1-22-2014

On the Bootstrap for Persistence Diagrams and Landscapes

Frederic Chazal
INRIA

Brittany Therese Fasy
Tulane University

Fabrizio Lecci
Carnegie Mellon University

Alessandro Rinaldo
Carnegie Mellon University, arinaldo@stat.cmu.edu

Aarti Singh
Carnegie Mellon University, aarti@cs.cmu.edu

See next page for additional authors

Follow this and additional works at: http://repository.cmu.edu/machine_learning

Published In
Modeling and Analysis of Information Systems, 20, 6, 111-120.
Authors
Frederic Chazal, Brittany Therese Fasy, Fabrizio Lecci, Alessandro Rinaldo, Aarti Singh, and Larry Wasserman

This article is available at Research Showcase @ CMU: http://repository.cmu.edu/machine_learning/106
On the Bootstrap for Persistence Diagrams and Landscapes

Frédéric Chazal¹, Brittany Terese Fasy², Fabrizio Lecci³, Alessandro Rinaldo³, Aarti Singh⁴, and Larry Wasserman³

¹INRIA Saclay
²Computer Science Department, Tulane University
³Department of Statistics, Carnegie Mellon University
⁴Machine Learning Department, Carnegie Mellon University

January 22, 2014

Abstract

Persistent homology probes topological properties from point clouds and functions. By looking at multiple scales simultaneously, one can record the births and deaths of topological features as the scale varies. In this paper we use a statistical technique, the empirical bootstrap, to separate topological signal from topological noise. In particular, we derive confidence sets for persistence diagrams and confidence bands for persistence landscapes.
Introduction

Persistent homology is a method for studying the homology at multiple scales simultaneously. Given a manifold $\mathbb{X}$ embedded in a metric space $\mathbb{Y}$, we consider a probability density function $p: \mathbb{Y} \to \mathbb{R}$, defined over $\mathbb{Y}$ but concentrated around $\mathbb{X}$; that is, the density is positive for a small neighborhood around $\mathbb{X}$ and very small over $\mathbb{Y} \setminus \mathbb{X}$. For the right scale parameter $t$, the superlevel set $p^{-1}([t, \infty))$ captures the homology of $\mathbb{X}$. The problem, however, is that $t$ is not known a priori. Persistent homology quantifies the topological changes of the superlevel sets with a multiset of points in the extended plane; we call this multiset the persistence diagram, and denote it by $\mathcal{P}$. Another way to represent the information contained in a persistence diagram is with the landscape function $\mathcal{L}: \mathbb{R} \to \mathbb{R}$, which can be thought of as a functional summary of $\mathcal{P}$; we define these concepts in Section 1.1.

Computationally, it may be difficult to compute $\mathcal{P}$ or $\mathcal{L}$ directly. Instead, we assume that $p$ corresponds to a probability distribution $\mathcal{P}$, from which we can sample. Given a sample of size $n$, we create an estimate of the probability density function $p_n$ using a kernel density estimate. As $n$ increases, $p_n$ approaches the true probability density. Given $n$ large enough, we compute the persistence diagram $\mathcal{P}_n$ and the landscape $\mathcal{L}_n$ corresponding to $p_n$.

Sometimes knowing the estimate of a persistence diagram or landscape is not enough. The bigger question is: How close is the estimated persistence diagram or landscape to the true one? We answer this question by constructing a confidence set for persistence diagrams and a confidence band for persistence landscapes.

A $(1 - \alpha)$-confidence interval for a parameter $\theta$ is an interval $[a, b]$ such that the probability $\mathbb{P}(\theta \in [a, b])$ is at least $1 - \alpha$. In our setting, we desire to find a confidence set for a persistence diagram $\mathcal{P}$. To do so, we compute an estimated diagram $\hat{\mathcal{P}}$ and an interval $[0, c]$ such that the bottleneck distance between $\mathcal{P}$ and $\hat{\mathcal{P}}$ is contained in $[0, c]$ with probability $1 - \alpha$. That is, we find a metric ball containing $\mathcal{P}$ with high probability.

In this paper, we present the bootstrap, a method for computing confidence intervals, and we apply it to persistence diagrams and landscapes. After briefly reviewing the necessary concepts from computational topology, we give the general technique of bootstrapping in statistics in Section 1.2. In Section 2, we apply the bootstrap to persistence diagrams and landscapes, providing a few examples of these confidence intervals. We conclude in Section 2.3 with a discussion of our ongoing research and open questions.

