five peaks and three valleys in this image, which in persistent homology correspond to five features in  $\text{PH}_0$  and three in  $\text{PH}_1$ . (c) Generating  $\{(X_i, Y_i)\}_{i=1}^{5000}$  i.i.d. observations from the model presented in (5.2).



**Figure 11.** (a) The “true” barcode of the persistent homology of the regression function  $f$  presented in Figure 10. (b) The estimated persistent homology  $\widehat{PH}_*^\varepsilon(f)$ , with  $n = 5000$ ,  $r = 0.1$ ,  $\varepsilon = 0.001$ , is very close to the true barcode. For visualization purposes, we left bars with length less than 0.05 out of the figure. In both the true and the estimated barcodes we observe five significant features in  $H_0$  and three in  $H_1$ , corresponding to the five peaks and three valleys in the graph of the function  $f$ .

### 5.3. Dataset related to spectral clustering

Spectral clustering uses spectral graph theory to cluster observations (see the review papers in [57,71]). It is mostly useful in cases where the clusters are not necessarily concentrated close to a single point, but have a more complicated shape (such as the data in Figure 12). We revisit a simulated example from the spectral clustering literature to illustrate how well we can recover the number of clusters and cluster features using our level sets approach. We generate  $n = 10,000$  points from three concentric circles (of radii 1, 2, 3) and added multivariate Gaussian noise with  $\sigma = 0.2$ . The result is presented in Figure 12(a). The topological features we wish to recover here



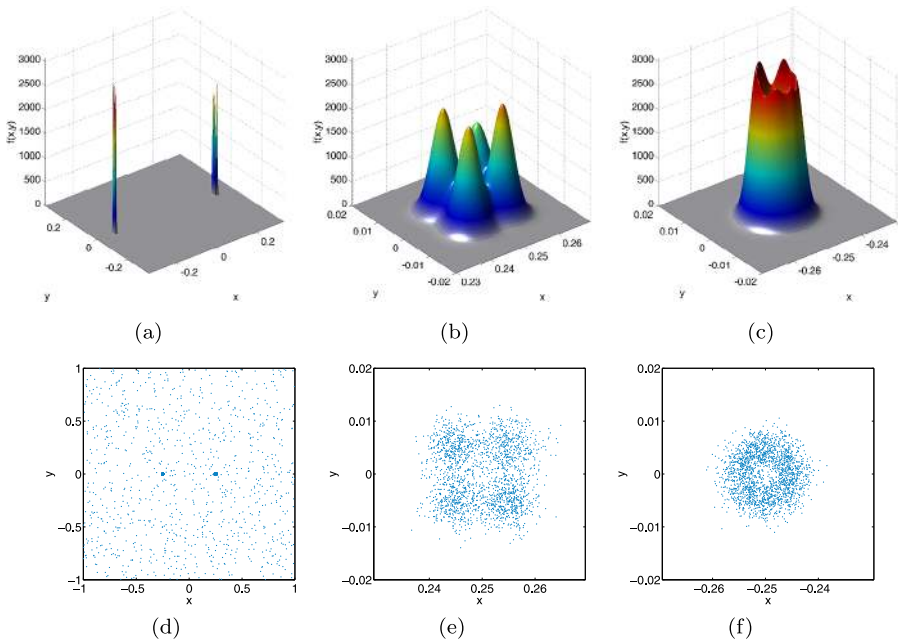
**Figure 12.** (a) A sample set generated from three concentric circles. (b) The barcode for  $\widehat{PH}_0^\varepsilon(f)$ , where we indeed observe three dominating components. (c) The barcode for  $\widehat{PH}_1^\varepsilon(f)$ , where we indeed observe three dominating cycles. The parameters used in this simulation are  $n = 10,000$ ,  $r = 0.125$ ,  $\varepsilon = 0.005$ .



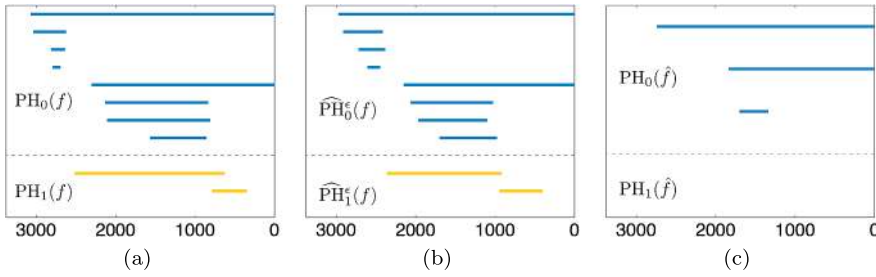
are the three connected components and the three cycles (spectral clustering would find the three connected components). The parameters we used are  $r = 0.125$ ,  $\varepsilon = 0.005$ . Figure 12(b) displays  $\widehat{\text{PH}}_0^\varepsilon(f)$ . Here we see that there are indeed three dominating features (bars that persist over a long period of time). The rest of the features are generated by the fluctuations in the estimated density function. Similarly, in Figure 12(c) we observe three dominating features as well, representing the three cycles in the data.

## 5.4. Hierarchical clustering

This example will be used to show how using our method we can capture features of a density function with hierarchical structure. Consider a probability density  $f$  on  $\mathbb{R}^2$  that consists of two concentrated densities that are far apart and centered at  $(\pm 0.25, 0)$ , see Figure 13(a). Once we zoom into the two densities we realize there is a finer structure in this problem. The density around  $(0.25, 0)$  is a mixture of four Gaussians that are very near each other, see Figure 13(b). The density around  $(-0.25, 0)$  is one density that looks like a volcano crater (made of a mixture of 100 Gaussians), see Figure 13(c). The result of this finer structure is that when we examine



**Figure 13.** A hierarchical density function. (a) The density function at a coarse level, consisting of two sharp peaks. (b) Zooming in on the density around  $(0.25, 0)$  we observe that this sharp peak actually consists of four adjacent peaks. (c) Zooming in on the density around  $(-0.25, 0)$  we observe that the peak has a crater-like structure with small fluctuation around the rim. (d)–(f) A sample of  $n = 5000$  points generated by  $f$ .



**Figure 14.** Estimating the persistent homology of the density function  $f$  presented in Figure 13. (a) The “true” barcode for the function  $f$ , that is,  $PH_*(f)$  (computed by sampling the density function on a fine grid). (b) The barcode computed from the estimator  $\widehat{PH}_*^\varepsilon(f)$ . The parameters used are  $n = 5000$ ,  $r = 0.001$ ,  $\varepsilon = 3.5$ . (c) The barcode computed for the kernel density estimator –  $PH_*(\hat{f})$ . The kernel parameters are the same as for  $\widehat{PH}_*^\varepsilon(f)$ , the grid size taken is  $500 \times 500$ . Note that the estimator  $\widehat{PH}_*^\varepsilon(f)$  gives a result that is very similar to the true barcode. In both cases there are five significant features in  $H_0$  and two significant features in  $H_1$ . The barcode for  $PH_*(\hat{f})$  only recover the coarse features, namely the two clusters, but completely ignores the finer structures. We note that for visualization purposes we filtered out the very small bars before drawing the barcodes here.

the persistence homology of  $f$  we expect to see: (1) five dominating features in  $PH_0$  – the four bumps on the right, and the entire volcano on the left, (2) two dominating features in  $PH_1$  – one coming from the cycle along the rim of the volcano, and another one from the cycle that surrounds the four bumps, (3) fluctuations on the rim will introduce features in  $PH_0(f)$  but these will have low persistence. We will show how we can accurately capture the homology of this hierarchical structure.

The barcode in Figure 14(a) displays the “true” persistent homology  $PH_*(f)$  that was computed by evaluating the function values directly on a very fine grid around the peaks. Looking at the barcode of  $PH_0(f)$ , we see two dominant features, with death time close to zero. These two features correspond to the two clusters represented by the peaks seen in Figure 13(a). The other three dominant features correspond to the three additional peaks we have in Figure 13(b). The rest of the bars (as well as other shorter bars we kept out of the figure for visualization purposes) correspond to the fluctuation along the rim of the crater in Figure 13(c). In  $PH_1(f)$ , we see exactly two features corresponding to the two cycles described above.

