The Annals of Statistics
2023, Vol. 51, No. 4, 1484–1509
https://doi.org/10.1214/23-AOS2277

BOOTSTRAPPING PERSISTENT BETTI NUMBERS AND OTHER STABILIZING STATISTICS

BY BENJAMIN ROYCRAFT[1,a], JOHANNES KREBS[2,c] AND WOLFGANG POLONIK[1,b]

[1]*Department of Statistics, University of California, Davis,* [a]*btroycraft@ucdavis.edu,* [b]*wpolonik@ucdavis.edu*
[2]*Department of Mathematics, KU Eichstätt-Ingolstadt,* [c]*johannes.krebs@ku.de*

We investigate multivariate bootstrap procedures for general stabilizing statistics, with specific application to topological data analysis. The work relates to other general results in the area of stabilizing statistics, including central limit theorems for geometric and topological functionals of Poisson and binomial processes in the critical regime, where limit theorems prove difficult to use in practice, motivating the use of a bootstrap approach. A smoothed bootstrap procedure is shown to give consistent estimation in these settings. Specific statistics considered include the persistent Betti numbers of Čech and Vietoris–Rips complexes over point sets in $\mathbb{R}^d$, along with Euler characteristics, and the total edge length of the k-nearest neighbor graph. Special emphasis is given to weakening the necessary conditions needed to establish bootstrap consistency. In particular, the assumption of a continuous underlying density is not required. Numerical studies illustrate the performance of the proposed method.

1. Introduction. In recent years, a multitude of topological statistics have been developed to describe and analyze the structure of data, achieving notable success. These methods have seen application in astrophysics [1, 36–38], cancer genomics [2, 9, 19], medical imaging [17], materials science [24], fluid dynamics [25], chemistry [46], and other wide ranging fields [44].

The use of simplicial complexes to summarize geometric and topological properties of data culminates in the techniques of persistent homology. Summary statistics based on persistent homology, persistent Betti numbers, persistence diagrams and derivatives thereof effectively extract essential topological properties from point cloud data. A broad introduction to the methods of topological data analysis can be found in [3, 5, 10, 14, 16, 20, 45].

While the use of such statistics has seen wide success, very little is currently known about the statistical properties of these topological summaries. An initial attempt at statistical analysis using persistent homology can be seen in [8], with the later introduction of persistence landscapes in [7]. Likewise, central limit theorems have been developed for persistence landscapes [12], Betti numbers [48] and persistent Betti numbers [22, 27] under a variety of asymptotic settings. However, the form of these results is insufficient to provide for valid confidence intervals.

For the construction of asymptotically valid confidence intervals, subsampling and bootstrap estimation have proven successful. In [21], various techniques are given for constructing confidence sets for persistence diagrams and derived statistics, including persistence diagrams generated from sublevel sets of the density function, as well as for the Čech and Vietoris–Rips complexes of data constrained to a manifold embedded in $\mathbb{R}^d$. In [12, 13], bootstrap consistency is established very generally for persistence landscapes drawn from independently generated point clouds in $\mathbb{R}^d$, assuming that the number of independent samples

Received March 2021; revised March 2023.
MSC2020 subject classifications. Primary 62F40; secondary 55N31, 62R40, 62H10, 62G05.
Key words and phrases. Betti numbers, bootstrap, Euler characteristic, random geometric complexes, stabilizing statistics, stochastic geometry, topological data analysis, persistent homology.

is allowed to grow. Finally, [11] considers subsampling for novel topological statistics in the multi-sample regime. Related results are found in [15, 31, 32].

However, even with these recent developments, the available techniques for constructing confidence sets using topological statistics remain severely limited. The bootstrap has proven one of the only effective tools; however, the theoretical properties of bootstrap estimation applied to topological statistics are not well understood. For the large-sample asymptotic regime, in particular, results are largely nonexistent.

The goal of this work is to provide the foundational theory for the bootstrap in this area. We use the smooth bootstrap, rather than a standard bootstrap for reasons described below. The validity of the smooth bootstrap in the multivariate setting is established, a key step toward an eventual process-level result. However, the latter remains a significant technical hurdle. While motivated primarily by application to topological data analysis, the results presented here apply much more generally over a class of *stabilizing* statistics. As defined in [33], a statistic *stabilizes* if the change in the function value induced by addition of new points to the underlying sample is at most locally determined. This concept has lead to many developments in topological data analysis [22, 27, 42, 47, 48] and geometric probability, as discussed in more detail below.

Our general result allows the analysis of large-sample asymptotic properties of the bootstrap applied to Betti numbers and Euler characteristics over Čech and Vietoris–Rips complexes directly, where the underlying point cloud is a sample drawn from a common distribution on $\mathbb{R}^d$. Another application is the convergence for the bootstrap applied to the total edge length of the k-nearest neighbor graph. Throughout this work, a special focus is given toward weakening the necessary assumptions compared to previous results. Specifically, the theorems presented here apply for distributions with unbounded support, unbounded density and possible discontinuities. We assume only a bound for an appropriate L_p-norm of the underlying sampling density.

The first half of this paper considers stabilizing statistics in general. Section 2 introduces the concept of stabilization, establishes intermediate technical results and presents our general bootstrap consistency theorem. The second half introduces the main topological and geometric statistics of interest to which the general theory is applied. In particular, Section 3 connects the general theory to persistent homology and related statistics. Toward this end, a short introduction to simplicial complexes and persistent homology is presented. Section 4 analyzes the stabilization properties of persistent Betti numbers and Euler characteristics for general classes of distance-based simplicial complexes. Bootstrap consistency is established for each, as well as for the total edge length of the k-nearest neighbor graph. Section 5 and Appendix A [40] present numerical studies (simulations and a real data application), demonstrating the finite-sample properties of the smoothed bootstrap applied to persistent Betti numbers. The source code is available at github.com/btroycraft/stabilizing_statistics_bootstrap [39]. The proof of all results can be found in Section 7 and Appendix B [40]. Appendices B and C [40] contain supporting results, including a characterization of L_p-norm consistency for the kernel density estimator, which is an interesting result in its own right.

2. Stabilizing statistics.

2.1. *Central limit theorems for stabilizing statistics.* Before proving bootstrap convergence, we give a brief overview of the existing work regarding stabilizing statistics. For the precise definitions used throughout this paper, see Section 2.2.

In the seminal work of [33], a stabilization property was first formally defined. In short, we say that a functional ψ defined on point sets in $\mathbb{R}^d$ *stabilizes* if the cost of adding an additional point, or a set of points, to the point cloud varies only on a bounded region. Specific

definitions differ by context (precise definitions are given below). Penrose and Yukich [33] use this concept to prove central limit theorems for certain types of geometric functionals, including the length of the k-nearest neighbor graph and the number of edges in the sphere of influence graph. This initial work distilled two properties key to showing central limit theorems for geometric functionals: a stabilization property and a moment bound.

In [33], the authors distinguish between two data generating regimes: A homogenous Poisson process over $\mathbb{R}^d$ and a binomial process, the latter being equivalent to an i.i.d. sample of fixed size from an appropriate probability distribution. Here, the functional under consideration is restricted to a bounded domain B_n of volume n, where n is allowed to increase. In this initial work, only homogenous Poisson processes and uniform binomial sampling are considered. In [34], a similar framework is used to establish laws of large numbers for graph-based functionals, including the number of connected components in the minimum spanning tree. Further quantitative refinements on the general central limit theorems for stabilizing statistics are shown in [28, 29] and [30].

As pertains to topological statistics, an initial central limit theorem for Betti numbers (see Section 3.2 for definitions) was shown in [48], establishing so-called "weak stabilization" for Betti numbers in the homogenous Poisson and uniform binomial sampling settings. There an alternative set-up is being used where the domain is kept fixed, while the filtration parameter is decreasing to zero. A similar result for persistent Betti numbers is given in [22].

Finally, [27] establishes multivariate central limit theorems for persistent Betti numbers under a flexible sampling setting. Here, a nonhomogeneous Poisson or binomial process is generated again over a growing domain with fixed filtration radii.

With these central limit theorem results, the stabilization property plays a central role in understanding the asymptotic behavior for wide classes of geometric and topological functionals. Unfortunately, as a reoccurring trend, explicit forms for the asymptotic normal distributions are unavailable or computationally intractable. In this work, it is shown how a smoothed bootstrap procedure allows for consistent estimation of these inaccessible limiting distributions, and thus for any subsequent inference derived therefrom.

2.2. *Stabilization.* Here, we extend and rephrase existing definitions found in [33, 34, 48] and [27] to provide a more general and consistent statistical framework. Let $\mathcal{X}$ denote the space consisting of multisets drawn from $\mathbb{R}^d$ with no accumulation points, with the further restriction that no point in a given multiset may be counted more than finitely often. Any locally-finite point process on $\mathbb{R}^d$ can be represented as a random element of $\mathcal{X}$. Let $\tilde{\mathcal{X}} \subset \mathcal{X}$ contain the finite multisets drawn from $\mathbb{R}^d$ and $\psi\colon \tilde{\mathcal{X}} \to \mathbb{R}$ be a measurable function. Furthermore, for $S, T \in \tilde{\mathcal{X}}$ define the *addition cost* of T to S as $D(S; T) := \psi(S \cup T) - \psi(S)$. When $T = \{z\}$ consists of a single point, we call

$$D_z(S) := \psi\big(S \cup \{z\}\big) - \psi(S)$$

an *add-one cost* or the *add-z cost*. Broadly, we say that ψ *stabilizes* if the addition cost of a given T varies only on a bounded region. Examples for functionals ψ of interest, such as persistent Betti numbers, Euler characteristics or the length of the k-nearest neighbor graph, are discussed in Sections 3.3 to 3.5.

In the preceding literature, the terms "strong" and "weak" stabilization are very often used, with precise definitions changing based on circumstance. In the interest of providing more explanatory and specific terminology, we propose the below definitions.

There, almost-sure and locally-determined almost-sure stabilization (see Definitions 2.4 and 2.6) correspond, respectively, to Definitions 3.1 and 2.1 in [33]. Here, we have generalized by accounting for possible measurability issues, however, the definitions are essentially equivalent. Let $B_z(r)$ denote the closed Euclidean ball centered at $z \in \mathbb{R}^d$ with radius r. For convenience, the dependence on ψ is implicit in each of the following.

DEFINITION 2.1 (Terminal addition cost). $D^{\infty}\colon \mathcal{X} \to \mathbb{R}$ is a *terminal addition cost* of $T \in \widetilde{X}$ centered at $z \in \mathbb{R}^d$ if

$$D^{\infty}(S;T) = \lim_{\ell\to\infty} D\big(S \cap B_z(\ell); T\big)$$

for any $S \in \mathcal{X}$ such that the limit exists.

For a finite multiset $S \in \tilde{\mathcal{X}}$, the terminal addition cost centered at $z \in \mathbb{R}^d$ is $D^{\infty}(S;T) = D(S;T)$, because no changes may occur once $S \cap B_z(a) = S$ for $a > 0$ sufficiently large. The same does not hold for infinite multisets, motivating a separate definition. In the special case where $T = \{z\}$ is a singleton containing the center point $z \in \mathbb{R}^d$, the notation D_z^{∞} may be used, and will appear throughout the remaining sections of the paper.

DEFINITION 2.2 (Stabilization in probability). For a given center point $z \in \mathbb{R}^d$, $T \in \tilde{\mathcal{X}}$ and point process $\mathbf{S}$ taking value in $\mathcal{X}$, ψ *stabilizes* on $\mathbf{S}$ *in probability* if there exists a terminal addition cost D^{∞} for ψ such that

$$\lim_{\ell\to\infty} \mathbb{P}^*\big[D\big(\mathbf{S} \cap B_z(\ell); T\big) \neq D^{\infty}(\mathbf{S};T)\big] = 0.$$

Here, $\mathbb{P}^*$ denotes the outer probability of a set. Stabilization is said to occur *in probability* because, for any sequence of nonnegative radii $(\ell_i)_{i\in\mathbb{N}}$ such that $\lim_{i\to\infty} \ell_i = \infty$, $D(\mathbf{S} \cap B_z(\ell_i); T) \xrightarrow{\mathrm{p}} D^{\infty}(\mathbf{S};T)$ whenever both quantities are measurable. D^{∞} is unique up to a null set in this case. Stabilization in probability is difficult to show directly for many functionals of interest. As such, we have the following.

DEFINITION 2.3 (Radius of stabilization). Given $T \in \tilde{\mathcal{X}}$, $\rho\colon \mathcal{X} \to [0,\infty]$ is a *radius of stabilization* for ψ centered at $z \in \mathbb{R}^d$ if, for any $S \in \mathcal{X}$ and $L \in \mathbb{R}$ such that $\rho(S) \le L < \infty$,

$$D\big(S \cap B_z(L); T\big) = D\big(S \cap B_z\big(\rho(S)\big); T\big).$$

Here, $D^{\infty}(S;T) := D(S \cap B_z(\rho(S)); T)$ is a valid terminal addition cost. Note that, in the case where $\lim_{\ell\to\infty} D(S \cap B_z(\ell); T)$ does not exist, $\rho(S) = \infty$ necessarily, with the stabilization criterion being satisfied vacuously. When $T = \{z\}$, we denote $\rho = \rho_z$.

In general, for any ψ, $T \in \tilde{\mathcal{X}}$ and center point $z \in \mathbb{R}^d$, there exists a unique minimal radius of stabilization, defined as the pointwise minimum over all such radii sharing the same center point z. This minimum exists because $\psi(S \cap B_z(\ell))$ is piecewise constant in ℓ, changing value only when a new point of S is added, and because S has no accumulation points.

DEFINITION 2.4 (Stabilization almost surely). For $\mathbf{S}$, a point process taking values in $\mathcal{X}$, ψ stabilizes on $\mathbf{S}$ *almost surely* if there exists a radius of stabilization $\rho\colon \mathcal{X} \to [0,\infty]$ for ψ centered at $z \in \mathbb{R}^d$ such that

$$\lim_{\ell\to\infty} \mathbb{P}^*\big[\rho(\mathbf{S}) > \ell\big] = 0.$$

Mirroring our previous terminology, we say stabilization occurs *almost surely* because, for any sequence of nonnegative radii $(\ell_i)_{i\in\mathbb{N}}$ such that $\ell_i \to \infty$, $D(\mathbf{S} \cap B_z(\ell_i); T) \xrightarrow{\text{a.s.}} D^{\infty}(\mathbf{S};T) = D(\mathbf{S} \cap B_z(\rho(\mathbf{S})); T)$ whenever both quantities are measurable. Here, we use outer probability, because a radius of stabilization may not be a measurable function, specifically in the case of the unique minimal radius. Almost sure stabilization implies stabilization in probability, as shown in the following.

PROPOSITION 2.5. *For* $\mathbf{S}$*, a simple point process taking values in* $\mathcal{X}$*, let* ψ *stabilize on* $\mathbf{S}$ *almost surely. Then* ψ *stabilizes on* $\mathbf{S}$ *in probability.*

For our proof techniques, it is often necessary to compare the stabilization properties of a function over a range of related point processes. For example, corresponding binomial and Poisson processes can be shown to have essentially equivalent local properties, while differing globally. This motivates the following.