1 Background

Before presenting our results, we review the necessary definitions and theorems from persistent homology. Then, we present the bootstrap. Due to space constraints, we cover the basics and provide references for a more detailed description.
1.1 Persistence Diagrams and Landscapes

Let $Y$ be a metric space, for example. Let $Y$ be a compact subspace of $\mathbb{R}^D$. Suppose we have a probability density function $p: Y \to \mathbb{R}$ concentrated in a neighborhood of a manifold $X \subseteq Y$. Persistent homology monitors the evolution of the generators of the homology groups of $p^{-1}([t, \infty))$, the superlevel sets of $p$, and assigns to each generator of these groups a birth time (or scale) $b$ and a death time $d$. The persistence diagram $P$ records each pair $(b, d)$ as the point $(b + d, b - d)$; that is, the $x$-coordinate is the mid-life of the homological feature and the $y$-coordinate is the half-life or half of the persistence of the feature. We refer the reader to Edelsbrunner and Harer [2010] for a more complete introduction to persistent homology.

Let $D_T$ be the space of positive, countable, $T$-bounded persistence diagrams; that is, for each point $(x, y) = (b + d, b - d) \in P$, we have $0 \leq d \leq b \leq T$ and there are a countable number of points for which $y > 0$. We note here that each point on the line $x = 0$ is included in the persistence diagram $P$ with infinite multiplicity. Letting $W_\infty(P_1, P_2)$ denote the bottleneck distance between diagrams $P_1$ and $P_2$, the space $(D_T, W_\infty)$ is a metric space. We then have the following stability result from Cohen-Steiner et al. [2007] and generalized in Chazal et al. [2012]:

**Theorem 1.1 (Stability Theorem).** Let $M$ be finitely triangulable. Let $f, g: M \to \mathbb{R}$ be two continuous functions. Then, the corresponding persistence diagrams $P_f$ and $P_g$ are well defined, and $W_\infty(P_f, P_g) \leq \|f - g\|_\infty$.

Bubenik [2012] introduced another representation called the persistence landscape, which is in one-to-one correspondence with persistence diagrams. A persistence landscape is a continuous, piecewise linear function $L: \mathbb{Z}^+ \times \mathbb{R} \to \mathbb{R}$. To define the persistence landscape function, we replace each persistence point $p = (x, y) = (b + d, b - d)$ with the triangle function

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

Notice that $p$ is itself on the graph of $t_p(z)$. We obtain an arrangement of curves by overlaying the graphs of the functions $\{t_p(z)\}_{p \in P}$; see Figure 1. The persistence landscape is defined formally as a walk through this arrangement:

$$L_P(k, z) = \max_{p \in P} t_p(z),$$

where $k_{\text{max}}$ is the $k$th maximum value in the set; in particular, $1_{\text{max}}$ is the usual maximum function. Observe that $L_P(k, z)$ is $1$-Lipschitz. For the ease of exposition,
Figure 1: The pink circles are the points in a persistence diagram. The cyan curve is the landscape $L(1, \cdot)$.

we will focus on $k = 1$ in this paper, using $L(z) = L_P(1, z)$. However, the ideas we present in Section 2.2 hold for $k > 1$. Our definition of the persistence landscape is equivalent to the original definition given in Bubenik [2012].

1.2 The Standard Bootstrap

Introduced in Efron [1979], the bootstrap is a general method for estimating standard errors and for computing confidence intervals. We focus on the latter in this paper, but refer the interested reader to Efron et al. [2001], Davison and Hinkley [1997], and Van der Vaart [2000] for more details on the versatility of the bootstrap.