We can compare the true barcode to the barcode generated by our estimator for  $PH_*(f)$  using  $\widehat{PH}_*^\varepsilon(f)$ . The parameters we used in the estimator are  $n = 5000$ ,  $r = 0.001$ ,  $\varepsilon = 3.5$ . The barcode for  $\widehat{PH}_*^\varepsilon(f)$  is presented in Figure 14(b). The global picture is very consistent with that of the true function. As expected our estimates have extra variation in the endpoints of the bars.

In Fasy *et al.* [37], an alternate approach is developed to estimate  $PH_*(f)$ . Their idea is to use a kernel density estimation to obtain an estimate  $\hat{f}_n$  of the density  $f$ . Then they compute the persistent homology of  $\hat{f}_n$ , denoted by  $PH_*(\hat{f}_n)$ . They are able to provide a theoretical bound on the bottleneck distance between  $PH_*(f)$  and  $PH_*(\hat{f}_n)$ . This result is similar in spirit to Theorem 3.7 in our paper. The main difference in their method versus our method is that they focus on getting a good estimate of the function values or ensuring  $\hat{f}_n \approx f(x)$  everywhere, whereas we compute  $\widehat{PH}_*^\varepsilon(f)$  by approximating the level sets directly.

In the case of a density function with hierarchical structure, these two approaches often have different empirical performance. In particular, we argue that the estimator  $\widehat{\text{PH}}_*^\varepsilon(f)$  is favorable to  $\text{PH}_*(\hat{f})$ . The crux of the argument in favor of computing  $\widehat{\text{PH}}_*^\varepsilon(f)$  is that in evaluating the fit of  $\hat{f}$  there is a resolution parameter of how fine in  $\mathbb{R}^2$  one measures  $f$ , which we denote as  $\Delta$  (in addition to the bandwidth parameter of the kernel  $-r$ ). The problem arises in that one needs to know what value of  $\Delta$  is small enough to capture fine structure in  $f$ . This raises two issues: (1) how to adaptively estimate  $\Delta$  from data and (2) taking a finer resolution parameter will result in an increase in the sample complexity of the inference problem. Our approach of directly estimating  $\widehat{\text{PH}}_*^\varepsilon(f)$  avoids these difficulties, since we only work with the original sample points rather than  $\hat{f}$ . In Figure 14(c), we present the barcode for  $\text{PH}_*(\hat{f})$ , computed using the same kernel, on a grid of size  $500 \times 500$  (i.e.,  $\Delta = 1/250$ ).

## 6. Conclusion

In this paper, we introduce a consistent estimator for the homology of level sets for both density and regression functions. We apply this procedure to infer the homology of a manifold from noisy observations, and infer the persistent homology of either density or regression functions. The conditions we require are weaker than previous results in this direction.

We view this work as an important step in closing the gap between topological data analysis and statistics. For topological data analysis, we provide a consistent estimator for the homology and persistent homology of spaces underlying random data. As future work, we will consider refinements of our analysis to obtain convergence rates and confidence intervals of the estimates. We suspect this will require more assumptions on the geometry of the underlying spaces. From a statistical perspective, this work suggests that topological summaries of density and regression functions are of interest and provide insights in statistical modeling. We suspect these characteristics or topological summaries will be very useful in classification or hypothesis testing problems, when the assumptions on different decision regions can be naturally captured by coarse geometry or topology.

## Appendix: Proofs

In this section we provide the proofs for Theorems 3.3, 3.6, 3.7, and 4.3.

### A.1. Some definitions and lemmas

Recall that

$$\mathcal{X}_n^L \triangleq \{X_i : \hat{f}_n(X_i) \geq L; 1 \leq i \leq n\}.$$

Our first step would be to assign some probabilistic quantification of the accuracy of the assignments  $\mathcal{X}_n^L$  with respect to  $D_L$ . We will do this by first defining two sets: the set  $D_{L,r}^\uparrow$  corresponds

to “inflating”  $D_L$  by a radius  $r$  and  $D_{L,r}^\downarrow$  corresponds to “deflating”  $D_L$  by a radius  $r$ . To define these sets, we first define the tube of radius  $r$  around the boundary of  $D_L$

$$\partial D_L^r = \bigcup_{x \in \partial D_L} B_r(x), \quad \partial D_L \text{ is the boundary of } D_L.$$

We then define  $D_{L,r}^\uparrow$  and  $D_{L,r}^\downarrow$  as follows

$$D_L^\uparrow(r) = D_L \cup \partial D_L^r, \quad D_L^\downarrow(r) = D_L \setminus \partial D_L^r.$$

Using these definitions the following lemma provides a bound on the false positive and false negative error of the set  $\mathcal{X}_n^L$  with respect to  $D_L$ .

**Lemma A.1.** *Assume that constraint (C1) on the kernel function holds and either condition (C2) holds for the regression case or in the density estimation case the density is bounded and tame. For every  $L > 0$ , and  $\varepsilon \in (0, L)$ , if  $r \rightarrow 0$  and  $nr^d \rightarrow \infty$ , then there exists a constant  $C_\varepsilon^*$  such that for  $n$  large enough we have*

$$\mathbb{P}(\exists X_i \notin D_{L-\varepsilon}^\uparrow(r) : \hat{f}_n(X_i) \geq L) \leq ne^{-C_\varepsilon^* nr^d}, \tag{A.1}$$

and

$$\mathbb{P}(\exists X_i \in D_{L+\varepsilon}^\downarrow(r) : \hat{f}_n(X_i) \leq L) \leq ne^{-C_\varepsilon^* nr^d}. \tag{A.2}$$

Equation (A.1) bounds the probability of a false-positive error, and equation (A.2) bounds the probability of a false-negative error. The value of  $C_\varepsilon^*$  is different for density estimation versus regression and is given by (3.7) and (3.8).

Next, recall that

$$\hat{D}_L(n, r) \triangleq U(\mathcal{X}_n^L, r).$$

We would like to prove that with a high probability this empirical set is sandwiched by two sets which should be “close” to  $D_L$ . The following lemma states the precise result.

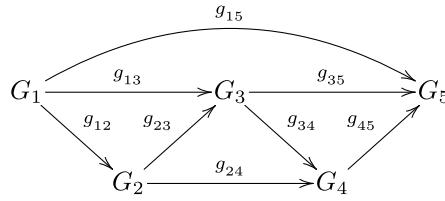
**Lemma A.2.** *For every  $L > 0$ , and  $\varepsilon \in (0, L)$ , if  $r \rightarrow 0$  and  $nr^d \rightarrow \infty$ , then for large enough  $n$  we have*

$$\mathbb{P}(D_{L+\varepsilon}^\downarrow(2r) \subset \hat{D}_L(n, r) \subset D_{L-\varepsilon}^\uparrow(2r)) \geq 1 - 3ne^{-C_\varepsilon^* nr^d},$$

In other words, our estimator  $\hat{D}_L(n, r)$  is sandwiched between the two non-random approximations of  $D_L$ .

The last ingredient we need for the proving the theorems is the following purely algebraic lemma.

**Lemma A.3.** Consider the following commutative diagram of groups,



(by “commutative” we mean that all paths with the same endpoints lead to the same result), and for every  $i, j$  define  $G_{ij} = \text{Im}(g_{ij}) \subset G_j$ .

If  $g_{35} : G_3 \rightarrow G_5$  is an isomorphism from  $G_3$  to  $G_{35}$ . Then the map  $g_{34} : G_3 \rightarrow G_4$  is an isomorphism from  $G_3$  to  $G_{24} \subset G_4$ . In particular, we have  $G_3 \cong G_{24}$ .

