Journal of Machine Learning Research 18 (2018) 1-40 Submitted 9/15; Revised 2/17; Published 4/18

Robust Topological Inference: Distance To a Measure and Kernel Distance

Frédéric Chazal FREDERIC.CHAZAL@INRIA.FR
Inria Saclay - Ile-de-France
Alan Turing Bldg, Office 2043
1 rue Honoré d'Estienne d'Orves
91120 Palaiseau, FRANCE

Brittany Fasy BRITTANY@CS.MONTANA.EDU
Computer Science Department
Montana State University
357 EPS Building
Montana State University
Bozeman, MT 59717

Fabrizio Lecci FABRZIO.LECCI@GMAIL.COM
New York, NY

Bertrand Michel BERTRAND.MICHEL@EC-NANTES.FR
Ecole Centrale de Nantes
Laboratoire de mathématiques Jean Leray
1 Rue de La Noe
44300 Nantes FRANCE

Alessandro Rinaldo ARINALDO@CMU.EDU
Department of Statistics
Carnegie Mellon University
Pittsburgh, PA 15213

Larry Wasserman LARRY@CMU.EDU
Department of Statistics
Carnegie Mellon University
Pittsburgh, PA 15213

Editor: Mikhail Belkin

Abstract

Let P be a distribution with support S. The salient features of S can be quantified with persistent homology, which summarizes topological features of the sublevel sets of the distance function (the distance of any point x to S). Given a sample from P we can infer the persistent homology using an empirical version of the distance function. However, the empirical distance function is highly non-robust to noise and outliers. Even one outlier is deadly. The distance-to-a-measure (DTM), introduced by Chazal et al. (2011), and the kernel distance, introduced by Phillips et al. (2014), are smooth functions that provide useful topological information but are robust to noise and outliers. Chazal et al. (2015) derived concentration bounds for DTM. Building on these results, we derive limiting distributions and confidence sets, and we propose a method for choosing tuning parameters.

Keywords: Topological data analysis, persistent homology, RKHS.

1. Introduction

Figure 1 shows three complex point clouds, based on a model used for simulating cosmology data. Visually, the three samples look very similar. Below the data plots are the persistence diagrams, which are summaries of topological features defined in Section 2. The persistence diagrams make it clearer that the third data set is from a different data generating process than the first two.

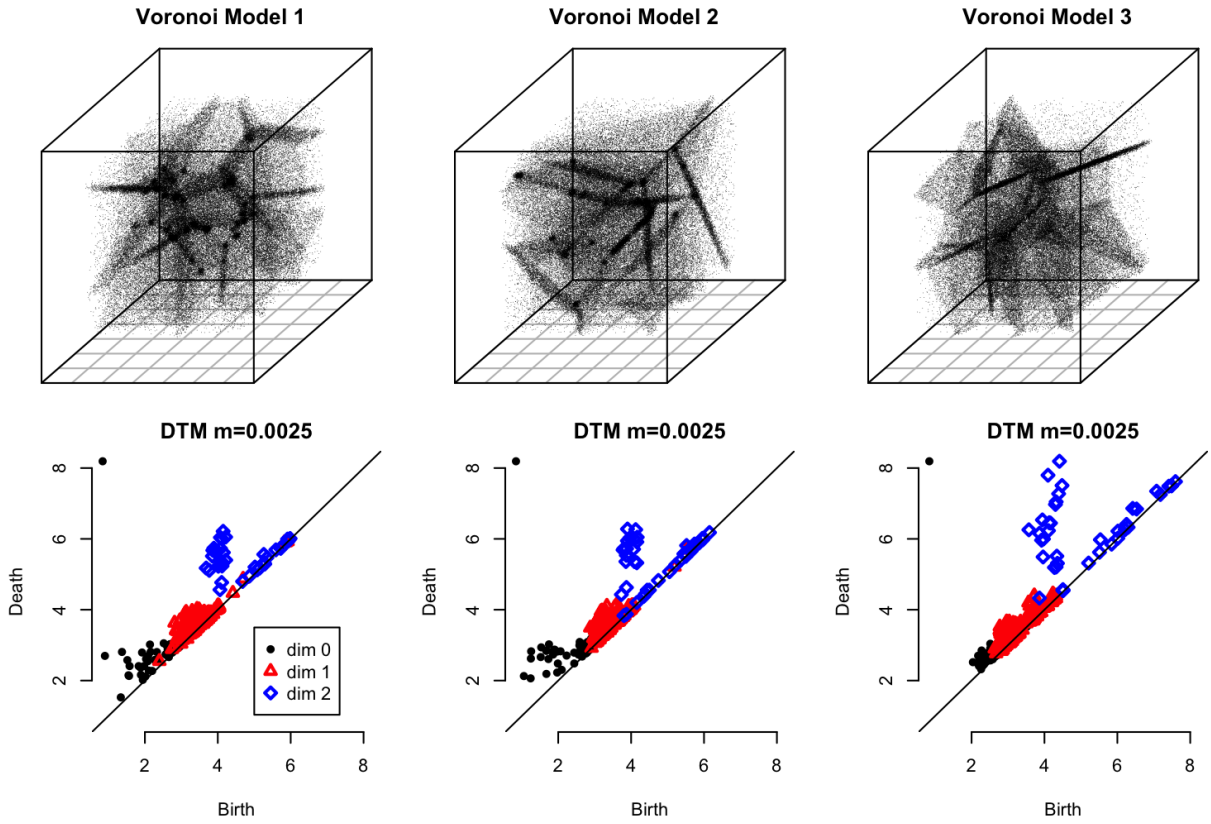

Figure 1: The first two datasets come from the same data generating mechanism. In the third one, the particles are more concentrated around the walls of the Voronoi cells. Although the difference is not clear from the scatterplots, it is evident from the persistence diagrams of the sublevel sets of the distance-to-measure functions. See Example 3 for more details on the Voronoi Models.

This is an example of how topological features can summarize structure in point clouds. The field of topological data analysis (TDA) is concerned with defining such topological features; see Carlsson (2009). When performing TDA, it is important to use topological measures that are robust to noise. This paper explores some of these robust topological measures.

Let P be a distribution with compact support $S \subset \mathbb{R}^d$. One way to describe the shape of S is by using homology. Roughly speaking, the homology of S measures the topological features of S, such as the connected components, the holes, and the voids. A more nuanced way to describe the shape of S is using persistent homology, which is a multiscale version of homology. To describe persistent homology, we begin with the distance function $\Delta_S \colon \mathbb{R}^d \to \mathbb{R}$ for S which is defined by

$$\Delta_S(x) = \inf_{y \in S} \|x - y\|. \tag{1}$$

The sublevel sets $L_t = \{x : \Delta_S(x) \leq t\}$ provide multiscale topological information about S. As t varies from zero to ∞, topological features — connected components, loops, voids — are born and die. Persistent homology quantifies the evolution of these topological features as a function of t. See Figure 2. Each point on the persistence diagram represents the birth and death time of a topological feature.

Given a sample $X_1, \ldots, X_n \sim P$, the empirical distance function is defined by

$$\widehat{\Delta}(x) = \min_{X_i} \|x - X_i\|. \tag{2}$$

If P is supported on S, and has a density bounded away from zero and infinity, then $\widehat{\Delta}$ is a consistent estimator of Δ_S, i.e., $\sup_x |\widehat{\Delta}(x) - \Delta_S(x)| \xrightarrow{P} 0$. However, if there are outliers, or noise, then $\widehat{\Delta}(x)$ is no longer consistent. Figure 3 (bottom) shows that a few outliers completely change the distance function. In the language of robust statistics, the empirical distance function has breakdown point zero.

A more robust approach is to estimate the persistent homology of the super-level sets of the density p of P. As long as P is concentrated near S, we expect the level sets of p to provide useful topological information about S. Specifically, some level sets of p are homotopic to S under weak conditions, and this implies that we can estimate the homology of S. Note that, in this case, we are using the persistent homology of the super-level sets of p, to estimate the homology of S. This is the approach suggested by Bubenik (2015), Fasy et al. (2014b) and Bobrowski et al. (2014). A related idea is to use persistent homology based on a kernel distance (Phillips et al., 2014). In fact, the sublevel sets of the kernel distance are a rescaling of the super-level sets of p, so these two ideas are essentially equivalent. We discuss this approach in Section 5.

A different approach, more closely related to the distance function, but robust to noise, is to use the *distance-to-a-measure (DTM)*, $\delta \equiv \delta_{P,m}$, from Chazal et al. (2011); see Section 2. An estimate $\widehat{\delta}$ of δ is obtained by replacing the true probability measure with the empirical probability measure P_n, or with a deconvolved version of the observed measure Caillerie et al. (2011). One then constructs a persistence diagram based on the sub-level sets of the DTM. See Figure 1. This approach is aimed at estimating the persistent homology of S. (The DTM also suggests new approaches to density estimation; see Biau et al. (2011).)

The density estimation approach and the DTM are both trying to probe the topology of S. But the former is using persistent homology to estimate the homology of S, while the DTM is directly trying to estimate the persistent homology of distance function of S. We discuss this point in detail in Section 9.1.

In this paper, we explore some statistical properties of these methods. In particular:

1. We show that $\sqrt{n}(\widehat{\delta}^2(x) - \delta^2(x))$ converges to a Gaussian process. (Theorem 5).
2. We show that the bootstrap provides asymptotically valid confidence bands for δ. This allows us to identify significant topological features. (Theorem 19).
3. We find the limiting distribution of a key topological quantity called the bottleneck distance. (Section 4.1).
4. We also show that, under additional assumptions, there is another version of the bootstrap—which we call the bottleneck bootstrap—that provides more precise inferences. (Section 6).
5. We show similar results for the kernel distance. (Section 5).
6. We propose a method for choosing the tuning parameter m for DTM and the bandwidth h for the kernel distance. (Section 7.1).
7. We show that the DTM and the kernel density estimator (KDE) both suffer from boundary bias and we suggest a method for reducing the bias. (Section 7.2).

Notation. $B(x, \epsilon)$ is a Euclidean ball of radius ϵ, centered at x. We define $A \oplus \epsilon = \bigcup_{x \in A} B(x, \epsilon)$, the union of ϵ-balls centered at points in A. If x is a vector then $||x||_\infty = \max_j |x_j|$. Similarly, if f is a real-valued function then $||f||_\infty = \sup_x |f(x)|$. We write $X_n \rightsquigarrow X$ to mean that X_n converges in distribution to X, and we use symbols like $c, C, \ldots,$ as generic positive constants.

Remark: The computing for the examples in this paper were done using the R package **TDA**. See Fasy et al. (2014a). The package can be downloaded from `http://cran.r-project.org/web/packages/TDA/index.html`.

Remark: In this paper, we discuss the DTM which uses a smoothing parameter m and the kernel density estimator which uses a smoothing bandwidth h. Unlike in traditional function estimation, we do not send these parameters to zero as n increases. In TDA, the topological features created with a fixed smoothing parameter are of interest. Thus, all the theory in this paper treats the smoothing parameters as being bounded away from 0. See also Section 4.4 in Fasy et al. (2014b). In Section 7.1, we discuss the choice of these smoothing parameters.

2. Background

In this section, we define several distance functions and distance-like functions, and we introduce the relevant concepts from computational topology. For more detail, we refer the reader to Edelsbrunner and Harer (2010).

2.1 Distance Functions and Persistent Homology

Let $S \subset \mathbb{R}^d$ be a compact set. The *homology* of S characterizes certain topological features of S, such as its connected components, holes, and voids. *Persistent homology* is a multiscale version of homology. Recall that the distance function Δ_S for S is

$$\Delta_S(x) = \inf_{y \in S} \|x - y\|. \tag{3}$$

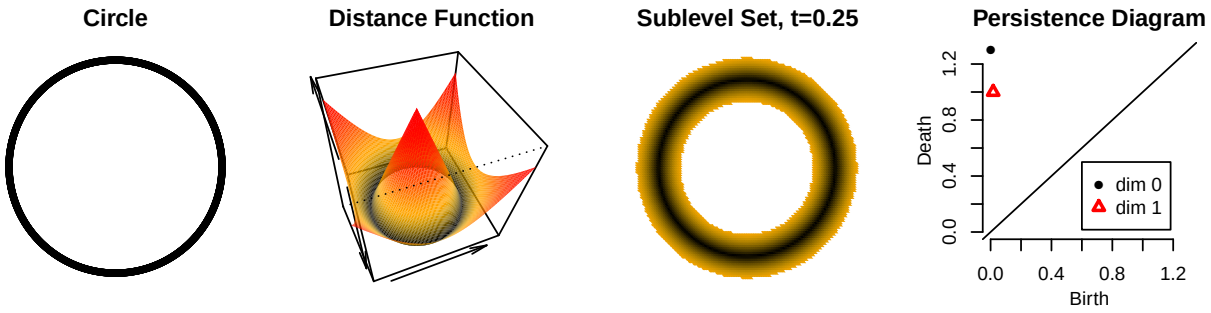

Figure 2: The left plot shows a one-dimensional curve. The second plot is the distance function. The third plot shows a typical sublevel set of the distance function. The fourth plot is the persistence diagram which shows the birth and death times of loops (triangles) and connected components (points) of the sublevel sets.

Let $L_t = \{x : \Delta_S(x) \leq t\}$. We will refer to the parameter t as "time."

Given the nested family of the sublevel sets of Δ_S, the topology of L_t changes as t increases: new connected components can appear, existing connected components can merge, cycles and cavities can appear or be filled, etc. Persistent homology tracks these changes, identifies *features* and associates an *interval* or *lifetime* (from t_{birth} to t_{death}) to them. For instance, a connected component is a feature that is born at the smallest t such that the component is present in L_t, and dies when it merges with an older connected component. Intuitively, the longer a feature persists, the more relevant it is.

A feature, or more precisely its lifetime, can be represented as a segment whose extremities have abscissae t_{birth} and t_{death}; the set of these segments is called the *barcode* of Δ_S. An interval can also be represented as a point in the plane with coordinates $(u, v) = (t_{\text{birth}}, t_{\text{death}})$. The set of points (with multiplicity) representing the intervals is called the *persistence diagram* of Δ_S. Note that the diagram is entirely contained in the half-plane above the diagonal defined by $u = v$, since death always occurs after birth. This diagram is well-defined for any compact set S (Chazal et al. (2012), Theorem 2.22). The most persistent features (supposedly the most important) are those represented by the points furthest from the diagonal in the diagram, whereas points close to the diagonal can be interpreted as (topological) noise.

Figure 2 shows a simple example. Here, the points on the circle are regarded as a subset of $\mathbb{R}^2$. At time zero, there is one connected component and one loop. As t increases, the loop dies.

Let S_1 and S_2 be compact sets with distance functions Δ_1 and Δ_2 and diagrams D_1 and D_2. The bottleneck distance between D_1 and D_2 is defined by

$$W_\infty(D_1, D_2) = \min_{g:\ D_1 \to D_2} \sup_{z \in D_1} \|z - g(z)\|_\infty, \tag{4}$$

where the minimum is over all bijections between D_1 and D_2. In words, the bottleneck distance is the maximum distance between the points of the two diagrams, after minimizing over all possible pairings of the points (including the points on the diagonals).

A fundamental property of persistence diagrams is their *stability*. According to the Persistence Stability Theorem (Cohen-Steiner et al. (2005); Chazal et al. (2012))

$$W_\infty(D_1, D_2) \leq ||\Delta_1 - \Delta_2||_\infty = H(S_1, S_2). \tag{5}$$

Here, H is the Hausdorff distance, namely,

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

where we recall that $A \oplus \epsilon = \bigcup_{x \in A} B(x, \epsilon)$. More generally, the definition of persistence diagrams and the above stability theorem do not restrict to distance functions but also extend to families of sublevel sets (resp. upper-level sets) of functions defined on $\mathbb{R}^d$ under very weak assumption. We refer the reader to Edelsbrunner and Harer (2010); Chazal et al. (2009, 2012) for a detailed exposition of the theory.

Given a sample $X_1, \ldots, X_n \sim P$, the empirical distance function is defined by

$$\widehat{\Delta}(x) = \min_{X_i} \|x - X_i\|. \tag{6}$$

Lemma 1 (Lemma 4 in Fasy et al., 2014b) *Suppose that P is supported on S, and has a density bounded away from zero and infinity. Then*

$$\sup_x |\widehat{\Delta}(x) - \Delta_S(x)| \xrightarrow{P} 0.$$

See also Cuevas and Rodríguez-Casal (2004). The previous lemma justifies using $\widehat{\Delta}$ to estimate the persistent homology of sublevel sets of Δ_S. In fact, the sublevel sets of $\widehat{\Delta}$ are just unions of balls around the observed data. That is,

$$L_t = \big\{x : \ \widehat{\Delta}(x) \leq t\big\} = \bigcup_{i=1}^{n} B(X_i, t).$$

The persistent homology of the union of the balls as t increases may be computed by creating a combinatorial representation (called a Cech complex) of the union of balls, and then applying basic operations from linear algebra (Edelsbrunner and Harer, 2010, Sections VI.2 and VII.1).