Let $X_1, \ldots, X_n$ be independent and identically distributed random variables taking values in the measure space $(\mathcal{X}, \mathcal{X}, P)$. Suppose we are interested in estimating the real-valued parameter $\theta$ corresponding to the distribution $P$ of the observation. We estimate $\theta$ using the statistic $\hat{\theta} = g(X_1, \ldots, X_n)$, which is some function of the data. For example, $\theta$ and $\hat{\theta}$ could be the population mean and the sample mean, respectively. The distribution of the difference $\hat{\theta} - \theta$ contains all the information that we need to construct a confidence interval of level $1 - \alpha$ for $\theta$; that is, an interval $[a, b]$ depending on the data $X_1, \ldots, X_n$ such that $\mathbb{P}(\theta \in [a, b]) \geq 1 - \alpha$. If we knew the cumulative distribution $F$ of $\hat{\theta} - \theta$, then the quantiles $F^{-1}(1 - \alpha/2)$ and $F^{-1}(\alpha/2)$ can be computed. Furthermore, setting $a = \hat{\theta} - F^{-1}(1 - \alpha/2)$ and $b = \hat{\theta} - F^{-1}(\alpha/2)$, we immediately obtain a $(1 - \alpha)$-confidence interval for $\theta$:

$$
\mathbb{P}(\theta \in [a, b]) = \mathbb{P}
\left(\hat{\theta} - \theta \leq F^{-1}
\left(\frac{\alpha}{2}\right)
\right) = 1 - \alpha.
$$

Unfortunately, the distribution of $\hat{\theta} - \theta$ depends on the unknown distribution $P$.

In the first step in the bootstrap procedure, we approximate the unknown $P$ with the empirical measure $P_n$ that puts mass $1/n$ at each $X_i$ in the sample. Let $X_1^*, \ldots, X_n^*$ be a sample of size $n$ from $P_n$. Equivalently, we can think of drawing $X_1^*, \ldots, X_n^*$ from $X_1, \ldots, X_n$ with replacement. We estimate the distribution $F(r)$ with the distribution $\hat{F}(r) = P_n(\hat{\theta}^* - \hat{\theta} \leq r)$, where $\hat{\theta}^* = g(X_1^*, \ldots, X_n^*)$. 


The distribution $\hat{F}$ is still not analytically computable, but can be approximated by simulation: for large $B$, obtain $B$ different values of $\hat{\theta}^*$ and approximate $\hat{F}(r)$, and hence $F(r)$, with $\hat{F}(r) = \frac{1}{B} \sum_{j=1}^{B} I(\hat{\theta}_j^* - \hat{\theta} \leq r)$. Since the quantiles of $\hat{F}$ approximate the quantiles of $F$, we define the estimated confidence interval as

$$C_n = \left[ \hat{\theta} - \tilde{F}_n^{-1}(1 - \alpha/2), \hat{\theta} - \tilde{F}_n^{-1}(\alpha/2) \right].$$

In summary, the standard bootstrap procedure is:

1. Compute the estimate $\hat{\theta} = g(X_1, \ldots, X_n)$.
2. Draw $X_1^*, \ldots, X_n^*$ from $P_n$ and compute $\hat{\theta}^* = g(X_1^*, \ldots, X_n^*)$.
3. Repeat the previous step $B$ times to obtain $\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*$.
4. Compute the quantiles of $\tilde{F}$ and construct the confidence interval $C_n$.

There are two sources of error in the Bootstrap procedure. We first approximate $F$ with $\hat{F}$ and then we estimate $\hat{F}$ by simulation. The second error can be made arbitrarily small, by choosing $B$ large enough. Therefore, this error is usually ignored in the theory of the bootstrap.

Formally, one has to show that $\sup_r \left| \tilde{F}(r) - F(r) \right| \rightarrow 0$, which implies that the confidence interval $C_n$, defined in (2), is asymptotically consistent at level $1 - \alpha$; that is, $\lim \inf_{n \rightarrow \infty} \mathbb{P}(\theta \in C_n) \geq 1 - \alpha$.

### 1.3 The Bootstrap Empirical Process

When a random variable is a function rather than a real value, the bootstrap procedure described above can be used to construct a confidence interval for the function evaluated at a particular element of the domain. Instead, we use the bootstrap empirical process, which can be used to find a confidence band for a function $h(t)$; that is, we find a pair of functions $a(t)$ and $b(t)$ such that the probability that $h(t) \in [a(t), b(t)]$ for all $t$ is at least $1 - \alpha$. We describe this technique below, but refer the reader to Van der Vaart and Wellner [1996] and Kosorok [2008] for more details.