### A.2. Proving the theorems

**Proof of Theorem 3.3.** Using Lemma A.2 for  $\hat{D}_{L+\varepsilon}(n, r)$  and  $\hat{D}_{L-\varepsilon}(n, r)$  we have that for  $n$  large enough

$$\begin{aligned}
 \mathbb{P}(D_{L+(3/2)\varepsilon}^\downarrow(2r) \subset \hat{D}_{L+\varepsilon}(n, r) \subset D_{L+(1/2)\varepsilon}^\uparrow(2r)) &\geq 1 - 3ne^{-C_{\varepsilon/2}^*nr^d}, \\
 \mathbb{P}(D_{L-\frac{1}{2}\varepsilon}^\downarrow(2r) \subset \hat{D}_{L-\varepsilon}(n, r) \subset D_{L-(3/2)\varepsilon}^\uparrow(2r)) &\geq 1 - 3ne^{-C_{\varepsilon/2}^*nr^d}.
 \end{aligned} \tag{A.3}$$

Since we assume  $L$  is  $\varepsilon$ -regular, if  $r$  is small enough, we have

$$D_{L+2\varepsilon} \subset D_{L+(3/2)\varepsilon}^\downarrow(2r) \subset D_{L+(1/2)\varepsilon}^\uparrow(2r) \subset D_L \subset D_{L-(1/2)\varepsilon}^\downarrow(2r) \subset D_{L-(3/2)\varepsilon}^\uparrow(2r) \subset D_{L-2\varepsilon},$$

and from (A.3) we conclude that

$$\mathbb{P}(D_{L+2\varepsilon} \subset \hat{D}_{L+\varepsilon}(n, r) \subset D_L \subset \hat{D}_{L-\varepsilon}(n, r) \subset D_{L-2\varepsilon}) \geq 1 - 6ne^{-C_{\varepsilon/2}^*nr^d}.$$

Denote

$$S_1 = D_{L+2\varepsilon}, \quad S_2 = \hat{D}_{L+\varepsilon}(n, r), \quad S_3 = D_L, \quad S_4 = \hat{D}_{L-\varepsilon}(n, r), \quad S_5 = D_{L-2\varepsilon},$$

and let  $G_i = H_*(S_i)$ . Since we assume that  $f(x)$  has no critical values in  $[L - 2\varepsilon, L + 2\varepsilon]$ , and using the notation of Lemma A.3 we have that the maps  $g_{13}, g_{35}$  and  $g_{15}$  induced by the inclusions  $S_1 \subset S_3 \subset S_5$  are all isomorphisms. If, in addition,  $S_1 \subset S_2 \subset S_3 \subset S_4 \subset S_5$ , then using Lemma A.3 we conclude that  $G_{24} \cong G_3$ . Observing that  $G_{24} = \text{Im}(t_*)$  (see (3.4)) we have that  $\text{Im}(t_*) \cong H_*(D_L)$  which completes the proof.  $\square$

**Proof of Theorem 3.6.** Recall that  $N_\varepsilon = \sup_{x \in \mathbb{R}^d} \lceil f(x)/2\varepsilon \rceil$ , and  $L_{\max} = 2\varepsilon N_\varepsilon$ . Let  $E$  be the event that for every  $1 \leq i \leq N_\varepsilon$  the following inclusion holds:

$$D_{L_{i-1}} = D_{L_i+2\varepsilon} \hookrightarrow \hat{D}_{L_i+\varepsilon}(n, r) \hookrightarrow D_{L_i} \hookrightarrow \hat{D}_{L_i-\varepsilon}(n, r) \hookrightarrow D_{L_i-2\varepsilon} = D_{L_{i+1}}. \tag{A.4}$$

Applying Lemma A.2 (as in the proof of Theorem 3.3)  $N_\varepsilon$  times we can show that if  $r \rightarrow 0$  and  $nr^d \rightarrow \infty$  then for  $n$  large enough

$$\mathbb{P}(E) \geq 1 - 3nN_\varepsilon e^{-C_\varepsilon^* nr^d}.$$

From here on we will assume that (A.4) is true for all  $1 \leq i \leq N_\varepsilon$ . Choosing  $i^*$  as

$$i^* \triangleq 1 + \min\{i \in \{1, \dots, N_\varepsilon\} : \hat{\beta}_m(L_i, \varepsilon; n) = 1\},$$

our goal is to show that  $[L_{i^*} - 2\varepsilon, L_{i^*} + 2\varepsilon] \subset (A, B)$ , and therefore the arguments used in the proof of Theorem 3.3 guarantee that  $\hat{H}_*(L_{i^*}, \varepsilon; n) \cong H_*(D_{L_{i^*}}) \cong H_*(\mathcal{M})$ .

Since  $\mathcal{M}$  is assumed to be connected, we have that  $\beta_0(\mathcal{M}) = 1$ , and by Poincaré duality (cf. [43,55]) we conclude that  $\beta_m(\mathcal{M}) = 1$ . If  $L_i \in (A, B)$  then from Definition 3.5 we have that  $D_{L_i} \simeq \mathcal{M}$  and thus  $\beta_m(D_{L_i}) = 1$  as well. On the other hand, if  $L_i > B$  then  $D_{L_i} \simeq \mathcal{M}'$  where  $\mathcal{M}'$  is a compact locally contractible proper subset of the  $\mathcal{M}$ . Using Proposition 3.46 in [43] we have that  $\beta_m(\mathcal{M}') = \beta_m(D_{L_i}) = 0$ .

Our requirement that  $L_{i-1} - L_i = 2\varepsilon$  and  $B - A \geq 8\varepsilon$  guarantees that there are at least four consecutive levels  $L_i$  such that  $L_i \in (A, B)$ . Let  $L_{i_1} > L_{i_2} > L_{i_3} > L_{i_4}$  be the first (highest) such levels. For  $k = 2, 3$  we have that  $[L_{i_k} - 2\varepsilon, L_{i_k} + 2\varepsilon] \subset (A, B)$ , and from the proof of Theorem 3.3 and the previous paragraph we conclude that  $\hat{\beta}_m(L_{i_k}, \varepsilon; n) = 1$ . For  $i_1$  however, it is not true that  $[L_{i_1} - 2\varepsilon, L_{i_1} + 2\varepsilon] \subset (A, B)$  and therefore,  $\hat{\beta}_m(L_{i_1}, \varepsilon; n)$  might be either zero or one. Finally, defining  $i^*$  the way we did,  $i^*$  might be either  $i_2$  or  $i_3$ . In both cases we have  $[L_{i^*} - 2\varepsilon, L_{i^*} + 2\varepsilon] \subset (A, B)$ , and that completes the proof.  $\square$

**Proof of Theorem 3.7.** Recall that  $\mathcal{D} = \{D_L\}_{L \in \mathbb{R}}$  is the continuous filtration of the (super) level sets of  $f$ , and  $\hat{\mathcal{D}}^\varepsilon = \{\hat{D}_{L_i}(n, r)\}_{i \in \mathbb{Z}}$  is a discrete approximation. To prove that the corresponding persistent homologies  $\text{PH}_*(f)$ ,  $\widehat{\text{PH}}_*^\varepsilon(f)$  satisfy

$$d_B(\widehat{\text{PH}}_*^\varepsilon(f), \text{PH}_*(f)) \leq 5\varepsilon,$$

we will use the language of  $\varepsilon$ -interleaving introduced in [18]. The first step would be to define a discrete version of the filtration  $\mathcal{D}$  given by

$$\mathcal{D}^\varepsilon \triangleq \{D_{L_i + \varepsilon}\}_{i \in \mathbb{Z}},$$

where  $L_i$  is defined in (3.10). Denote the persistent homology of  $\mathcal{D}^\varepsilon$  by  $\text{PH}_*^\varepsilon(f)$ . Since  $\mathcal{D}^\varepsilon$  is a discrete approximation of the continuous filtration  $\mathcal{D}$ , with step size  $2\varepsilon$ , the maximum difference between  $\text{PH}_*(f)$  and  $\text{PH}_*^\varepsilon(f)$  would be the step size, and thus we have

$$d_B(\text{PH}_*^\varepsilon(f), \text{PH}_*(f)) \leq 2\varepsilon.$$

To prove the theorem, it is therefore enough to show that with a high probability we have  $d_B(\widehat{\text{PH}}_*^\varepsilon(f), \text{PH}_*^\varepsilon(f)) \leq 3\varepsilon$ .