DEFINITION 2.6 (Locally determined radius of stabilization). A radius of stabilization ρ centered at $z \in \mathbb{R}^d$ is *locally determined* if for any $S, S' \in \mathcal{X}$,

$$S' \cap B_z\big(\rho(S)\big) = S \cap B_z\big(\rho(S)\big) \quad \Longrightarrow \quad \rho\big(S'\big) = \rho(S).$$

With the local-determination criterion from Definition 2.6, we can assure that stabilization must occur simultaneously on any two point processes, which are locally equivalent. As in the nonlocally-determined case, there exists a unique minimal locally-determined radius of stabilization.

PROPOSITION 2.7. *For* $\mathcal{R}$*, the set of locally-determined radii of stabilization for* ψ *centered at* $z \in \mathbb{R}^d$*, let* $\rho^*\colon \mathcal{X} \to [0, \infty]$ *such that* $\rho^*(S) = \inf_{\rho \in \mathcal{R}} \rho(S)$*. Then* ρ^* *is a locally determined radius of stabilization for* ψ *centered at* z*.*

2.3. *Technical results.* Let F and G be distributions on $\mathbb{R}^d$ with densities $f := \mathrm{d}F/\mathrm{d}\lambda$ and $g := \mathrm{d}G/\mathrm{d}\lambda$, respectively, where λ is the Lebesgue measure on $\mathbb{R}^d$. F will be used to refer to a fixed central distribution, whereas G may be arbitrary. Let $(X_i)_{i\in\mathbb{N}} \overset{\text{iid}}{\sim} F$. Then for any $n \in \mathbb{N}$ define the binomial point process $\mathbf{X}_n := \{X_i\}_{i=1}^n$, with $X' \sim F$ independent of the $(X_i)_{i\in\mathbb{N}}$. Similarly, given $(Y_i)_{i\in\mathbb{N}} \overset{\text{iid}}{\sim} G$, for any $n \in \mathbb{N}$ let $\mathbf{Y}_n := \{Y_i\}_{i=1}^n$ denote the corresponding binomial process, and let $Y' \sim G$ be independent of the $(Y_i)_{i\in\mathbb{N}}$.

Let $\|\cdot\|_p^p := \int_{\mathbb{R}^d} |\cdot|^p \,\mathrm{d}\lambda$. We will use the following moment assumptions, where the addition cost D is based on a measurable function $\psi\colon \tilde{\mathcal{X}} \to \mathbb{R}$:

(E1) For some $p \geq 2$,

$$\lim_{\delta\to 0}\lim_{k\to\infty} \sup_{g\colon \|g-f\|_p \leq \delta} \sup_{n\in\mathbb{N}} \mathbb{E}\big[D_{\sqrt[d]{n}Y'}\big(\sqrt[d]{n}\mathbf{Y}_n\big)^2 \mathbb{1}\big\{\big|D_{\sqrt[d]{n}Y'}\big(\sqrt[d]{n}\mathbf{Y}_n\big)\big| > k\big\}\big] = 0.$$

(E2) There exist some $R, U \in \mathbb{R}_+$ and $u \geq 0$ such that for any $S \in \tilde{\mathcal{X}}$ and $y \in \mathbb{R}^d$,

$$\big|D_y(S)\big| \leq U\big(1 + \#\big\{S \cap B_y(R)\big\}^u\big).$$

(E1) is primary, describing a moment bound that holds uniformly in the sample size n and distribution G, within a neighborhood of the central data distribution F. Alternatively, (E1) represents a form of uniform integrability. However, the form of (E1) follows purely from technical necessity. A strictly stronger, but more concrete moment condition is as follows.

STATEMENT 2.8. For some $p \geq 2$, there exist $a > 2$ and $\delta > 0$ such that

$$\sup_{g\colon \|g-f\|_p \leq \delta} \sup_{n\in\mathbb{N}} \mathbb{E}\big[\big|D_{\sqrt[d]{n}Y'}\big(\sqrt[d]{n}\mathbf{Y}_n\big)\big|^a\big] < \infty.$$

It may be shown that (E1) follows immediately from Statement 2.8 and Hölder's inequality. Statement 2.8 is closely related to the "uniform bounded moments" condition, Definition 2.2

in [33]. In the context of the topological statistics considered later in this work, (E1) is primarily useful for proof purposes, and is instead established via the intermediate (E2) (see Lemma 2.9), which directly relates the addition cost to a local count within the underlying point set. However, as will be seen with the case of the k-nearest neighbor graph (Corollary 4.7), there exist useful statistics which cannot be deterministically controlled via (E2), and the more general probabilistic condition must be established directly.

LEMMA 2.9. *Let ψ satisfy* (E2). *Then the following hold*:

1. *If $\|f\|_{\max\{2u+1,2\}} < \infty$, then ψ satisfies* (E1).
2. *If $\|f\|_{\max\{p,2\}} < \infty$ for some $p > 2u+1$, then ψ satisfies Statement 2.8.*

Next, we formulate the required stabilization conditions. Recall that $\mathbf{X}_n$ and $\mathbf{Y}_n$ denote the binomial point process with densities f and g, respectively, where f is the fixed density. Furthermore, $X' \sim f$, $Y' \sim g$ are independent of $\mathbf{X}_n$ and $\mathbf{Y}_n$, respectively.

(S1) There exists a sequence $(\epsilon_\delta)_{\delta \geq 0}$ such that $\lim_{\delta \to 0} \epsilon_\delta \to 0$ and

$$\lim_{\delta \to 0} \sup_{g\colon \|g-f\|_2 \leq \delta} \sup_{n \in \mathbb{N}} \mathbb{P}\left[D_{\sqrt[d]{n}Y'}\left(\left(\sqrt[d]{n}\mathbf{Y}_n\right) \cap B_{\sqrt[d]{n}Y'}\left(\sqrt[d]{\frac{\epsilon_\delta}{\delta}} \right) \right) \neq D_{\sqrt[d]{n}Y'}\left(\sqrt[d]{n}\mathbf{Y}_n\right) \right] = 0.$$

(S2) There exist locally-determined radii of stabilization $(\rho_z)_{z \in \mathbb{R}^d}$ for ψ satisfying

$$\lim_{\ell \to \infty} \sup_{n \in \mathbb{N}} \mathbb{P}^*\left[\rho_{\sqrt[d]{n}X'}\left(\sqrt[d]{n}\mathbf{X}_n\right) > \ell\right] = 0.$$

(S1) and (S2) can be summarized as uniform stabilization conditions, either in probability or almost surely. (S1) mainly serves to weaken the necessary conditions providing for bootstrap consistency. We have the following lemma linking (S1) and (S2).

LEMMA 2.10. *Let ψ satisfy* (S2). *Then if $\|f\|_2 < \infty$, ψ satisfies* (S1).

The quantities appearing in (S1) and (S2) can often be greatly simplified. For example, if ψ is translation-invariant, given a radius of stabilization ρ_0 and addition cost D_0 centered at the origin, corresponding quantities can be constructed for any other center point $z \in \mathbb{R}^d$ via translation.

The next lemma provides a convenient tool for "de-Poissonizing" a locally-determined radius of stabilization. Often it is easier to first establish stabilization properties on a homogeneous Poisson process than on a binomial process directly, and Lemma 2.11 allows us to extend Poisson results to the binomial setting, for instance as is required for Lemma 4.1 and Corollary 4.7. Let $\mathbf{P}_\lambda$ denote a homogeneous Poisson process on $\mathbb{R}^d$ with intensity λ.

LEMMA 2.11. *Let ψ be translation-invariant with a locally-determined radius of stabilization ρ_0 and $\|f\|_2 < \infty$. Suppose that for any given $a, b, \delta > 0$ there exist some $L < \infty$ and measurable $A \subset \mathcal{X}$ such that*

$$\rho_0^{-1}\big((L, \infty]\big) \subseteq A \quad \text{and} \quad \sup_{\lambda \in [a,b]} \mathbb{P}[\mathbf{P}_\lambda \in A] \leq \delta.$$

Then for any $\delta > 0$ there exist some $n_0 \in \mathbb{N}$ and $L < \infty$ such that

$$\sup_{n \geq n_0} \mathbb{P}^*\left[\rho_0\left(\sqrt[d]{n}(\mathbf{X}_n - X')\right) > L\right] \leq \delta.$$

Note that the conclusion of Lemma 2.11 is not the same as (S1), only applying for $n \geq n_0$. Some extra effort is required for the conclusion to hold for all $n \in \mathbb{N}$. We come now to an important proposition, the main supporting result for our general bootstrap consistency theorem, Theorem 2.13.

PROPOSITION 2.12. *Let ψ satisfy* (S1) *and* (E1) *with $\|f\|_p < \infty$. Then there exists a coupling between $(X_i)_{i\in\mathbb{N}}$ and $(Y_i)_{i\in\mathbb{N}}$ depending on G such that*

$$\sup_{n\in\mathbb{N}} \operatorname{Var}\left[\frac{1}{\sqrt{n}}\left(\psi\left(\sqrt[d]{n}\mathbf{Y}_n\right) - \psi\left(\sqrt[d]{n}\mathbf{X}_n\right)\right)\right] \leq \gamma\left(\|g - f\|_p\right),$$

where the rate function $\gamma : \mathbb{R}_+ \to \mathbb{R}_+$ is increasing and depends only on f and p such that $\lim_{\delta\to 0}\gamma(\delta) = 0$.

The proof of this result is provided in Section 7. For any two distributions $\mathcal{L}_1$ and $\mathcal{L}_2$ on $\mathbb{R}$, we may define the 2-Wasserstein distance between $\mathcal{L}_1$ and $\mathcal{L}_2$ as

$$W_2(\mathcal{L}_1, \mathcal{L}_2) := \sqrt{\inf_{U\sim\mathcal{L}_1, V\sim\mathcal{L}_2} \mathbb{E}\left[(U - V)^2\right]},$$

where it is assumed that U and V follow a joint distribution with marginals $\mathcal{L}_1$ and $\mathcal{L}_2$. For $\mathcal{L}$ denoting the law or distribution of a random variable, the variance given in the conclusion of Proposition 2.12 is an upper bound for

$$W_2^2\left(\mathcal{L}\left\{n^{-\frac{1}{2}}\left(\psi\left(\sqrt[d]{n}\mathbf{X}_n\right) - \mathbb{E}\left[\psi\left(\sqrt[d]{n}\mathbf{X}_n\right)\right]\right)\right\}, \mathcal{L}\left\{n^{-\frac{1}{2}}\left(\psi\left(\sqrt[d]{n}\mathbf{Y}_n\right) - \mathbb{E}\left[\psi\left(\sqrt[d]{n}\mathbf{Y}_n\right)\right]\right)\right\}\right).$$

Consequently, Proposition 2.12 shows that this W_2-distance can be made arbitrarily small uniformly over a neighborhood of distributions around F. An appropriately smoothed empirical distribution falls within such a small neighborhood with high probability, given sufficiently large sample sizes.

Furthermore, it can be seen that Proposition 2.12 extends directly to finite sums. Given any $(A_i)_{i=1}^k$ and $(B_i)_{i=1}^k$, we have that $\operatorname{Var}[\sum_{i=1}^k A_i - \sum_{i=1}^k B_i] \leq k\sum_{i=1}^k \operatorname{Var}[A_i - B_i]$. Thus, if the conclusion of Proposition 2.12 holds for any finite set of functions, $(\psi_i)_{i=1}^k$, it also holds for $\sum_{i=1}^k \psi_i$, with rate depending on the worst case ψ_i.

It should be noted that (S1) is slightly stronger than necessary to establish Proposition 2.12. As stated, $D_{\sqrt[d]{n}Y'}((\sqrt[d]{n}\mathbf{Y}_n) \cap B_{\sqrt[d]{n}Y'}(l_\epsilon))$ itself is compared to the terminal add-one cost $D_{\sqrt[d]{n}Y'}(\sqrt[d]{n}\mathbf{Y}_n)$. As could be useful for some statistics, it is only required that an appropriate bound displays the desired stabilization property, see the provided proof for details.

2.4. *Smoothed bootstrap.* The bootstrap is an estimation technique used to construct approximate confidence intervals for a given population parameter. In cases where asymptotic approximations for the sampling distribution of a statistic are inconvenient or unavailable, bootstrap estimation provides a general tool for constructing approximate confidence intervals. Bootstrap estimation is well studied in the statistical literature, an introduction being provided in [35]. In this section, we will show consistency for a smoothed bootstrap procedure in estimating the limiting distribution of a standardized stabilizing statistic in the multivariate setting. We describe the general bootstrap procedure below.

Let $\mathbf{X}_n = \{X_i\}_{i=1}^n \overset{\text{iid}}{\sim} F$. We estimate the sampling distribution of

$$\frac{1}{\sqrt{n}}\left(\psi\left(\sqrt[d]{n}\mathbf{X}_n\right) - \mathbb{E}\left[\psi\left(\sqrt[d]{n}\mathbf{X}_n\right)\right]\right)$$

using a plug-in estimator $\hat{F}_n$ for the underlying data distribution F. In the standard nonparametric bootstrap, we estimate F by the empirical distribution, giving probability to each unique value of $\mathbf{X}_n$ proportional to the number of repetitions. We have the bootstrap statistic

$$\frac{1}{\sqrt{m}}\big(\psi\big(\sqrt[d]{m}\mathbf{X}_m^*\big) - \mathbb{E}\big[\psi\big(\sqrt[d]{m}\mathbf{X}_m^*\big)|\mathbf{X}_n\big]\big),$$

where $\mathbf{X}_m^* = \{X_i^*\}_{i=1}^m \overset{\text{iid}}{\sim} \hat{F}_n$, conditional on $\mathbf{X}_n$. The sampling distribution of the bootstrap analog provides an estimate for the distribution of the original statistic, which in the ideal case converges to the truth in the large-sample limit. Confidence intervals for $\mathbb{E}[\psi(\sqrt[d]{n}\mathbf{X}_n)]$ are then constructed from the bootstrap distribution and $\psi(\sqrt[d]{n}\mathbf{X}_n)$.

However, as will be seen in Section 4.1, for some classes of topological statistics the standard bootstrap may not directly replicate the correct sampling distribution asymptotically. Consequently, we instead estimate F by a smoothed distributional approximation. Such a smoothed bootstrap procedure can be shown to provide consistent estimation, even when the standard nonparametric bootstrap may fail.

To define the smoothed bootstrap sampling procedure outlined here, recall first that F has a density $f := \mathrm{d}F/\mathrm{d}\lambda$. Let $\hat{f}_n$ be a given estimator of f derived from $\mathbf{X}_n$ and $\hat{F}_n$ the corresponding probability distribution. Conditional on $\mathbf{X}_n$, we draw bootstrap samples $\mathbf{X}_m^*$ independently from $\hat{F}_n$. A particular choice of $\hat{f}_n$ is given by a kernel density estimator (KDE). For a kernel function $Q\colon \mathbb{R} \to \mathbb{R}$ and bandwidth $h > 0$, the KDE of $f(x)$ based on the sample $(X_i)_{i=1}^n$ is $\hat{f}_{n,h}(x) := 1/(nh^d)\sum_{i=1}^n Q((x - X_i)/h)$.