An empirical process is a stochastic process based on a random sample. Let $X_1, \ldots, X_n$ be independent and identically distributed random variables taking values in the measure space $(X, \mathcal{X}, P)$. For a measurable function $f : X \rightarrow \mathbb{R}$, we denote $Pf = \int f dP$ and $P_nf = \int f dP_n = n^{-1} \sum_{i=1}^{n} f(X_i)$. By the law of large numbers $P_n f$ converges almost surely to $Pf$. Given a class $\mathcal{F}$ of measurable functions, we define the empirical process $\mathcal{G}_n$ indexed by $\mathcal{F}$ as

$$\{ \mathcal{G}_n f \}_{f \in \mathcal{F}} = \left\{ \sqrt{n} (P_n f - Pf) \right\}_{f \in \mathcal{F}}.$$
Example 1.2. If $\mathcal{F} = \{I(x \leq t)\}_{t \in \mathbb{R}}$, then $\{P_nf\}_{f \in \mathcal{F}} = \{n^{-1} \sum_{i=1}^{n} I(X_i \leq t)\}_{t \in \mathbb{R}}$, which is the empirical distribution function seen as a stochastic process indexed by $t$. Furthermore, we have $\{G_n f\}_{f \in \mathcal{F}} = \{n^{-1/2} \sum_{i=1}^{n} I(X_i \leq t) - P(X_i \leq t)\}_{t \in \mathbb{R}}$, which is the empirical distribution function seen as a stochastic process indexed by $t$.

Definition 1.3. A class $\mathcal{F}$ of measurable functions $f : X \to \mathbb{R}$ is called $P$-Donsker if the process $\{G_n f\}_{f \in \mathcal{F}}$ converges in distribution to a limit process in the space $\ell^\infty(\mathcal{F})$, where $\ell^\infty(\mathcal{F})$ is the collection of all bounded functions $f : X \to \mathbb{R}$.

The limit process is a Gaussian process $\mathcal{G}$ with zero mean and covariance function $E \mathcal{G}f \mathcal{G}g := Pfg - PfPg$; this process is known as a Brownian Bridge.

Let $P^*_n f = n^{-1} \sum_{i=1}^{n} f(X^*_i)$ where $\{X^*_1, \ldots, X^*_n\}$ is a bootstrap sample from $P_n$, the measure that puts mass $1/n$ on each element of the sample $\{X_1, \ldots, X_n\}$. The bootstrap empirical process $G^*_n$ indexed by $\mathcal{F}$ is defined as

$$\{G^*_n f\}_{f \in \mathcal{F}} = \{\sqrt{n}(P^*_n f - P_n f)\}_{f \in \mathcal{F}}.$$

Theorem 1.4 (Theorem 2.4 in Giné and Zinn [1990]). $\mathcal{F}$ is $P$-Donsker if and only if $G^*_n$ converges in distribution to $\mathcal{G}$ in $\ell^\infty(\mathcal{F})$.

In words, Theorem 1.4 states that $\mathcal{F}$ is $P$-Donsker if and only if the bootstrap empirical process converges in distribution to the limit process $\mathcal{G}$ given in Definition 1.3.

Suppose we are interested in constructing a confidence band of level $1 - \alpha$ for $\{Pf\}_{f \in \mathcal{F}}$, where $\mathcal{F}$ is $P$-Donsker. Let $\hat{\theta} = \sup_{f \in \mathcal{F}} |G_n f|$. We proceed as follows:

1. Draw $X^*_1, \ldots, X^*_n$ from $P_n$ and compute $\hat{\theta}^* = \sup_{f \in \mathcal{F}} |G^*_n f|$.
2. Repeat the previous step $B$ times to obtain $\hat{\theta}^*_1, \ldots, \hat{\theta}^*_B$.
3. Compute $q_\alpha = \inf \left\{ q : \frac{1}{B} \sum_{j=1}^{B} I(\hat{\theta}^*_j \geq q) \leq \alpha \right\}$.
4. For $f \in \mathcal{F}$ define the confidence band $C_n(f) = \left[ P_n f - \frac{q_\alpha}{\sqrt{n}}, P_n f + \frac{q_\alpha}{\sqrt{n}} \right]$.

A consequence of Theorem 1.4 is that, for large $n$ and $B$, the interval $[0, q_\alpha]$ has coverage $1 - \alpha$ for $\hat{\theta}$ and the band $C_n(f)_{f \in \mathcal{F}}$ has coverage $1 - \alpha$ for $\{Pf\}_{f \in \mathcal{F}}$.