Let  $E$  be the event that we have the following sequence of inclusions:

$$D_{L_0+\varepsilon} \searrow \hat{D}_{L_0}(n, r) \nearrow D_{L_1+\varepsilon} \searrow \hat{D}_{L_1}(n, r) \nearrow D_{L_2+\varepsilon} \searrow \hat{D}_{L_2}(n, r) \nearrow \dots \quad (\text{A.5})$$

Applying Lemma A.2  $N_\varepsilon$  times we can show that if  $n$  is large enough

$$\mathbb{P}(E) \geq 1 - 3nN_\varepsilon e^{-C_\varepsilon^* / 2nr^d}.$$

Using the notation in [18] (A.5) implies that  $\mathcal{D}^\varepsilon$  and  $\hat{\mathcal{D}}^\varepsilon$  are *weakly  $\varepsilon$ -interleaving*. Denoting the persistent homology of  $\hat{\mathcal{D}}^\varepsilon$  by  $\widehat{\text{PH}}_*^\varepsilon(f)$ , using Theorem 4.3 in [18] yields

$$d_B(\widehat{\text{PH}}_*^\varepsilon(f), \text{PH}_*^\varepsilon(f)) \leq 3\varepsilon. \quad (\text{A.6})$$

This completes the proof. □

**Proof of Theorem 4.3.** Consider the following sequence of simplicial complexes,

$$C_{L\pm\varepsilon}(n, r) \hookrightarrow R_{L\pm\varepsilon}(n, r) \hookrightarrow C_{L\pm\varepsilon}(n, \sqrt{2}r).$$

This sequence induces the following sequence in homology

$$H_*(C_{L\pm\varepsilon}(n, r)) \rightarrow H_*(R_{L\pm\varepsilon}(n, r)) \rightarrow H_*(C_{L\pm\varepsilon}(n, \sqrt{2}r)),$$

or equivalently,

$$H_*(\hat{D}_{L\pm\varepsilon}(n, r)) \rightarrow H_*(R_{L\pm\varepsilon}(n, r)) \rightarrow H_*(\hat{D}_{L\pm\varepsilon}(n, \sqrt{2}r)). \quad (\text{A.7})$$

From the proof of Lemma A.2 (see (A.16),(A.17)) we have that

$$\begin{aligned} \mathbb{P}(D_{L+(3/2)\varepsilon}^\downarrow(2r) \not\subset \hat{D}_{L+\varepsilon}(n, r)) &\leq 2ne^{-C_\varepsilon^* / 2nr^d}, \\ \mathbb{P}(D_{L-(1/2)\varepsilon}^\downarrow(2r) \not\subset \hat{D}_{L-\varepsilon}(n, r)) &\leq 2ne^{-C_\varepsilon^* / 2nr^d}, \\ \mathbb{P}(\hat{D}_{L+\varepsilon}(n, \sqrt{2}r) \not\subset D_{L+(1/2)\varepsilon}^\uparrow(2\sqrt{2}r)) &\leq ne^{-C_\varepsilon^* 2^{d/2} / 2nr^d}, \\ \mathbb{P}(\hat{D}_{L-\varepsilon}(n, \sqrt{2}r) \not\subset D_{L-(3/2)\varepsilon}^\uparrow(2\sqrt{2}r)) &\leq ne^{-C_\varepsilon^* 2^{d/2} / 2nr^d}. \end{aligned}$$

Therefore, for  $n$  large enough we have

$$\begin{aligned} \mathbb{P}(D_{L+(3/2)\varepsilon}^\downarrow(2r) \subset \hat{D}_{L+\varepsilon}(n, r) \subset \hat{D}_{L+\varepsilon}(n, \sqrt{2}r) \subset D_{L+(1/2)\varepsilon}^\uparrow(2\sqrt{2}r)) &\geq 1 - 3ne^{C_\varepsilon^* / 2nr^d}, \\ \mathbb{P}(D_{L-(1/2)\varepsilon}^\downarrow(2r) \subset \hat{D}_{L-\varepsilon}(n, r) \subset \hat{D}_{L-\varepsilon}(n, \sqrt{2}r) \subset D_{L-(3/2)\varepsilon}^\uparrow(2\sqrt{2}r)) &\geq 1 - 3ne^{C_\varepsilon^* / 2nr^d}. \end{aligned}$$

Since we assume that all the levels we study are  $\varepsilon$ -regular, if  $r$  is small enough we can order them in the following way

$$\begin{aligned} D_{L+2\varepsilon} &\subset D_{L+(3/2)\varepsilon}^\downarrow(2r) \subset D_{L+(1/2)\varepsilon}^\uparrow(2\sqrt{2}r) \subset D_L \subset D_{L-(1/2)\varepsilon}^\downarrow(2r) \\ &\subset D_{L-(3/2)\varepsilon}^\uparrow(2\sqrt{2}r) \subset D_{L-2\varepsilon}. \end{aligned}$$

Combining that with (A.7), we conclude that with a high probability we have the following sequence in homology (induced by composing inclusion maps),

$$\begin{array}{ccccccc} \star & H_*(D_{L+2\varepsilon}) & \rightarrow & H_*(D_{L+(3/2)\varepsilon}^\downarrow(2r)) & \rightarrow & H_*(\hat{D}_{L+\varepsilon}(n, r)) & \\ & & & & & \downarrow & \\ & & & & & H_*(R_{L+\varepsilon}(n, r)) & \star \\ & & & \leftarrow & H_*(D_{L+(1/2)\varepsilon}^\uparrow(2\sqrt{2}r)) & \leftarrow & H_*(\hat{D}_{L+\varepsilon}(n, \sqrt{2}r)) \\ \star & H_*(D_L) & & & & \downarrow & \\ & & \rightarrow & H_*(D_{L-(1/2)\varepsilon}^\downarrow(2r)) & \rightarrow & H_*(\hat{D}_{L-\varepsilon}(n, r)) & \\ & & & & & \downarrow & \\ & & & & & H_*(R_{L-\varepsilon}(n, r)) & \star \\ & & & & & \downarrow & \\ \star & H_*(D_{L-2\varepsilon}) & \leftarrow & H_*(D_{L-(3/2)\varepsilon}^\uparrow(2\sqrt{2}r)) & \leftarrow & H_*(\hat{D}_{L-\varepsilon}(n, \sqrt{2}r)) & \end{array}$$

Taking out the spaces marked in  $\star$  we have

$$H_*(D_{L+2\varepsilon}) \rightarrow H_*(R_{L+\varepsilon}(n, r)) \rightarrow H_*(D_L) \rightarrow H_*(R_{L-\varepsilon}(n, r)) \rightarrow H_*(D_{L-2\varepsilon}).$$

Since  $f(x)$  has no critical values in  $[L - 2\varepsilon, L + 2\varepsilon]$ , using Lemma A.3 completes the proof.  $\square$

### A.3. Proving the lemmas

One of the main probability tools we use is Bernstein's inequality [33], basically a law of large numbers bound. If  $Z_1, \dots, Z_n$  are i.i.d., with  $\mathbb{E}\{Z_i\} = 0$ ,  $\text{Var}(Z_i) = \sigma^2$  such that  $|Z_i| \leq M$  almost surely, then

$$\mathbb{P}\left(\sum_{i=1}^n Z_i \geq t\right) \leq \exp\left(-\frac{t^2/2}{n\sigma^2 + Mt/3}\right). \tag{A.8}$$

**Proof of Lemma A.1 (Density estimation).** To reconstruct the level sets of the density, we will use a kernel density estimator. Recall that the kernel function  $K : \mathbb{R}^d \rightarrow \mathbb{R}$  we use satisfies the following:

- $\text{supp}(K) \subset B_1(0)$ ,
- $K(x) \in [0, 1]$ , and  $K(0) = 1$ ,
- $\int K(\xi) d\xi = C_K$ , for some  $C_K \in (0, 1)$ .