However, as soon as there is noise or outliers, the empirical distance function becomes useless, as illustrated in Figure 3. More specifically, suppose that

$$P = \pi R + (1 - \pi)(Q \star \Phi_\sigma), \tag{7}$$

where $\pi \in [0, 1]$, R is an outlier distribution (such as a uniform on a large set), Q is supported on S, $\star$ denotes convolution, and Φ_σ is a compactly supported noise distribution with scale parameter σ.

Recovering the persistent homology of Δ_S exactly (or even the homology of S) is not possible in general since the problem is under-identified. But we would still like to find a function that is similar to the distance function for S. The empirical distance function fails miserably even when π and σ are small. Instead, we now turn to the DTM.

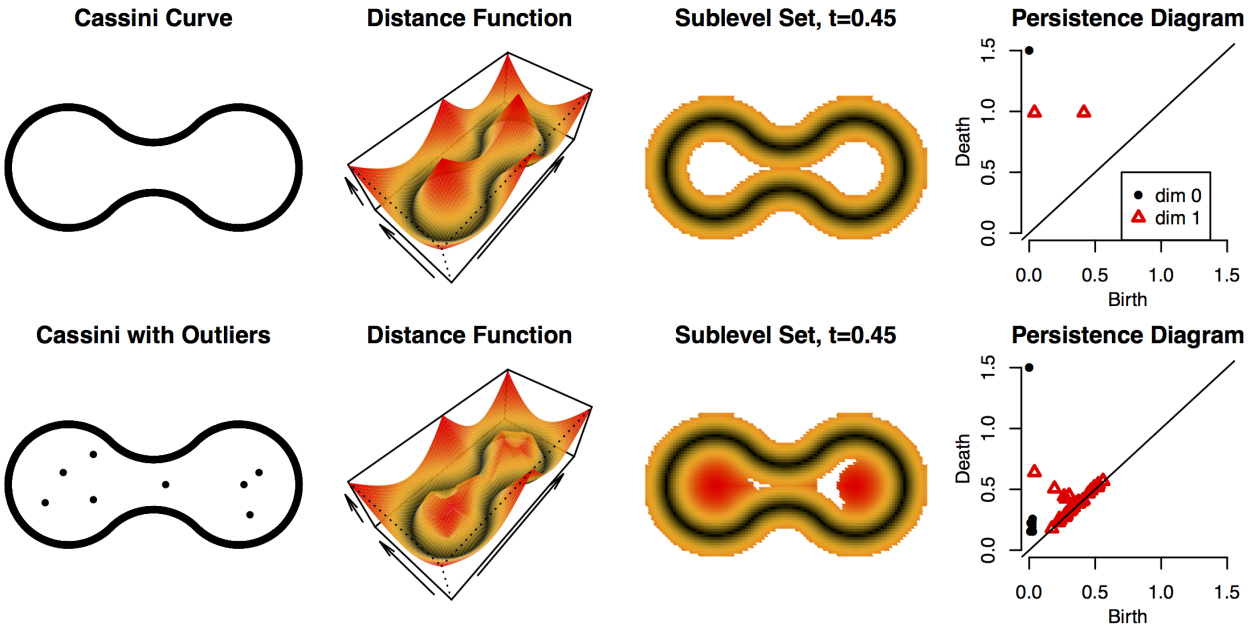

Figure 3: Top: data on the Cassini curve, the distance function $\widehat{\Delta}$, a typical sublevel set $\{x : \widehat{\Delta}(x) \leq t\}$ and the resulting persistence diagram. Bottom: the effect of adding a few outliers. Note that the distance function and persistence diagram are dramatically different.

2.2 Distance to a Measure

Given a probability measure P, for $0 < m < 1$, the *distance-to-measure (DTM)* at resolution m (Chazal et al., 2011) is defined by

$$\delta(x) \equiv \delta_{P,m}(x) = \sqrt{\frac{1}{m}\int_0^m (G_x^{-1}(u))^2 du}, \tag{8}$$

where $G_x(t) = P(\|X - x\| \leq t)$. Alternatively, the DTM can be defined using the cdf of the squared distances, as in the following lemma:

Lemma 2 (Chazal et al., 2015) *Let $F_x(t) = P(\|X - x\|^2 \leq t)$. Then*

$$\delta_{P,m}^2(x) = \frac{1}{m}\int_0^m F_x^{-1}(u) du.$$

Proof For any $0 < u < 1$,

$$\begin{aligned}\left[G_x^{-1}(u)\right]^2 &= \inf\left\{t^2 : G_x(t) \geq u\right\} = \inf\left\{t^2 : P(\|X - x\| \leq t) \geq u\right\} \\ &= \inf\left\{t : P(\|X - x\|^2 \leq t) \geq u\right\} = \inf\left\{t : F_x(t) \geq u\right\} = F_x^{-1}(u).\end{aligned}$$

Therefore

$$\delta_{P,m}^2(x) = \frac{1}{m}\int_0^m (G_x^{-1}(u))^2 du = \frac{1}{m}\int_0^m F_x^{-1}(u) du.$$

■

Given a sample $X_1, \ldots, X_n \sim P$, let P_n be the probability measure that puts mass $1/n$ on each X_i. It is easy to see that the distance to the measure P_n at resolution m is

$$\widehat{\delta}^2(x) \equiv \delta_{P_n,m}^2(x) = \frac{1}{k}\sum_{X_i \in N_k(x)} \|X_i - x\|^2, \tag{9}$$

where $k = \lceil mn \rceil$ and $N_k(x)$ is the set containing the k nearest neighbors of x among $X_1, \ldots, X_n$. We will use $\widehat{\delta}$ to estimate δ.

Now we summarize some important properties of the DTM, all of which are proved in Chazal et al. (2011) and Buchet et al. (2013). First, recall that the *Wasserstein distance of order p* between two probability measures P and Q is given by

$$W_p(P,Q) = \inf_J \left(\int \|x - y\|^p \, dJ(x,y) \right)^{1/p}, \tag{10}$$

where the infimum is over all joint distributions J for (X, Y) such that $X \sim P$ and $Y \sim Q$. We say that P satisfies the (a,b)-*condition* if there exist $a, b > 0$ such that, for every x in the support of P and every $\epsilon > 0$,

$$P\Big(B(x,\epsilon)\Big) \geq a\epsilon^b. \tag{11}$$

This means that the support does not have long, thin components.

Theorem 3 (Properties of DTM) *The following properties hold:*

1. *The distance to measure is 1-Lipschitz: for any probability measure P on $\mathbb{R}^d$ and any $(x, x') \in \mathbb{R}^d$,*
$$|\delta_{P,m}(x) - \delta_{P,m}(x')| \leq \|x - x'\|.$$
2. *If Q satisfies (11) and is supported on a compact set S, then*
$$\sup_x |\delta_{Q,m}(x) - \Delta_S(x)| \leq a^{-1/b} m^{1/b}. \tag{12}$$
In particular, $\sup_x |\delta_{Q,m}(x) - \Delta_S(x)| \to 0$ as $m \to 0$.
3. *If P and Q are two distributions, then*
$$\sup_x |\delta_{P,m}(x) - \delta_{Q,m}(x)| \leq \frac{1}{\sqrt{m}} W_2(P,Q). \tag{13}$$
4. *If Q satisfies (11) and is supported on a compact set S and P is another distribution (not necessarily supported on S), then*
$$\sup_x |\delta_{P,m}(x) - \Delta_S(x)| \leq a^{-1/b} m^{1/b} + \frac{1}{\sqrt{m}} W_2(P,Q) \tag{14}$$
Hence, if $m \asymp W_2(P,Q)^{2b/(2+b)}$, then $\sup_x |\delta_{P,m}(x) - \Delta_S(x)| = O(W_2(P,Q)^{2/(2+b)})$.

5. Let D_P be the diagram from $\delta_{P,m}$ and let D_Q be the diagram from $\delta_{Q,m}$, then

$$W_\infty(D_P, D_Q) \leq ||\delta_{P,m} - \delta_{Q,m}||_\infty. \tag{15}$$

For any compact set $A \subset \mathbb{R}^d$, let $r(A)$ denotes the radius of the smallest enclosing ball of A centered at zero:

$$r(A) = \inf\{r > 0 : A \subset B(0, r)\}.$$

We conclude this section by bounding the distance between the diagrams $D_{\delta_{P,m}}$ and D_{Δ_S}.

Lemma 4 (Comparison of Diagrams) *Let $P = \pi R + (1-\pi)(Q \star \Phi_\sigma)$ where Q is supported on S and satisfies (11), R is uniform on a compact set $A \subset \mathbb{R}^d$ and $\Phi_\sigma = N(0, \sigma^2 I)$. Then,*

$$W_\infty\left(D_{\delta_{P,m}}, D_{\Delta_S}\right) \leq a^{-1/b} m^{1/b} + \frac{\pi\sqrt{r(A)^2 + 2r(S)^2 + 2\sigma^2} + \sigma}{\sqrt{m}}.$$

Proof We first apply the stability theorem and parts 4 and 5 of the previous result:

$$W_\infty\left(D_{\delta_{P,m}}, D_{\Delta_S}\right) \leq a^{-1/b} m^{1/b} + \frac{1}{\sqrt{m}} W_2(P, Q).$$

The term $W_2(P, Q)$ can be upper bounded as follows:

$$W_2(P, Q) \leq W_2(P, Q \star \Phi_\sigma) + W_2(Q \star \Phi_\sigma, Q)$$

These two terms can be bounded with simple transport plans. Let Z be a Bernoulli random variable with parameter π. Let X and Y be random variables with distributions R and $Q \star \Phi_\sigma$. We take these three random variables to be independent. Then, the random variable V defined by $V = ZX + (1-Z)Y$ has for distribution the mixture distribution P. By definition of W_2, one has

$$\begin{aligned} W_2^2(P, Q \star \Phi_\sigma) &\leq \mathbb{E}\left(\|V - Y\|^2\right) \\ &\leq \mathbb{E}\left(|Z|^2\right) \mathbb{E}\left(\|X - Y\|^2\right), \end{aligned}$$

by definition of V and by independence of Z and $X - Y$. Next, we have $\mathbb{E}\left(\|X\|^2\right) \leq r(A)^2$ and $\mathbb{E}\left(\|Y\|^2\right) \leq 2[r(S)^2 + \sigma^2]$. Thus

$$W_2^2(P, Q \star \Phi_\sigma) \leq \pi^2\left(r(A)^2 + 2r(S)^2 + 2\sigma^2\right).$$

It can be checked in a similar way that $W_2(Q \star \Phi_\sigma, Q) \leq \sigma$ (see for instance the proof of Proposition 1 in Caillerie et al. (2011)) and the Lemma is proved. ■

Remark: Note that when π and σ are small (and m tends to 0) we see that the diagrams $D_{\delta_{P,m}}$ and D_{Δ_S} are close.

3. Limiting Distribution of the Empirical DTM

In this section, we find the limiting distribution of $\widehat{\delta}$ and we use this to find confidence bands for $\delta(x)$. We start with the pointwise limit.

Let $\delta(x) \equiv \delta_{P,m}(x)$ and $\widehat{\delta}(x) \equiv \delta_{P_n,m}(x)$, as defined in the previous section.

Theorem 5 (Convergence to Normal Distribution) *Let P be some distribution in $\mathbb{R}^d$. For some fixed x, assume that F_x is differentiable at $F_x^{-1}(m)$, for $m \in (0,1)$, with positive derivative $F_x'(F_x^{-1}(m))$. Then we have*

$$\sqrt{n}(\widehat{\delta}^2(x) - \delta^2(x)) \rightsquigarrow N(0, \sigma_x^2), \tag{16}$$

where

$$\sigma_x^2 = \frac{1}{m^2} \int_0^{F_x^{-1}(m)} \int_0^{F_x^{-1}(m)} [F_x(s \wedge t) - F_x(s)F_x(t)]\, ds\, dt.$$

Remark 6 *Note that assuming that F_x is differentiable is not a strong assumption. According to the Lebesgue differentiation theorem on $\mathbb{R}$, it will be satisfied as soon as the push forward measure of P by the function $\|x - \cdot\|^2$ is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}$.*

Proof From Lemma 2,

$$\delta^2(x) = \frac{1}{m}\int_0^m (G_x^{-1}(t))^2 dt = \frac{1}{m}\int_0^m F_x^{-1}(t) dt$$

where $G_x(t) = \mathbb{P}(\|X - x\| \leq t)$ and $F_x(t) = \mathbb{P}(\|X - x\|^2 \leq t)$. So

$$\sqrt{n}(\widehat{\delta}^2(x) - \delta^2(x)) = \frac{1}{m}\int_0^m \sqrt{n}[\widehat{F}_x^{-1}(t) - F_x^{-1}(t)] dt. \tag{17}$$

First suppose that $\widehat{F}_x^{-1}(m) > F_x^{-1}(m)$. Then, by integrating "horizontally" rather than

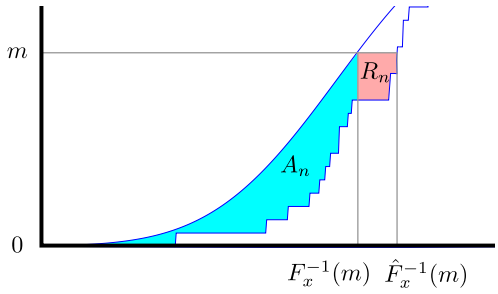

Figure 4: The integral of (17) can be decomposed into two parts, A_n and R_n.

"vertically", we can split the integral into two parts, as illustrated in Figure 4:

$$\begin{aligned}\frac{1}{m}\int_0^m \sqrt{n}[\widehat{F}_x^{-1}(t) - F_x^{-1}(t)] dt &= \frac{1}{m}\int_0^{F_x^{-1}(m)} \sqrt{n}[F_x(t) - \widehat{F}_x(t)] dt + \frac{1}{m}\int_{F_x^{-1}(m)}^{\widehat{F}_x^{-1}(m)} \sqrt{n}[m - \widehat{F}_x(t)] dt \\ &\equiv A_n(x) + R_n(x) \end{aligned} \tag{18}$$

Next, it can be easily checked that (18) is also true when $\widehat{F}_x^{-1}(m) < F_x^{-1}(m)$ if we take $\int_a^b f(u)du := -\int_b^a f(u)du$ when $a > b$. Now, since F_x is differentiable at m, we have that $\left|F_x^{-1}(m) - \widehat{F}_x^{-1}(m)\right| = O_P(1/\sqrt{n})$, see for instance Corollary 21.5 in van der Vaart (2000). According to the DKW (Dvoretsky-Kiefer-Wolfowitz) inequality we have that $\sup_t \left|F_x(t) - \widehat{F}_x(t)\right| = O_P(\sqrt{1/n})$ and thus

$$|R_n| \leq \frac{\sqrt{n}}{m} \left|F_x^{-1}(m) - \widehat{F}_x^{-1}(m)\right| \sup_t \left|F_x(t) - \widehat{F}_x(t)\right| = o_P(1).$$

Next, note that $\sqrt{n}[F_x(t) - \widehat{F}_x(t)] \rightsquigarrow \mathbb{B}(t)$, where $\mathbb{B}(t)$ is a Gaussian process with covariance function $[F_x(s \wedge t) - F_x(s)F_x(t)]$ (See, for example, van der Vaart and Wellner (1996)). By taking the integral, which is a bounded operator, we have that

$$A_n \rightsquigarrow \int_0^{F_x^{-1}(m)} \mathbb{B}(t)\, dt \stackrel{d}{=} N(0, \sigma_x^2),$$

where

$$\sigma_x^2 = \frac{1}{m^2} \int_0^{F_x^{-1}(m)} \int_0^{F_x^{-1}(m)} [F_x(s \wedge t) - F_x(s)F_x(t)]\, ds\, dt.$$

■

Now, we consider the functional limit of the distance to measure, on a compact domain $\mathcal{X} \subset \mathbb{R}^d$. The functional convergence of the DTM requires assumptions on the regularity of the quantile functions F_x^{-1}. We say that $\omega_x : (0,1) \to \mathbb{R}^+$ is a *modulus of (uniform) continuity* of F_x^{-1} if, for any $u \in (0,1)$,

$$\sup_{(m,m') \in (0,1)^2,\, |m'-m|<u} |F_x^{-1}(m') - F_x^{-1}(m)| \leq \omega_x(u),$$

with $\lim_{u \to 0} \omega_x(u) = \omega_x(0) = 0$. We say that $\omega_{\mathcal{X}} : (0,1) \to \mathbb{R}^+$ is an *uniform modulus of continuity* for the family of quantiles functions $(F_x^{-1})_{\mathcal{X}}$ if, for any $u \in (0,1)$ and any $x \in \mathcal{X}$,

$$\sup_{(m,m') \in (0,1)^2,\, |m'-m|<u} |F_x^{-1}(m') - F_x^{-1}(m)| \leq \omega_{\mathcal{X}}(u),$$

with $\lim_{u \to 0} \omega_{\mathcal{X}}(u) = \omega_{\mathcal{X}}(0) = 0$. When such modulus of continuity ω exists, note that it always can be chosen nondecreasing and this allows us to consider its generalized inverse ω^{-1}.