In practice, when Q corresponds to a probability density, the KDE allows for convenient sampling, as is required in later computational steps. Generating a sample following $\hat{f}_{n,h}$ is equivalent to first drawing from the empirical distribution on $\mathbf{X}_n$, then adding independent noise following the distribution defined by Q, scaled first by the bandwidth h. Other density estimators, including those using higher-order kernels, may not facilitate efficient sampling. However, the theory established here supports the use of any density estimator, which meets the required convergence criteria, implementation difficulties aside. More complicated data-dependent estimators are also possible, falling under a similar sampling framework. See Sections 5 and Appendix A [40] for specifics on density estimation as pertains to this work from a practical perspective. In algorithmic form, the bootstrap procedure for producing a nominal level-γ confidence interval for $\mathbb{E}[\psi(\sqrt[d]{n}\mathbf{X}_n)]$ is as follows:

Smoothed Bootstrap Procedure

1: Given $\mathbf{X}_n = \{X_1, \ldots, X_n\} \sim F$ and $\psi(\mathbf{X}_n)$
2: Generate $\mathbf{X}_{m,1}^*, \ldots, \mathbf{X}_{m,B}^* \sim \hat{F}_n$
3: Calculate $\{\frac{1}{\sqrt{m}}(\psi(\sqrt[d]{m}\mathbf{X}_{m,\ell}^*) - \frac{1}{B}\sum_{j=1}^B \psi(\sqrt[d]{m}\mathbf{X}_{m,j}^*))\}_{\ell=1}^B$
4: Calculate sample quantiles $q(\alpha_1)$, $q(1-\alpha_2)$ such that $\gamma = 1 - \alpha_1 - \alpha_2$
return $(\psi(\sqrt[d]{n}\mathbf{X}_n) - \sqrt{n}q(1-\alpha_2), \psi(\sqrt[d]{n}\mathbf{X}_n) - \sqrt{n}q(\alpha_1))$

Similar algorithms can be used to produce simultaneous coverage sets for multivariate statistics. We now present our main result; the theorem establishes consistency for a smoothed bootstrap in the multivariate setting. The result is given for a vector $\vec{\psi}$ of stabilizing statistics. In the context of the topological statistics introduced in Section 3, this can be the persistent Betti numbers or Euler characteristic evaluated at different filtration parameters or

feature dimensions. Given a probability distribution F on $\mathbb{R}^d$ with density $f := \mathrm{d}F/\mathrm{d}\lambda$, let $(X_i)_{i\in\mathbb{N}} \overset{\text{iid}}{\sim} F$, with $\mathbf{X}_n := \{X_i\}_{i=1}^n$ for any $n \in \mathbb{N}$. $\hat{f}_n$ denotes an estimate of f such that each of the relevant quantities are measurable, and Ψ is a limiting multivariate distribution.

THEOREM 2.13. *Suppose $\vec{\psi}\colon \tilde{\mathcal{X}} \to \mathbb{R}^k$ has component functions $\psi_j\colon \tilde{\mathcal{X}} \to \mathbb{R}$, $1 \le j \le k$ satisfying* (E1) *and* (S1) *with $\|f\|_2 < \infty$. Furthermore, let $\hat{f}_n$ be such that $\|\hat{f}_n - f\|_p \to 0$ in probability (resp., a.s.) as $n \to \infty$. For any $m \in \mathbb{N}$, we have a corresponding bootstrap sample $\mathbf{X}_m^* = \{X_i^*\}_{i=1}^m \overset{\text{iid}}{\sim} \hat{F}_n|\mathbf{X}_n$. Then for any sequence $(m_n)_{n\in\mathbb{N}}$ such that $\lim_{n\to\infty} m_n = \infty$,*

$$\frac{1}{\sqrt{n}}\big(\vec{\psi}\big(\sqrt[d]{n}\mathbf{X}_n\big) - \mathbb{E}\big[\vec{\psi}\big(\sqrt[d]{n}\mathbf{X}_n\big)\big]\big) \xrightarrow{\mathrm{d}} \Psi$$

if and only if

$$\frac{1}{\sqrt{m_n}}\big(\vec{\psi}\big(\sqrt[d]{m_n}\mathbf{X}_{m_n}^*\big) - \mathbb{E}\big[\vec{\psi}\big(\sqrt[d]{m_n}\mathbf{X}_{m_n}^*\big)|\mathbf{X}_n\big]\big) \xrightarrow{\mathrm{d}} \Psi \quad \textit{in probability (resp., a.s.).}$$

Theorem 2.13 establishes the asymptotic validity of bootstrap estimation for a range of stabilizing statistics with only very mild conditions on the underlying density. However, it should be noted that further restrictions on the density and density estimate may be required to satisfy Statement 2.8 and (S1); see Corollary 4.7 for example. Proposition C.1 [40] considers the convergence of $\|\hat{f}_{n,h_n} - f\|_p$ for $p \ge 2$, either in probability or almost surely. This result is outside the main contribution of this paper, but is interesting in its own right. Notably, no conditions are placed on the density f except $\|f\|_p < \infty$.

As a point of caution, it is known that kernel density estimators suffer from a curse of dimensionality. The convergence properties of the density estimator $\hat{f}_n$ appear implicitly within the necessary assumptions for Theorem 2.13. In particular, diminishing performance can be expected in higher dimensions, as shown by the provided simulations of Section 5.

The above result holds for any choice of m_n such that $\lim_{n\to\infty} m_n = \infty$, and is stated as such for the sake of generality. In practical application, $m_n = n$ is standard, and will be used throughout the simulation and data analysis sections of this paper. However, given that the computational complexity of ψ often grows quickly with n, using a smaller m_n could prove more feasible from a computational perspective.

Strictly speaking, convergence to a limiting distribution is not required for the bootstrap to provide asymptotically valid confidence intervals. Proposition 2.12 gives that, with high probability, the smoothed bootstrap and true sampling distributions become close in 2-Wasserstein distance. For $\Psi_n := (\vec{\psi}(\sqrt[d]{n}\mathbf{X}_n) - \mathbb{E}[\vec{\psi}(\sqrt[d]{n}\mathbf{X}_n)])/\sqrt{n}$, provided that the cumulative distribution functions F_{Ψ_n} has the property

$$\lim_{\|\delta\|\to 0} \limsup_{n\to\infty} \sup_{x\in\mathbb{R}^d} \big|F_{\Psi_n}(x+\delta) - F_{\Psi_n}(x)\big| = 0,$$

it can be shown that confidence intervals constructed from the bootstrap statistic still achieve the stated confidence level with high probability, given a sufficiently large sample. Convergence to a continuous limiting CDF is just one way to satisfy this condition. However, this extension is unavailable for the topological statistics considered here, since the distributional behavior of the finite sample statistics is currently very poorly understood.

In the later sections, we will show that the necessary moment and stabilization conditions for Theorem 2.13 are satisfied for several specific statistics of interest, chiefly the Euler characteristic and persistent Betti numbers for a class of simplicial complexes.

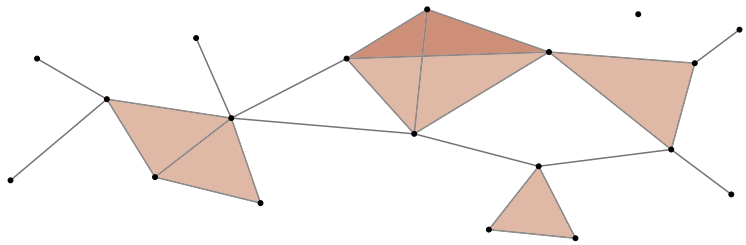

FIG. 1. *Visualization of a simplicial complex. Simplices up to dimension $q = 2$ are included, represented by vertices ($q = 0$), edges ($q = 1$) and triangles ($q = 2$), respectively.*

3. Simplicial complexes and persistence homology.

3.1. *Simplicial complexes.* Given a vertex multiset S, each subset $\sigma = \{x_{i_1}, \ldots, x_{i_{q+1}}\} \subset S$ is called a *q-dimensional simplex* (over S), or simply a q-simplex. An *abstract simplicial complex* K over S is a collection of simplices, such that (i) $\{x\} \in K$ for all $x \in S$, and (ii) if $\sigma \in K$ and $\tau \subset \sigma$ then $\tau \in K$. Notice that geometrically a k-dimensional simplex τ with $\tau \subset \sigma$ can be thought of as a face of σ, meaning that with every simplex all of its faces are included in the complex, and also all the faces of its faces. A filtration of simplicial complexes $\mathcal{K} = \{K^r\}_{r \in \mathbb{R}}$ is a collection of simplicial complexes with $K^r \subseteq K^t$ for $r < t$. For a given simplicial complex K, K_q denotes the subset of K consisting of all q-simplices $\{v_1, \ldots, v_{q+1}\} \subset V$, consisting of $q + 1$ vertices. A graph or network is a simplicial complex consisting of only 1-simplices (edges) and 0-simplices (vertices). A visualization of a simplicial complex can be found in Figure 1, including features of dimension up to $q = 2$.

We will be looking at simplicial complexes constructed over point clouds $S \subset \mathbb{R}^d$. The two prime examples are the Čech and Vietoris–Rips complexes:

$$K^r_{\mathrm{C}}(S) = \{\sigma \subseteq S \colon \exists z \in \mathbb{R}^d \text{ s.t. } \|z - x\| \leq r \ \forall x \in \sigma\},$$
$$K^r_{\mathrm{VR}}(S) = \{\sigma \subseteq S \colon \|x - y\| \leq 2r \ \forall x, y \in \sigma\}.$$

Each of these complexes summarizes the geometric and topological properties within a given point cloud S. The Vietoris–Rips complex can be considered a "completion" of the Čech complex, insomuch that the Vietoris–Rips complex is the largest simplicial complex with the same edge set as the Čech complex. While the primary motivation for the results given here is application to the Čech and Vietoris–Rips complexes, our main results apply for a range of possible complexes. For example, for computational reasons it is often convenient to limit the number of simplices present within the final complex. As such, we have two approximations, the alpha complex and its completion,

$$K^r_{\alpha}(S) = \{\sigma \subseteq S \colon \exists z \in \mathbb{R}^d \text{ s.t. } \|z - x\| \leq r \text{ and } \|z - x\| \leq \|z - y\| \ \forall x \in \sigma \ \forall y \in S\},$$
$$K^r_{\alpha^*}(S) = \{\sigma \subseteq S \colon \{x, y\} \in K^r_{\alpha}(S) \ \forall x, y \in \sigma\}.$$

These complexes avoid adding simplices between disparate points, controlling the total size of the complex. It has been shown that the alpha and Čech complexes share equivalent homology groups. However, for the completion, denoted here as the alpha* complex, there is no such relationship. The alpha complex is a subcomplex of the Čech complex as well as the Delaunay complex,

$$K_{\mathrm{D}}(S) = \{\sigma \subseteq S \colon \exists z \in \mathbb{R}^d \text{ s.t. } \|z - x\| \leq z - y \ \forall x \in \sigma \ \forall y \in S\}.$$

3.2. *Persistent homology.* Now, of chief interest are the topological properties for a given simplicial complex. Both the Čech and Vietoris–Rips complexes reflect the structure present within an underlying point cloud. As such, the topology of each provides an effective summary statistic for describing the structural properties of a data set in $\mathbb{R}^d$. The following provides a short introduction to homology and persistence homology as used in topological data analysis.

Define $C(K)$ to be the free Abelian group generated by the simplices in K. Elements of $C(K)$ are sums of the form $\sum_{i\in I} a_i\sigma_i$, where $\sigma_i \in K$ and, for the purpose of this paper, the coefficients a_i are drawn from the two-element field $\mathbb{F}_2 = \{0, 1\}$. Thus, $C(K)$ is a vector space. $C(K)$ is equipped with a linear boundary operator $\partial\colon C(K) \to C(K)$ where $\partial(\{x_1, \ldots, x_{q+1}\}) = \sum_{i=1}^{q}(-1)^i\{x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_{q+1}\}$. As a fundamental property, $\partial \circ \partial = 0$. With coefficients in $\mathbb{F}_2$, the boundary of a simplex reduces to the sum of all its faces. $C_q(K) = C(K_q)$ is the subspace spanned by the q-simplices of K, with the image of $C_q(K)$ under ∂ lying in $C_{q-1}(K)$. $\partial_q\colon C_q(K) \to C_{q-1}(K)$ denotes the restriction of ∂ to $C_q(K)$.

We now construct the homology groups of K. Let $Z(K) = \ker(\partial)$ be the subspace of $C(K)$ containing the cycles, those elements whose boundary under ∂ is 0. $Z_q(K) = Z(K_q) = \ker(\partial_q)$ is the restriction of $Z(K)$ to dimension q. Let $B(K) = \operatorname{im}(\partial)$ denote the subspace of boundaries in $C(K)$. $B_q(K) = B(K_q) = \operatorname{im}(\partial_{q+1})$ is the subspace consisting of the boundaries of elements in $C_{q+1}(K)$, lying in $C_q(K)$.

The *homology groups* are given by $H_q(K) := Z_q(K)/B_q(K)$, the cycles Z_q in dimension q modulo the boundaries B_q. In words, the elements of the homology groups represent "holes" within the simplicial complex, shown by closed loops whose interior is not filled by other elements in the complex. These homology groups provide a topological summary of the structure in the simplicial complex K. As stated previously, because we assume field coefficients for $C(K)$, each homology group is also a vector space. The *Betti numbers* of the complex represent the degree or dimension of each homology space. We denote the qth Betti number of K by $\beta_q(K) = \dim(Z_q(K)/B_q(K)) = \dim(Z_q(K)) - \dim(B_q(K))$. Moving forward, Betti numbers and their like will be of primary interest.

Homology provides a topological invariant constructed from a single simplicial complex. For a filtration of nested simplicial complexes, *persistent homology* provides more detail. Given a filtration $\mathcal{K} = \{K^r\}_{r\in\mathbb{R}}$, the homology groups for each complex, $H_q(K^r)$, are defined. However, due to the nested structure of the filtration, simplices are shared across complexes, and thus there exists a natural inclusion map between homology spaces. Cycles in $Z_q(K^r)$ are also cycles in $Z_q(K^t)$ if $r < t$. The boundary spaces behave similarly. For a given equivalence class, $x + B_q(K^r) \in H_q(K^r)$, $x + B_q(K^r) \to x + B_q(K^t)$ specifies the inclusion map from $H_q(K^r)$ to $H_q(K^t)$.