In this case, our kernel estimator is

$$\hat{f}_n(x) = \frac{\sum_{i=1}^n K_r(x - X_i)}{C_K n r^d},$$

where  $K_r(x) = K(x/r)$ . We start by proving (A.1). Using a simple union bound we have

$$\begin{aligned} \mathbb{P}(\exists X_i \notin D_{L-\varepsilon}^\uparrow(r) : \hat{f}_n(X_i) \geq L) &\leq n \mathbb{P}(X_1 \in (D_{L-\varepsilon}^\uparrow(r))^c : \hat{f}_n(X_1) \geq L) \\ &= n \int_{(D_{L-\varepsilon}^\uparrow(r))^c} f_X(x) \mathbb{P}(\hat{f}_n(X_1) \geq L \mid X_1 = x) dx. \end{aligned} \quad (\text{A.9})$$

Next,

$$\begin{aligned} \mathbb{P}(\hat{f}_n(X_1) \geq L \mid X_1 = x) &= \mathbb{P}\left(K_r(0) + \sum_{i=2}^n K_r(x - X_i) \geq LC_K n r^d\right) \\ &= \mathbb{P}\left(\sum_{i=2}^n Z_i \geq n(LC_K r^d - p_r(x)) + p_r(x) - 1\right), \end{aligned} \quad (\text{A.10})$$

where

$$p_r(x) \triangleq \mathbb{E}\{K_r(x - X_i)\},$$

and  $Z_i = K_r(x - X_i) - p_r(x)$  are independent variables with  $\mathbb{E}\{Z_i\} = 0$ . Note that  $p_r(x) \in [0, 1]$  since  $K_r(x) \in [0, 1]$ . Also, since  $x \in (D_{L-\varepsilon}^\uparrow(r))^c$ , we have that

$$p_r(x) = \int_{B_r(x)} f(\xi) K_r(x - \xi) d\xi \leq (L - \varepsilon) C_K r^d, \quad (\text{A.11})$$

and therefore from (A.10) we have,

$$\mathbb{P}(\hat{f}_n(X_1) \geq L \mid X_1 = x) \leq \mathbb{P}\left(\sum_{i=2}^n Z_i \geq \varepsilon C_K n r^d - 1\right). \quad (\text{A.12})$$

We would like to apply the inequality in (A.8) for  $t = \varepsilon C_K n r^d - 1$ . Note that  $|Z_i| \leq 1$ , and also that

$$\text{Var}(Z_i) \leq \mathbb{E}\{K_r^2(x - X_i)\} \leq p_{\max} C_K r^d.$$

Therefore, we have

$$\begin{aligned} \mathbb{P}(\hat{f}_n(X_1) \geq L \mid X_1 = x) &\leq \exp\left(-\frac{t^2/2}{(n-1)p_{\max} C_K r^d + t/3}\right) \\ &= \exp\left(-\frac{t/2}{t^{-1}(n-1)p_{\max} C_K r^d + 1/3}\right). \end{aligned}$$

Since  $nr^d \rightarrow \infty$ , we have

$$\frac{(1/2)t(nr^d)^{-1}}{t^{-1}(n-1)p_{\max}C_K r^d + 1/3} \rightarrow \frac{3\varepsilon^2 C_K}{6f_{\max} + 2\varepsilon} > \frac{\varepsilon^2 C_K}{3p_{\max} + \varepsilon}.$$

Thus, for  $n$  large enough we have

$$\mathbb{P}(\hat{f}_n(X_1) \geq L \mid X_1 = x) \leq e^{-C_\varepsilon^* nr^d},$$

where

$$C_\varepsilon^* = \frac{\varepsilon^2 C_K}{3p_{\max} + \varepsilon}. \tag{A.13}$$

Which completes the proof of (A.1)

To prove (A.2) we start the same way, and similarly to (A.12) we have,

$$\mathbb{P}(\hat{f}_n(X_1) \leq L \mid X_1 = x) \leq \mathbb{P}\left(\sum_{i=2}^n Z_i \leq -\varepsilon C_K nr^d\right),$$

where we used the fact that  $x \in D_{L+\varepsilon, r}^\downarrow$ , and therefore we have  $(L + \varepsilon)C_K r^d \leq p_r(x) \leq 1$ . Thus, to complete the proof we should use (A.8) for the variables  $(-Z_i)$  and  $t = \varepsilon C_K nr^d$ . Similarly to the proof above, we then have that

$$\mathbb{P}(\hat{f}_n(X_1) \leq L \mid X_1 = x) \leq e^{-C_\varepsilon^* nr^d},$$

which completes the proof. □

**Proof of Lemma A.1 (Kernel regression).** Recall that in the kernel regression model, we have a set of pairs  $(X_1, Y_1), \dots, (X_n, Y_n)$ , where the pairs are i.i.d.,  $X_i \in \mathbb{R}^d, Y_i \in \mathbb{R}$ , and they have a common density function  $f_{X,Y} : \mathbb{R}^d \times \mathbb{R} \rightarrow \mathbb{R}$ . Our estimation target is the conditional expectation

$$f(x) = \mathbb{E}\{Y \mid X = x\}.$$

The estimator we use is given by

$$\hat{f}_n(x) = \frac{\sum_{i=1}^n Y_i K_r(x - X_i)}{\sum_{i=1}^n K_r(x - X_i)},$$

where the assumptions on  $K_r$  are the same as above. In addition we have the following assumptions:

- $f_X$  has a compact support –  $\text{supp}(f)$ .
- $p_{\min} \triangleq \inf_{x \in \text{supp}(f)} f_X(x) > 0$ ,
- $|Y_i| \leq Y_{\max}$  almost surely, for some non-random value  $Y_{\max} > 0$ .

We start by proving (A.1). We use the union bound again to have

$$\begin{aligned} & \mathbb{P}(\exists X_i \notin D_{L-\varepsilon}^\uparrow(r) : \hat{f}_n(X_i) \geq L) \\ & \leq n \int_{(D_{L-\varepsilon}^\uparrow(r))^c} \int_{\mathbb{R}} f_{X,Y}(x, y) \mathbb{P}(\hat{f}_n(X_1) \geq L \mid X_1 = x, Y_1 = y) dy dx. \end{aligned} \tag{A.14}$$

Note that writing  $\hat{f}_n(x) \geq L$  is equivalent to

$$\sum_{i=1}^n Y_i K_r(x - X_i) \geq \sum_{i=1}^n L K_r(x - X_i).$$

Using the fact that  $x \in (D_{L-\varepsilon}^\uparrow(r))^c$ , similar derivations to the ones used for density functions can be applied to show that

$$\begin{aligned} \mathbb{P}(\hat{f}_n(X_1) \geq L \mid X_1 = x, Y_1 = y) & \leq \mathbb{P}\left(\sum_{i=2}^n Z_i \geq \varepsilon(n-1)p_r(x) + L - y\right) \\ & \leq \mathbb{P}\left(\sum_{i=2}^n Z_i \geq \varepsilon f_{\min} C_K(n-1)r^d + L - y\right), \end{aligned}$$

where here

$$Z_i \triangleq (Y_i - f(X_i))K_r(x - X_i) - \varepsilon(K_r(x - X_i) - p_r(x)),$$

and  $p_r(x) = \mathbb{E}\{K_r(x - X_i)\}$ , and we used the fact that  $p_r(x) \geq p_{\min} C_K r^d$ . We would like to use Bernstein's inequality to bound this probability. First, denote

$$\begin{aligned} Z_i^{(1)} &= (Y_i - f(X_i))K_r(x - X_i), \\ Z_i^{(2)} &= \varepsilon(K_r(x - X_i) - p_r(x)). \end{aligned}$$