One may ask if the existence of the uniform modulus of continuity over a compact domain $\mathcal{X}$ is a strong assumption or not. To answer this issue, let us introduce the following assumption:

$(H_{\omega,\mathcal{X}})$: for any $x \in \mathcal{X}$, the push forward measure P_x of P by $\|x - \cdot\|^2$ is supported on a finite closed interval.

Note that Assumption $(H_{\omega,\mathcal{X}})$ is not very strong. For instance it is satisfied for a measure P supported on a compact and connected manifold, with P_x absolutely continuous for the Hausdorff measure on P. The following Lemma derives from general results on quantile functions given in Bobkov and Ledoux (2014) (see their Appendix A); the lemma shows that a uniform modulus of continuity for the quantiles exists under Assumption $(H_{\omega,\mathcal{X}})$.

Lemma 7 (Existence of Uniform Modulus of Continuity) *Let $\mathcal{X}$ be a compact domain and let P be a measure with compact support in $\mathbb{R}^d$. Assume that Assumption $(H_{\omega,\mathcal{X}})$ is satisfied. Then there exists a uniform modulus of continuity for the family of quantile functions F_x^{-1} over $\mathcal{X}$.*

Proof Let $x \in \mathcal{X}$. According to Propositions A.7 and A.12 in Bobkov and Ledoux (2014), Assumption $(H_{\omega,\mathcal{X}})$ is equivalent to assuming the existence of a uniform modulus of continuity of F_x^{-1} (it tends to zero at zero). We can then define ω_x on $(0,1)$ by

$$u \in (0,1) \mapsto \omega_x(u) := \sup_{(m,m')\in(0,1)^2\,,\,|m'-m|<u} \left|F_x^{-1}(m') - F_x^{-1}(m)\right|.$$

According to Lemma 8, we have that for any $(x,x') \in \mathcal{X}^2$:

$$\left|F_{x'}^{-1}(m) - F_x^{-1}(m)\right| \leq C\|x'-x\|, \tag{19}$$

where C only depends on P and $\mathcal{X}$. According to (19), for any $(m,m') \in (0,1)^2$, and for any $(x,x') \in \mathcal{X}^2$:

$$\left|F_x^{-1}(m') - F_x^{-1}(m)\right| \leq \left|F_{x'}^{-1}(m') - F_{x'}^{-1}(m)\right| + 2C\|x'-x\|.$$

By taking the supremum over the m and the m' such that $|m'-m| < u$, it yields:

$$\omega_x(u) \leq \omega_{x'}(u) + 2C\|x'-x\|,$$

and $x \mapsto \omega_x(u)$ is thus Lipschitz at any u. For any $u \in (0,1)$, let

$$\omega_{\mathcal{X}}(u) := \sup_{x\in\mathcal{X}} \omega_x(u),$$

which is finite because the function $x \mapsto \omega_x(u)$ is continuous on the compact $\mathcal{X}$ for any $u \in (0,1)$. We only need to prove that $\omega_{\mathcal{X}}$ is continuous at 0. Let $(u_n) \in (0,1)^{\mathbb{N}}$ be a decreasing sequence to zero. Since $\omega_{\mathcal{X}}$ is a non decreasing function, $\omega_{\mathcal{X}}(u_n)$ has a limit. For any $n \in \mathbb{N}$, there exists a point $x_n \in \mathcal{X}$ such that $\omega_{\mathcal{X}}(u_n) = \omega_{x_n}(u_n)$. Let $x_{\phi(n)}$ be a subsequence which converges to $\bar{x} \in \mathcal{X}$. According to (19),

$$\begin{aligned}
\omega_{\mathcal{X}}(u_{\phi(n)}) &\leq \left|\omega_{x_{\phi(n)}}(u_{\phi(n)}) - \omega_{\bar{x}}(u_{\phi(n)})\right| + \left|\omega_{\bar{x}}(u_{\phi(n)})\right| \\
&\leq C\left\|x_{\phi(n)} - \bar{x}\right\| + \left|\omega_{\bar{x}}(u_{\phi(n)})\right|
\end{aligned}$$

which gives that $\omega_{\mathcal{X}}(u_{\phi(n)})$ and $\omega_{\mathcal{X}}(u_n)$ both tend to zero because $\omega_{\bar{x}}$ is continuous at zero. Thus $\omega_{\mathcal{X}}$ is continuous at zero and the Lemma is proved. ■

We will also need the the following result, which shows that on any compact domain $\mathcal{X}$, the function $x \mapsto F_x^{-1}(m)$ is Lipschitz. For a domain $\mathcal{X} \in \mathbb{R}^d$, a probability P and a level m, we introduce the quantity $q_{P,\mathcal{X}}(m) \in \bar{\mathbb{R}}$, defined by

$$q_{P,\mathcal{X}}(m) := \sup_{x\in\mathcal{X}} F_x^{-1}(m).$$

Lemma 8 (Lipschitz Lemma) *Let P be a measure on $\mathbb{R}^d$ and let $m \in (0,1)$. Then, for any $(x, x') \in \mathbb{R}^d$,*

$$\left| \sqrt{F_{x'}^{-1}(m)} - \sqrt{F_x^{-1}(m)} \right| \leq \|x' - x\|.$$

Moreover, if $\mathcal{X}$ is a compact domain in $\mathbb{R}^d$, then $q_{P,\mathcal{X}}(m) < \infty$ and for any $(x, x') \in \mathcal{X}^2$:

$$\left| F_{x'}^{-1}(m) - F_x^{-1}(m) \right| \leq 2\sqrt{q_{P,\mathcal{X}}(m)}\ \|x' - x\|.$$

Proof Let $(x, a) \in \mathbb{R}^2$, note that

$$B\left(x, \sqrt{F_x^{-1}(m)}\right) \subseteq B\left(x + a, \sqrt{F_x^{-1}(m)} + \|a\|\right),$$

which implies

$$m = \mathbb{P}\left[B\left(x, \sqrt{F_x^{-1}(m)}\right)\right] \leq \mathbb{P}\left[B\left(x + a, \sqrt{F_x^{-1}(m)} + \|a\|\right)\right].$$

Therefore $\sqrt{F_{x+a}^{-1}(m)} \leq \sqrt{F_x^{-1}(m)} + \|a\|$. Similarly,

$$m = \mathbb{P}\left[B\left(x + a, \sqrt{F_{x+a}^{-1}(m)}\right)\right] \leq \mathbb{P}\left[B\left(x, \sqrt{F_{x+a}^{-1}(m)} + \|a\|\right)\right],$$

which implies $\sqrt{F_x^{-1}(m)} \leq \sqrt{F_{x+a}^{-1}(m)} + \|a\|$.

Let $\mathcal{X}$ be a compact domain of $\mathbb{R}^d$, then according to the previous result for some fixed $x \in \mathcal{X}$ and for any $x' \in \mathcal{X}$, $\sqrt{F_{x'}^{-1}(m)} \leq \|x' - x\| + \sqrt{F_x^{-1}(m)}$ which is bounded on $\mathcal{X}$. The last statement follows from the fact that $|x - y| = |\sqrt{x} - \sqrt{y}|\ |\sqrt{x} + \sqrt{y}|$. ■

We are now in position to state the functional limit of the distance to measure to the empirical measure.

Theorem 9 (Functional Limit) *Let P be a measure on $\mathbb{R}^d$ with compact support. Let $\mathcal{X}$ be a compact domain on $\mathbb{R}^d$ and $m \in (0,1)$. Assume that there exists a uniform modulus of continuity $\omega_{\mathcal{X}}$ for the family $(F_x^{-1})_{\mathcal{X}}$. Then $\sqrt{n}(\widehat{\delta}^2(x) - \delta^2(x)) \rightsquigarrow \mathbb{B}(x)$ for a centered Gaussian process $\mathbb{B}(x)$ with covariance kernel*

$$\kappa(x, y) = \frac{1}{m^2} \int_0^{F_x^{-1}(m)} \int_0^{F_y^{-1}(m)} \left(\mathbb{P}\left[B(x, \sqrt{t}) \cap B(y, \sqrt{s})\right] - F_x(t) F_y(s) \right) ds\ dt.$$

Remark 10 *Note that the functional limit is valid for any value of $m \in (0,1)$. A local version of this result could be also proposed by considering the (local) modulii of continuity of the quantile functions at m. For the sake of clarity, we prefer to give a global version.*

Remark 11 *By the delta method (as described in Section 4) a similar result holds for $\sqrt{n}(\widehat{\delta}(x) - \delta(x))$ as long as $\inf_x \delta(x) > 0$.*

Proof In the proof of Theorem 5 we showed that $\sqrt{n}(\widehat{\delta}^2(x) - \delta^2(x)) = A_n(x) + R_n(x)$ where

$$A_n(x) = \frac{1}{m}\int_0^{F_x^{-1}(m)} \sqrt{n}[F_x(t) - \widehat{F}_x(t)]dt$$
$$R_n(x) = \frac{1}{m}\int_{F_x^{-1}(m)}^{\widehat{F}_x^{-1}(m)} \sqrt{n}[m - \widehat{F}_x(t)]dt.$$

First, we show that $\sup_{x\in\mathcal{X}} |R_n(x)| = o_P(1)$. Then we prove that $A_n(x)$ converges to a Gaussian process.

Note that $|R_n(x)| \leq \frac{\sqrt{n}}{m}|S_n(x)||T_n(x)|$ where

$$S_n(x) = \left|F_x^{-1}(m) - \widehat{F}_x^{-1}(m)\right|, \quad T_n(x) = \sup_t \left|F_x(t) - \widehat{F}_x(t)\right|.$$

Let $\xi_i \sim$Uniform (0,1), for $i = 1, \ldots, n$ and let H_n be their empirical distribution function. Define $k = mn$. Then $\widehat{F}_x^{-1}(m) \stackrel{d}{=} F_x^{-1}(\xi_{(k)}) = F_x^{-1}\left(H_n^{-1}(m)\right)$, where $\xi_{(k)}$ is the kth order statistic. Thus, for any $m > 0$ and any $x \in \mathcal{X}$:

$$\begin{aligned}
\mathbb{P}\left(|S_n(x)| > \epsilon\right) &= \mathbb{P}\left(|F_x^{-1}(H_n^{-1}(m)) - F_x^{-1}(m)| > \epsilon\right) \\
&\leq \mathbb{P}\left(\omega_{\mathcal{X}}(|m - H_n^{-1}(m)|) > \epsilon\right) \\
&\leq \mathbb{P}\left(|m - H_n^{-1}(m)| > \omega_{\mathcal{X}}^{-1}(\epsilon)\right) \\
&\leq 2\exp\left\{-\frac{n\left[\omega_{\mathcal{X}}^{-1}(\epsilon)\right]^2}{m}\frac{1}{1 + \frac{2\omega_{\mathcal{X}}^{-1}(\epsilon)}{3m}}\right\} \qquad (20)
\end{aligned}$$

In the last line we used inequality 1 page 453 and Point (12) of Proposition 1 page 455 of Shorack and Wellner (2009). Note that $\omega_{\mathcal{X}}^{-1}(\epsilon) > 0$ for any $\varepsilon > 0$ because $\omega_{\mathcal{X}}$ is assumed to be continuous at zero by definition.

Fix $\varepsilon > 0$. There exists an absolute constant $C_{\mathcal{X}}$ such that there exists an integer $N \leq C_{\mathcal{X}}\varepsilon^{-d}$ and N points $(x_1, \ldots, x_N)$ laying in $\mathcal{X}$ such that $\bigcup_{j=1\ldots N} B_j \supseteq \mathcal{X}$, where $B_j = B(x_j, \varepsilon)$. Now, we apply Lemma 8 with P, and with P_n and we find that for any $x \in B_j$:

$$\left|F_x^{-1}(m) - F_{x_j}^{-1}(m)\right| \leq 2\sqrt{q_{P,\mathcal{X}}(m)}\,\varepsilon \quad \text{and} \quad \left|\widehat{F}_x^{-1}(m) - \widehat{F}_{x_j}^{-1}(m)\right| \leq 2\sqrt{q_{P_n,\mathcal{X}}(m)}\,\varepsilon.$$

Thus, for any $x \in B_j$,

$$\begin{aligned}
\left|F_x^{-1}(m) - \widehat{F}_x^{-1}(m)\right| &\leq \left|F_x^{-1}(m) - F_{x_j}^{-1}(m)\right| + \left|F_{x_j}^{-1}(m) - \widehat{F}_{x_j}^{-1}(m)\right| + \left|\widehat{F}_{x_j}^{-1}(m) - \widehat{F}_x^{-1}(m)\right| \\
&\leq 2\left[\sqrt{q_{P,\mathcal{X}}(m)} + \sqrt{q_{P_n,\mathcal{X}}(m)}\right]\varepsilon + |F_{x_j}^{-1}(m) - \widehat{F}_{x_j}^{-1}(m)| \\
&\leq C\varepsilon + |F_{x_j}^{-1}(m) - \widehat{F}_{x_j}^{-1}(m)| \qquad (21)
\end{aligned}$$

where C is a positive constant which only depends on $\mathcal{X}$ and P. Using a union bound together with (20) , we find that

$$P\left(\sup_{x\in\mathcal{X}}|S_n(x)|>2C\varepsilon\right)\leq P\left(\sup_{j=1\dots N}|S_n(x_j)|>C\varepsilon\right)$$
$$\leq 2C_{\mathcal{X}}\varepsilon^{-d}\exp\left\{-\frac{n\left[\omega_{\mathcal{X}}^{-1}(C\epsilon)\right]^2}{m}\frac{1}{1+\frac{2\omega_{\mathcal{X}}^{-1}(C\epsilon)}{3m}}\right\}.$$

Thus, $\sup_{x\in\mathcal{X}}|S_n(x)|=o_P(1)$. Then

$$\begin{aligned}\sup_{x\in\mathcal{X}}|T_n(x)| &= \sup_{x\in\mathcal{X}}\sup_t|\widehat{F}_x(t)-F_x(t)|\\ &= \sup_{x\in\mathcal{X}}\sup_t\left|\mathbb{P}_n\left(B(x,\sqrt{t}\right)-\mathbb{P}\left(B(x,\sqrt{t}\right)\right|\\ &\leq \sup_{B\in\mathcal{B}_d}|\mathbb{P}_n(B)-\mathbb{P}(B)|=O_P\left(\sqrt{\frac{d}{n}}\right)\end{aligned}\tag{22}$$

where $\mathcal{B}_d$ is the set of balls in $\mathbb{R}^d$ and we used the Vapnik-Chervonenkis theorem. Finally, we obtain that

$$\sup_{x\in\mathcal{X}}|R_n(x)|\leq\frac{\sqrt{n}}{m}\sup_{x\in\mathcal{X}}|S_n(x)|\sup_{x\in\mathcal{X}}|T_n(x)|=o_P(1).\tag{23}$$

Since $\sup_{x\in\mathcal{X}}|R_n(x)|=o_P(1)$, it only remains to prove that the process A_n converges to a Gaussian process.

Now, we consider the process A_n on $\mathcal{X}$. Let us denote $\nu_n:=\sqrt{n}(P_n-P)$ the empirical process. Note that

$$A_n(x)=\frac{1}{m}\nu_n\left(\int_0^{F_x^{-1}(m)}I_{\|x-X\|^2\leq t}dt\right)=\frac{1}{m}\nu_n\left(f_x\right)$$

where $f_x(y):=\left[F_x^{-1}(m)-\|x-y\|^2\right]\wedge 0$. For any $(x,x')\in\mathcal{X}$ and any $y\in\mathbb{R}^d$, we have

$$\begin{aligned}|f_x(y)-f_x(y)| &\leq |F_x^{-1}(m)-F_{x'}^{-1}(m)|+\|x-x'\|\left[\|x\|+\|x'\|+2\|y\|\right]\\ &\leq 2\left[r(\mathcal{X})+\|y\|+\sqrt{q_{P,\mathcal{X}}(m)}\right]\|x-x'\|\end{aligned}$$

Since P is compactly supported, then the collection of functions $(f_x)_{x\in\mathcal{X}}$ is P-Donsker (see for instance 19.7 in van der Vaart (2000)) and $A_n(x)\rightsquigarrow\mathbb{B}(x)$ for a centered Gaussian process $\mathbb{B}(x)$ with covariance kernel

$$\begin{aligned}\kappa(x,y) &= \mathrm{Cov}(A_n(x),A_n(y))=\mathbb{E}[A_n(x)A_n(y)]\\ &= \frac{1}{m^2}\int_0^{F_x^{-1}(m)}\int_0^{F_y^{-1}(m)}\mathbb{E}\left[\left(\widehat{F}_x(t)-F_x(t)\right)\left(\widehat{F}_y(s)-F_y(s)\right)\right]\,ds\,dt\\ &= \frac{1}{m^2}\int_0^{F_x^{-1}(m)}\int_0^{F_y^{-1}(m)}\left(\mathbb{P}\left[B(x,\sqrt{t})\cap B(y,\sqrt{s})\right]-F_x(t)F_y(s)\right)ds\,dt.\end{aligned}$$

■

4. Hadamard Differentiability and The Bootstrap

In this section, we use the bootstrap to get a confidence band for δ. Define c_α by

$$\mathbb{P}(\sqrt{n}||\widehat{\delta} - \delta||_\infty > c_\alpha) = \alpha.$$

Let $X_1^*, \ldots, X_n^*$ be a sample from the empirical measure P_n and let $\widehat{\delta}^*$ be the corresponding empirical DTM. The bootstrap estimate $\widehat{c}_\alpha$ is defined by

$$\mathbb{P}(\sqrt{n}||\widehat{\delta}^* - \widehat{\delta}||_\infty > \widehat{c}_\alpha \mid X_1, \ldots, X_n) = \alpha.$$

As usual, $\widehat{c}_\alpha$ can be approximated by Monte Carlo. Below we show that this bootstrap is valid. It then follows that

$$\mathbb{P}\left(||\delta - \widehat{\delta}||_\infty < \frac{\widehat{c}_\alpha}{\sqrt{n}}\right) \to 1 - \alpha.$$

A different approach to the bootstrap is considered in Section 6.