If a given element $\tilde{x} \in H_q(K^r)$ maps to $\tilde{y} \in H_q(K^t)$ upon inclusion, with $\tilde{y} \neq B_q(K^t)$, we say that $\tilde{x}$ represents a persistent cycle across the filtration. Essentially, the same underlying element is reflected in the homology groups over a range of simplicial complexes. The collection of homology groups and inclusion maps form a *persistence module*. A wide body of work exists on the properties of these persistence modules; see [49] for an introduction. For any cycle feature in the filtration, there is a well-defined death time, being the smallest parameter level for which the given element lies in the kernel. The Betti numbers of a filtration form a function in the filtration parameter, r. We use the notation $\beta_q^r(\mathcal{K}) := \beta_q(K^r)$. The Betti numbers in this context count the number of persistent features extant at r.

It is a fundamental theorem of persistent homology that a sufficiently well-behaved persistence module can be represented by a *persistence diagram*. A diagram $\mathcal{D}(\mathcal{K})$ is a multiset in $\mathbb{R}^2 \times \mathbb{N}_0$ of points (b, d, q). Each point represents a single persistent feature in the module. b denotes the birth time of the feature, being the smallest parameter level for which that feature

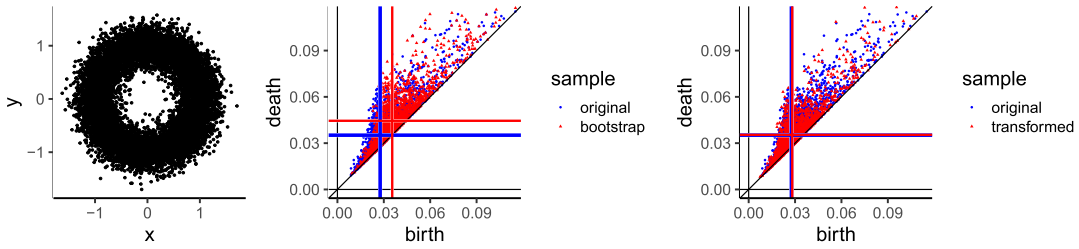

FIG. 2. Left: *The original data set of size* $n = 10{,}000$, *from which a single standard bootstrap sample is drawn.* Middle: *Persistence diagrams for both the original and bootstrap samples, along with lines denoting the median birth and death in each diagram. The asymptotic bias discussed in Section* 4.1 *can be clearly seen.* Right: *Persistence diagrams after application of a multiplicative correction factor of* $\sqrt{1-e^{-1}} \approx 0.795$ *to the bootstrap sample. Note that the median birth and death values correspond after this transformation is applied.*

is represented in the homology groups. Likewise, d gives the death time, and q the dimension of the feature. The collection of persistent features represented by the diagram are a basis for the corresponding persistence module.

The persistence diagram is a simple summary statistic, which condenses the complex topological information present within a filtration. An example of a persistence diagram is shown in Figure 2.

3.3. *Persistent Betti numbers.* We arrive at the main focus of this section. For $r \le s$, define the *persistent homology groups* of a filtration $\mathcal{K} = \{K^r\}_{r\in\mathbb{R}}$ as

$$H_q^{r,s}(\mathcal{K}) := Z_q(K^r)/\big(B_q(K^s) \cap Z_q(K^r)\big).$$

Nonzero elements in this group represent features born at or before time r, which persist until at least time s. The dimension of these spaces gives the *persistent Betti numbers*,

$$\begin{aligned}\beta_q^{r,s}(K) &:= \dim\big(Z_q(K^r)/B_q(K^s) \cap Z_q(K^r)\big)\\ &= \dim\big(Z_q(K^r)\big) - \dim\big(B_q(K^s) \cap Z_q(K^r)\big).\end{aligned}$$

Persistent Betti numbers are in one-to-one correspondence with the respective persistence diagram. Here, $\beta_q^{r,s}(\mathcal{K})$ counts the number of points in $\mathcal{D}(\mathcal{K})$ of feature dimension q falling within $(-\infty, r] \times (s, \infty]$. When $s = r$, we recover the regular Betti numbers, $\beta_q^{r,r}(\mathcal{K}) = \beta_q(K^r)$. An important result for persistent Betti numbers is given in the following lemma.

LEMMA 3.1 (Geometric lemma (Lemma 2.11 [22])). *Let* $\mathcal{J} = \{J^r\}_{r\in\mathbb{R}}$ *and* $\mathcal{K} = \{K^r\}_{r\in\mathbb{R}}$ *be filtrations of simplicial complexes with* $J^r \subseteq K^r$ *for all* $r \in \mathbb{R}$. *Then*

$$\begin{aligned}\big|\beta_q^{r,s}(\mathcal{K}) - \beta_q^{r,s}(\mathcal{J})\big| &\le \max\big\{\#\big\{K_q^r \setminus J_q^r\big\}, \#\big\{K_{q+1}^s \setminus J_{q+1}^s\big\}\big\}\\ &\le \#\big\{K_q^r \setminus J_q^r\big\} + \#\big\{K_{q+1}^s \setminus J_{q+1}^s\big\}.\end{aligned}$$

The geometric Lemma 3.1 relates the change in persistent Betti numbers between two filtrations to the additional simplices gained moving between them. As a brief explanation of the lemma, simplices can be divided into two classes, positive and negative. For two simplicial complexes $J \subset K$, if we imagine adding the additional q-simplices in K to J one by one, a positive q-simplex will increase the dimension of Z_q by one, and a negative q-simplex will increase the dimension of B_{q-1} by one. Either change can affect the persistent Betti numbers. This dichotomy is a basic result from persistent homology (see [6]). The bound given in the geometric lemma describes a worst case, when all q-simplices at time r are positive or all $(q+1)$-simplices at time s are negative. The geometric lemma will be critical moving forward, as it allows us to control the change in persistent Betti numbers by counting appropriate simplices.

3.4. *Euler characteristic.* For a given simplicial complex K, the *Euler characteristic* is defined as

$$\chi(K) := \sum_{k=0}^{\infty} (-1)^k \#\{K_k\}.$$

Provided there is an $m \in \mathbb{N}$ such that the Betti numbers $\beta_q(K)$ are 0 for all $q > m$ (as in (D4) holds), it can be shown that the Euler characteristic has the following identity with the Betti numbers:

$$\chi(K) = \sum_{k=0}^{\infty} (-1)^k \beta_k(K).$$

This relationship with the Betti numbers makes the Euler characteristic an important topological invariant in its own right. Applications of the Euler characteristic and derivatives may be found in [36, 38, 43].

3.5. *k-nearest neighbor graph.* The k-nearest neighbor graph $\mathcal{K}_{\mathrm{NN},k}$ of a vertex set S connects each point $x \in S$ with the k closest vertices to x within $\mathbf{S} \setminus x$. This graph may either be directed or undirected. $\mathcal{K}_{\mathrm{NN},k}$ is commonly used to analyze the clustering structure of a point cloud. Let the total length of the edges in this graph be denoted by $l_{\mathrm{NN},k}$. The total length of the k-nearest neighbor graph, when suitably scaled, provides a measure of the average local "density," or concentration of the points in S. In Section 4.5, we will show bootstrap consistency for $l_{\mathrm{NN},k}$ within the stabilization framework.

4. Bootstrapping topological statistics.

4.1. *Nonparametric bootstrap.* In this section, we will argue that the standard nonparametric bootstrap may fail to reproduce the correct sampling distribution asymptotically when applied to common topological statistics.

For a wide class of simplicial complexes built over point sets in $\mathbb{R}^d$, the corresponding persistence diagram is unaffected by the inclusion of repeated points within the vertex set. This behavior holds for both the Vietoris–Rips and Čech complexes, defined in Section 3.1. In the case of the Čech complex, this phenomenon is seen most directly. The Čech complex under the Euclidean metric is homologically equivalent to a union of closed balls centered on the corresponding vertex points in $\mathbb{R}^d$. Additional repetitions within the set of vertex points do not affect the union, and thus to do change the derived persistence diagram.

In cases like this where repetitions may be ignored in statistic calculations, the standard bootstrap behaves effectively like a subsampling technique. The size of a given subsample is random, equal to the number of unique points present in the corresponding bootstrap sample.

Given a random sample $\mathbf{X}_n = \{X_1, \ldots, X_n\}$, it can be shown using elementary arguments that a given bootstrap sample $\mathbf{X}_n^*$ of size n from the empirical distribution over $\mathbf{X}_n$ is expected to contain $n(1 - (1 - 1/n)^n) \approx (1 - e^{-1})n \approx 0.632n$ unique points. As such, $\mathbf{X}_n^*$ behaves similar to a sample of size $0.632n$, but is not scaled accordingly within the statistic $(\beta_q^{r,s}(\sqrt[d]{n}\mathbf{X}_n^*) - \mathbb{E}[\beta_q^{r,s}(\sqrt[d]{n}\mathbf{X}_n^*)|\mathbf{X}_n])/\sqrt[d]{n}$. This discrepancy in scaling introduces a non-negligible asymptotic bias. The effect is illustrated in Figure 2 for the Vietoris–Rips complex.

Furthermore, the standard nonparametric bootstrap results in a fundamentally different point process limit at small scales when compared to the original data generating mechanism. For the original sample, when $\mathbf{X}_n$ is drawn from a distribution with density f, the shifted and rescaled sample $\sqrt[d]{n}(\mathbf{X}_n - z)$ approaches a homogeneous Poisson process $\mathbf{P}_z$ with

intensity $f(z)$. From the preceding stabilization literature ([27, 33]), this limiting local point process drives the asymptotic sampling distribution of $(\beta_q^{r,s}(\sqrt[d]{n}\mathbf{X}_n) - \mathbb{E}[\beta_q^{r,s}(\sqrt[d]{n}\mathbf{X}_n)])/\sqrt[d]{n}$. Considering the large-sample behavior of $\sqrt[d]{n}(\mathbf{X}_n^* - z)|\mathbf{X}_n$, the smoothed bootstrap sampling procedure described in Section 2.4 can be shown to reproduce the same local Poisson process $\mathbf{P}_z$ asymptotically.

However, the same is not true for the standard bootstrap when repeated points are ignored. In this case, $\sqrt[d]{n}(\mathbf{X}_n^* - z)|\mathbf{X}_n$ is restricted to the discrete set $\sqrt[d]{n}(\mathbf{X}_n - z)$, and thus cannot reproduce $\mathbf{P}_z$, whose domain is $\mathbb{R}^d$. For this case, we describe the resulting point process limit $\mathbf{Q}_z$ in two steps. First, a homogenous Poisson process $\mathbf{P}_z$ is generated, representing $\sqrt[d]{n}(\mathbf{X}_n - z)$. Defined conditionally, $\mathbf{Q}_z$ is a random subset of $\mathbf{P}_z$ such that $\mathbb{P}[x \in \mathbf{Q}_z|\mathbf{P}_z] = 1 - e^{-1} \approx 0.632$, considering each point $x \in \mathbf{P}_z$ independently. We have $\sqrt[d]{n}(\mathbf{X}_n^* - z) \xrightarrow{\mathrm{d}} \mathbf{Q}_z$.

This difference in local behavior, combined with the asymptotic bias effect illustrated earlier, is a strong indicator that $(\beta_q^{r,s}(\sqrt[d]{n}\mathbf{X}_n^*) - \mathbb{E}[\beta_q^{r,s}(\sqrt[d]{n}\mathbf{X}_n^*)]|\mathbf{X}_n)/\sqrt[d]{n}$ and $(\beta_q^{r,s}(\sqrt[d]{n}\mathbf{X}_n) - \mathbb{E}[\beta_q^{r,s}(\sqrt[d]{n}\mathbf{X}_n)])/\sqrt[d]{n}$ likely do not share a weak limit. A technical treatment is omitted here, and is outlined merely to justify the use of our smoothed bootstrap procedure in place of the standard method. The smoothed bootstrap procedure provides for bootstrap consistency (Corollaries 4.3 and 4.4), and in the following sections we consider only this approach.

4.2. *General conditions for simplicial complexes*. The results presented in the following sections apply for a range of simplicial complexes constructed over point clouds in $\mathbb{R}^d$. Here, we will explain the specific conditions used, and for which common simplicial complexes they apply. Let K be a function taking as input $S \in \tilde{\mathcal{X}}(\mathbb{R}^d)$, giving as output a simplicial complex with vertices in S. For a given simplex σ, let the set diameter be $\mathrm{diam}(\sigma)$. We have the following conditions:

(K1) For any $S \in \tilde{\mathcal{X}}(\mathbb{R}^d)$ and $z \notin S$, $K(S) \subseteq K(S \cup \{z\})$. Furthermore, $\sigma \in K(S \cup \{z\}) \setminus K(S)$ only if $z \in \sigma$.

(K2) For any $S \in \tilde{\mathcal{X}}(\mathbb{R}^d)$ and $z \in \mathbb{R}^d$, $\sigma \in K(S)$ only if $\sigma - z \in K(S - z)$.

(D1) There exists $\phi < \infty$ such that for any $S \in \tilde{\mathcal{X}}(\mathbb{R}^d)$, $\sigma \in K(S)$ only if $\mathrm{diam}(\sigma) \leq \phi$.

(D2) There exists $\phi < \infty$ such that for any $S \in \tilde{\mathcal{X}}(\mathbb{R}^d)$ and $z \in \mathbb{R}^d$, $\sigma \in K(S \cup \{z\}) \triangle K(S)$ only if $\sigma \subset B_z(\phi)$.

(D3) There exists an $\eta > 0$ such that for any $S \in \tilde{\mathcal{X}}(\mathbb{R}^d)$ and $x \in Z(K(S))$, $\mathrm{diam}(x) \leq \eta$ only if $x \in B(K(S))$.

(D4) There exists an $m \in \mathbb{N}$ such that for any $k > m$ and $S \in \tilde{\mathcal{X}}(\mathbb{R}^d)$, $Z_k(K(S)) = B_k(K(S))$.

(K1) means that the addition of a new point will not change the existing complex, only add new simplices. Furthermore, any new simplices gained must contain the added point as a vertex. (K2) gives that the complex is essentially translation invariant. (D1) sets a maximum diameter for any simplex in the complex. (D2) gives that the influence of a new point on the complex is confined to a local region around that point, within a fixed diameter. This condition allows for both the addition and removal of simplices from the complex, but only within the prescribed radius. It can be easily shown that if (D2) holds for ϕ, (D1) holds for 2ϕ. Conversely, if both (K1) and (D1) hold for ϕ, (D2) also holds for ϕ. Finally, (D3) gives that no small loops can exist with unfilled interiors, and (D4) gives that all Betti numbers are 0 in sufficiently high feature dimensions.

Now, let $\mathcal{K} = (K^r)_{r \in \mathbb{R}}$ be a function taking as input $S \in \tilde{\mathcal{X}}(\mathbb{R}^d)$, giving as output a filtration of simplicial complexes with vertices in S. As a slight abuse, we will often refer to the function $\mathcal{K}$ as a filtration of simplicial complexes, even though it is a function defining more than a single filtration, depending on the underlying point cloud. We say that a given

condition is satisfied for $\mathcal{K}$ if it is satisfied by K^r for any $r \in \mathbb{R}$. In the cases of (D1), (D2) and (D3), ϕ and η may depend on r as increasing functions $\phi\colon \mathbb{R} \to [0, \infty)$ and $\eta\colon \mathbb{R} \to [0, \infty)$.