Then it is easy to show that  $\mathbb{E}\{Z_i^{(1)}\} = \mathbb{E}\{Z_i^{(2)}\} = \mathbb{E}\{Z_i^{(1)}Z_i^{(2)}\} = 0$ , which implies that  $Z_i^{(1)}$  and  $Z_i^{(2)}$  are uncorrelated, and therefore

$$\sigma^2 = \text{Var}(Z_i) = \text{Var}(Z_i^{(1)}) + \text{Var}(Z_i^{(2)}).$$

Also, it is easy to show that

$$\text{Var}(Z_i^{(1)}) = \mathbb{E}\{\text{Var}(Y_i \mid X_i)K_r^2(x - X_i)\}.$$

Therefore, we have:

- $\text{Var}(Z_i^{(1)}) \leq Y_{\max}^2 \mathbb{E}\{K_r^2(x - X_i)\} \leq Y_{\max}^2 C_K p_{\max} r^d,$
- $\text{Var}(Z_i^{(2)}) \leq \varepsilon^2 \mathbb{E}\{K_r^2(x - X_i)\} \leq \varepsilon^2 C_K p_{\max} r^d,$

- and almost surely:

$$|Z_i| \leq |Y_i| + |f(X_i)| + \varepsilon(1 + p_r(x)) \leq 2Y_{\max} + \varepsilon(1 + C_K p_{\max} r^d) < 2(Y_{\max} + \varepsilon).$$

Using Bernstein's inequality (A.8), for  $t = \varepsilon f_{\min} C_K (n-1)r^d + L - y$ , we then have

$$\begin{aligned} & \mathbb{P}(\hat{f}_n(X_1) \geq L \mid X_1 = x, Y_1 = y) \\ & \leq \exp\left(-\frac{t/2}{t^{-1}(Y_{\max}^2 + \varepsilon^2)C_K p_{\max}(n-1)r^d + (2/3)(Y_{\max} + \varepsilon)}\right). \end{aligned}$$

Since  $nr^d \rightarrow \infty$ , we have that

$$\begin{aligned} & \frac{(1/2)t(nr^d)^{-1}}{t^{-1}(Y_{\max}^2 + \varepsilon^2)C_K p_{\max}(n-1)r^d + (2/3)(Y_{\max} + \varepsilon)} \\ & \rightarrow \frac{3\varepsilon^2 p_{\min}^2 C_K}{6(Y_{\max}^2 + \varepsilon^2)p_{\max} + 4\varepsilon p_{\min}(Y_{\max} + \varepsilon)} \\ & > \frac{\varepsilon^2 p_{\min}^2 C_K}{3(Y_{\max}^2 + \varepsilon^2)p_{\max} + 2\varepsilon p_{\min}(Y_{\max} + \varepsilon)}. \end{aligned}$$

Thus, for  $n$  large enough we have

$$\mathbb{P}(\hat{f}_n(X_1) \geq L \mid X_1 = x, Y_1 = y) \leq e^{-C_\varepsilon^* nr^d},$$

where

$$C_\varepsilon^* = \frac{\varepsilon^2 p_{\min}^2 C_K}{3(Y_{\max}^2 + \varepsilon^2)p_{\max} + 2\varepsilon p_{\min}(Y_{\max} + \varepsilon)}. \quad (\text{A.15})$$

Putting this back into (A.14) completes the proof of (A.1). The proof of (A.2) is similar, with some adjustments, and we omit it here.  $\square$

To prove Lemma A.2 we need the following lemma.

**Lemma A.4.** *If  $nr^d \rightarrow \infty$ , then*

$$\mathbb{P}(D_{L+\varepsilon}^\downarrow(2r) \not\subset \hat{D}_L(n, r)) \leq 2ne^{-C_\varepsilon^* nr^d},$$

where  $C_\varepsilon^*$  is the same as in Lemma A.1.

**Proof.** Note that in both cases (density estimation and kernel regression) we have that the set  $D_{L+\varepsilon}^\downarrow(2r)$  is bounded. Let  $\delta \in (0, 1)$ , and let  $\mathcal{S} \subset D_{L+\varepsilon}^\downarrow(2r)$  be a finite set of points satisfying that for every  $x \in D_{L+\varepsilon}^\downarrow(2r)$  there exists  $s \in \mathcal{S}$  such that  $\|x - s\| \leq \delta r$ . Then there exists a constant  $c > 0$  such that we can construct  $\mathcal{S}$  with  $|\mathcal{S}| \leq c(\delta r)^{-d}$  points. Note that if there is  $x \in D_{L+\varepsilon}^\downarrow(2r)$

that is not covered by the balls of radius  $r$ , it necessarily means that there is  $s \in \mathcal{S}$  that is not covered by the balls of radius  $(1 - \delta)r$ . Therefore,

$$\begin{aligned} \mathbb{P}(D_{L+\varepsilon}^\downarrow(2r) \not\subset \hat{D}_L(n, r)) &\leq \mathbb{P}(\exists s \in \mathcal{S} : B_{(1-\delta)r}(s) \cap \mathcal{X}_n^L = \emptyset) \\ &= \mathbb{P}(\exists s \in \mathcal{S} : B_{(1-\delta)r}(s) \cap \mathcal{X}_n^L = \emptyset; D_{L+\varepsilon}^\downarrow(r) \cap \mathcal{X}_n \subset \mathcal{X}_n^L) \\ &\quad + \mathbb{P}(\exists s \in \mathcal{S} : B_{(1-\delta)r}(s) \cap \mathcal{X}_n^L = \emptyset; D_{L+\varepsilon}^\downarrow(r) \cap \mathcal{X}_n \not\subset \mathcal{X}_n^L) \\ &\leq \mathbb{P}(\exists s \in \mathcal{S} : B_{(1-\delta)r}(s) \cap \mathcal{X}_n = \emptyset) + \mathbb{P}(D_{L+\varepsilon}^\downarrow(r) \cap \mathcal{X}_n \not\subset \mathcal{X}_n^L), \end{aligned}$$

where the last inequality is due to the fact that for every two events  $A, B$  we have  $\mathbb{P}(A \cap B) \leq \mathbb{P}(A)$ . In other words the event of not covering  $D_{L+\varepsilon}^\downarrow(2r)$  might occur for two different reasons. Either the original sample (before filtering)  $\mathcal{X}_n$  does not cover  $D_{L+\varepsilon}^\downarrow(2r)$  (the first term), or our filtering method got rid of too many points (second term). The second term can be bounded using Lemma A.1. For the first term we have

$$\begin{aligned} \mathbb{P}(\exists s \in \mathcal{S} : B_{(1-\delta)r}(s) \cap \mathcal{X}_n = \emptyset) &\leq \sum_{s \in \mathcal{S}} \mathbb{P}(B_{(1-\delta)r}(s) \cap \mathcal{X}_n = \emptyset) \\ &= \sum_{s \in \mathcal{S}} (1 - F(B_{(1-\delta)r}(s)))^n \\ &\leq \sum_{s \in \mathcal{S}} e^{-nF(B_{(1-\delta)r}(s))}, \end{aligned}$$

where  $F(A) = \int_A f_X(x) dx$ . For the density estimation,  $s \in D_{L+\varepsilon}^\downarrow(2r)$  implies that

$$F(B_{(1-\delta)r}(s)) \geq (L + \varepsilon)(1 - \delta)^d \omega_d r^d \geq L(1 - \delta)^d \omega_d r^d.$$

For the kernel regression model, we have that

$$F(B_{(1-\delta)r}(s)) \geq p_{\min}(1 - \delta)^d \omega_d r^d.$$

Thus, if we choose  $C_1 = c\delta^{-d}$ , and

$$C_2 = \begin{cases} L(1 - \delta)^d \omega_d, & \text{density estimation,} \\ p_{\min}(1 - \delta)^d \omega_d, & \text{kernel regression,} \end{cases}$$

we have that

$$\mathbb{P}(\exists s \in \mathcal{S} : B_{(1-\delta)r}(s) \cap \mathcal{X}_n = \emptyset) \leq C_1 r^{-d} e^{-C_2 n r^d}.$$

From Lemma A.1 we know that

$$\mathbb{P}(D_{L+\varepsilon}^\downarrow(r) \cap \mathcal{X}_n \not\subset \mathcal{X}_n^L) \leq n e^{-C_\varepsilon^* n r^d}.$$

Note that for both models we have that  $C_\varepsilon^* < C_2$  (see (A.13), (A.15)), and also that  $r^{-d} = o(n)$ . Therefore the latter probability is necessarily the dominant one in the bound we have. This completes the proof.  $\square$

**Proof of Lemma A.2.** If  $nr^d \rightarrow \infty$ , then by Lemma A.4 we have

$$\mathbb{P}(D_{L+\varepsilon}^\downarrow(2r) \not\subset \hat{D}_L(n, r)) \leq 2ne^{-C_\varepsilon^*nr^d}. \tag{A.16}$$

In addition, from Lemma A.1 we have

$$\mathbb{P}(\hat{D}_L(n, r) \not\subset D_{L-\varepsilon}^\uparrow(2r)) \leq \mathbb{P}(\mathcal{X}_n^L \cap (D_{L-\varepsilon}^\uparrow(r))^c \neq \emptyset) \leq ne^{-C_\varepsilon^*nr^d}. \tag{A.17}$$

Using the union bound completes the proof.  $\square$

The last piece of the puzzle is the proof of the algebraic Lemma A.3.