To prepare for our next result, let $\mathcal{B}$ denote the class of all closed Euclidean balls in $\mathbb{R}^d$ and let $\mathbb{B}$ denote the P-Brownian bridge on $\mathcal{B}$, i.e. the centered Gaussian process on $\mathcal{B}$ with covariance function $\kappa(B, C) = P(B \cap C) - P(B)P(C)$, with B, C in $\mathcal{B}$. We will denote with $\mathbb{B}_x(r)$ the value of $\mathbb{B}$ at $B(x, r)$, the closed ball centered at $x \in \mathbb{R}^d$ and with radius $r > 0$.

Theorem 12 (Bootstrap Validity) *Let P be a measure on $\mathbb{R}^d$ with compact support S, $m \in (0, 1)$ be fixed and $\mathcal{X}$ be a compact domain in $\mathbb{R}^d$. Assume that $F_{P,x} = F_x$ is differentiable at $F_x^{-1}(m)$ and that there exist a constant $C > 0$ such that for all small $\eta \in \mathbb{R}$,*

$$\sup_{x \in \mathcal{X}} \left|F_x\left(F_x^{-1}(m)\right) - F_x(F_x^{-1}(m) + \eta)\right| < \epsilon \quad \textit{implies} \quad |\eta| < C\epsilon. \tag{24}$$

for all $x \in \mathcal{X}$. Then, $\sup_{x \in \mathcal{X}} \sqrt{n} \left| \left(\widehat{\delta}^*(x)\right)^2 - \left(\widehat{\delta}(x)\right)^2 \right|$ converges in distribution to

$$\sup_{x \in \mathcal{X}} \left| \frac{1}{m} \int_0^{F_x^{-1}(m)} \mathbb{B}_x(u) du \right|$$

conditionally given $X_1, X_2, \ldots$, in probability.

We will establish the above result using the functional delta method, which entails showing that the distance to measure function is Hadamard differentiable at P. In fact, the proof further shows that the process

$$x \in \mathcal{X} \mapsto \sqrt{n}\left(\delta^2(x) - \widehat{\delta}^2(x)\right),$$

converges weakly to the Gaussian process

$$x \in \mathcal{X} \mapsto -\frac{1}{m} \int_0^{F_x^{-1}(m)} \mathbb{B}_x(u) du.$$

Remark 13 *This result is consistent with the result established in Theorem 9, but in order to establish Hadamard differentiability, we use a slightly different assumption. Theorem 9 is proved by assuming a uniform modulus of continuity on the quantile functions F_x^{-1} whereas in Theorem 12 a uniform lower bound on the derivatives is required. These two assumptions are consistent: they both say that F_x^{-1} is well-behaved in a neighborhood of m for all x. However, the condition used in Theorem 12 is stronger than the condition used in Theorem 9.*

Proof [Proof of Theorem 12] Let us first give the definition of Hadamard differentiability, for which we refer the reader to, e.g., Section 3.9 of van der Vaart and Wellner (1996). A map ϕ from a normed space $(\mathcal{D}, \|\cdot\|_{\mathcal{D}})$ to a normed space $(\mathcal{E}, \|\cdot\|_{\mathcal{E}})$ is Hadamard differentiable at the point $x \in \mathcal{D}$ if there exists a continuous linear map $\phi'_x : \mathcal{D} \to \mathcal{E}$ such that

$$\left\| \frac{\phi(x + th_t) - \phi(x)}{t} - \phi'_x(h) \right\|_{\mathcal{E}} \to 0, \tag{25}$$

whenever $\|h_t - h\|_{\mathcal{D}} \to 0$ as $t \to 0$.

We also recall the functional delta method (see, e.g. van der Vaart and Wellner, 1996, Theorem 3.9.4): suppose that T_n takes values in $\mathcal{D}$, $r_n \to \infty$, $r_n(T_n - \theta) \rightsquigarrow T$, and suppose that ϕ is Hadamard differentiable at θ. Then $r_n(\phi(T_n) - \phi(\theta)) \rightsquigarrow \phi'_\theta(T)$. Moreover, by Theorem 3.9.11 of van der Vaart and Wellner (1996) the bootstrap has the same limit. More precisely, given $X_1, X_2, \ldots$, we have that $r_n(\phi(T_n^*) - \phi(T_n))$ converges conditionally in distribution to $\phi'_\theta(T)$, in probability. This implies the validity of the bootstrap confidence sets.

We define $\mathcal{M}$ to be the space of finite, σ-finite signed measures on $(\mathbb{R}^d, \mathcal{B}^d)$ supported on the compact set S and the mapping $\|\cdot\|_{\mathcal{B}} \colon \mathcal{M} \mapsto \mathbb{R}$ given by

$$\|\mu\|_{\mathcal{B}} = \sup_{B \in \mathcal{B}} |\mu(B)|, \quad \mu \in \mathcal{M}.$$

In Lemma 14 we show that this is a normed space.

For our purposes, instead of using $\mathcal{M}$ it will be convenient to work with the equivalent space of the evaluations of all $\mu \in \mathcal{M}$ over the balls $\mathcal{B}$. Formally, let $\ell^\infty(\mathcal{B})$ denote the normed space of bounded functions on $\mathcal{B}$ equipped with the supremum norm. Then, by Lemma 14, the mapping from $\mathcal{M}$ into $\ell^\infty(\mathcal{B})$ given by

$$\mu \mapsto (\mu \colon \mathcal{B} \to [0, 1]) \tag{26}$$

is a bijection on its image, which we will denote by $\mathcal{D}$. By definition, the supremum norm on $\mathcal{D}$ is exactly the norm $\|\cdot\|_{\mathcal{B}}$, so that $\mathcal{D} \subset \ell^\infty(\mathcal{B})$ equipped with the supremum norm is a normed space. With a slight abuse of notation, we will identify measures in $\mathcal{M}$ with the corresponding points in $\mathcal{D}$ and write $\mu \in \mathcal{D}$ to denote the signed measure μ corresponding to the point $\{\mu(B), B \in \mathcal{B}\}$ in $\mathcal{D}$.

The advantage of using the space $\mathcal{D}$ instead of $\mathcal{M}$ is that the convergence of the empirical process $(\sqrt{n}(\mathbb{P}_n(B) - P(B)) : \ B \in \mathcal{B})$ to the Brownian bridge takes place in $\mathcal{D}$, as required by the delta-method for the bootstrap (see van der Vaart and Wellner, 1996, Theorem 3.9.).

For a signed measure μ in $\mathcal{M}$, $x \in \mathbb{R}^d$ and $r > 0$, we set $F_{\mu,x}(r) = \mu\left(B(x, \sqrt{r})\right)$. Notice if P is a probability measure, then $F_{P,x}$ is the c.d.f. of the univariate random variable

$\|X - x\|^2$, with $X \sim P$. For a general $\mu \in \mathcal{M}$, $F_{\mu,x}$ is a cadlag function, though not monotone. For any $m \in \mathbb{R}$, μ in $\mathcal{M}$ and $x \in \mathbb{R}^d$, set

$$F^{-1}_{\mu,x}(m) = \inf\Big\{r > 0 \colon \mu(B(x, \sqrt{r})) \geq m\Big\},$$

where the infimum over the empty set is define to be ∞. If P is a probability measure and $m \in (0, 1)$ then $F^{-1}_{P,x}(m)$ is just the m-th quantile of the random variable $\|X - x\|^2$, $X \sim P$.

Fix a $m \in (0, 1)$ and let $\mathcal{M}_m = \mathcal{M}_m(\mathcal{X})$ denote the subset of $\mathcal{M}$ consisting of all finite signed measure μ such that, there exists a value of $r > 0$ for which $\inf_{x\in\mathcal{X}} \mu\left(B(x, \sqrt{r})\right) \geq m$. Thus, for any $\mu \in \mathcal{M}_m$ and $x \in \mathcal{X}$, $F^{-1}_{\mu,x}(m) < \infty$. Let $\mathcal{D}_m$ be the image of $\mathcal{M}_m$ by the mapping (26).

Let $\mathcal{E}$ the set of bounded, real-valued function on $\mathcal{X}$, a normed space with respect to the sup norm. Finally, we define $\phi \colon \mathcal{D}_m \to \mathcal{E}$ to be the mapping

$$\mu \in \mathcal{D}_m \mapsto \phi(\mu)(x) = F^{-1}_{\mu,x}(m) - \frac{1}{m}\int_0^{F^{-1}_{\mu,x}(m)} F_{\mu,x}(u)du, \quad x \in \mathcal{X} \tag{27}$$

Notice that if P is a probability measure, simple algebra shows that $\phi(P)(x)$ is the square value of the distance to measure of P at the point x, i.e. $\delta_p^2(x)$; see Figure 5.

Below we will show that, for any probability measure P, the mapping (27) is Hadamard differentiable at P.

For an arbitrary $Q \in \mathcal{D}$, let $\{Q_t\}_{t>0} \subset \mathcal{D}$ be a sequence of signed measure such that $\lim_{t\to 0} \|Q_t - Q\|_{\mathcal{B}} = 0$ and such that $P + tQ_t \in \mathcal{D}_m$ for all t. Sequences of this form exist: since $\|tQ_t\|_{\mathcal{B}} \to 0$ as $t \to 0$, for any arbitrary $0 < \epsilon < 1 - m$ and all t small enough,

$$\inf_{x\in\mathcal{X}} (P + tQ_t)\left(B\left(x, F^{-1}_{P,x}(m + \epsilon)\right)\right) \geq m + \epsilon/2.$$

By the boundedness of $\mathcal{X}$ and compactness of S, this implies that there exists a number $r > 0$ such that

$$\inf_x (P + tQ_t)\left(B\left(x, r\right)\right) \geq m,$$

so the image of $P + tQ_t$ by (27) is an element of $\mathcal{E}$ (i.e. it is a bounded function).

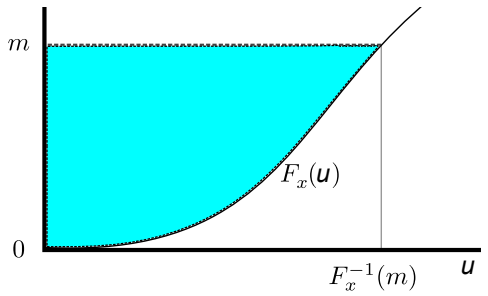

Figure 5: The integral $\int_0^m F^{-1}_{P,x}(u)du$ is equivalent to $mF^{-1}_{P,x}(m) - \int_0^{F^{-1}_{P,x}(m)} F_{P,x}(u)du$.

For sake of readability, below we will write $F_{x,t}$ and $F^{-1}_{x,t}(m)$ for $F_{P+tQ_t,x}$ and $F^{-1}_{P+tQ_t,x}(m)$, respectively, and F_x for $F_{P,x}$. Also, for each $x \in \mathcal{X}$ and $z \in \mathbb{R}_+$ we the set $\mathcal{A}_{x,z} = \{y : \|y - x\|^2 \leq z\}$ and $F_{x,t}(z) = (P + tQ_t)(\mathcal{A}_{x,z})$.

Thus,

$$\phi(P)(x) = \delta_P^2(x) = F_x^{-1}(m) - \frac{1}{m}\int_0^{F_x^{-1}(m)} F_x(u)du$$

and

$$\phi(P + tQ_t)(x) = F_{x,t}^{-1}(m) - \frac{1}{m}\int_0^{F_{x,t}^{-1}(m)} F_{x,t}(u)du.$$

Some algebra show that, for any x,

$$\frac{\phi(P)(x) - \phi(P + tQ_t)(x)}{t} = \frac{F_x^{-1}(m) - F_{x,t}^{-1}(m)}{t} - \frac{A(x,t)}{mt}, \tag{28}$$

where

$$A(x,t) = \begin{cases} \int_0^{F_x^{-1}(m)} [F_x(u) - F_{x,t}(u)]\, du - \int_{F_x^{-1}(m)}^{F_{x,t}^{-1}(m)} F_{x,t}(u)du & \text{if } F_x^{-1}(m) \le F_{x,t}^{-1}(m) \\ \int_0^{F_x^{-1}(m)} [F_x(u) - F_{x,t}(u)]\, du + \int_{F_{x,t}^{-1}(m)}^{F_x^{-1}(m)} F_{x,t}(u)du & \text{if } F_x^{-1}(m) > F_{x,t}^{-1}(m). \end{cases}$$

To demonstrate Hadamard differentiability (see 25), we will prove that, as $t \to 0$, the expression in (28), as a bounded function of $x \in \mathcal{X}$, will converge in $\mathcal{E}$ to the bounded function

$$x \in \mathcal{X} \mapsto -\frac{1}{m}\int_0^{F_x^{-1}(m)} Q(\mathcal{A}_{x,u})du.$$

Towards that end, we have, for all t and any $x \in \mathcal{X}$,

$$\begin{aligned} \frac{A(x,t)}{t} &= \frac{1}{t}\left[\int_0^{F_x^{-1}(m)} tQ_t(\mathcal{A}_{x,u})du - \int_{F_x^{-1}(m)}^{F_{x,t}^{-1}(m)} (P + tQ_t)(\mathcal{A}_{x,u})du\right] \\ &= \int_0^{F_x^{-1}(m)} Q_t(\mathcal{A}_{x,u})du - \frac{1}{t}\int_{F_x^{-1}(m)}^{F_{x,t}^{-1}(m)} P(\mathcal{A}_{x,u})du - \int_{F_x^{-1}(m)}^{F_{x,t}^{-1}(m)} (Q_t)(\mathcal{A}_{x,u})du \\ &\equiv A_1(x,t) - A_2(x,t) - A_3(x,t), \end{aligned}$$

where, for $a < b$, we write $\int_b^a = -\int_a^b$.

To handle the three terms appearing in the last display, we use Lemma 15 below which shows that $\sup_{x\in\mathcal{X}} \left|m - F_x\left(F_{x,t}^{-1}(m)\right)\right| = O(t)$ and that $\sup_{x\in\mathcal{X}} |F_x^{-1}(m) - F_{x,t}^{-1}(m)| = O(t)$ as as $t \to 0$.

We now analyze the terms $A_1(x,t)$, $A_2(x,t)$ and $A_3(x,t)$ separately.

- **Term $A_1(x,t)$.** As $t \to 0$, $Q_t \to Q$ and, uniformly in $x \in \mathcal{X}$ and $z > 0$, $|Q_t(\mathcal{A}_{x,z})| \le |Q_t(S)| = |Q(S)| + o(1) < \infty$. Furthermore, $\sup_{x\in\mathcal{X}} F_x^{-1}(m) < \infty$ by compactness of $\mathcal{X}$ and S. Therefore, using the dominated convergence theorem,

$$\lim_{t\to 0}\sup_{x\in\mathcal{X}}\left|\frac{A_1(x,t)}{m} - \frac{1}{m}\int_0^{F_x^{-1}(m)} Q(\mathcal{A}_{x,u})du\right| = 0. \tag{29}$$

- **Term** $A_2(x,t)$**.** Since $P(\mathcal{A}_{x,u})$ is non-decreasing in u for all x, we have

$$\frac{F_{x,t}^{-1}(m) - F_x^{-1}(m)}{t} \times \min\left\{m, F_x(F_{x,t}^{-1}(m))\right\} \leq A_2(x,t) \leq \frac{F_{x,t}^{-1}(m) - F_x^{-1}(m)}{t} \times \max\left\{m, F_x(F_{x,t}^{-1}(m))\right\}.$$

  Using (32), we conclude that

$$\lim_{t\to 0} \sup_{x\in\mathcal{X}} \left| \frac{F_x^{-1}(m) - F_{x,t}^{-1}(m)}{t} - \frac{A_2(x,t)}{m} \right| = 0. \tag{30}$$

- **Term** $A_3(x,t)$**.** Finally, since $|Q_t(S)| \leq |Q(S)| + o(1)$ as $t \to 0$ and using (35), we obtain

$$\sup_x |A_3(x,t)| = O\Big(\sup_x \left|F_{x,t}^{-1}(m) - F_x^{-1}(m)\right| \Big) = o(1) \tag{31}$$

  as $t \to 0$.