It can be shown that all of (K1)–(D3) are satisfied for both the Vietoris–Rips complex in $\mathbb{R}^d$ using $\phi(r) = \eta(r) = 2r$ and for the Čech complex using $\phi(r) = 2r$, $\eta(r) = r$. The functions for the Čech complex are established via interleaving with the VR complex; see [18] for details. The same functions apply for the alpha complex in $\mathbb{R}^d$ and its completion $\mathcal{K}_{\alpha^*}$, with the notable exception that (K1) is violated. Finally, it is known that (D4) is satisfied by the alpha, Čech and Delaunay complexes in $\mathbb{R}^d$ for $m = d - 1$.

While covering a wide class of distance-based simplicial complexes, there are several complexes used in practice that may fail to satisfy any or all of these. For example, the addition of a new point to the Delaunay complex, Gabriel graph, witness complex or k-nearest neighbor graph can both add and remove simplices, violating (K1). Furthermore, there is not any limit on the simplex diameter within any of these complexes, violating (D1). Likewise, the addition of a single point can alter simplices at arbitrarily large distances, violating (D2). As a special note, it is common in practice to consider the intersection of the Vietoris–Rips and Delaunay complexes, which unfortunately may violate all the assumptions here. It is unclear if an extension or special consideration could be made to incorporate these complexes.

4.3. *Stabilization of persistent Betti numbers.* To apply the general bootstrap theorem, we first require a technical lemma establishing a locally-determined radius of stabilization for persistent Betti numbers. The result given applies for general classes of simplicial complexes constructed over subsets of $\mathbb{R}^d$, under the same conditions stated in Section 4.2. The following apply for a probability distribution F on $\mathbb{R}^d$ with density f and a filtration of simplicial complexes $\mathcal{K} = \{K^r\}_{r\in\mathbb{R}}$. We have the following.

LEMMA 4.1. *Let* $\|f\|_2 < \infty$ *and* $\mathcal{K}$ *satisfy* (K2), (D2) *and* (D3). *Then* $\beta_q^{r,s}(\mathcal{K})$ *satisfies* (S2) *for any* $r \in \mathbb{R}$, $s \in \mathbb{R}$ *and* $q \geq 0$.

4.4. *Bootstrap results for persistence homology.* Here, we present the main applied results of this paper. Each is derived from Theorem 2.13 and the stabilization lemma for persistent Betti numbers (Lemma 4.1). For given vectors of birth and death times, $\vec{r} = (r_i)_{i=1}^k$ and $\vec{s} = (s_i)_{i=1}^k$, let $\beta_q^{\vec{r},\vec{s}} = (\beta_q^{r_i,s_i})_{i=1}^k$ denote the multivariate function whose components are the persistent Betti numbers evaluated at each pair of birth and death times. For a vector of filtration times $\vec{r} = (r_i)_{i=1}^k$, let $\chi^{\vec{r}}$ denote the function giving the Euler characteristic at each time r_i, with $\chi^{\vec{r}} := (\chi(K^{r_i}))_{i=1}^k$.

The following apply for a given multivariate statistic $\vec{\psi}$ and a probability distribution F on $\mathbb{R}^d$ with density $f := \mathrm{d}F/\mathrm{d}\lambda$ such that $\|f\|_p < \infty$ for some specified $p \geq 2$. $\hat{F}_n$ is a random distributional estimate with density $\hat{f}_n := \mathrm{d}\hat{F}_n/\mathrm{d}\lambda$ such that $\|\hat{f}_n - f\|_p \to 0$ in probability (or almost surely). $(X_i)_{i\in\mathbb{N}} \overset{\text{iid}}{\sim} F$, and $\mathbf{X}_n := \{X_i\}_{i=1}^n$ for any $n \in \mathbb{N}$. $(X_{n,i}^*)_{i\in\mathbb{N}} \overset{\text{iid}}{\sim} \hat{F}_n$ is a conditionally independent sequence of bootstrap samples drawn from $\hat{F}_n$, and $\mathbf{X}_{n,m} := \{X_{i,n}^*\}_{i=1}^m$ for any $m, n \in \mathbb{N}$. Ψ denotes a limiting multivariate distribution, and $(m_n)_{n\in\mathbb{N}}$ is any sequence such that $\lim_{n\to\infty} m_n = \infty$. Recalling the conclusion of Theorem 2.13, we have the following.

STATEMENT 4.2.

$$\frac{1}{\sqrt{n}}\big(\vec{\psi}\big(\sqrt[d]{n}\mathbf{X}_n\big) - \mathbb{E}\big[\vec{\psi}\big(\sqrt[d]{n}\mathbf{X}_n\big)\big]\big) \xrightarrow{\mathrm{d}} \Psi$$

if and only if

$$\frac{1}{\sqrt{m_n}}\big(\vec{\psi}\big(\sqrt[d]{m_n}\mathbf{X}_{n,m_n}^*\big) - \mathbb{E}\big[\vec{\psi}\big(\sqrt[d]{m_n}\mathbf{X}_{n,m_n}^*\big)\,|\,\hat{F}_n\big]\big) \xrightarrow{\mathrm{d}} \Psi \quad \text{in probability (resp., a.s.).}$$

For cases with a corresponding central limit theorem, Ψ is a limiting normal distribution. For each of the following, $\mathcal{K} = (K^r)_{r\in\mathbb{R}}$ is a filtration of simplicial complexes.

COROLLARY 4.3 (Persistent Betti numbers). *Let $q \geq 0$ and $p \geq 2q+3$. Let $\mathcal{K}$ satisfy* (K1), (K2), (D1) *and* (D3). *Then for any given $\vec{r}$, $\vec{s}$, Statement 4.2 holds for $\beta_q^{\vec{r},\vec{s}}$.*

COROLLARY 4.4 (Persistent Betti numbers—Alt.). *Let $q \geq 0$ and $p \geq 2q+5$. Let $\mathcal{K}$ satisfy* (K2), (D2) *and* (D3). *Then for any given $\vec{r}$, $\vec{s}$, Statement 4.2 holds for $\beta_q^{\vec{r},\vec{s}}$.*

The only differences between the above corollaries are the conditions satisfied by the underlying simplicial complex and the necessary norm bound on the density. The corresponding results for Betti numbers follow as special cases of Corollaries 4.3 and 4.4, when the given birth and death parameters are equal ($\beta_q^{\vec{r}} = \beta_q^{\vec{r},\vec{r}}$). Also, although the statements of Corollaries 4.3 and 4.4 are given in terms of a fixed feature dimension q, a direct extension exists if $q = q_i$ is allowed to differ for each (r_i, s_i). The form as given shows the dependence of the density norm assumption on the chosen feature dimension.

Note, throughout this work, including Corollaries 4.3 and 4.4, the asymptotic regime we consider consists of a fixed statistic and a rescaled underlying sample. However, for the persistent Betti numbers of the Čech and Vietoris–Rips complexes, we can equivalently shift the scaling factor from the sample to the filtration parameters $(\vec{r}_n, \vec{s}_n) = (\vec{r}/\sqrt[d]{n}, \vec{s}/\sqrt[d]{n})$.

The higher value of p required in Corollary 4.4 compared to Corollary 4.3 can be explained intuitively based on the assumptions used. For the persistent Betti numbers, the main quantity controlling convergence is the expected number of simplices altered or introduced when a new data point is added to the sample. (D2) ensures that these simplices fall within a small ball around the new data point. The stated density norm conditions control the expected number of points, and by extension possible simplices, that can lie within that small ball. Introducing (K1) further controls the number of possible simplices, and allows for a weakening of the necessary norm condition. (K1) requires that, as the sample grows by a single point, any additional simplices must contain the new point as a vertex, and no deletion of simplices is possible. This means that every added simplex has one less "free" vertex, and a weaker norm condition is required for control. The same intuition applies whenever (K1) is assumed.

In the specific case of the alpha complex, both of the above Corollaries 4.3 and 4.4 apply. While the alpha complex does not satisfy (K1), it has equal persistent Betti numbers to the Čech complex, which does. Thus, the weaker conditions of Corollary 4.3 are sufficient in this unique case.

COROLLARY 4.5 (Euler characteristic). *Let $m < \infty$ and $p \geq 2m+3$. Let $\mathcal{K}$ be a filtration of simplicial complexes satisfying* (K1), (K2), (D1), (D3) *and* (D4). *Then for any given $\vec{r}$, Statement 4.2 holds for $\chi^{\vec{r}}$.*

COROLLARY 4.6 (Euler characteristic—Alt.). *Let $m < \infty$ and $p \geq 2m+5$. Let $\mathcal{K}$ be a filtration of simplicial complexes satisfying* (K2), (D2), (D3) *and* (D4). *Then for any given $\vec{r}$, Statement 4.2 holds for $\chi^{\vec{r}}$.*

It is suspected that some of the simplicial complex assumptions can be relaxed in the persistent Betti number and Euler characteristic cases, but the extent to which this is possible is still unknown. Specifically, Corollary 4.3 requires a translation-invariant simplicial complex (K2), along with the elimination of small loops via (D3). Furthermore, (D4) is necessary for the Euler characteristic to grow polynomially, as in (E2). See [26] for an analysis of the Euler characteristic where this assumption can be relaxed.

To strengthen Corollaries 4.3–4.6 with rates, we require more specific knowledge about the convergence to G of the original statistic. For persistent Betti numbers in the multivariate setting, general central limit theorems have been shown in [27], but little is known at this time with regards to rates of convergence. Proposition 2.12 and Statement 2.8 together do allow for rates of convergence in 2-Wasserstein distance between the bootstrap and true sampling distributions for finite sample sizes, but is phrased in terms of a tail probability for the radius of stabilization. See the proofs of Corollaries 4.3–4.6 for details. For persistent Betti numbers, the tail behavior of the radius of stabilization is poorly understood. Owing to these difficulties, we may only conclude consistency of the smoothed bootstrap for the functions considered.

4.5. *Bootstrap results for k-nearest neighbor graphs.* In the following, let $\mathcal{D}_{\gamma,r_0}(C)$ be the class of distributions G with support on a bounded $C \subset \mathbb{R}^d$ such that $\int_{B_x(r)} \mathrm{d}G \geq \gamma r^d$ for all $r \leq r_0$ and $x \in C$ for some $\gamma > 0$. This set of criteria is widely used and sometimes known as the "standard asssumption" for probability measures (see [14, 33]).

COROLLARY 4.7 (Total edge length of the k-nearest neighbor graph). *Let $p > 2$. Furthermore, let $F \in \mathcal{D}_{\gamma,r_0}(C)$ and $\mathbb{1}\{\hat{F}_n \in \mathcal{D}_{\gamma,r_0}(C)\} \to 1$ in probability (resp., a.s.). Then Statement 4.2 holds for $l_{\mathrm{NN},k}$.*

The conditions of Corollary 4.7 are in particular satisfied when C is known and convex, with f bounded below on C by a positive constant, provided further that $\|\hat{f}_n - f\|_\infty \to 0$ in probability (resp., a.s.). We include this final result to demonstrate the utility of stabilization as a general tool for proving bootstrap convergence theorems outside of topological data analysis. The k-nearest neighbor graph does not fall under the general simplicial complex conditions provided in Section 4.2, thus special treatment is needed to show the required stabilization and moment conditions. Here, we rely on previous results from the literature; see [33] for stabilization results and the corresponding central limit theorem.

5. Simulation study. In this section, we present the results of a series of simulations illustrating the finite-sample properties of the smoothed bootstrap applied to persistent Betti numbers $\beta_q^{r,s}$ of the Vietoris–Rips complex constructed over point sets in $\mathbb{R}^d$. Precise definitions and an introduction to the properties of these statistics may be found in Section 3. Source code is available at github.com/btroycraft/stabilizing_statistics_bootstrap [39].

We investigate the coverage probability of bootstrap confidence intervals on the expected persistent Betti numbers $\mathbb{E}[\beta_q^{r,s}(\sqrt[d]{n}\mathbf{X}_n)]$ for a variety of feature dimensions, sample sizes, data generating mechanisms and bandwidth selectors. Table 1 lists brief descriptions of the

TABLE 1

Description of densities or distributions considered for the simulation study of Section 5. For the distributions based on manifolds, we first draw uniformly from the manifold, then apply the prescribed additive noise. Detailed explanations of the distributions considered, along with precise definitions are provided in Appendix D [40]

Label	Description
F_1	Rotationally symmetric in $\mathbb{R}^2$, finite L_8 norm
F_2	Rotationally symmetric in $\mathbb{R}^2$, finite L_2 norm, infinite L_8 norm
F_3	$\mathbb{S}^1$ embedded in $\mathbb{R}^2$, additive Gaussian noise
F_4	Uniformly distributed over $B_0(1)$ in $\mathbb{R}^3$, additive Gaussian noise
F_5	5 clusters in $\mathbb{R}^3$, additive exponential noise
F_6	$\mathbb{S}^2$ embedded in $\mathbb{R}^5$, additive Cauchy noise
F_7	Flat figure-8 embedded in $\mathbb{R}^{10}$, additive Gaussian noise

TABLE 2

Coverage proportions for 95% smoothed bootstrap confidence intervals on the mean persistent Betti numbers; coverage is estimated using $N = 1000$ independent base samples with $B = 500$ bootstrap samples each. True mean persistent Betti numbers are estimated using a large ($N = 100{,}000$) number of independent samples from the true distribution. For each case, the values from top to bottom: Coverage proportions using "Hpi.diag," "Hlscv.diag," "Hscv.diag" and "bw.silv" bandwidth selectors, respectively (see Section 5)

	F_1	F_2	F_3	F_4	F_5	F_6	F_7	F_4	F_5	F_6	F_7
Distr.				$q = 1$				$q = 2$			
r	4.94	5.20	3.03	1.92	0.30	1.78	1.28	2.96	0.39	2.71	1.46
s	5.36	5.60	3.28	2.12	0.31	1.91	1.32	3.04	0.40	2.80	1.47
$n = 100$	0.896	0.965	0.921	0.859	0.954	0.19		0.908	0.705	0.038	
	0.931	0.959	0.914	0.809	0.941	0.133		0.903	0.604	0.045	
	0.903	0.97	0.91	0.859	0.927	0.049		0.902	0.363	0.002	
	0.359	0.931	0.942	0.864	0	0	0.656	0.902	0	0	0.045
$n = 200$	0.908	0.971	0.94	0.898	0.942	0.159		0.878	0.795	0.125	
	0.92	0.972	0.946	0.891	0.923	0.106		0.872	0.707	0.074	
	0.888	0.975	0.959	0.906	0.892	0.06		0.908	0.277	0.031	
	0.299	0.954	0.903	0.899	0	0	0.766	0.882	0	0	0.537
$n = 300$	0.9	0.971	0.926	0.921	0.94	0.183		0.854	0.906	0.225	
	0.94	0.971	0.938	0.896	0.94	0.087		0.854	0.917	0.072	
	0.913	0.971	0.94	0.896	0.922	0.054		0.855	0.964	0.074	
	0.283	0.956	0.925	0.906	0	0	0.835	0.856	0	0	0.508
$n = 400$	0.918	0.961	0.947	0.934	0.96	0.175		0.851	0.883	0.259	
	0.927	0.951	0.938	0.92	0.955	0.063		0.839	0.88	0.076	
	0.908	0.976	0.933	0.924	0.939	0.062		0.863	0.958	0.099	
	0.266	0.961	0.909	0.922	0.114	0	0.891	0.859	0	0	0.584

data distributions considered. For more detailed explanations, see Appendix D [40]. Simulation results are given in Table 2. For the persistent Betti numbers, a single choice of (r, s) was made for each combination of distribution and feature dimension, chosen to lie within the main body of features in the corresponding persistence diagram. For computational reasons, only feature dimensions $q = 1$ and $q = 2$ are considered.