2 Applications of the Bootstrap

In this section, we apply the bootstrap from the previous section to persistence diagrams, as well as to persistence landscapes.
2.1 Persistence Diagrams

Let $X_1, \ldots, X_n$ be a sample from the distribution $P$, supported on a smooth manifold $\mathcal{X} \subset \mathbb{R}^D$. Let $p_h(x) = \int_{\mathcal{X}} \frac{1}{h^D} K \left( \|x - u\|_h \right) \, dP(u)$, where $K : \mathbb{R} \to \mathbb{R}$ is an integrable function satisfying $\int K(u) \, du = 1$ and $K(u)$ is nonnegative for all $u$; thus $p_h$ is a probability distribution. The function $K$ is called a kernel and the parameter $h > 0$ is its bandwidth. Then $p_h$ is the density of the probability measure $P_h$ which is the convolution $P_h = P \ast \mathbb{K}_h$ where $\mathbb{K}_h(A) = h^{-D} \mathbb{K}(h^{-1} A)$ and $\mathbb{K}(A) = \int_A K(t) \, dt$. $P_h$ is a smoothed version of $P$.

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

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

notice that if $X_i$ are fixed, then $\hat{p}_h$ is a probability distribution. Let $\hat{\mathcal{P}}_h$ be the corresponding persistence diagram. We wish to find a confidence set for $\mathcal{P}_h$, i.e., an interval $[0, c_n]$ such that $\limsup_{n \to \infty} \mathbb{P}(W_\infty(\hat{\mathcal{P}}_h, \mathcal{P}_h) \in [0, c_n]) = 1 - \alpha$. From Theorem 1.3 (Stability), it suffices to find $c_n$ such that $\limsup_{n \to \infty} \mathbb{P}(\|\hat{p}_h - p_h\|_\infty > c_n) \leq \alpha$.

To find $c_n$, we use the bootstrap. Let $\mathcal{F} = \{ f_x(u) = \frac{1}{n} \int \frac{K}{h^D} K \left( \frac{\|x - u\|}{h} \right) \}_{x \in \mathcal{X}}$. Using the notation of Section 1.3 it follows that $P f_x = p_h(x)$, $P_n f_x = \hat{p}_h(x)$ and $\hat{\theta} = \sup_{f \in \mathcal{F}} |G_n f| = \sqrt{n} \| \hat{p}_h - p_h \|_\infty$. The approximated $1 - \alpha$ quantile $q_\alpha$ can be obtained through simulation, i.e., $q_\alpha = \inf \{ q : \frac{1}{B} \sum_{j=1}^{B} I(\sqrt{n} \| \hat{p}_n^j - \hat{p}_n \| \geq q) \leq \alpha \}$, where $p_n^j(x)$ denotes the probability distribution corresponding to the $j^{th}$ bootstrap sample. The following result holds under suitable regularity conditions on the kernel $K$ for which $\mathcal{F}$ is Donsker; see [Giné and Guillou 2002].

**Theorem 2.1** (Lemma 15 in [Balakrishnan et al., 2013]). We have that

$$\limsup_{n \to \infty} \mathbb{P}(\sqrt{n} \| \hat{p}_h - p_h \|_\infty > q_\alpha) \leq \alpha.$$

By the Stability Theorem, we conclude: $\lim_{n \to \infty} \mathbb{P}(\sqrt{n} \| \hat{p}_h - p_h \|_\infty > \frac{q_\alpha}{\sqrt{n}}) \leq \alpha$.

**Example 2.2** (Torus). We embed the torus $\mathbb{S}^1 \times \mathbb{S}^1$ in $\mathbb{R}^3$ and we use the rejection sampling algorithm of [Diaconis et al., 2013] ($R = 1.5, r = 0.8$) to sample 10,000 points uniformly from the torus. Then, we compute the persistence diagram $\hat{\mathcal{P}}_h$ using the Gaussian kernel with bandwidth $h = 0.25$ and use the bootstrap to construct the 0.95% confidence interval $[0, 0.01]$ for $W_\infty(\hat{\mathcal{P}}_h, \mathcal{P}_h)$; see Figure 2. Notice that the confidence set correctly captures the topology of the torus. That is, only the points representing real features of the torus are significantly far from the horizontal axis.
2.2 Landscapes

Let the diagrams $\mathcal{P}_1, \ldots, \mathcal{P}_n$ be a sample from the distribution $P$ over the space of persistence diagrams $\mathcal{D}_T$. Thus, by definition, we have $x + y \leq T < \infty$ and $0 \leq y \leq T/2$ for all $(x,y) \in \bigcup_i \mathcal{P}_i$.