Therefore, from (28), (29), (30), and (31),

$$\lim_{t\to 0} \sup_x \left| \frac{\phi(P)(x) - \phi(P + tQ_t)(x)}{t} + \frac{1}{m} \int_0^{F_x^{-1}(m)} Q(\mathcal{A}_{x,u}) du \right| = 0,$$

which shows that

$$x \in \mathcal{X} \mapsto -\frac{1}{m} \int_0^{F_x^{-1}(m)} Q(\mathcal{A}_{x,u}) du$$

is the Hadamard derivative of δ^2 at P.

The statement of the theorem now follows from an application of Theorem 3.9.11 in van der Vaart and Wellner (1996) and the fact that, since $\mathcal{B}$ is a Donsker class, the empirical process $(\sqrt{n}(\mathbb{P}_n(B) - P(B)) : B \in \mathcal{B})$ converges to the Brownian bridge $\mathbb{B}$ on $\mathcal{B}$ with covariance kernel $\kappa(B,C) = P(B \cap C) - P(B)P(C)$. ■

Lemma 14 (Normed Space) *The pair* $(\mathcal{M}, \|\cdot\|_{\mathcal{B}})$ *is a normed space.*

Proof It is clear that $\mathcal{M}$ is closed under addition and scalar multiplication, and so it is a linear space. We then need to show that the mapping $\|\cdot\|_{\mathcal{B}}$ is a norm. It is immediate to see that it absolutely homogeneous and satisfies the triangle inequality: for any μ and ν in $\mathcal{M}$ and $c \in \mathbb{R}$, $\|c\mu\|_{\mathcal{B}} = |c|\|\mu\|_{\mathcal{B}}$ and $\|\mu + \nu\|_{\mathcal{B}} \leq \|\mu\|_{\mathcal{B}} + \|\nu\|_{\mathcal{B}}$. It remains to prove that $\|\mu\|_{\mathcal{B}} = 0$ if and only if μ is identically zero, i.e. $\mu(A) = 0$ for all Borel sets A. One direction is immediate: if $\|\mu\|_{\mathcal{B}} > 0$, then there exists a ball B such that $\mu(B) \neq 0$, so that $\mu \neq 0$. For the other direction, assume that $\mu \in \mathcal{M}$ is such that $\|\mu\|_{\mathcal{B}} = 0$. By the Jordan decomposition, μ can be represented as the difference of two singular, non-negative finite measures: $\mu = \mu_+ - \mu_-$. The condition $\mu(B) = 0$ for all $B \in \mathcal{B}$ is equivalent to $\mu_+(B) = \mu_-(B)$ for all $B \in \mathcal{B}$. We will show that this further implies that the supports of μ_+ and μ_-, denoted with S_+ and S_- respectively, are both empty, and therefore that μ is identically zero. Indeed, recall that the support of a Borel measure λ over a topological space $\mathbb{X}$ is the set of points $x \in \mathbb{X}$ all of whose open neighborhoods have positive λ-measure.

In our setting this is equivalent to the set of points in $\mathbb{R}^d$ such that all open balls centered at those points have positive measure, which in turn is equivalent to the set of points such that all closed balls centered at those points have positive measure. Therefore, using the fact that $\mu^+(B) = \mu^-(B)$, for all $B \in \mathcal{B}$,

$$S_+ = \left\{x \in \mathbb{R}^d \colon \mu_+(B(x,r)) > 0, \forall r > 0\right\} = \left\{x \in \mathbb{R}^d \colon \mu_-(B(x,r)) > 0, \forall r > 0\right\} = S_-.$$

where $B(X,r) = \{y \in \mathbb{R}^d \colon \|y - x\| \leq r\}$. It then follows that S_+ and S_- must be empty, for otherwise μ_+ and μ_- would be mutually singular, non-zero measures with the same support, a contradiction. ■

Lemma 15 *Under the assumptions of the theorem and as $t \to 0$,*

$$\sup_{x \in \mathcal{X}} \left| m - F_x\left(F_{x,t}^{-1}(m)\right)\right| = O(t), \tag{32}$$

and

$$\sup_{x \in \mathcal{X}} |F_x^{-1}(m) - F_{x,t}^{-1}(m)| = O(t). \tag{33}$$

Proof Set $\mathcal{A}_{x,t} = \{y \colon \|y - x\|^2 \leq F_{x,t}^{-1}(m)\}$ and let γ_t be any positive, decreasing function of t such that $\gamma_t = o(t)$ as $t \to 0$. Then, for all small enough t and all $x \in \mathcal{X}$, the set $\mathcal{A}_{x,t,\gamma_t} = \{y \colon \|y - x\|^2 \leq F_{x,t}^{-1}(m) - \gamma_t\}$ is non-empty and

$$F_x\left(F_{x,t}^{-1}(m) - \gamma_t\right) + tQ_t(\mathcal{A}_{x,t,\gamma_t}) \leq m \leq F_x\left(F_{x,t}^{-1}(m)\right) + tQ_t(\mathcal{A}_{x,t}), \tag{34}$$

because $P(\mathcal{A}_{x,t}) = F_x(F_{x,t}^{-1}(m))$ and $P(\mathcal{A}_{x,t,\gamma}) = F_x(F_{x,t}^{-1}(m) - \gamma_t)$. Rearranging and using the bound $\sup_{x \in \mathcal{X}} F_x(F_{x,t}^{-1}(m) - \gamma_t) = \sup_{x \in \mathcal{X}} F_x(F_{x,t}^{-1}(m)) + O(\gamma_t)$, which holds for all small enough t, we obtain that, for all such values of t,

$$\sup_{x \in \mathcal{X}} \left| m - F_x\left(F_{x,t}^{-1}(m)\right)\right| \leq t \sup_{x \in \mathcal{X}} \max\left\{|Q_t(\mathcal{A}_{x,t})|, |Q_t(\mathcal{A}_{x,t,\gamma_t})|\right\} + o(t) = tO\left(|Q(S)|\right),$$

since $|Q_t(S)| = |Q(S)| + o(1)$ as $t \to 0$. This establishes (32). Next, by the monotonicity of F_x for each $x \in \mathcal{X}$ and the facts – both implied by (24) – that $m = F_x(F_x^{-1}(m))$ and $\inf_{x \in \mathcal{X}} F_x'(F_x^{-1}(m))$ is bounded away from 0, (32) yields that

$$\lim_{t \to 0} \sup_{x \in \mathcal{X}} |F_x^{-1}(m) - F_{x,t}^{-1}(m)| = 0. \tag{35}$$

Combining the last display with the bound

$$\sup_{x \in \mathcal{X}} \left| F_x\left(F_x^{-1}(m)\right) - F_x\left(F_{x,t}^{-1}(m)\right)\right| = tO\left(|Q(S)|\right)$$

and assumption (24) again, we obtain that

$$\sup_{x \in \mathcal{X}} \left|F_x^{-1}(m) - F_{x,t}^{-1}(m)\right| \leq CtO(|Q(S)|) = O(t),$$

for all t small enough, where C is the constant in (24). This completes the proof of (33). ■

4.1 Significance of Topological Features

Fasy et al (2014) showed how to use the bootstrap to test the significance of a topological feature. They did this for distance functions and density estimators but the same idea works for DTM as we now explain. We assume in this section that the support of the distribution is contained in a compact set. The supremum norm refers to the supremum over this set.

Given a feature with birth and death time (u, v), we will say that the feature is significant if $|v - u| > 2c_\alpha/\sqrt{n}$ where c_α is defined by

$$\mathbb{P}(\sqrt{n}\|\widehat{\delta}(x) - \delta(x)\|_\infty > c_\alpha) = \alpha.$$

In particular, c_α can be estimated from the bootstrap as we showed in the previous section. Specifically, define $\widehat{c}_\alpha$ by

$$\mathbb{P}(\sqrt{n}\|\widehat{\delta}^*(x) - \widehat{\delta}(x)\|_\infty > \widehat{c}_\alpha | X_1, \ldots, X_n) = \alpha.$$

Then $\widehat{c}_\alpha$ is a consistent estimate of c_α.

To see why this makes sense, let $\mathcal{D}$ be the set of all persistence diagrams. Let $D \equiv D_\delta$ be the true diagram and let $\widehat{D} \equiv D_{\widehat{\delta}}$ be the estimated diagram. Let

$$\mathcal{C}_n = \left\{ E \in \mathcal{D} : \; W_\infty(\widehat{D}, E) \leq \frac{\widehat{c}_\alpha}{\sqrt{n}} \right\}.$$

Then

$$\mathbb{P}(D \in \mathcal{C}_n) = \mathbb{P}\left(W_\infty(D, \widehat{D}) \leq \frac{\widehat{c}_\alpha}{\sqrt{n}} \right) \geq \mathbb{P}(\sqrt{n}||\widehat{\delta}(x) - \delta(x)||_\infty \leq \widehat{c}_\alpha) \to 1 - \alpha$$

as $n \to \infty$. Now $|v - u| > 2\widehat{c}_\alpha/\sqrt{n}$ if and only if the feature cannot be matched to the diagonal for any diagram in $\mathcal{C}$. (Recall that the diagonal corresponds to features with zero lifetime.)

We can visualize the significant features by putting a band of size $2c_\alpha/\sqrt{n}$ around the diagonal of $\widehat{D}$. See Figure 6.

5. Theory for Kernels

In this section, we consider an alternative to the DTM, namely, kernel based methods. This includes the kernel distance and the kernel density estimator.

Phillips et al. (2014) suggest using the kernel distance for topological inference. Given a kernel $K(x, y)$, the kernel distance between two probability measures P and Q is

$$D_K(P, Q) = \sqrt{\iint K(x, y) dP(x) dP(y) + \iint K(x, y) dQ(x) dQ(y) - 2 \iint K(x, y) dP(x) dQ(y)}.$$

It can be shown that $D_K(P, Q) = \|\mu_P - \mu_Q\|$ for vectors μ_P and μ_Q in an appropriate reproducing kernel Hilbert space (RKHS). Such distances are popular in machine learning; see Sriperumbudur et al. (2009), for example.

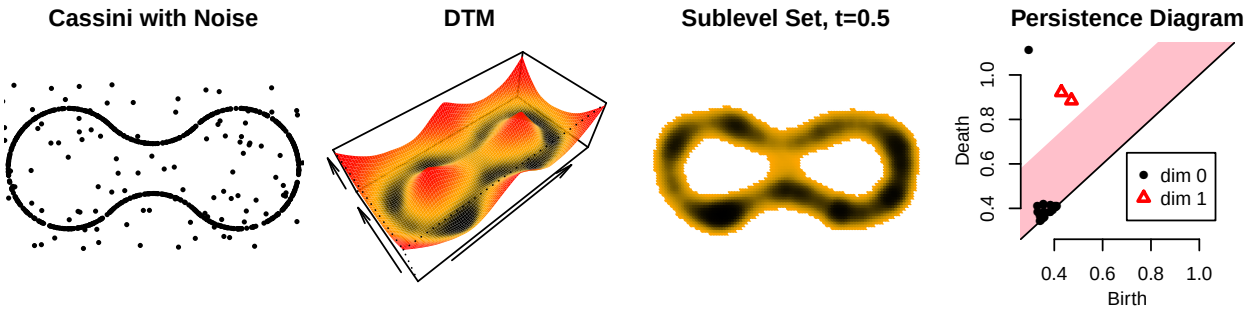

Figure 6: The left plot shows a sample from the Cassini curve together with a few outliers. The second plot is the empirical DTM. The third plot is one sub-level set of the DTM. The last plot is the persistence diagram. Points not in the shaded band are significant features. Thus, this method detects one significant connected component and two significant loops in the sublevel set filtration of the empirical DTM function.

Given a sample $X_1, \ldots, X_n \sim P$, let P_n be the probability measure that puts mass $1/n$ on each X_i. Let ϑ_x be the Dirac measure that puts mass one on x. Phillips et al. (2014) suggest using the discrete kernel distance

$$\widehat{D}_K(x) \equiv D_K(P_n, \vartheta_x) = \sqrt{\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} K(X_i, X_j) + K(x,x) - \frac{2}{n}\sum_{i=1}^{n} K(x, X_i)} \tag{36}$$

for topological inference. This is an estimate of the population quantity

$$D_K(x) \equiv D_K(P, \vartheta_x) = \sqrt{\iint K(z,y)dP(z)dP(y) + K(x,x) - 2\int K(x,y)dP(y)}.$$

The most common choice of kernel is the Gaussian kernel $K(x,y) \equiv K_h(x,y) = \exp\left(-\frac{\|x-y\|^2}{2h^2}\right)$, which has one tuning parameter h. We recall that, in topological inference, we generally do not let h tend to zero. The reason is that topological features can be detected with $h > 0$ and keeping h bounded way from 0 reduces the variance of the estimator. See the related discussion in Section 4.4 of Fasy et al. (2014b).

Recall that the kernel density estimator is defined by

$$\widehat{p}_h(x) = \frac{1}{n(\sqrt{2\pi}h)^d}\sum_i K(x, X_i).$$

Let $p_h(x) = \mathbb{E}[\widehat{p}_h(x)]$. We see that

$$\begin{aligned}
\widehat{D}_K^2(x) &= h^d \left(\frac{(\sqrt{2\pi})^d}{n} \sum_i \widehat{p}_h(X_i) + h^{-d} K(0,0) - 2(\sqrt{2\pi})^d \widehat{p}_h(x) \right) \\
&= h^d \left(\frac{(\sqrt{2\pi})^d}{n} \sum_i [\widehat{p}_h(X_i) - p_h(X_i)] + \frac{(\sqrt{2\pi})^d}{n} \sum_i p_h(X_i) + h^{-d} K(0,0) - 2(\sqrt{2\pi})^d \widehat{p}_h(x) \right) \\
&= (\sqrt{2\pi})^d h^d (c + o_P(1)) + O_P\left(\sqrt{\frac{\log n}{n}} \right) + K(0,0) - 2(\sqrt{2\pi} h)^d \widehat{p}_h(x).
\end{aligned}$$

Here, we used the fact that $n^{-1} \sum_{i=1}^n p_h(X_i) = c + o_P(1)$ and $||\widehat{p}_h - p_h||_\infty = O_P(\sqrt{\log n / n})$ where $c = \int p_h p$.

We see that up to small order terms, the sublevel sets of $D_K(x)$ are a rescaled version of the super-level sets of the kernel density estimator. Hence, the kernel distance approach and the density estimator approach are essentially the same, up to a rescaling. However, D_K^2 has some nice properties; see Phillips et al. (2014).

The limiting properties of $\widehat{D}_K^2(x)$ follow immediately from well-known properties of kernel density estimators. In fact, the conditions needed for $\widehat{D}_K^2$ are weaker than for the DTM.

Theorem 16 (Limiting Behavior of Kernel Distance) *We have that*

$$\sqrt{n}(\widehat{D}_K^2 - D_K^2) \rightsquigarrow \mathbb{B},$$

where $\mathbb{B}$ is a Brownian bridge. The bootstrap version converges to the same limit, conditionally almost surely.

The proof of the above theorem is based on the aforementioned equivalence of D_K to the rescaled density function and the well known fact that $\sqrt{n}(\widehat{p}_h(x) - p_h(x))$ converges to a Brownian bridge. This theorem justifies using the bootstrap to construct L_∞ bands for $p_h = \mathbb{E}(\widehat{p}_h)$ or D_K.

As we mentioned before, for topological inference, we keep the bandwidth h fixed. Thus, it is important to keep in mind that we view $\widehat{p}_h$ as an estimate of $p_h(x) = \mathbb{E}[\widehat{p}_h(x)] = \int K_h(x,u) dP(u)$.

6. The Bottleneck Bootstrap

More precise inferences can be obtained by directly bootstrapping the persistence diagram. Let $X_1^*, \ldots, X_n^*$ be as before a sample from the empirical measure P_n and let $\widehat{D}^*$ be the (random) persistence diagram defined on this point cloud. Define $\widehat{t}_\alpha$ by

$$\mathbb{P}(\sqrt{n} W_\infty(\widehat{D}^*, \widehat{D}) > \widehat{t}_\alpha \,|\, X_1, \ldots, X_n) = \alpha. \tag{37}$$

The quantile $\widehat{t}_\alpha$ can be estimated by Monte Carlo. We then use a band of size $2\widehat{t}_\alpha$ on the diagram D.