We consider four data-driven bandwidth selectors. First are the "Hpi.diag" (plug-in), "Hlscv.diag" (least-squares cross-validation) and "Hscv.diag" (smoothed cross-validation) selectors from the *ks* package in R. Each of these selectors are available for data dimension up to $d = 6$. Last, we consider Silverman's rule of thumb (see [41]) via "bw.silv" from the *kernelboot* package in R, which accepts data in any dimension.

For the two cross-validation selectors, note that a bandwidth is not always selected, throwing errors on some data sets. To accommodate the automatic setting of this simulation study, any error-producing data sets were simply rejected for each of these cases.

There is a noticeable drop-off in coverage as the data dimension increases. This is expected, as the kernel density estimator is known to suffer from a "curse of dimensionality." Furthermore, there is a similar decrease for increasing feature dimension, as well. This is also expected, because the W_2-convergence rate bounds of Proposition 2.12 are slower with increased feature dimension (see Appendix B.5 [40] for details.)

For distribution F_6, which exhibits heavy tails, performance generally is very low, due to poor performance of the underlying selectors. It is likely that performance will suffer generally in the presence of heavy-tailed data when using one of these common selectors which does not account for the prevalence of outliers in heavy-tailed data. Furthermore, error due to boundary effects can be expected for highly discontinuous distributions, for example, as

seen in the case of F_4 in feature dimension $q = 2$. Here, the probability mass near the support boundary is more highly spread after smoothing than in the original distribution, leading to an upward shift in the scale of the associated topological features and a corresponding bias in the persistent Betti numbers.

The coverage proportion is generally smaller than the nominal level of 95%. Therefore, it is recommended to use a larger than desired level, especially for limited sample sizes. In terms of general performance, we recommend any of "Hpi.diag," "Hlscv.diag" or "Hscv.diag." These selectors provide the most consistent coverage, and effectively replicate the nominal 95% level in many cases, especially for the largest sample size $n = 400$. Silverman's rule performs badly in several cases, and should only be used in the absence of better alternatives.

A real data application using Galaxy data can be found in Appendix A [40].

6. Discussion. In this work, we have shown the large-sample consistency of multivariate bootstrap estimation for a range of stabilizing statistics. This includes the persistent Betti numbers, the Euler characteristic and the total edge length of the k-nearest neighbor graph. However, many open questions still remain.

In Section 4.1, it was argued that the standard nonparametric bootstrap may fail to directly reproduce the correct sampling distribution asymptotically for topological statistics like the persistent Betti numbers. However, there remains the possibility that a corrected version of the standard bootstrap could provide for consistency. As discussed in Section 4.1, standard bootstrap sampling results in a fundamentally different point-process limit at small scales. Previous stabilization results primarily consider Poisson and related processes, meaning a full theoretical treatment of the standard bootstrap would likely require reconstructing much of the previous stabilization and central limit theorem results for the alternative limiting process.

The results for the smoothed bootstrap presented here apply only in the multivariate setting, the obvious extension being to stochastic processes. Essential to a process-level result concerning the persistent Betti numbers would be a convenient tail bound for the radius of stabilization, which is yet unavailable. In the case of persistent Betti numbers, there is a strong relationship between the persistent Betti function and an empirical CDF in two dimensions. As such, there is much established theory in that regard, which may be applied once stochastic equicontinuity is established.

In practice, it is common that data comes not from a density in $\mathbb{R}^d$, but instead from a manifold. It is suspected that a version of the results in this paper could apply in the manifold setting. However, this requires a bootstrap that adapts to a possibly unknown manifold structure, similar to that found in [23]. Combined with the inherent challenges of working with manifolds, this extension presents many technical hurdles. Alternatively, variance estimation using subsampling or the jackknife may provide for consistent variance estimation in cases where the support manifold of the data distribution is not known a priori. Subsampling in this data context was initially developed in [21]; however, the resulting confidence sets are given for the persistence diagram of the manifold, as opposed to the expected persistent Betti number considered in this work. Furthermore, these confidence sets are conservative, stemming from their reliance on the so-called "stability theorem." However, beyond establishing consistency for these procedures, several factors require theoretical consideration, including both the choice of subsample size and rates of convergence.

Furthermore, in this work we have shown only consistency for bootstrap estimation to a common limiting distribution. The rates of convergence in the 2-Wasserstein distance regarding the persistent Betti numbers rely on the unknown tail properties of the corresponding radius of stabilization. Quantifying these tail properties is a challenging open problem, and seems to be a key step toward an eventual rate calculation, as well as the previously mentioned process-level result.

Finally, there are several statistics of interest, including those based on the Delaunay complex, which do not fit into the specific frameworks provided here. It may be that these statistics still satisfy Theorem 2.13 in the general case, by techniques other than those provided here.

7. Technical results. PROOF OF LEMMA 2.11. We refer to Appendix B.1 [40] for a reference list of the general inequalities used here. Define two independent sets of random variables $(U_i)_{i=1}^{\infty} \overset{\text{iid}}{\sim} F$ and $(U_i^*)_{i=1}^{\infty} \overset{\text{iid}}{\sim} F$. For $N \sim \text{Pois}(n)$, denote by $\mathbf{P}_n$ the Poisson process given by $\{U_i\}_{i=1}^{N}$, which has intensity nf over $\mathbb{R}^d$, where $f := \mathrm{d}F/\mathrm{d}\lambda$. We will couple this Poisson process to $\mathbf{X}_n$. $\{U_i\}_{i=1}^{N\wedge n} \cup \{U_i^*\}_{i=1}^{(n-N)^+}$ has the same distribution as $\mathbf{X}_n$, thus we assume that the two random variables are equal almost surely. For a given random variable U_i or U_i^* and $L > 0$, conditional on X' the probability of falling within $B_{X'}(L/\sqrt[d]{n})$ is $\int_{B_{X'}(L/\sqrt[d]{n})} f \,\mathrm{d}\lambda \le (V_d L^d/n)\,\mathrm{M}\, f(X')$. Here, V_d is the volume of a unit ball in $\mathbb{R}^d$ and M is the Hardy–Littlewood maximal operator such that $\mathrm{M}\, f(x) := \sup_{r\in\mathbb{R}_+} (1/V_d r^d) \int_{B_x(r)} f \,\mathrm{d}\lambda$. Via the strong type Hardy–Littlewood maximal inequality, there exists a universal constant $C_2 < \infty$ such that the unconditional probability is bounded by

$$\begin{aligned}\frac{V_d L^d}{n} \int (\mathrm{M}\, f) f \,\mathrm{d}\lambda &\le \frac{V_d L^d}{n} \|\mathrm{M}\, f\|_2 \|f\|_2 \\ &\le \frac{C_2 V_d L^d}{n} \|f\|_2^2.\end{aligned}$$

The expected number of points within $B_{X'}(L/\sqrt[d]{n})$ that contribute to $\mathbf{P}_n \triangle \mathbf{X}_n$ is then at most

$$\begin{aligned}\mathbb{E}\left[|N-n| \frac{C_2 V_d L^d}{n} \|f\|_2^2\right] &\le \frac{C_2 V_d L^d}{n} \|f\|_2^2 \sqrt{\text{Var}[N]} \\ &= \frac{C_2 V_d L^d}{\sqrt{n}} \|f\|_2^2.\end{aligned}$$

This expectation bounds the probability that $\mathbf{X}_n$ and $\mathbf{P}_n$ differ within $B_{X'}(L/\sqrt[d]{n})$. For sufficiently large n, this bound can be made arbitrarily small.

Next, we will couple the Poisson process $\mathbf{P}_n$ with a conditionally homogeneous approximation. Let $\mathbf{T}$ be a homogeneous Poisson process on $\mathbb{R}^d \times \mathbb{R}_+$ with unit intensity. The point process given by $\{U_i \text{ s.t. } (U_i, T_i) \in \mathbf{T}, T_i \le nf(U_i)\}$ is a nonhomogeneous Poisson process with intensity nf. Without loss of generality, this process is assumed to equal $\mathbf{P}_n$ almost surely. Define the point process $\mathbf{H}_n := \{U_i \text{ s.t. } (U_i, T_i) \in \mathbf{T} \text{ and } T_i \le nf(X')\}$.

Conditional on X', $\mathbf{H}_n$ is a homogeneous Poisson process with intensity $nf(X')$. The number of observations within $B_{X'}(L/\sqrt[d]{n})$ that contribute to $\mathbf{P}_n \triangle \mathbf{H}_n$ follows a Poisson distribution with parameter $n\int_{B_{X'}(L/\sqrt[d]{n})} |f - f(X')| \,\mathrm{d}\lambda$. Removing the conditioning on X', the expected number is $n\int\int_{B_x(L/\sqrt[d]{n})} |f(y) - f(x)| f(x) \,\mathrm{d}y \,\mathrm{d}x \ge \mathbb{P}[(\mathbf{P}_n \triangle \mathbf{H}_n) \cap B_{X'}(L/\sqrt[d]{n}) \ne \varnothing]$. We will show that this quantity can be made arbitrarily small. Consider C, the set of Lebesgue points of f. For $C_{\gamma,R} :=$
$\bigcap_{r<R}\{x \in \mathbb{R}^d \text{ s.t. } (1/V_d r^d) \int_{B_x(r)} |f - f(x)| \,\mathrm{d}\lambda \le \gamma\}$, we have $C^c = \bigcup_{\gamma>0} \bigcap_{R>0} C_{\gamma,R}^c$. Thus, for any $\gamma > 0$, by the Lebesgue differentiation theorem $\lambda(\bigcap_{R>0} C_{\gamma,R}^c) \le \lambda(C^c) = 0$ for any $\gamma > 0$. Also, $(1/V_d r^d) \int_{B_x(r)} |f - f(x)| \,\mathrm{d}\lambda$ is continuous in r, thus via the separability of $\mathbb{R}^d$ C_γ may be expressed using only countably many sets. By continuity of measure,

$\lim_{\gamma\to 0}|C_\gamma^c| = 0$. We have

$$\begin{aligned}
&n\int\int_{B_x(L/\sqrt[d]{n})} |f(y)-f(x)|f(x)\,\mathrm{d}y\,\mathrm{d}x \\
&\quad\le \gamma V_d L^d + n\int_{C_\gamma^c}\int_{B_x(L/\sqrt[d]{n})} |f(y)-f(x)|f(x)\,\mathrm{d}y\,\mathrm{d}x \\
&\quad= \gamma V_d L^d + \int_{B_0(L/\sqrt[d]{n})}\int_{C_\gamma^c} |f(x+t)-f(x)|f(x)\,\mathrm{d}x\,\mathrm{d}t \\
&\quad\le V_d L^d\Big(\gamma + 2\|f\|_2\sqrt{\int_{C_\gamma^c} f^2\,\mathrm{d}\lambda}\Big).
\end{aligned}$$

This quantity does not depend on n. Because $\|f\|_2 < \infty$, by the dominated convergence theorem this bound goes to 0 as $\gamma \to 0$. Combining with the previous steps, we have coupled $\mathbf{X}_n$ and $\mathbf{H}_n$ to be equal with arbitrarily high probability.

By assumption, for any given $a, b, \delta > 0$, L may be chosen so that there exists $A \subset \mathcal{X}$ such that for any homogenous Poisson process $\mathbf{Q}_\lambda$ on $\mathbb{R}^d$ with intensity $\lambda \in [a, b]$ we have $\rho_0^{-1}((L,\infty]) \subseteq A$ and $\mathbb{P}[\mathbf{Q}_\lambda \in A] \le \delta$.

Because $\sqrt[d]{n}(\mathbf{H}_n - X')$ is a conditionally homogeneous Poisson process, we have

$$\begin{aligned}
&\mathbb{P}^*\big[\rho_0\big(\sqrt[d]{n}(\mathbf{H}_n - X')\big) > L \text{ and } f(X') \in [a,b]\big] \\
&\quad\le \mathbb{P}\big[\sqrt[d]{n}(\mathbf{H}_n - X') \in A \text{ and } f(X') \in [a,b]\big] \\
&\quad= \mathbb{E}\big[\mathbb{P}\big[\sqrt[d]{n}(\mathbf{H}_n - X') \in A \text{ and } f(X') \in [a,b] | X'\big]\big] \\
&\quad\le \delta\,\mathbb{P}\big[f(X') \in [a,b]\big] \le \delta.
\end{aligned}$$

Combining the pieces, we have that

$$\begin{aligned}
&\mathbb{P}^*\big[\rho_0\big(\sqrt[d]{n}(\mathbf{X}_n - X')\big) > L\big] \\
&\quad\le \mathbb{P}^*\big[\rho_0\big(\sqrt[d]{n}(\mathbf{H}_n - X')\big) > L \text{ and } f(X') \in [a,b]\big] \\
&\qquad + \mathbb{P}\big[(\mathbf{X}_n \triangle \mathbf{P}_n) \cap B_{X'}\big(L/\sqrt[d]{n}\big) \ne \varnothing\big] \\
&\qquad + \mathbb{P}\big[(\mathbf{P}_n \triangle \mathbf{H}_n) \cap B_{X'}\big(L/\sqrt[d]{n}\big) \ne \varnothing\big] \\
&\qquad + \mathbb{P}\big[f(X') < a\big] + \mathbb{P}\big[f(X') > b\big].
\end{aligned}$$

Choose $a, \delta \to 0$ and $b \to \infty$. Because our previous choice of L is possibly unbounded, let $\gamma \to 0$ and $n \to \infty$ so that the entire expression goes to 0. The result follows. □

PROOF OF PROPOSITION 2.12. Our proof technique is inspired by that of Proposition 5.4 [27]. We expand using a martingale difference sequence. Let $\{(X_i, Y_i)\}_{i=1}^\infty$ be i.i.d. such that $\{X_i\}_{i=1}^\infty \overset{\text{iid}}{\sim} F$ and $\{Y_i\}_{i=1}^\infty \overset{\text{iid}}{\sim} G$. For $f := \mathrm{d}F/\mathrm{d}\lambda$ and $g := \mathrm{d}G/\mathrm{d}\lambda$, denote $\delta_1 := \|g - f\|_1$, $\delta_2 := \|g - f\|_2$ and $\delta_p := \|g - f\|_p \le \delta$. By Proposition B.1 [40], there are increasing functions $\xi_1 \colon \mathbb{R}_+ \to \mathbb{R}_+$ and $\xi_2 \colon \mathbb{R}_+ \to \mathbb{R}_+$ depending only on f and p such that $\lim_{\alpha\to 0}\xi_1(\alpha) = \lim_{\alpha\to 0}\xi_2(\alpha) = 0$, $\delta_1 \le \xi_1(\delta_p) \le \xi_1(\delta)$ and $\delta_2 \le \xi_2(\delta_p) \le \xi_2(\delta)$.