**Proof of Lemma A.3.** We need to show that  $g_{34}$  is injective and that  $\text{Im}(g_{34}) = G_{24}$ .

1. The assumption that  $g_{35}$  is an isomorphism from  $G_3$  to  $G_{15}$  implies that  $g_{35}$  is injective. Since  $g_{35} = g_{45} \circ g_{34}$  we have that  $g_{34}$  is injective as well.
2. Since (a)  $g_{15} : G_1 \rightarrow G_{15}$  is surjective, (b)  $g_{35} : G_3 \rightarrow G_{15}$  is an isomorphism, and (c)  $g_{15} = g_{35} \circ g_{13}$ , we conclude that  $g_{13} : G_1 \rightarrow G_3$  is surjective. Since  $g_{13} = g_{23} \circ g_{12}$ , we have that  $g_{23}$  is surjective as well.  
 Finally, since (a)  $\text{Im}(g_{24}) = G_{24}$ , (b)  $g_{24} = g_{34} \circ g_{23}$ , and (c)  $g_{23} : G_2 \rightarrow G_3$  is surjective, we have that  $\text{Im}(g_{34}) = G_{24}$  as well.

$\square$

## Acknowledgements

The authors would like to thank: Robert Adler, Paul Bendich, Ulrich Bauer, Ezra Miller and Andrew Nobel for many useful discussions. We also wish to thank Frédéric Chazal, Larry Wasserman and the anonymous referees for very useful comments on previous revisions of this paper.

Omer Bobrowski gratefully acknowledges the support of AFOSR: FA9550-10-1-0436, and NSF DMS-1127914. Sayan Mukherjee is pleased to acknowledge support from grants AFOSR: FA9550-10-1-0436, and NSF: CCF-1049290. Jonathan E. Taylor was supported by the AFOSR, Grant 113039.

## References

- [1] Adams, H., Tausz, A. and Vejdemo-Johansson, M. (2014). javaPlex: A research software package for persistent (co) homology. In *Mathematical Software–ICMS 2014* 129–136. Springer, Berlin.
- [2] Adler, R.J., Bobrowski, O., Borman, M.S., Subag, E. and Weinberger, S. (2010). Persistent homology for random fields and complexes. In *Borrowing Strength: Theory Powering Applications – a Festschrift for Lawrence D. Brown. Inst. Math. Stat. Collect.* **6** 124–143. Beachwood, OH: IMS. MR2798515

- [3] Adler, R.J., Bobrowski, O. and Weinberger, S. (2014). Crackle: The homology of noise. *Discrete Comput. Geom.* **52** 680–704. [MR3279544](#)
- [4] Bafllo, A. (2003). Total error in a plug-in estimator of level sets. *Statist. Probab. Lett.* **65** 411–417. [MR2039885](#)
- [5] Bafllo, A., Cuesta-Albertos, J.A. and Cuevas, A. (2001). Convergence rates in nonparametric estimation of level sets. *Statist. Probab. Lett.* **53** 27–35. [MR1843338](#)
- [6] Bafllo, A., Cuevas, A. and Justel, A. (2000). Set estimation and nonparametric detection. *Canad. J. Statist.* **28** 765–782. [MR1821433](#)
- [7] Balakrishnan, S., Rinaldo, A., Sheehy, D., Singh, A. and Wasserman, L. (2012). Minimax rates for homology inference. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics* 64–72. La Palma, Canary Islands.
- [8] Balakrishnan, S., Rinaldo, A., Singh, A. and Wasserman, L. (2013). Tight lower bounds for homology inference. Available at [arXiv:1307.7666](#).
- [9] Bendich, P., Wang, B. and Mukherjee, S. (2012). Local homology transfer and stratification learning. In *Proceedings of the Twenty-Third Annual ACM-SIAM Symposium on Discrete Algorithms* 1355–1370. New York: ACM. [MR3205297](#)
- [10] Blumberg, A.J., Gal, I., Mandell, M.A. and Pancia, M. (2014). Robust statistics, hypothesis testing, and confidence intervals for persistent homology on metric measure spaces. *Found. Comput. Math.* **14** 745–789. [MR3230014](#)
- [11] Bobrowski, O. and Adler, R.J. (2014). Distance functions, critical points, and the topology of random Čech complexes. *Homology, Homotopy Appl.* **16** 311–344. [MR3280987](#)
- [12] Bobrowski, O. and Mukherjee, S. (2015). The topology of probability distributions on manifolds. *Probab. Theory Related Fields* **161** 651–686. [MR3334278](#)
- [13] Borsuk, K. (1948). On the imbedding of systems of compacta in simplicial complexes. *Fund. Math.* **35** 217–234. [MR0028019](#)
- [14] Bubenik, P. (2015). Statistical topological data analysis using persistence landscapes. *J. Mach. Learn. Res.* **16** 77–102. [MR3317230](#)
- [15] 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. Providence, RI: Amer. Math. Soc. [MR2730741](#)
- [16] Bubenik, P. and Scott, J.A. (2014). Categorification of persistent homology. *Discrete Comput. Geom.* **51** 600–627. [MR3201246](#)
- [17] Carlsson, G. (2009). Topology and data. *Bull. Amer. Math. Soc. (N.S.)* **46** 255–308. [MR2476414](#)
- [18] Chazal, F., Cohen-Steiner, D., Glisse, M., Guibas, L.J. and Oudot, S.Y. (2009). Proximity of persistence modules and their diagrams. In *Proceedings of the Twenty-Fifth Annual Symposium on Computational Geometry* 237–246. ACM, New York.
- [19] Chazal, F., Cohen-Steiner, D. and Lieutier, A. (2009). A sampling theorem for compact sets in Euclidean space. *Discrete Comput. Geom.* **41** 461–479. [MR2486371](#)
- [20] Chazal, F., Guibas, L.J., Oudot, S.Y. and Skraba, P. (2011). Scalar field analysis over point cloud data. *Discrete Comput. Geom.* **46** 743–775. [MR2846177](#)
- [21] Chazal, F., Guibas, L.J., Oudot, S.Y. and Skraba, P. (2013). Persistence-based clustering in Riemannian manifolds. *J. ACM* **60** Art. 41, 38. [MR3144911](#)
- [22] Chung, M.K., Bubenik, P. and Kim, P.T. (2009). Persistence diagrams of cortical surface data. In *Information Processing in Medical Imaging* 386–397. Springer, Berlin.
- [23] Cohen-Steiner, D., Edelsbrunner, H. and Harer, J. (2007). Stability of persistence diagrams. *Discrete Comput. Geom.* **37** 103–120. [MR2279866](#)
- [24] Cuevas, A. (1990). On pattern analysis in the nonconvex case. *Kybernetes* **19** 26–33. [MR1084947](#)