In the following, we show that $\widehat{t}_\alpha$ consistently estimates the population value t_α defined by

$$\mathbb{P}(\sqrt{n}W_\infty(\widehat{D}, D) > t_\alpha) = \alpha. \tag{38}$$

The reason why the bottleneck bootstrap can lead to more precise inferences than the functional bootstrap from the previous section is that the functional bootstrap uses the fact that $W_\infty(\widehat{D}, D) \leq ||\widehat{\delta} - \delta||_\infty$ and finds an upper bound for $||\widehat{\delta} - \delta||_\infty$. But in many cases the inequality is not sharp so the confidence set can be very conservative. Moreover, we can obtain different critical values for different dimensions (connected components, loops, voids, ...) and so the inferences are tuned to the specific features we are estimating. See Figure 7.

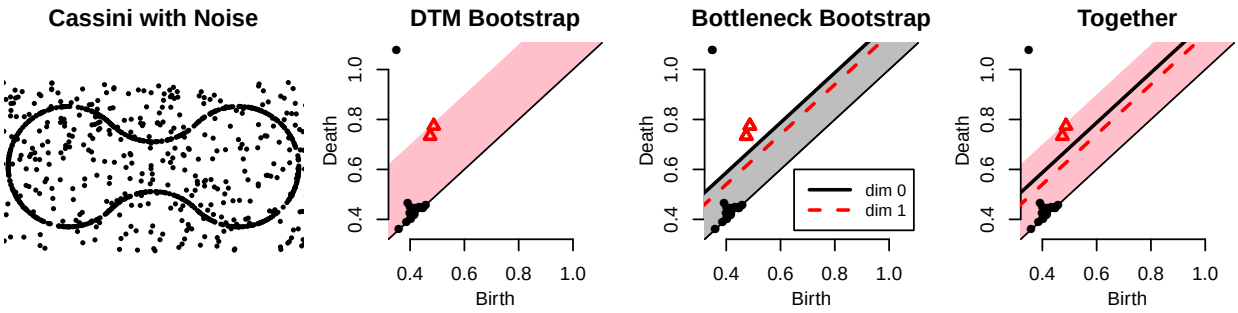

Figure 7: The left plot shows a sample from the Cassini curve together with a few outliers. The second plot shows the DTM persistence diagram with a 95% confidence band constructed using the method of Section 4. The third plot shows the same persistence diagram with two 95% confidence bands constructed using the bottleneck bootstrap with zero-dimensional features and one-dimensional features. The fourth plot shows the three confidence bands at the same time. In Section 8, we use this compact form to show multiple confidence bands.

Although the bottleneck bootstrap can be used with either the DTM or the KDE, we shall only prove its validity for the KDE. First, we need the following result. For any function p, let $g = \nabla p$ denote its gradient and let $H = \nabla^2 p$ denotes its Hessian. We say that x is a critical point if $g(x) = (0, \ldots, 0)^T$. We then call $p(x)$ a critical value. A function is Morse if the Hessian is non-degenerate at each critical point. The Morse index of a critical point x is the number of negative eigenvalues of $H(x)$.

Lemma 17 (Stability of Critical Points) *Let p be a density with compact support S. Assume that S is a d-dimensional compact submanifold of $\mathbb{R}^d$ with boundary. Assume p is a Morse function with finitely many, distinct, critical values with corresponding critical points $c_1, \ldots, c_k$. Also assume that p is of class $\mathcal{C}^2$ on the interior of S, continuous and differentiable with non vanishing gradient on the boundary of S. Then, there exist $\epsilon_0 > 0$ and $c > 0$ such that for all $0 < \epsilon < \epsilon_0$, there exists $\eta \geq c\epsilon$ such that, for any density q with*

support S satisfying

$$\sup_x |p(x) - q(x)| < \eta, \qquad \sup_x |\nabla p(x) - \nabla q(x)| < \eta, \qquad \sup_x |\nabla^2 p(x) - \nabla^2 q(x)| < \eta,$$

q is a Morse function with exactly k critical points $c'_1, \ldots, c'_k$ say, and, after a suitable re-labeling of indices,

$$\max_j ||c_j - c'_j|| \leq \epsilon.$$

Moreover, c_j and c'_j have the same Morse index.

Proof This lemma is a consequence of classical stability properties of Morse functions. First, from Theorem 5.31, p.140 in Banyaga and Hurtubise (2004) and Proposition II.2.2, p.79 in Golubitsky and Guillemin (1986), there exists $\epsilon_1 > 0$ such that if q is at distance less than ϵ_1 in the $\mathcal{C}^2$ topology (i.e. such that the sup-norm of $p - q$ and its first and second derivatives are bounded by ϵ_1) then q is a Morse function. Moreover, there exist two diffeomorphisms $h : \mathbb{R} \to \mathbb{R}$ and $\phi : S \to S$ such that $q = h \circ p \circ \phi$. As the notion of critical point and of index are invariant by diffeomorphism, p and q have the same number of critical points with same index. More precisely, the critical points of q are the points $c'_i = \phi^{-1}(c_i)$.

Now let $\epsilon > 0$ be small enough such that $2\epsilon < \min_{i \neq j} \|c_i - c_j\|$, and for any $i \neq j$, $p(B(c_i, \epsilon)) \cap p(B(c_j, \epsilon)) = \emptyset$. Then $\eta_1 = \eta_1(\epsilon) = \min_{i \neq j} d(p(B(c_i, \epsilon)), p(B(c_j, \epsilon)))$ where $d(A, B) = \min_{a \in A, b \in B} |a - b|$ and $\eta_2 = \eta_2(\epsilon) = \inf\{\|\nabla p(x)\| : x \in S \setminus \cup_{i=1}^k B(c_i, \epsilon)\}$ are both positive. If q satisfies the assumptions of the lemma for any $0 < \eta \leq \min(\eta_1, \eta_2)$, then the critical values of q have to be in $\cup_i p(B(c_i, \epsilon))$ and the critical points c'_i have to be in $\cup_i B(c_i, \epsilon)$.

More precisely, notice that since p is a Morse function, for ϵ small enough, $\eta_2 = O(\epsilon)$, and, for any $i \in \{1, \cdots, k\}$, the Taylor series of ∇p about c_j yields

$$\nabla p(x) = H_i(x - c_i) + \|x - c_i\| r(x - c_i),$$

where $r(z) \to 0$ as $\|z\| \to 0$ and H_i is the Hessian of p at c_i. Let $\lambda_{\min}$ be the smallest absolute eigenvalue of the Hessians at all the critical points. Since p is a Morse function, the matrix H_i is full rank and $\lambda_{\min}$ is positive. As a consequence, for all $x \in S \setminus \cup_{i=1}^k B(c_i, \epsilon)$ and ϵ small enough, $\|\nabla p(x)\| \geq \frac{\lambda_{\min}}{2}\epsilon$. Since η_1 is a non-increasing function of ϵ, we have that, for ϵ small enough, $\eta = \eta_2 \geq \frac{\lambda_{\min}}{2}\epsilon$.

To conclude the proof of the lemma, we need to prove that each ball $B(c_i, \epsilon)$ contains exactly one critical point of q. Indeed, for $t \in [0, 1]$, the functions $q_t(x) = p(x) + t(q(x) - p(x))$ are Morse functions satisfying the same properties as q. Now, since each c_i is a non-degenerate point of p, it follows from the continuity of the critical points (see, e.g. Prop. 4.6.1 in Demazure (2013)) that, restricting ϵ if necessary, there exist smooth functions $c_i : [0, 1] \to S$, $c_i(0) = c_i, c_i(1) = c'_i$ such that $c_i(t)$ is the unique critical point of q_t in $B(c_i, \epsilon)$. Moreover, since all the q_t are Morse functions and since the Hessian of q_t at $c_i(t)$ is a continuous function of t, then for any $t \in [0, 1]$, $c_i(t)$ is a non-degenerate critical point of q_t with same index as c_i. ■

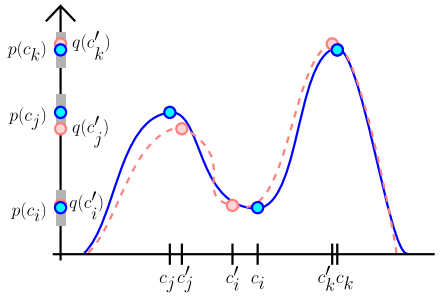

Figure 8: This figure illustrates the assumptions of Lemma 18. The functions p and q are shown in solid blue and dashed pink, respectively. The grey regions on the y-axis represent the sets $p(c) \pm b$ for critical points c of p.

Consider now two smooth functions such that the critical points are close, as illustrated in Figure 8. Next we show that, in this circumstance, the bottleneck distance takes a simple form.

Lemma 18 (Critical Distances) *Let p and q be two Morse functions as in Lemma 17, with finitely many critical points $C = \{c_1, \ldots, c_k\}$ and $C' = \{c'_1, \ldots, c'_k\}$ respectively. Let D_p and D_q be the persistence diagrams from the upper level set (i.e. super level sets) filtrations of p and q respectively and let $a = \min_{i \neq j} |p(c_i) - p(c_j)|$ and $b = \max_j |p(c_j) - q(c'_j)|$. If $b \leq a/2 - ||p - q||_\infty$ and $a/2 > 2||p - q||_\infty$, then $W_\infty(D_p, D_q) = b$.*

Proof The topology of the upper level sets of the Morse functions p and q only changes at critical values (Theorem 3.1 in Milnor (1963)). As a consequence the non-diagonal points of D_p (resp. D_q) have their coordinates among the set $\{p(c_1), \ldots, p(c_k)\}$ (resp. $\{q(c'_1), \ldots, p(c'_k)\}$) and each $p(c_i)$ is the coordinate of exactly one point in D_p. Moreover, the pairwise distances between the points of D_p are lower bounded by a and all non-diagonal points of D_p are at distance at least a from the diagonal. From the persistence stability theorem Cohen-Steiner et al. (2005); Chazal et al. (2012), $W_\infty(D_p, D_q) \leq ||p - q||_\infty$. Since $a > 4||p - q||_\infty$ and $a \geq 2b + 2||p - q||_\infty$, the (unique) optimal matching realizing the bottleneck distance $W_\infty(D_p, D_q)$ is such that if $(p(c_i), p(c_j)) \in D_p$ then it is matched to the point $(q(c'_i), q(c'_j))$ which thus have to be in D_q. It follows that $W_\infty(D_p, D_q) = b$. ■

Now we establish the limiting distribution of $\sqrt{n} W_\infty(\widehat{D}, D)$.

Theorem 19 (Limiting Distribution) *Let $p_h(x) = \mathbb{E}[\widehat{p}_h(x)]$, where $\widehat{p}_h(x)$ is the Kernel Density Estimator evaluated in x. Assume that p_h is a Morse function with two uniformly bounded continuous derivatives and finitely many critical points $c = \{c_1, \ldots, c_k\}$. Let D be the persistence diagram of the upper level sets of p_h and let $\widehat{D}$ be the diagram of upper level sets of $\widehat{p}_h$. Then*

$$\sqrt{n} W_\infty(\widehat{D}, D) \rightsquigarrow ||Z||_\infty$$

where $Z = (Z_1, \ldots, Z_k) \sim N(0, \Sigma)$ *and*

$$\Sigma_{jk} = \int K_h(c_j, u) K_h(c_k, u) dP(u) - \int K_h(c_j, u) dP(u) \int K_h(c_k, u) dP(u).$$

Proof Let $\widehat{c} = \{\widehat{c}_1, \widehat{c}_2, \ldots\}$ be the set of critical points of $\widehat{p}_h$. Let g and H be the gradient and Hessian of p_h. Let $\widehat{g}$ and $\widehat{H}$ be the gradient and Hessian of $\widehat{p}_h$. By a standard concentration of measure argument (and recalling that the support is compact), for any $\eta > 0$ there is an event $A_{n,\eta}$ such that, on $A_{n,\eta}$,

$$\sup_x ||\widehat{p}_h^{(i)}(x) - p_h^{(i)}(x)|| < \eta \tag{39}$$

for $i = 0, 1, 2$, and $\mathbb{P}(A_{n,\eta}^c) \leq e^{-nc\eta^2}$. This is proved for $i = 0$ in Rao (1983), Giné and Guillou (2002), Yukich (1985), and the same proof gives the results for $i = 1, 2$. It follows that $\sup_x ||g(x) - \widehat{g}(x)|| = O_P(1/\sqrt{n})$ and $\sup_x ||H(x) - \widehat{H}(x)||_{\max} = O_P(1/\sqrt{n})$.

For η smaller than a fixed value η_0, we can apply Lemma 17, we get that on $A_{n,\eta}$, $\widehat{c}$ and c have the same number of elements and can be indexed so that

$$\max_{j=1,\ldots,k} \|\widehat{c}_j - c_j\| \quad \leq \quad \frac{\eta}{C}$$

where C is the same constant is in Lemma 17. We then take $\eta_n := \sqrt{\frac{\log n}{n}}$ and we consider the events $A_n := A_{n,\eta_n}$. Then, for n large enough, on A_n we get

$$\max_{j=1,\ldots,k} \|\widehat{c}_j - c_j\| = O\left(\sqrt{\frac{\log n}{n}}\right)$$

whereas $P(A_n^c) = o(1)$. In the following, we thus can restrict to A_n.

The critical values of p_h are $v = (v_1 \equiv p_h(c_1), \ldots, v_k \equiv p_h(c_k))$ and the critical values of $\widehat{p}_h$ are $\widehat{v} = (\widehat{v}_1 \equiv \widehat{p}_h(\widehat{c}_1), \ldots, \widehat{v}_k \equiv \widehat{p}_h(\widehat{c}_k))$. Now we use Lemma 18 to conclude that $W_\infty(\widehat{D}, D) = \max_j \|\widehat{v}_j - v_j\|_\infty$ for n large enough. Hence,

$$W_\infty(\widehat{D}, D) = \max_{j=1,\ldots,k} |\widehat{p}_h(\widehat{c}_j) - p_h(c_j)|.$$

Then, using a Taylor expansion, for each j,

$$\widehat{p}_h(\widehat{c}_j) = \widehat{p}_h(c_j) + (\widehat{c}_j - c_j)^T \widehat{g}(c_j) + O(||\widehat{c}_j - c_j||^2).$$

Since $g(c_j) = (0, \ldots, 0)$ we can write the last equation as

$$\widehat{p}_h(\widehat{c}_j) = \widehat{p}_h(c_j) + (\widehat{c}_j - c_j)^T (\widehat{g}(c_j) - g(c_j)) + O(||\widehat{c}_j - c_j||^2).$$

So,

$$\begin{aligned}
\sqrt{n}(\widehat{v}_j - v_j) &= \sqrt{n}(\widehat{p}_h(\widehat{c}_j) - p_h(c_j)) \\
&= \sqrt{n}(\widehat{p}_h(c_j) - p_h(c_j)) + \sqrt{n}(\widehat{c}_j - c_j)^T (\widehat{g}(c_j) - g(c_j)) + O(||\widehat{c}_j - c_j||^2) \\
&= \sqrt{n}(\widehat{p}_h(c_j) - p_h(c_j)) + \sqrt{n}(\widehat{c}_j - c_j)^T (\widehat{g}(c_j) - g(c_j)) + o(1/\sqrt{n}).
\end{aligned}$$

For the second term, note that $\sqrt{n}(\widehat{c}_j - c_j) = O(\log n)$ and $(\widehat{g}(c_j) - g(c_j)) = O_P(1/\sqrt{n})$. So

$$\sqrt{n}(\widehat{v}_j - v_j) = \sqrt{n}(\widehat{p}_h(c_j) - p_h(c_j)) + o_P(1).$$

Therefore,

$$\sqrt{n}W_\infty(\widehat{D}, D) = \sqrt{n}\max_j |\widehat{v}_j - v_j| = \max_j |\sqrt{n}(\widehat{p}_h(c_j) - p_h(c_j))| + o_P(1).$$

By the multivariate Berry-Esseen theorem (Bentkus, 2003),

$$\sup_A |\mathbb{P}(\sqrt{n}(\widehat{p}_h(c) - p_h(c)) \in A) - \mathbb{P}(Z \in A)| \leq \frac{C_1}{\sqrt{n}}$$

where the supremum is over all convex sets $A \in \mathbb{R}^k$, $C_1 > 0$ depends on k and the third moment of $h^{-d}K(x - X/h)$ (which is finite since h is fixed and the support is compact), $Z = (Z_1, \ldots, Z_k) \sim N(0, \Sigma)$ and

$$\Sigma_{jk} = \int K_h(c_j, u)K_h(c_k, u)dP(u) - \int K_h(c_j, u)dP(u) \int K_h(c_k, u)dP(u).$$

Hence,

$$\sup_t \left| \mathbb{P}\Big(\max_j |\sqrt{n}(\widehat{p}_h(c_j) - p_h(c_j))| \leq t\Big) - \mathbb{P}(||Z||_\infty \leq t) \right| \leq \frac{C_1}{\sqrt{n}}.$$

By Lemma 18, $W_\infty(\widehat{D}, D) = \max_j |\widehat{v}_j - v_j|$. The result follows. ■

Let

$$\widehat{F}_n(t) = \mathbb{P}(\sqrt{n}W_\infty(\widehat{D}, D) \leq t).$$

Let $X_1^*, \ldots, X_n^* \sim P_n$ where P_n is the empirical distribution. Let $\widehat{D}^*$ be the diagram from $\widehat{p}_h^*$ and let

$$\widehat{F}_n(t) = \mathbb{P}\Big(\sqrt{n}W_\infty(\widehat{D}^*, \widehat{D}) \leq t \;\Big|\; X_1, \ldots, X_n\Big)$$

be the bootstrap approximation to F_n.