Each pair (X_i, Y_i) can be identically coupled such that $\mathbb{P}[X_i \ne Y_i] = \frac{1}{2}\delta_1$. Specifically, conditional on the event $\{X_i \ne Y_i\}$, X_i and Y_i follow the respective marginal densities $2(f-g)^+/\delta_1$ and $2(g-f)^+/\delta_1$. For each $j \in \mathbb{N}$, define $\mathbf{X}_j := \{X_i\}_{i=1}^j$, $\mathbf{Y}_j := \{Y_i\}_{i=1}^j$, and $\mathcal{F}_j := \sigma(\{\mathbf{X}_j, \mathbf{Y}_j\})$, where σ signifies the generated σ-algebra. Likewise, $\mathcal{F}_0 := \{\Omega, \varnothing\}$.

For (X', Y') an independent copy of the (X_i, Y_i), let

$$\mathbf{X}'_{n,j} := \{X_1, \dots, X_{j-1}, X', X_{j+1}, \dots X_n\},$$
$$\mathbf{Y}'_{n,j} := \{Y_1, \dots, Y_{j-1}, Y', Y_{j+1}, \dots, Y_n\}.$$

We apply the condensed notation $H_n(\mathbf{S}, \mathbf{T}) := \psi(\sqrt[d]{n}\mathbf{S}) - \psi(\sqrt[d]{n}\mathbf{T})$. Using the pairwise orthogonality of a martingale difference sequence and the conditional version of Jensen's inequality,

$$\begin{aligned}
\operatorname{Var}\left[\frac{1}{\sqrt{n}} H_n(\mathbf{X}_n, \mathbf{Y}_n)\right] &= \frac{1}{n}\mathbb{E}\left[\left|\sum_{j=1}^{n} \mathbb{E}\big[H_n(\mathbf{X}_n, \mathbf{Y}_n)|\mathcal{F}_j\big] - \mathbb{E}\big[H_n(\mathbf{X}_n, \mathbf{Y}_n)|\mathcal{F}_{j-1}\big]\right|^2\right] \\
&= \frac{1}{n}\mathbb{E}\left[\left|\sum_{j=1}^{n} \mathbb{E}\big[H_n(\mathbf{X}_n, \mathbf{Y}_n) - H_n(\mathbf{X}'_{n,j}, \mathbf{Y}'_{n,j})|\mathcal{F}_j\big]\right|^2\right] \\
&= \frac{1}{n}\sum_{j=1}^{n} \mathbb{E}\big[\mathbb{E}\big[H_n(\mathbf{X}_n, \mathbf{Y}_n) - H_n(\mathbf{X}'_{n,j}, \mathbf{Y}'_{n,j})|\mathcal{F}_j\big]^2\big] \\
&\le \mathbb{E}\big[\big|H_n(\mathbf{X}_n, \mathbf{Y}_n) - H_n(\mathbf{X}'_{n,\ell}, \mathbf{Y}'_{n,\ell})\big|^2\big].
\end{aligned} \tag{1}$$

The above holds for any $1 \le \ell \le n$. We have an upper bound for (1) of

$$\begin{aligned}
&2\,\mathbb{E}\big[\big|H_n(\mathbf{X}_n \cup X', \mathbf{X}_n) - H_n(\mathbf{Y}_n \cup Y', \mathbf{Y}_n)\big|^2\big] \\
&\quad + 2\,\mathbb{E}\big[\big|H_n(\mathbf{X}_n \cup X', \mathbf{X}'_{n,j}) - H_n(\mathbf{Y}_n \cup Y', \mathbf{Y}'_{n,j})\big|^2\big] \\
&= 4\,\mathbb{E}\big[\big|H_n(\mathbf{X}_n \cup X', \mathbf{X}_n) - H_n(\mathbf{Y}_n \cup Y', \mathbf{Y}_n)\big|^2\big].
\end{aligned} \tag{2}$$

We will decompose the expectation in (2) using the stabilization of ψ. For $L > 0$, we have the following events:

$$\begin{aligned}
A &:= \{Y' = X'\}, \\
B &:= \{\mathbf{Y}_n \cap B_{X'}(L/\sqrt[d]{n}) = \mathbf{X}_n \cap B_{X'}(L/\sqrt[d]{n})\}, \\
C_{X*} &:= \{D_{\sqrt[d]{n}X'}((\sqrt[d]{n}\mathbf{X}_n) \cap B_{\sqrt[d]{n}X'}(L)) = D_{\sqrt[d]{n}X'}(\sqrt[d]{n}\mathbf{X}_n)\}, \\
C_{Y*} &:= \{D_{\sqrt[d]{n}Y'}((\sqrt[d]{n}\mathbf{Y}_n) \cap B_{\sqrt[d]{n}Y'}(L)) = D_{\sqrt[d]{n}Y'}(\sqrt[d]{n}\mathbf{Y}_n)\}.
\end{aligned}$$

Note when all four are satisfied, $H_n(\mathbf{X}_n \cup X', \mathbf{X}_n) = H_n(\mathbf{Y}_n \cup Y', \mathbf{Y}_n)$. Let $C_X \subseteq C_{X*}$ and $C_Y \subseteq C_{Y*}$ be measurable with $\mathbb{P}[C_X^c] = \mathbb{P}^*[C_{X*}^c]$ and $\mathbb{P}[C_Y^c] = \mathbb{P}^*[C_{Y*}^c]$. Then for any $k \ge 0$,

$$\begin{aligned}
&\mathbb{E}\big[\big|H_n(\mathbf{X}_n \cup X', \mathbf{X}_n) - H_n(\mathbf{Y}_n \cup Y', \mathbf{Y}_n)\big|^2\big] \\
&\quad = \mathbb{E}\big[\big|H_n(\mathbf{X}_n \cup X', \mathbf{X}_n) - H_n(\mathbf{Y}_n \cup Y', \mathbf{Y}_n)\big|^2 \mathbb{1}\{A^c \cup B^c \cup C_X^c \cup C_Y^c\}\big] \\
&\quad \le 2\big(\mathbb{E}\big[H_n(\mathbf{X}_n \cup X', \mathbf{X}_n)^2 \mathbb{1}\{A^c \cup B^c \cup C_X^c \cup C_Y^c\}\big] \\
&\qquad + \mathbb{E}\big[H_n(\mathbf{Y}_n \cup Y', \mathbf{Y}_n)^2 \mathbb{1}\{A^c \cup B^c \cup C_X^c \cup C_Y^c\}\big]\big) \\
&\quad \le 2\big(\mathbb{E}\big[H_n(\mathbf{X}_n \cup X', \mathbf{X}_n)^2 \mathbb{1}\{|H_n(\mathbf{X}_n \cup X', \mathbf{X}_n)| > k\}\big] \\
&\qquad + k^2\,\mathbb{P}\big[A^c \cup B^c \cup C_X^c \cup C_Y^c\big] \\
&\qquad + \mathbb{E}\big[H_n(\mathbf{Y}_n \cup Y', \mathbf{Y}_n)^2 \mathbb{1}\{|H_n(\mathbf{Y}_n \cup Y', \mathbf{Y}_n)| > k\}\big] \\
&\qquad + k^2\,\mathbb{P}\big[A^c \cup B^c \cup C_X^c \cup C_Y^c\big]\big)
\end{aligned} \tag{3}$$

Because (E1) holds, it suffices to show that each of A^c, B^c, C_X^c and C_Y^c can be made to occur with small probability, uniformly in G and n. Then $k = k_\delta$ may be chosen such that $\lim_{\delta\to 0} k_\delta = \infty$ and (3) goes to 0 uniformly.

For A^c, this is satisfied because $\mathbb{P}[X' \neq Y'] \leq \frac{1}{2}\xi_1(\delta)$. Consider B^c next. The sample pairs, which contribute to $\mathbf{X}_n \cap B_{X'}(L/\sqrt[d]{n})$ but not $\mathbf{Y}_n \cap B_{X'}(L/\sqrt[d]{n})$, are those (X_i, Y_i) for which $X_i \neq Y_i$ and either $\|X_i - X'\| \leq L/\sqrt[d]{n}$ or $\|Y_i - X'\| \leq L/\sqrt[d]{n}$. Conditional on X', their count follows a binomial distribution with expectation at most $n\int_{B_{X'}(L/\sqrt[d]{n})} |g - f|\,\mathrm{d}\lambda \leq V_d L^d\,\mathrm{M}|g - f|(X')$. Here, M is the Hardy–Littlewood maximal operator. Removing the conditioning on X', the expected count is at most

$$V_d L^d \int (\mathrm{M}|g - f|) f\,\mathrm{d}\lambda \leq V_d L^d \|\mathrm{M}|g - f|\|_2 \|f\|_2 \leq C_2 V_d L^d \xi_2(\delta)\|f\|_2.$$

The above follows from the strong type Hardy–Littlewood maximal inequality for some constant $C_2 < \infty$. This final expression provides an upper bound on $\mathbb{P}[B^c]$. Then from (S1), there exists a choice $L = L_\delta$ such that $\lim_{\delta\to 0} L_\delta^d \xi_2(\delta) = 0$ and each of $\mathbb{P}[B^c] \to 0$, $\mathbb{P}[C_X^c] \to 0$, and $\mathbb{P}[C_Y^c] \to 0$, uniform in n and G. The result follows. □

PROOF OF THEOREM 2.13. Let any bounded, Lipschitz function $v\colon \mathbb{R}^k \to \mathbb{R}$ be given. Then for some $M > 0$, v is bounded within $[-M, M]$ with a Lipschitz constant of M. For any $m \in \mathbb{N}$, define the functional V_m as follows. We use the condensed notation $\vec{H}_{m,j}(\mathbf{S}) := (\psi_j(\sqrt[d]{m}\mathbf{S}) - \mathbb{E}[\psi_j(\sqrt[d]{m}\mathbf{S})])/\sqrt{m}$ with $\vec{H}_m := (H_{m,j})_{j=1}^k$. For a probability distribution G on $\mathbb{R}^d$, let $(Y_i)_{i\in\mathbb{N}} \overset{\text{iid}}{\sim} G$ and $\mathbf{Y}_m := \{Y_i\}_{i=1}^m$. Define $V_m(G) := \mathbb{E}[v(\vec{H}_m(\mathbf{Y}_m))]$.

First, assuming that $\vec{H}_n(\mathbf{X}_n) \xrightarrow{\text{d}} \Psi$, we have $\lim_{n\to\infty} V_n(F) = \int v\,\mathrm{d}\Psi$. Now, let $(X_i')_{i\in\mathbb{N}} \overset{\text{iid}}{\sim} F$ be independent of $\hat{F}_n$ and define $\mathbf{X}_m' := \{X_i'\}_{i=1}^m$ for any $m \in \mathbb{N}$. Via Proposition 2.12 and Chebyshev's inequality, for any $\varepsilon > 0$ we have almost surely that

$$\begin{aligned}
V_{m_n}(\hat{F}_n) &= \mathbb{E}\big[v\big(\vec{H}_{m_n}(\mathbf{X}^*_{n,m_n})\big)|\mathbf{X}_n\big] \\
&= \mathbb{E}\big[v\big(\vec{H}_{m_n}(\mathbf{X}^*_{n,m_n})\big)\,\mathbb{1}\{\|\vec{H}_{m_n}(\mathbf{X}^*_{n,m_n}) - \vec{H}_{m_n}(\mathbf{X}'_{m_n})\| \leq \varepsilon\}|\mathbf{X}_n\big] \\
&\quad + \mathbb{E}\big[v\big(\vec{H}_{m_n}(\mathbf{X}^*_{n,m_n})\big)\,\mathbb{1}\{\|\vec{H}_{m_n}(\mathbf{X}^*_{n,m_n}) - \vec{H}_{m_n}(\mathbf{X}'_{m_n})\| > \varepsilon\}|\mathbf{X}_n\big] \\
&\leq \mathbb{E}\big[v\big(\vec{H}_{m_n}(\mathbf{X}'_{m_n})\big)\big] + M\varepsilon \\
&\quad + M\sum_{j=1}^k \mathbb{P}\Big[\big|H_{m_n,j}(\mathbf{X}^*_{n,m_n}) - H_{m_n,j}(\mathbf{X}'_{m_n})\big| > \frac{\varepsilon}{\sqrt{k}}|\mathbf{X}_n\Big] \\
&\leq \mathbb{E}\big[v\big(\vec{H}_{m_n}(\mathbf{X}'_{m_n})\big)\big] + M\varepsilon + \frac{Mk}{\varepsilon^2}\sum_{j=1}^k \gamma_j(\|\hat{f}_n - f\|_p).
\end{aligned}$$

Here, for each $j \in \{1, \ldots, k\}$, $\gamma_j\colon \mathbb{R}_+ \to \mathbb{R}_+$ is as given in Proposition 2.12 applied to ψ_j. Similarly, almost surely

$$V_{m_n}(\hat{F}_n) \geq \mathbb{E}\big[v\big(\vec{H}_{m_n}(\mathbf{X}'_{m_n})\big)\big] - M\varepsilon - \frac{Mk}{\varepsilon^2}\sum_{i=1}^k \gamma_j(\|\hat{f}_n - f\|_p).$$

Because $\|\hat{f}_n - f\|_p \to 0$, we have that the lower bound for $V_{m_n}(\hat{F}_n)$ converges to $\int_{\mathbb{R}} v\,\mathrm{d}\Psi - M\varepsilon$ and the upper bound converges to $\int_{\mathbb{R}} v\,\mathrm{d}\Psi + M\varepsilon$, either in probability or almost surely, depending on assumptions. Since this holds for any $\varepsilon > 0$, we have that $V_{m_n}(\hat{F}_n) \to \int_{\mathbb{R}} v\,\mathrm{d}\Psi$ in probability (or a.s.).