Let $\mathcal{L}_1, \ldots, \mathcal{L}_n$ be the landscape functions corresponding to $\mathcal{P}_1, \ldots, \mathcal{P}_n$. That is, $\mathcal{L}_i(t) = \mathcal{L}_{\mathcal{P}_i}(1,t)$, as defined in \cite{1}. We define the mean landscape $\mu(t) = \mathbb{E}_P[\mathcal{L}_i(t)]$, and the empirical mean landscape $\bar{\mathcal{L}}_n(t) = \frac{1}{n} \sum_{i=1}^n \mathcal{L}_i(t)$. In this section, we show that the process $\sqrt{n} (\bar{\mathcal{L}}_n(t) - \mu(t))$ converges to a Gaussian process, so that we may use the procedure given in Section 1.3.

Let $\mathcal{F} = \{ f_t : 0 \leq t \leq T \}$, where $f_t : \mathcal{D} \to \mathbb{R}$ is defined by $f_t(\mathcal{P}) = \mathcal{L}_{\mathcal{P}}(1,t)$. We note here that $f_t(\mathcal{P}) = 0$ if $t \notin (0,T)$. We can now write $\sqrt{n} (\bar{\mathcal{L}}_n(t) - \mu(t))$ as an empirical process indexed by $t \in [0,T]$

$$\sqrt{n} (\bar{\mathcal{L}}_n(t) - \mu(t)) = \sqrt{n} \left( \frac{1}{n} \sum_{i=1}^n \mathcal{L}_i(t) - \mu(t) \right) = \sqrt{n} (P_n f_t - Pf_t) \equiv G_n f_t.$$  

We note that the constant function $F(\mathcal{P}) = T/2$ is a measurable envelope for $\mathcal{F}$.

Given a probability measure $Q$ over $\mathcal{F}$, let $\|f - g\|_{Q,2} = \sqrt{\int |f - g|^2 dQ}$ and let $N(\mathcal{F}, L_2(Q), \varepsilon)$ be the covering number of $\mathcal{F}$, that is, the size of the smallest $\varepsilon$-net in this metric.

Lemma 2.3 (Theorem 2.5 in \cite{Kosorok2008}). Let $\mathcal{F}$ be a class of measurable functions satisfying $\int_0^1 \sqrt{\log \sup_Q N(\mathcal{F}, L_2(Q), \varepsilon \| F \|_{Q,2})} d\varepsilon < \infty$, where $F$ is a measurable
envelope of $\mathcal{F}$ and the supremum is taken over all finitely discrete probability measures $Q$ with $\|F\|_{Q,2} > 0$. If $PF^2 < \infty$, then $\mathcal{F}$ is $P$-Donsker.

**Theorem 2.4** (Weak Convergence of Landscapes). Let $\mathcal{G}$ be a Brownian bridge with covariance function $\kappa(t,u) = \int f_t(P)f_u(P)dP(P) - \int f_t(P)dP(P)\int f_u(P)dP(P)$. Then, $\mathcal{G}_n$ converges in distribution to $\mathcal{G}$.

*Proof.* Since persistence landscapes are $1$-Lipschitz, we have $\|f_t - f_u\|_{Q,2} \leq |t - u|$. Construct a regular grid $0 \equiv t_0 < t_1 < \cdots < t_N \equiv T$, where $t_{j+1} - t_j = \varepsilon T/2$. We claim that $\{f_{t_j} : 1 \leq j \leq N\}$ is an $(\varepsilon T/2)$-net for $\mathcal{F}$: choose $f_t \in \mathcal{F}$; then there is a $j$ so that $t_j \leq t \leq t_{j+1}$ and $\|f_{t_{j+1}} - f_t\|_{Q,2} \leq |t_{j+1} - t_j| = \varepsilon T/2$. The fact that $\{f_{t_j} : 1 \leq j \leq N\}$ is an $(\varepsilon T/2)$-net implies $\sup_Q N(\mathcal{F},L_2(Q),\varepsilon\|F\|_2) \leq 2/\varepsilon$. Hence, $\int_0^1 \sqrt{\log \sup_Q N(\mathcal{F},L_2(Q),\varepsilon\|F\|_2)}d\varepsilon < \infty$. $F = T/2$ is trivially square-integrable. By Lemma 2.3, $\mathcal{G}_n$ converges in distribution to $\mathcal{G}$. \hfill \square