- [25] Cuevas, A., Febrero, M. and Fraiman, R. (2000). Estimating the number of clusters. *Canad. J. Statist.* **28** 367–382. [MR1792055](#)
- [26] 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](#)
- [27] Cuevas, A. and Fraiman, R. (1997). A plug-in approach to support estimation. *Ann. Statist.* **25** 2300–2312. [MR1604449](#)
- [28] Cuevas, A. and Rodríguez-Casal, A. (2004). On boundary estimation. *Adv. in Appl. Probab.* **36** 340–354. [MR2058139](#)
- [29] de Silva, V. and Ghrist, R. (2007). Coverage in sensor networks via persistent homology. *Algebr. Geom. Topol.* **7** 339–358. [MR2308949](#)
- [30] Deprins, D. and Simar, L. (1983). On Farrell measures of technical efficiency. *Recherches Économiques de Louvain/Louvain Economic Review* 123–137. Available at <http://www.jstor.org/stable/40723714>.
- [31] Deprins, D., Simar, L. and Tulkens, H. (2006). Measuring labor-efficiency in post offices. In *Public Goods, Environmental Externalities and Fiscal Competition* 285–309. Springer, New York.
- [32] 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](#)
- [33] Dudley, R.M. (2002). *Real Analysis and Probability*, 2nd ed. Cambridge: Cambridge Univ. Press.
- [34] Edelsbrunner, H. and Harer, J. (2008). Persistent homology – a survey. In *Surveys on Discrete and Computational Geometry. Contemp. Math.* **453** 257–282. Providence, RI: Amer. Math. Soc. [MR2405684](#)
- [35] Edelsbrunner, H. and Harer, J.L. (2010). *Computational Topology: An Introduction*. Providence, RI: Amer. Math. Soc. [MR2572029](#)
- [36] Farrell, M.J. (1957). The measurement of productive efficiency. *J. R. Stat. Soc., A* **120** 253–290.
- [37] Fasy, B.T., Lecci, F., Rinaldo, A., Wasserman, L., Balakrishnan, S. and Singh, A. (2014). Confidence sets for persistence diagrams. *Ann. Statist.* **42** 2301–2339. [MR3269981](#)
- [38] Friedman, J. (1998). Computing Betti numbers via combinatorial Laplacians. *Algorithmica* **21** 331–346. [MR1622290](#)
- [39] Ghrist, R. (2008). Barcodes: The persistent topology of data. *Bull. Amer. Math. Soc. (N.S.)* **45** 61–75. [MR2358377](#)
- [40] Grenander, U. (1981). *Abstract Inference*. New York: Wiley. [MR0599175](#)
- [41] Hartigan, J.A. (1975). *Clustering Algorithms*. New York: Wiley. [MR0405726](#)
- [42] Hartigan, J.A. (1987). Estimation of a convex density contour in two dimensions. *J. Amer. Statist. Assoc.* **82** 267–270. [MR0883354](#)
- [43] Hatcher, A. (2002). *Algebraic Topology*. Cambridge: Cambridge Univ. Press. [MR1867354](#)
- [44] Kahle, M. (2011). Random geometric complexes. *Discrete Comput. Geom.* **45** 553–573. [MR2770552](#)
- [45] Kahle, M. and Meckes, E. (2013). Limit theorems for Betti numbers of random simplicial complexes. *Homology, Homotopy Appl.* **15** 343–374. [MR3079211](#)
- [46] Koltchinskii, V.I. (2000). Empirical geometry of multivariate data: A deconvolution approach. *Ann. Statist.* **28** 591–629. [MR1790011](#)
- [47] Korostelëv, A.P. and Tsybakov, A.B. (1993). *Minimax Theory of Image Reconstruction. Lecture Notes in Statistics* **82**. New York: Springer. [MR1226450](#)
- [48] MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proc. Fifth Berkeley Sympos. Math. Statist. and Probability (Berkeley, Calif., 1965/66)* (L.M. Le Cam and J. Neyman, eds.) 281–297. Berkeley, CA: Univ. California Press. [MR0214227](#)
- [49] Mileyko, Y., Mukherjee, S. and Harer, J. (2011). Probability measures on the space of persistence diagrams. *Inverse Probl.* **27** 124007, 22. [MR2854323](#)



- [50] Mischaikow, K. and Nanda, V. (2013). Morse theory for filtrations and efficient computation of persistent homology. *Discrete Comput. Geom.* **50** 330–353. [MR3090522](#)
- [51] Molchanov, I.S. (1991). Empirical estimation of quantiles of distributions of random closed sets. *Theory Probab. Appl.* **35** 594–600.
- [52] Molchanov, I.S. (1998). A limit theorem for solutions of inequalities. *Scand. J. Stat.* **25** 235–242. [MR1614288](#)
- [53] Müller, D.W. (1992). The excess mass approach in statistics. *Beiträge zur Statistik, Univ. Heidelberg*, 3.
- [54] Müller, D.W. and Sawitzki, G. (1991). Excess mass estimates and tests for multimodality. *J. Amer. Statist. Assoc.* **86** 738–746. [MR1147099](#)
- [55] Munkres, J.R. (1984). *Elements of Algebraic Topology*. Menlo Park, CA: Addison-Wesley Company. [MR0755006](#)
- [56] Nadaraya, E.A. (1964). On estimating regression. *Theory Probab. Appl.* **9** 141–142.
- [57] Ng, A.Y., Jordan, M.I. and Weiss, Y. (2002). On spectral clustering: Analysis and an algorithm. *Adv. Neural Inf. Process. Syst.* **2** 849–856.
- [58] 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](#)
- [59] Niyogi, P., Smale, S. and Weinberger, S. (2011). A topological view of unsupervised learning from noisy data. *SIAM J. Comput.* **40** 646–663. [MR2810909](#)
- [60] Nolan, D. (1991). The excess-mass ellipsoid. *J. Multivariate Anal.* **39** 348–371. [MR1147127](#)
- [61] Perea, J.A. and Harer, J. (2013). Sliding windows and persistence: An application of topological methods to signal analysis. *Found. Comput. Math.* **15** 1–40.
- [62] Phillips, J.M., Wang, B. and Zheng, Y. (2013). Geometric inference on kernel density estimates. Available at [arXiv:1307.7760](#).
- [63] Polonik, W. (1995). Measuring mass concentrations and estimating density contour clusters – an excess mass approach. *Ann. Statist.* **23** 855–881. [MR1345204](#)
- [64] Robins, V., Meiss, J.D. and Bradley, E. (1998). Computing connectedness: An exercise in computational topology. *Nonlinearity* **11** 913–922. [MR1632633](#)
- [65] Robins, V., Meiss, J.D. and Bradley, E. (2000). Computing connectedness: Disconnectedness and discreteness. *Phys. D* **139** 276–300. [MR1753085](#)
- [66] Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis. Monographs on Statistics and Applied Probability* **26** London: Chapman & Hall. [MR0848134](#)
- [67] Sneath, P.H.A. and Sokal, R.R. (1973). *Numerical Taxonomy: The Principles and Practice of Numerical Classification*. San Francisco, CA: W.H. Freeman and Co. [MR0456594](#)
- [68] Tsybakov, A.B. (1997). On nonparametric estimation of density level sets. *Ann. Statist.* **25** 948–969. [MR1447735](#)
- [69] Turner, K., Mileyko, Y., Mukherjee, S. and Harer, J. (2014). Fréchet means for distributions of persistence diagrams. *Discrete Comput. Geom.* **52** 44–70. [MR3231030](#)
- [70] Tyron, R.C. (1939). *Cluster Analysis*. Oxford, UK: Edwards Bros.
- [71] von Luxburg, U. (2007). A tutorial on spectral clustering. *Stat. Comput.* **17** 395–416. [MR2409803](#)
- [72] Walther, G. (1997). Granulometric smoothing. *Ann. Statist.* **25** 2273–2299. [MR1604445](#)
- [73] Watson, G.S. (1964). Smooth regression analysis. *Sankhya, Ser. A* **26** 359–372. [MR0185765](#)
- [74] Zomorodian, A. and Carlsson, G. (2005). Computing persistent homology. *Discrete Comput. Geom.* **33** 249–274. [MR2121296](#)

Received December 2014 and revised April 2015