Next we show that the bootstrap quantity $F_n(t)$ converges to the same limit as $F_n(t)$.

Corollary 20 *Assume the same conditions as the last theorem. Then,*

$$\sup_t |\widehat{F}_n(t) - F_n(t)| \xrightarrow{P} 0.$$

Proof The proof is essentially the same as the proof of Theorem 19 except that $\widehat{p}_h$ replaces p_h and $\widehat{p}_h^*$ replaces $\widehat{p}_h$. Using the same notations as in the proof of Theorem 19, we note that on the set A_n, for n larger than a fixed value n_0, the function $\widehat{p}_h$ is a Morse function with two uniformly bounded continuous derivatives and finitely many critical points $\widehat{c} = \{\widehat{c}_1, \ldots, \widehat{c}_k\}$. We can restrict the analysis to the sequence of events A_n since $P(A_n)$ tends

to zero. Assuming that A_n is satisfied, using the same argument as in Theorem 19, we get that:

$$\sup_t \left| \mathbb{P}\Big(\max_j |\sqrt{n}(\widehat{p}_h^*(\widehat{c}_j) - \widehat{p}_h(\widehat{c}_j))| \le t \;\Big|\; X_1, \ldots, X_n \Big) - \mathbb{P}(||\widetilde{Z}||_\infty \le t) \right| \le \frac{C_2^*}{\sqrt{n}}$$

where $\widetilde{Z} \sim N(0, \widehat{\Sigma})$ with

$$\widehat{\Sigma}_{jk} = \frac{1}{n}\sum_i K_h(\widehat{c}_j, X_i) K_h(\widehat{c}_k, X_i) - \frac{1}{n}\sum_i K_h(\widehat{c}_j, X_i)\, \frac{1}{n}\sum K_h(\widehat{c}_k, X_i)$$

and C_2^* depends on the empirical third moments of $h^{-d}K((x-X^*)/h)$. There exists an upper bound C_2 on C_2^* that only depends on K and P. Since $\max_{j,k} |\widehat{\Sigma}_{j,k} - \Sigma_{j,k}| = O_P(\log n/\sqrt{n})$ and $\max_j \|\widehat{c}_j - c_j\| = O_P(\log n/\sqrt{n})$, we conclude that

$$\sup_t |\mathbb{P}(||\widetilde{Z}||_\infty \le t) - \mathbb{P}(||Z||_\infty \le t)| = O_P\left(\frac{\log n}{\sqrt{n}}\right).$$

Then

$$\begin{aligned}\sup_t |\widehat{F}_n(t) - F_n(t)| &\le \sup_t |\widehat{F}_n(t) - \mathbb{P}(||\widetilde{Z}||_\infty \le t)| \\ &\quad + \sup_t |\mathbb{P}(||\widetilde{Z}||_\infty \le t) - \mathbb{P}(||Z||_\infty \le t)| + \sup_t |F_n(t) - \mathbb{P}(||Z||_\infty \le t)| = O_P\left(\frac{\log n}{\sqrt{n}}\right).\end{aligned}$$

The result follows. ∎

7. Extensions

In this section, we discuss how to deal with three issues that can arise: choosing the parameters, correcting for boundary bias, and dealing with noisy data.

7.1 A Method for Choosing the Smoothing Parameter

An unsolved problem in topological inference is how to choose the smoothing parameter m (or h). Guibas et al. (2013) suggested tracking the evolution of the persistence of the homological features as the tuning parameter varies. Here we make this method more formal, by selecting the parameter that maximizes the total amount of significant persistence.

Let $\ell_1(m), \ell_2(m), \ldots,$ be the lifetimes of the features at scale m. Let $c_\alpha(m)/\sqrt{n}$ be the significance cutoff at scale m. We define two quantities that measure the amount of significant information using parameter m:

$$N(m) = \#\left\{ i : \; \ell(i) > \frac{c_\alpha(m)}{\sqrt{n}} \right\}, \quad S(m) = \sum_i \left[\ell_i - \frac{c_\alpha(m)}{\sqrt{n}} \right]_+ .$$

These measures are small when m is small since $c_\alpha(m)$ is large. On the other hand, they are small when m is large since then all the features are smoothed out. Thus we have a kind of topological bias-variance trade-off. We choose m to maximize $N(m)$ or $S(m)$. The same idea can be applied to the kernel distance and kernel density estimator. See the example in Figure 9.

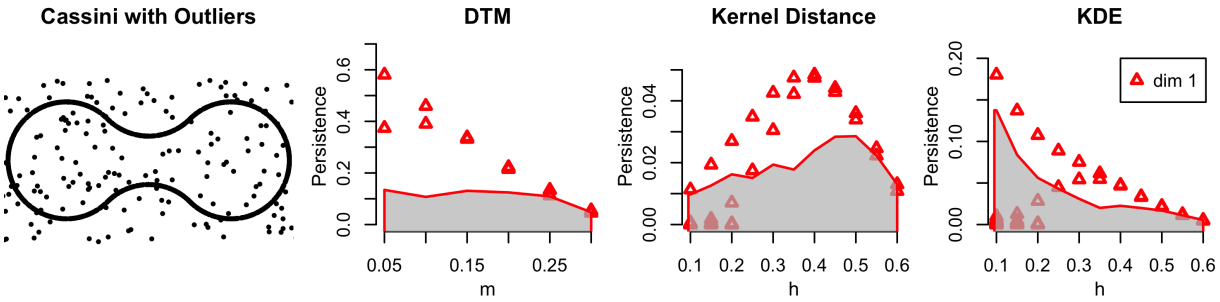

Figure 9: Max Persistence Method with Bottleneck Bootstrap Bands for 1-dimensional features. DTM: $\text{argmax}_m N(m) = \{0.05, 0.10, 0.15, 0.20\}$, $\text{argmax}_m S(m) = 0.05$; Kernel Distance: $\text{argmax}_h N(h) = \{0.25, 0.30, 0.35, 0.40, 0.45, 0.50\}$, $\text{argmax}_h S(h) = 0.35$; KDE: $\text{argmax}_h N(h) = \{0.25, 0.30, 0.35, 0.40, 0.45, 0.50\}$, $\text{argmax}_h S(h) = 0.3$ The plots show how to choose the smoothing parameters to maximize the number of significant features. The red triangles are the lifetimes of the features versus the tuning parameter.The red line is the significance cutoff.

7.2 Boundary Bias

It is well known that kernel density estimators suffer from boundary bias. For topological inference, this bias manifests itself in a particular form and the same problem affects the DTM. Consider Figure 10. Because of the bounding box, many of the loops are incomplete. The result is that, using either the DTM or the KDE we will miss many of the loops.

There is a large literature on reducing boundary bias in the kernel density estimation literature. Perhaps the simplest approach is to reflect the data around the boundaries (see for example Schuster (1958)). But there is a simpler fix for topological inference: we merely need to close the loops at the boundary. This can be done by adding points uniformly around the boundary.

7.3 Two Methods for Improving Performance

We can improve the performance of all the methods if we can mitigate the outliers and noise. Here we suggest two methods to do this. We focus on the kernel density estimator.

First, a simple method to reduce the number of outliers is to truncate the density, that is, we eliminate $\{X_i : \widehat{p}(X_i) < t\}$ for some threshold t. Then we re-estimate the density.

Secondly, we sharpen the data as described in Choi and Hall (1999) and Hall and Minnotte (2002). The idea of sharpening is to move each data point X_i slightly in the direction of the gradient $\nabla\widehat{p}(X_i)$ and then re-estimate the density. The authors show that this reduces the bias at peaks in the density which should make it easier to find topological features. It can be seen that the sharpening method amounts to running one or more steps if the mean-shift algorithm. This is a gradient ascent which is intended to find modes of the density estimator. Given a point x, we move x to

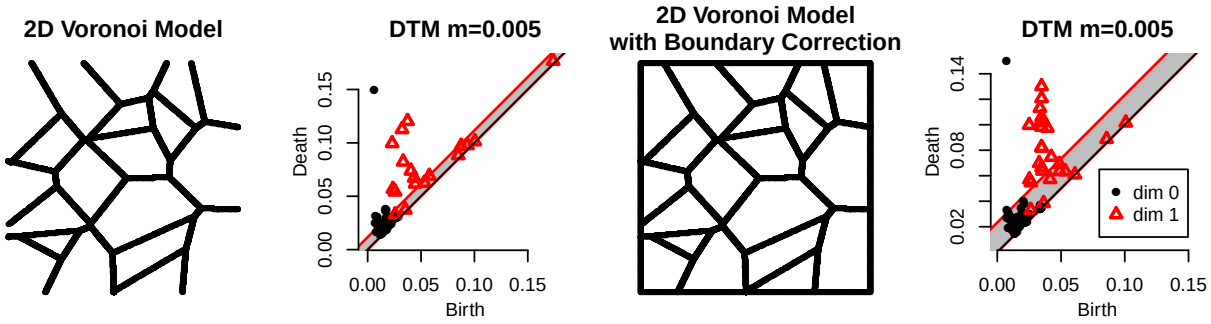

Figure 10: First: 10,000 points sampled from a 2D Voronoi model with 20 nuclei. Second: the corresponding persistence diagram of sublevel sets of the distance to measure function. Note that only 9 loops are detected as significant. Third: 2,000 points have been added on the boundary of the square delimiting the Voronoi model. Fourth: now the corresponding persistence diagram shows 16 significant loops.

$$\frac{\sum_i X_i K_h(x, X_i)}{\sum_i K_h(x, X_i)},$$

which is simply the local average centered at x. For data sharpening, we do one (or a few) iterations of this to each data point X_i. Then the density is re-estimated.
In fact, we could also use the subspace constrained mean shift algorithm (SCMS) which moves points towards ridges of the density; see Ozertem and Erdogmus (2011).
Figure 11 shows these methods applied to a simple example.

8. Examples

Example 1 (Noisy Grid) *The data in Figure 12 are 10,000 data points on a 2D grid. We add Gaussian noise plus 1,000 outliers and compute the persistence diagrams of Kernel Density Estimator, Kernel distance, and Distance to Measure. The pink bands show 95% confidence sets obtained by bootstrapping the corresponding functions. The black lines show 95% confidence bands obtained with the bottleneck bootstrap for dimension 0, while the red lines show 95% confidence bands obtained with the bottleneck bootstrap for dimension 1. The Distance to Measure, which is less sensitive to the density of the points, correctly captures the topology of the data. The Kernel Distance and KDE find some extra significant connected component, corresponding to high density regions at the intersection of the grid.*

Example 2 (Soccer) *Figure 13 shows the field position of two soccer players. The data come from body-sensor traces collected during a professional soccer game in late 2013 at the Alfheim Stadium in Tromso, Norway. The data are sampled at 20 Hz. See Pettersen et al. (2014). Although the data is a function observed over time, we treat it as a point cloud. Points on the boundary of the field have been added to avoid boundary bias. The DTM*

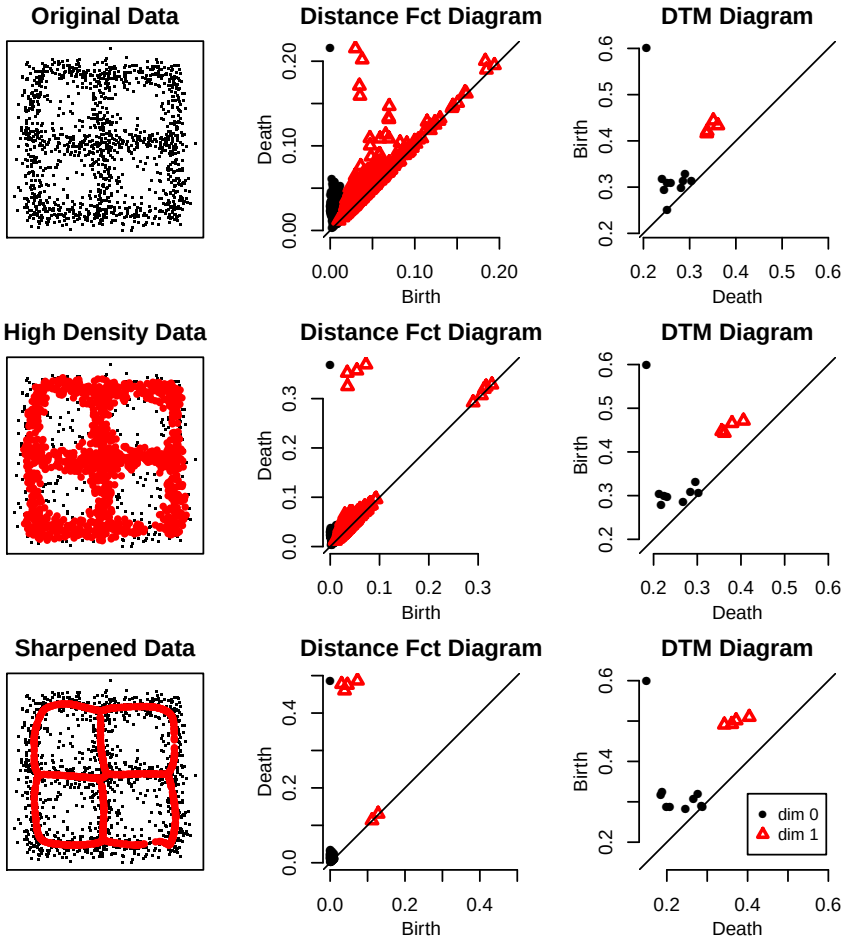

Figure 11: Top: 1,300 points sampled along a 2×2 grid with Gaussian noise; the diagram of the distance function shows many loops due to noise. Middle: the red points are the high density data (density > 0.15); the corresponding diagram of the distance function correctly captures the 4 loops, plus a few features with short lifetime. Bottom: the red points represent the sharpened high density data; now most of the noise in the corresponding diagram is eliminated. Note that the diagram of the distance to measure function does a good job with the original data. The bottom left plot shows a slight improvement, in the sense that the persistence of the 4 loops has increased.

captures the difference between the two players: the defender leaves one big portion of the filed uncovered (1 significant loop in the persistence diagram), while the midfielder does not cover the 4 corners (4 significant loops). Nonetheless, the Kernel distance, which is more sensible to the density of these points, fails to detect significant topological features.

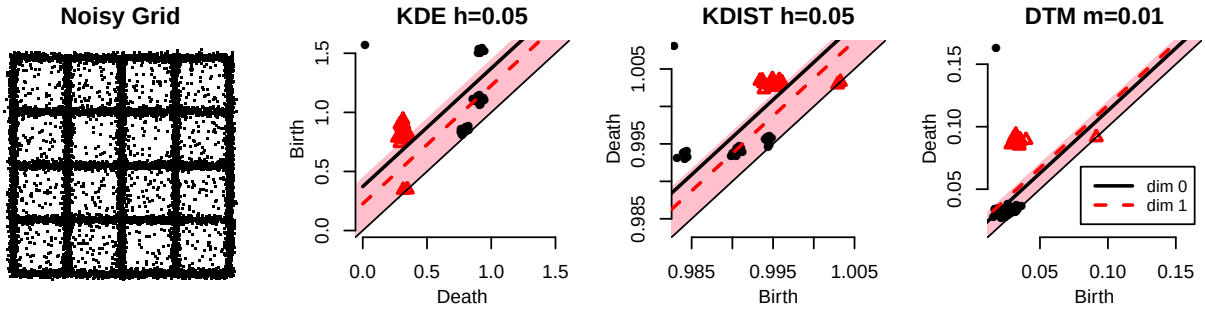

Figure 12: 10,000 data points on a 2D grid and the corresponding persistence diagrams of Kernel Density Estimator, Kernel distance, and Distance to Measure. For more details see Example 1.