Now we will show the converse direction. By similar arguments, for any $\varepsilon > 0$ we have

$$\begin{aligned} V_{m_n}(F) &= \mathbb{E}\big[\mathbb{E}\big[\upsilon\big(\vec{H}_{m_n}(\mathbf{X}'_{m_n})\big)|\mathbf{X}_n\big]\big] \\ &\le \mathbb{E}\big[\mathbb{E}\big[\upsilon\big(\vec{H}_{m_n}(\mathbf{X}^*_{n,m_n})\big)|\mathbf{X}_n\big]\big] + M\varepsilon \\ &\quad + M\,\mathbb{E}\left[\min\left\{\frac{k}{\varepsilon^2}\sum_{i=1}^{k}\gamma_j(\|\hat{f}_n - f\|_p), 1\right\}\right] \\ V_{m_n}(F) &\ge \mathbb{E}\big[\mathbb{E}\big[\upsilon\big(\vec{H}_{m_n}(\mathbf{X}^*_{n,m_n})\big)|\mathbf{X}_n\big]\big] - M\varepsilon \\ &\quad - M\,\mathbb{E}\left[\min\left\{\frac{k}{\varepsilon^2}\sum_{i=1}^{k}\gamma_j(\|\hat{f}_n - f\|_p), 1\right\}\right]. \end{aligned}$$

Each expectation involves only bounded variables, thus the lower bound tends to $\int_{\mathbb{R}} \upsilon \,\mathrm{d}\Psi - M\varepsilon$ and the upper bound to $\int_{\mathbb{R}} \upsilon \,\mathrm{d}\Psi + M\varepsilon$, assuming either $\mathbb{E}[\upsilon(\vec{H}_{m_n}(\mathbf{X}^*_{n,m_n}))|\mathbf{X}_n] \to \int_{\mathbb{R}} \upsilon \,\mathrm{d}\Psi$ in probability or almost surely. Since this holds for any $\varepsilon > 0$, we have that $\lim_{n\to\infty} V_{m_n}(F) = \int_{\mathbb{R}} \upsilon \,\mathrm{d}\Psi$. Since our initial choice of υ was arbitrary, the desired result follows. □

Acknowledgments. Thank you to the Associate Editor, Editor and reviewers for their helpful comments and thorough examination of this work.

Funding. Benjamin Roycraft was partially supported by the National Science Foundation (NSF), grant number DMS-1148643. Johannes Krebs was partially supported by the German Research Foundation (DFG), grant number KR-4977/2-1. Wolfgang Polonik was partially supported by the National Science Foundation (NSF), grant number DMS-2015575.

Funding for the Sloan Digital Sky Survey IV has been provided by the Alfred P. Sloan Foundation, the U.S. Department of Energy Office of Science and the Participating Institutions. SDSS-IV acknowledges support and resources from the Center for High-Performance Computing at the University of Utah. The SDSS website is www.sdss.org.

SUPPLEMENTARY MATERIAL

Supplement to "Bootstrapping persistent Betti numbers and other stabilizing statistics" (DOI: 10.1214/23-AOS2277SUPP; .pdf). A data application is made to a cosmic web data set from the Sloan Digital Sky Survey (SDSS) [4]. Proofs for results in the main paper are included. Furthermore, we give extended results beyond those presented in the main text. Specifically, we make a more detailed examination of the 2-Wasserstein distance in Proposition 2.12 for all statistics considered, and provide a proof for the L_p-convergence of kernel density estimation. Specific generating functions for the simulation study of Section 5 are also provided.

REFERENCES

[1] ADLER, R. J., AGAMI, S. and PRANAV, P. (2017). Modeling and replicating statistical topology and evidence for CMB nonhomogeneity. *Proc. Natl. Acad. Sci. USA* **114** 11878–11883. MR3725115 https://doi.org/10.1073/pnas.1706885114

[2] ARSUAGA, J., BORRMAN, T., CAVALCANTE, R., GONZALEZ, G. and PARK, C. (2015). Identification of copy number aberrations in breast cancer subtypes using persistence topology. *Microarrays* **4** 339–369.

[3] BISCIO, C. A. N., CHENAVIER, N., HIRSCH, C. and SVANE, A. M. (2020). Testing goodness of fit for point processes via topological data analysis. *Electron. J. Stat.* **14** 1024–1074. MR4067816 https://doi.org/10.1214/20-EJS1683

[4] BLANTON, M. R., BERSHADY, M. A., ABOLFATHI, B., ALBARETI, F. D., ALLENDE PRIETO, C., ALMEIDA, A., ALONSO-GARCÍA, J., ANDERS, F., ANDERSON, S. F. et al. (2017). Sloan digital sky survey IV: Mapping the milky way, nearby galaxies, and the distant universe. *Astron. J.* **154** 28.

[5] BOBROWSKI, O. and MUKHERJEE, S. (2015). The topology of probability distributions on manifolds. *Probab. Theory Related Fields* **161** 651–686. MR3334278 https://doi.org/10.1007/s00440-014-0556-x

[6] BOISSONNAT, J.-D., CHAZAL, F. and YVINEC, M. (2018). *Geometric and Topological Inference*. *Cambridge Texts in Applied Mathematics*. Cambridge Univ. Press, Cambridge. MR3837127 https://doi.org/10.1017/9781108297806

[7] BUBENIK, P. (2015). Statistical topological data analysis using persistence landscapes. *J. Mach. Learn. Res.* **16** 77–102. MR3317230

[8] BUBENIK, P. and KIM, P. T. (2007). A statistical approach to persistent homology. *Homology, Homotopy Appl.* **9** 337–362. MR2366953

[9] CAMARA, P. G., ROSENBLOOM, D. I. S., EMMETT, K. J., LEVINE, A. J. and RABADAN, R. (2016). Topological data analysis generates high-resolution, genome-wide maps of human recombination. *Cell Syst.* **3** 83–94. https://doi.org/10.1016/j.cels.2016.05.008

[10] CHAZAL, F. and DIVOL, V. (2019). The density of expected persistence diagrams and its kernel based estimation. *J. Comput. Geom.* **10** 127–153. MR4088069 https://doi.org/10.20382/jocg.v10i2a7

[11] CHAZAL, F., FASY, B., LECCI, F., MICHEL, B., RINALDO, A. and WASSERMAN, L. (2015). Subsampling methods for persistent homology. *Proc. 32nd Int. Conf. Mach. Learn.* **37** 2143–2151.

[12] CHAZAL, F., FASY, B. T., LECCI, F., RINALDO, A., SINGH, A. and WASSERMAN, L. (2015). On the bootstrap for persistence diagrams and landscapes. *Model. Anal. Inf. Syst.* **20** 111–120.

[13] CHAZAL, F., FASY, B. T., LECCI, F., RINALDO, A. and WASSERMAN, L. (2015). Stochastic convergence of persistence landscapes and silhouettes. *J. Comput. Geom.* **6** 140–161. MR3323391 https://doi.org/10.20382/jocg.v6i2a8

[14] CHAZAL, F. and MICHEL, B. (2021). An introduction to topological data analysis: Fundamental and practical aspects for data scientists. *Front. Artif. Intell.* **4** 667963. https://doi.org/10.3389/frai.2021.667963

[15] CHEN, Y.-C., WANG, D., RINALDO, A. and WASSERMAN, L. (2015). Statistical analysis of persistence intensity functions. ArXiv Preprint. Available at arXiv:1510.02502.

[16] CHUNG, Y.-M. and LAWSON, A. (2022). Persistence curves: A canonical framework for summarizing persistence diagrams. *Adv. Comput. Math.* **48** 6. MR4368950 https://doi.org/10.1007/s10444-021-09893-4

[17] CRAWFORD, L., MONOD, A., CHEN, A. X., MUKHERJEE, S. and RABADÁN, R. (2020). Predicting clinical outcomes in glioblastoma: An application of topological and functional data analysis. *J. Amer. Statist. Assoc.* **115** 1139–1150. MR4143455 https://doi.org/10.1080/01621459.2019.1671198

[18] DE SILVA, V. and GHRIST, R. (2007). Coverage in sensor networks via persistent homology. *Algebr. Geom. Topol.* **7** 339–358. MR2308949 https://doi.org/10.2140/agt.2007.7.339

[19] DEWOSKIN, D., CLIMENT, J., CRUZ-WHITE, I., VAZQUEZ, M., PARK, C. and ARSUAGA, J. (2010). Applications of computational homology to the analysis of treatment response in breast cancer patients. *Topology Appl.* **157** 157–164. MR2556091 https://doi.org/10.1016/j.topol.2009.04.036

[20] EDELSBRUNNER, LETSCHER and ZOMORODIAN (2002). Topological persistence and simplification. *Discrete Comput. Geom.* **28** 511–533.

[21] FASY, B. T., LECCI, F., RINALDO, A., WASSERMAN, L., BALAKRISHNAN, S. and SINGH, A. (2014). Confidence sets for persistence diagrams. *Ann. Statist.* **42** 2301–2339. MR3269981 https://doi.org/10.1214/14-AOS1252

[22] HIRAOKA, Y., SHIRAI, T. and TRINH, K. D. (2018). Limit theorems for persistence diagrams. *Ann. Appl. Probab.* **28** 2740–2780. MR3847972 https://doi.org/10.1214/17-AAP1371

[23] KIM, J., SHIN, J., RINALDO, A. and WASSERMAN, L. (2019). Uniform convergence rate of the kernel density estimator adaptive to intrinsic volume dimension. In *International Conference on Machine Learning* 3398–3407. PMLR.

[24] KRAMAR, M., GOULLET, A., KONDIC, L. and MISCHAIKOW, K. (2013). Persistence of force networks in compressed granular media. *Phys. Rev., E* **87**.

[25] KRAMÁR, M., LEVANGER, R., TITHOF, J., SURI, B., XU, M., PAUL, M., SCHATZ, M. F. and MISCHAIKOW, K. (2016). Analysis of Kolmogorov flow and Rayleigh-Bénard convection using persistent homology. *Phys. D* **334** 82–98. MR3545971 https://doi.org/10.1016/j.physd.2016.02.003

[26] KREBS, J., ROYCRAFT, B. and POLONIK, W. (2021). On approximation theorems for the Euler characteristic with applications to the bootstrap. *Electron. J. Stat.* **15** 4462–4509. MR4312855 https://doi.org/10.1214/21-ejs1898

[27] KREBS, J. T. and POLONIK, W. (2019). On the asymptotic normality of persistent Betti numbers. ArXiv Preprint. Available at arXiv:1903.03280.

[28] LACHIÈZE-REY, R., PECCATI, G. and YANG, X. (2022). Quantitative two-scale stabilization on the Poisson space. *Ann. Appl. Probab.* **32** 3085–3145. MR4474528 https://doi.org/10.1214/21-aap1768
[29] LACHIÈZE-REY, R., SCHULTE, M. and YUKICH, J. E. (2019). Normal approximation for stabilizing functionals. *Ann. Appl. Probab.* **29** 931–993. MR3910021 https://doi.org/10.1214/18-AAP1405
[30] LAST, G., PECCATI, G. and SCHULTE, M. (2016). Normal approximation on Poisson spaces: Mehler's formula, second order Poincaré inequalities and stabilization. *Probab. Theory Related Fields* **165** 667–723. MR3520016 https://doi.org/10.1007/s00440-015-0643-7
[31] OWADA, T. (2018). Limit theorems for Betti numbers of extreme sample clouds with application to persistence barcodes. *Ann. Appl. Probab.* **28** 2814–2854. MR3847974 https://doi.org/10.1214/17-AAP1375
[32] OWADA, T. and ADLER, R. J. (2017). Limit theorems for point processes under geometric constraints (and topological crackle). *Ann. Probab.* **45** 2004–2055. MR3650420 https://doi.org/10.1214/16-AOP1106
[33] PENROSE, M. D. and YUKICH, J. E. (2001). Central limit theorems for some graphs in computational geometry. *Ann. Appl. Probab.* **11** 1005–1041. MR1878288 https://doi.org/10.1214/aoap/1015345393
[34] PENROSE, M. D. and YUKICH, J. E. (2003). Weak laws of large numbers in geometric probability. *Ann. Appl. Probab.* **13** 277–303. MR1952000 https://doi.org/10.1214/aoap/1042765669
[35] POLITIS, D. N., ROMANO, J. P. and WOLF, M. (1999). *Subsampling*. *Springer Series in Statistics*. Springer, New York. MR1707286 https://doi.org/10.1007/978-1-4612-1554-7
[36] PRANAV, P., ADLER, R. J., BUCHERT, T., EDELSBRUNNER, H., JONES, B. J. T., SCHWARTZMAN, A., WAGNER, H. and VAN DE WEYGAERT, R. (2019). Unexpected topology of the temperature fluctuations in the cosmic microwave background. *Astron. Astrophys.* **627** A163.
[37] PRANAV, P., EDELSBRUNNER, H., VAN DE WEYGAERT, R., VEGTER, G., KERBER, M., JONES, B. J. T. and WINTRAECKEN, M. (2016). The topology of the cosmic web in terms of persistent Betti numbers. *Mon. Not. R. Astron. Soc.* **465** 4281–4310.
[38] PRANAV, P., VAN DE WEYGAERT, R., VEGTER, G., JONES, B. J. T., ADLER, R. J., FELDBRUGGE, J., PARK, C., BUCHERT, T. and KERBER, M. (2019). Topology and geometry of Gaussian random fields I: On Betti numbers, Euler characteristic, and Minkowski functionals. *Mon. Not. R. Astron. Soc.* **485** 4167–4208.
[39] ROYCRAFT, B. (2021). github.com/btroycraft/stabilizing_statistics_bootstrap.
[40] ROYCRAFT, B., KREBS, J. and POLONIK, W. (2023). Supplement to "Bootstrapping persistent Betti numbers and other stabilizing statistics." https://doi.org/10.1214/23-AOS2277SUPP
[41] SILVERMAN, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. *Monographs on Statistics and Applied Probability*. CRC Press, London. MR0848134 https://doi.org/10.1007/978-1-4899-3324-9
[42] TRINH, K. D. (2019). On central limit theorems in stochastic geometry for add-one cost stabilizing functionals. *Electron. Commun. Probab.* **24** 76. MR4049088 https://doi.org/10.1214/19-ecp279
[43] TURNER, K., MUKHERJEE, S. and BOYER, D. M. (2014). Persistent homology transform for modeling shapes and surfaces. *Inf. Inference* **3** 310–344. MR3311455 https://doi.org/10.1093/imaiai/iau011
[44] ULMER, M., ZIEGELMEIER, L. and TOPAZ, C. M. (2019). A topological approach to selecting models of biological experiments. *PLoS ONE* **14** 1–18.
[45] WASSERMAN, L. (2018). Topological data analysis. *Annu. Rev. Stat. Appl.* **5** 501–535. MR3774757 https://doi.org/10.1146/annurev-statistics-031017-100045
[46] XIA, K., FENG, X., TONG, Y. and WEI, G. W. (2014). Persistent homology for the quantitative prediction of fullerene stability. *J. Comput. Chem.* **36** 408–422.
[47] YOGESHWARAN, D. and ADLER, R. J. (2015). On the topology of random complexes built over stationary point processes. *Ann. Appl. Probab.* **25** 3338–3380. MR3404638 https://doi.org/10.1214/14-AAP1075
[48] YOGESHWARAN, D., SUBAG, E. and ADLER, R. J. (2017). Random geometric complexes in the thermodynamic regime. *Probab. Theory Related Fields* **167** 107–142. MR3602843 https://doi.org/10.1007/s00440-015-0678-9
[49] ZOMORODIAN, A. and CARLSSON, G. (2005). Computing persistent homology. *Discrete Comput. Geom.* **33** 249–274. MR2121296 https://doi.org/10.1007/s00454-004-1146-y