Now that we have shown that $\mathcal{G}_n$ converges to a Gaussian process, we can follow the procedure outlined in Section 1.3. Let $P_n$ be the empirical measure that puts mass $1/n$ at each diagram $\mathcal{P}_i$. We draw $\mathcal{P}_1^*,\ldots,\mathcal{P}_n^*$ from $P_n$ and construct the corresponding landscapes $L_1^*,\ldots,L_n^*$. Let $\bar{\mathcal{L}}_n^*$ be the empirical mean and $\hat{\theta}^* = \sup_{t \in \mathbb{R}}|\sqrt{n}(\mathcal{L}_n^*(t) - \bar{\mathcal{L}}_n(t))|$. Repeating this $B$ times, we obtain $\hat{\theta}_1^*,\ldots,\hat{\theta}_B^*$, and we compute the quantile $q_\alpha$.

**Theorem 2.5** (Confidence Band for Persistent Landscapes). The interval $C_n(t) = \left[\bar{\mathcal{L}}_n(t) - \frac{q_\alpha}{\sqrt{n}}, \bar{\mathcal{L}}_n(t) + \frac{q_\alpha}{\sqrt{n}}\right]$, is a confidence band for $\mu(t)$:

$$\lim_{n \to \infty} \mathbb{P} (\mu(t) \in C_n(t) \text{ for all } t) \geq 1 - \alpha.$$

**Example 2.6** (Circles). Given the nine circles of radii 0.4 and 0.3, shown in Figure 3, we obtain a sample $X_1,\ldots,X_{100}$ as follows: first, choose a circle $C_i$ uniformly at random, then sample a point iid from $C_i$. Let $\mathcal{P}$ be the (Betti 1) persistence diagram corresponding to the Rips filtration for the sample, and $\mathcal{L}$ be the landscape corresponding to $\mathcal{P}$. We repeat this 50 times to obtain diagrams $\mathcal{P}_1,\ldots,\mathcal{P}_{50}$ and landscapes $L_1,\ldots,L_{50}$. Then, we use the bootstrap procedure to obtain the quantile $q_\alpha = 0.234$. Together with $\bar{\mathcal{L}}_{50}$, this gives us an approximated 95% confidence band for $\mu(t) = \mathbb{E}_P(\mathcal{L}_i(t))$. On the right of Figure 3, we show the empirical mean landscape $\bar{\mathcal{L}}_{50}$ with the 95% confidence band for $\mu(t)$.

---

2Note that, since in this example we are using sublevel sets, the role of birth and death in the definitions of section 1.1 is inverted. The death time $d$ is greater than the birth time $b$. 


2.3 Discussion

In this paper, we have described the bootstrap as it applies to persistence diagrams and landscapes. The purpose of this paper was to introduce the bootstrap and the bootstrap empirical process to topologists. In a related paper [Balakrishnan et al. 2013], aimed towards a statistical audience, we derive the convergence rates for the technique presented in Section 2.1 as well as present three other methods for computing confidence sets for persistence diagrams.

The persistence landscape can be thought of as a summary function of a persistence diagram. The bootstrap method that we presented in Section 2.2 trivially generalizes to handle all landscapes $L(k, t)$. Furthermore, we need not limit the scope of this method to landscape functions. In a future paper, we plan to investigate other meaningful summary functions as well as the convergence rates for the techniques presented in Section 2.2.

We have demonstrated how the bootstrap works for two examples, given in Figure 2 and Figure 3. Part of our ongoing research is investigating applications for these confidence intervals; in particular, we are applying it to real (rather than simulated) data sets. One can use the confidence intervals for hypothesis testing, but an open question is how to determine the power of such a test.

Acknowledgement

The authors would like to thank Sivaraman Balakrishnan for his insightful discussions.
References