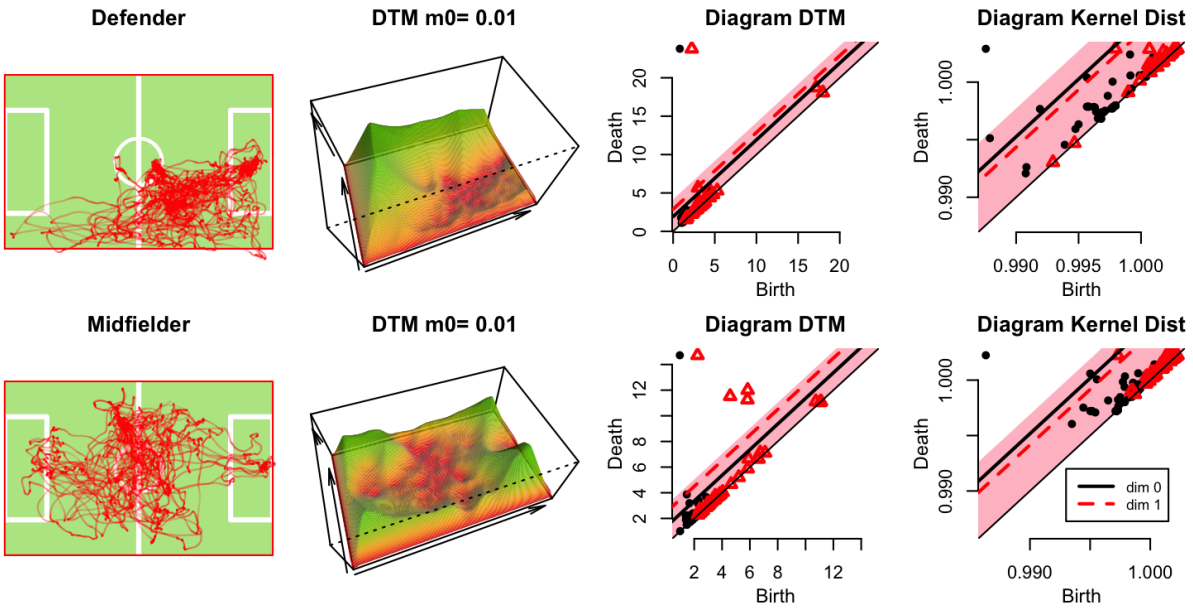

Figure 13: Top: data for a defender. We show the DTM, the digram for the DTM and the digram for the kernel distance. Bottom: same but for a midfielder. The midfielder data has more loops.

Example 3 (Voronoi Models) *Given k points (nuclei) $\{z_1, \ldots, z_k\} \subset \mathbb{R}^3$, let the Voronoi region R_k be $R_k = \left\{x \in \mathbb{R}^3 : \|x - z_k\| \leq \|x - x_j\| \text{ for all } j \neq k\right\}$. The Voronoi regions $R_1, \ldots, R_k$ partition the space, forming what is known as the Voronoi diagram. A face is formed by the intersection of 2 adjacent Voronoi regions; a line is formed at the intersection of two faces and a node is formed at the intersection of two or more lines.*

We will sample points around the the nodes, lines and faces that are formed at the intersection of the Voronoi regions. A Voronoi wall model is a sampling scheme that returns points within or around the Voronoi faces. Similarly, by sampling points exclusively around the lines or exclusively around the nodes, we can construct Voronoi filament models and Voronoi cluster models.

These models were introduced by Icke and van de Weygaert (1991) to mimic key features of cosmological data; see also van de Weygaert et al. (2011).

In this example we generate data from filament models and wall models using the basic definition of Voronoi diagram, computed on a fine grid in $[0, 50]^3$*. We also add random Gaussian noise. See Figure 14: the first two rows show 100K particles concentrated around the filaments of 8 and 64 Voronoi cells, respectively. The last two rows show 100K particles concentrated around the walls of 8 and 64 Voronoi cells. 60K points on the boundary of the boxes have been added to mitigate boundary bias. For each model we present the persistent diagrams of the distance function, distance to measure and kernel density estimator. We chose the smoothing parameters by maximizing the quantity* $S(\cdot)$*, defined in Section 7.1.*

The diagrams illustrate the evolution of the filtrations for the three different functions: at first, the connected components appear (black points in the diagrams); then they merge forming loops (red triangle), that eventually evolve into 3D voids (blue squares).

The persistence diagrams of the three functions allow us to distinguish the different models (see Figure 1 for a less trivial example) and the confidence bands, generated using the bootstrap method of Section 4.1, allow us to separate the topological signal from the topological noise. In general, the DTM performs better than the KDE, which is more affected by the high density of points around the nodes and filaments. For instance, this is very clear in the third row of Figure 14. The DTM diagram correctly captures the topology of the Voronoi wall model with 8 nuclei: one connected component and 8 voids are significant, while the remaining homological features fall into the band and are classified as noise.

9. Discussion

In this paper, we showed how the DTM and KDE can be used for robust topological inference. Further, we showed how to use the bootstrap to identify topological features that are distinguishable from noise. We conclude by discussing two issues: comparing DTM and KDE, and using persistent homology versus selecting a single level set.

9.1 Comparison of DTM and Kernel Distance

The DTM and the KDE have the same broad aim: to provide a means for extracting topological features from data. However, these two methods are really focused on different goals. Consider again the model $P = \pi R + (1 - \pi)(Q \star \Phi_\sigma)$ and let S be the support of Q. As before, we assume that S is a "small set" meaning that either it has dimension $k < d$ or that it is full dimensional but has small Lebesgue measure. When π and σ are small, the persistent homology of the upper level sets of the density p will be dominated by features corresponding to the homology of S. In other words, we are using the persistent homology of $\{p > t\}$ to learn about the homology of S. In contrast, the DTM is aimed at estimating the persistent homology of S. Both are useful, but they have slightly different goals.

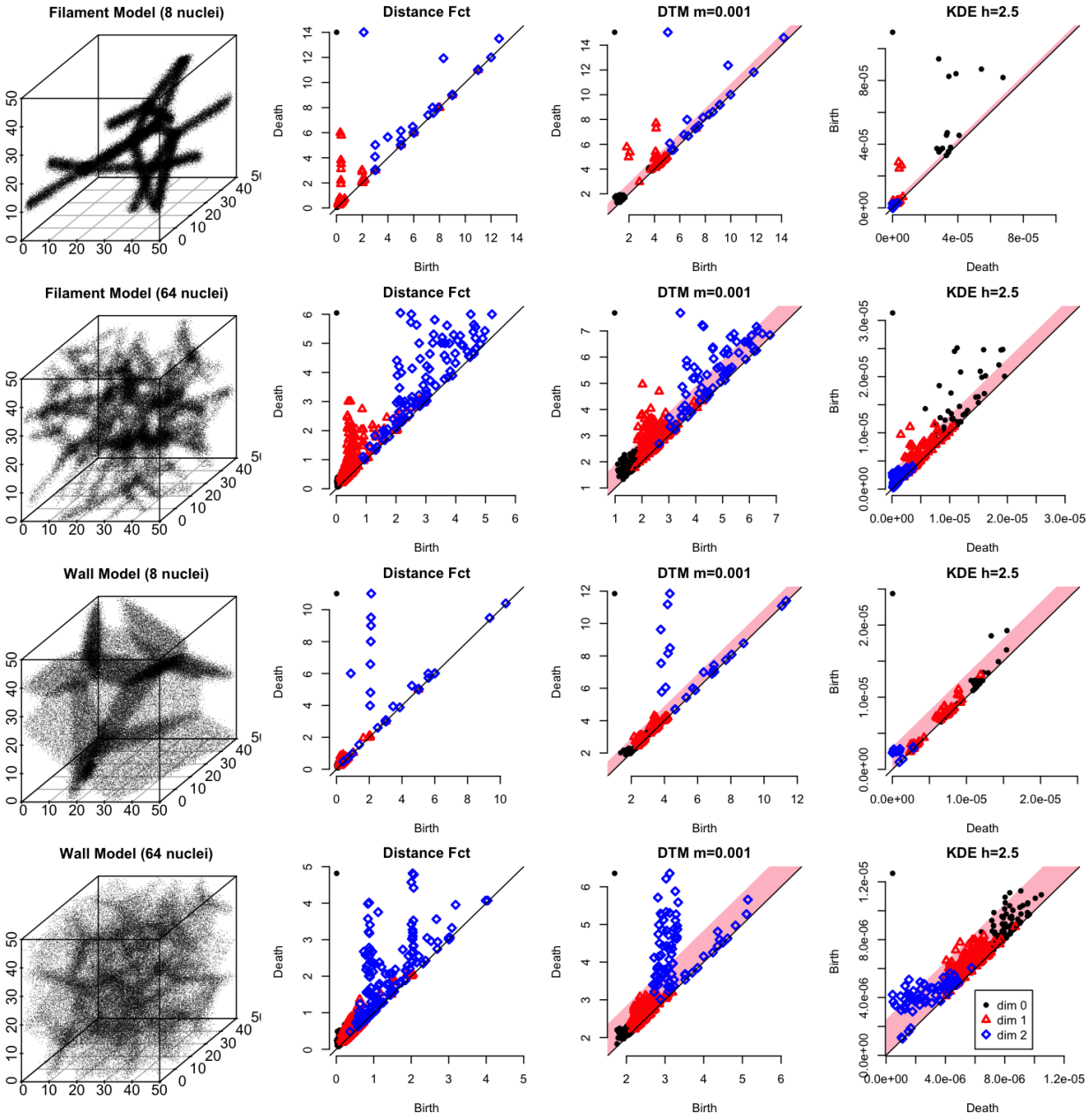

Figure 14: Data from four Voronoi foam models. In each case we show the diagrams of the distance function, the DTM and the KDE. A boundary correction was included.

This also raises the intriguing idea of extracting more information from both the KDE and DTM by varying more parameters. For example, if we look at the sets $\{p_h > t\}$ for fixed t but varying h, we get information very similar to that of the DTM. Conversely, for

the DTM, we can vary the tuning parameter m. There are many possibilities here which we will investigate in future work.

9.2 Persistent Homology Versus Choosing One Level Set

We have used the persistent homology of the upper level sets $\{\widehat{p}_h > t\}$ to probe the homology of S. This is the approach used in Bubenik (2015) and Phillips et al. (2014).

Bobrowski et al (2014) suggest a different approach. They select a particular level set $\{p > t\}$ and they form a robust estimate of the homology of this one level set. They have a data-driven method for selecting t. (This approach is only one part of the paper. They also consider persistent homology.)

They make two key assumptions. The first is that there exists $A < B$ such that $\{p > t\}$ is homotopic to S for all $A < t < B$. (If two sets are homotopic, then they have the same homology.) This is a very reasonable assumption. In the mixture model $P = \pi R + (1 - \pi)(Q \star \Phi_\sigma)$ this assumption will be satisfied when S is a small set and when π and σ are small. In this case, persistent homology will also work well: the dominant features in the persistence diagram will correspond to the homology of S.

Bobrowski et al (2014) make an additional assumption. They assume that the dimension k of S is known and that the rank of the k^{th} homology group is 0 for all $t > B$. This assumption is critical for their approach to choosing a single level set. Currently, it is not clear how strong this assumption is. In future work, we plan to compare the robustness of the single-level approach versus persistent homology.

9.3 Future Work

Lastly, we would like to mention that several issues deserve future attention. In particular, the methods we discussed for choosing the tuning parameters, for mitigating boundary bias and for sharpening the data, all deserve further investigation.

In a companion paper we will show how the ideas presented in this work can be used to develop hypothesis tests for comparing point clouds.

Acknowledgments

The authors are grateful to Jérome Dedecker for pointing out the key decomposition (18) of the DTM. The authors also would like to thank Jessi Cisewski and Jisu Kim for their comments and two referees for helpful suggestions. We would like to acknowledge support for this project from ANR-13-BS01-0008, NSF CAREER Grant DMS 1149677, Air Force Grant FA95500910373 and NSF Grant DMS-0806009.

References

A. Banyaga and D. Hurtubise. *Lectures on Morse Homology*. Kluwer Academic Publishers, 2004.

Vidmantas Bentkus. On the dependence of the berry–esseen bound on dimension. *Journal of Statistical Planning and Inference*, 113(2):385–402, 2003.

Gérard Biau, Frédéric Chazal, David Cohen-Steiner, Luc Devroye, Carlos Rodriguez, et al. A weighted k-nearest neighbor density estimate for geometric inference. *Electronic Journal of Statistics*, 5:204–237, 2011.

S. Bobkov and M. Ledoux. One-dimensional empirical measures, order statistics and Kantorovich transport distances. *Preprint*, 2014.

Omer Bobrowski, Sayan Mukherjee, and Jonathan Taylor. Topological consistency via kernel estimation. *arXiv preprint arXiv:1407.5272*, 2014.

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

Mickaël Buchet, Frédéric Chazal, Steve Y Oudot, and Donald R Sheehy. Efficient and robust topological data analysis on metric spaces. *arXiv preprint arXiv:1306.0039*, 2013.

Claire Caillerie, Frédéric Chazal, Jérôme Dedecker, and Bertrand Michel. Deconvolution for the wasserstein metric and geometric inference. *Electronic Journal of Statistics*, 5: 1394–1423, 2011.

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

F. Chazal, D. Cohen-Steiner, M. Glisse, L.J. Guibas, and S.Y. Oudot. Proximity of persistence modules and their diagrams. In *SCG*, pages 237–246, 2009. ISBN 978-1-60558-501-7. doi: http://doi.acm.org/10.1145/1542362.1542407.

F. Chazal, V. de Silva, M. Glisse, and S. Oudot. The structure and stability of persistence modules. *arXiv preprint arXiv:1207.3674*, 2012.

Frédéric Chazal, David Cohen-Steiner, and Quentin Mérigot. Geometric inference for probability measures. *Foundations of Computational Mathematics*, 11(6):733–751, 2011.

Frédéric Chazal, Pascal Massart, and Bertrand Michel. Rates of convergence for robust geometric inference. Technical report, ArXiv preprint 1505.07602, 2015.

D. Cohen-Steiner, H. Edelsbrunner, and J. Harer. Stability of persistence diagrams. In *SCG*, pages 263–271, 2005.

Antonio Cuevas and Alberto Rodríguez-Casal. On boundary estimation. *Advances in Applied Probability*, 36(2):340–354, 2004.

M. Demazure. *Bifurcations and Catastrophes: Geometry of Solutions to Nonlinear Problems*. Springer-Verlag, 2013.

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

Brittany Terese Fasy, Jisu Kim, Fabrizio Lecci, and Clement Maria. Introduction to the R package TDA. *arXiv preprint arXiv: 1411.1830*, 2014a.

Brittany Terese Fasy, Fabrizio Lecci, Alessandro Rinaldo, Larry Wasserman, Sivaraman Balakrishnan, and Aarti Singh. Confidence sets for persistence diagrams. *The Annals of Statistics*, 42(6):2301–2339, 2014b.

Evarist Giné and Armelle Guillou. Rates of strong uniform consistency for multivariate kernel density estimators. In *Annales de l'Institut Henri Poincare (B) Probability and Statistics*, volume 38, pages 907–921. Elsevier, 2002.

M. Golubitsky and V. Guillemin. *Stable Mappings and Their Singularities*. Springer-Verlag, 1986.

Leonidas Guibas, Dmitriy Morozov, and Quentin Mérigot. Witnessed k-distance. *Discrete & Computational Geometry*, 49(1):22–45, 2013.

Vincent Icke and Rien van de Weygaert. The galaxy distribution as a voronoi foam. *Quarterly Journal of the Royal Astronomical Society*, 32:85–112, 1991.

J/. Milnor. *Morse Theory*. Number 51. Princeton University Press, 1963.

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

Svein Arne Pettersen, Dag Johansen, Håvard Johansen, Vegard Berg-Johansen, Vamsidhar Reddy Gaddam, Asgeir Mortensen, Ragnar Langseth, Carsten Griwodz, Håkon Kvale Stensland, and Pål Halvorsen. Soccer video and player position dataset. In *Proceedings of the 5th ACM Multimedia Systems Conference*, pages 18–23. ACM, 2014.

Jeff M. Phillips, Bei Wang, and Yan Zheng. Geometric inference on kernel density estimates. *arXiv preprint arXiv:1307.7760*, 2014.

B. L. S. Prakasa Rao. *Nonparametric Functional Estimation*. Probability and Mathematical Statistics. Academic Press, Orlando, FL, 1983.

E. Schuster. Incorporating support constraints into nonparametric estimators of densities. *Communications in Statistics*, A(14):1123–1136, 1958.

Galen R Shorack and Jon A Wellner. *Empirical processes with applications to statistics*, volume 59. SIAM, 2009.

Bharath K Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Gert RG Lanckriet, and Bernhard Schölkopf. Kernel choice and classifiability for rkhs embeddings of probability distributions. In *NIPS*, pages 1750–1758, 2009.

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

Aad W. van der Vaart. *Asymptotic Statistics*, volume 3. Cambridge UP, 2000.

Aad W. van der Vaart and Jon A. Wellner. *Weak Convergence*. Springer, 1996.

JE Yukich. Laws of large numbers for classes of functions. *Journal of multivariate analysis*, 17(3):245–260, 1985.