DOI: 10.1111/cgf.15142

Eurographics Symposium on Geometry Processing 2024
R. Hu and S. Lefebvre
(Guest Editors)

COMPUTER GRAPHICS *forum*
Volume 43 (2024), Number 5

Stability for Inference with Persistent Homology Rank Functions

Qiquan Wang[1], Inés García-Redondo[1,2], Pierre Faugère[3], Gregory Henselman-Petrusek[4] and Anthea Monod[1]

[1]Department of Mathematics, Imperial College London, UK
[2]London School of Geometry and Number Theory, University College London, UK
[3]Department of Mathematics, ENS Lyon, France
[4]Pacific Northwest National Laboratory, Richland, Washington, USA

Abstract

Persistent homology barcodes and diagrams are a cornerstone of topological data analysis that capture the "shape" of a wide range of complex data structures, such as point clouds, networks, and functions. However, their use in statistical settings is challenging due to their complex geometric structure. In this paper, we revisit the persistent homology rank function, which is mathematically equivalent to a barcode and persistence diagram, as a tool for statistics and machine learning. Rank functions, being functions, enable the direct application of the statistical theory of functional data analysis (FDA)—a domain of statistics adapted for data in the form of functions. A key challenge they present over barcodes in practice, however, is their lack of stability—a property that is crucial to validate their use as a faithful representation of the data and therefore a viable summary statistic. In this paper, we fill this gap by deriving two stability results for persistent homology rank functions under a suitable metric for FDA integration. We then study the performance of rank functions in functional inferential statistics and machine learning on real data applications, in both single and multiparameter persistent homology. We find that the use of persistent homology captured by rank functions offers a clear improvement over existing non-persistence-based approaches.

CCS Concepts

• ***Mathematics of computing*** → *Algebraic topology;* • ***Theory of computation*** → *Computational geometry;* • ***General and reference*** → *Performance;*

1. Introduction

Topological data analysis (TDA) leverages theory from algebraic topology to computational and data analytic settings, and has enjoyed great success in applications in many fields, including biology [ESR16, CW17, CMW18], medicine [CMC*20, BRI*17], physics [dSG07], economics [Gid17], and motion planning [BGK15, VAB13], to name a few. A cornerstone methodology of TDA is *persistent homology*, which produces a summary statistic of data. Its widespread applicability stems from its flexibility to adapt to a variety of complex data structures; its interpretability within the scientific domains where data arise; and its stability, which is the focus of this work. Stability provides a notion of faithfulness of the topological representation of the input data, guaranteeing that a bounded perturbation of the input data results in a bounded perturbation of their topological representation captured by persistent homology. Stability is a crucial property that validates the use of persistent homology in real data applications.

Persistent homology has its roots in various constructions and thus has various representations; the most well-known is the *persistence diagram* or equivalently, the *barcode*. A less-used representation is the *rank function*, initially proposed in the early 1990s [Fro92] as the *size function*. Size functions were used as a mathematical tool for shape and image analysis in computer vision and pattern recognition [VUFF93,FL99,LF02,BCF*08,DFLM09], and were reinterpreted in algebraic terms using a correspondence between size functions and formal series [FL01] (see [BDFF*08] for a thorough survey on the theory of size functions). Such algebraic topological formulations of rank functions directly coincide with parallel concepts from persistent homology.

Although rank functions were proposed before persistence diagrams and barcodes, and they are in fact equivalent to them, their use has been comparatively restricted due to a the difficulty in establishing stability results and a less comprehensive understanding of their effectiveness in practical data scenarios. Nevertheless, with the current active research interest in *multiparameter persistent homology*, rank functions are again becoming increasingly relevant, as they inherently and directly adapt to this higher dimensional framework where persistence diagrams and barcodes do not [CZ07]. Additionally, unlike persistence diagrams and barcodes, rank functions, due to their structural form as functions, are naturally amenable to *functional data analysis* (FDA), a robust statistical field focused on analyzing data taking the form of functions, curves, and surfaces. This adaptability to FDA (see [Ram05, Ram02] for a thorough introduction to the field and its applications) offers a rich statistical toolkit not directly accessible with persistence diagrams and barcodes. FDA methods have been previously used in TDA by [CMC*20], who perform Gaussian process regression using a summary statistic constructed from a dynamic version of the Euler characteristic, which is an alternative topological invariant and distinct from persistence barcodes and diagrams. Principal component analysis (PCA) for rank functions—an important dimension reduction technique in descriptive, rather than inferential, statistics—has also been studied in [RT16].

In our work, we aim to study the performance of rank functions in inference tasks, which move beyond the descriptive analysis by [RT16] and, within the field of statistics, are arguably significantly more challenging: where descriptive statistics studies properties observed in a single sample of data, inferential statistics aims to impute information and provide guarantees on the possibly infinite, unobserved population from observed data. To validate our findings, however, we need to first establish suitable stability properties of rank functions. Here, "suitable" means that we will need to establish stability under a metric conducive to the application of FDA methods. Once this is achieved, we then study the performance of rank functions in both single- and multiparameter persistent homology in inferential machine learning tasks on real data applications. We find a clear improvement in performance using rank functions compared to existing methods that do not incorporate persistent homology, and other persistence-based methods.

The remainder of this paper is organized as follows. In Section 2, we provide a background and literature review on rank functions and discuss their relation to barcodes and persistence diagrams, as well as various metrics associated to these representations and their implications on stability. In Section 3, we present two stability results for rank functions with respect to a suitable metric for FDA implementation. These stability guarantees motivate the applications of rank functions in inferential tasks using FDA to real data presented in Sections 4 and 5. We close with a discussion in Section 6 on our findings and propose directions for future research. Proofs and further theoretical details are given in Appendix A.

2. Preliminaries: Persistent Homology

In this section, we review the essential background and existing literature to the theory of persistent homology and its metrics, foundational to our work.

2.1. Persistence Modules and Rank Functions

The algebraic object that is the central focus of persistent homology theory is the *persistence module*, a functor mapping from a poset category to the category of vector spaces, $M : (\mathsf{P}, \leq) \to \mathsf{Vec}$, also written as $M \in \mathsf{Vec}^{(\mathsf{P},\leq)}$ [BS14, BdSS15, KM21]. Unless otherwise specified, we assume that Vec is the category of finite dimensional vector spaces and work with *pointwise finite dimensional* (p.f.d.) persistence modules.

Arguably, the most relevant example is the module of persistent homology for a finite simplicial complex, first introduced by [ELZ02] and obtained as follows. Consider a *filtration*, i.e., a diagram $F \in \mathsf{Simp}^{(\mathbb{R},\leq)}$ such that $F_x := F(x)$ is a finite simplicial complex for each $x \in \mathbb{R}$, and such that for any other $y \in \mathbb{R}$ with $x \leq y$, $F(x \leq y)$ is an inclusion $F_x \subset F_y$. A common example is the Vietoris–Rips filtration [Vie27], which for a metric space (X,d) and a finite subset $S \subset X$ is denoted as $\mathrm{VR}(S) = \{\mathrm{VR}_t(S) : t \in [0,+\infty)]\}$. The simplicial complex at filtration value $t \in \mathbb{R}$ is defined as the family of all simplices of diameter less or equal than t that can be formed with the finite set S as set of vertices. An example of a Vietoris–Rips filtration is given in Figure 1.

Given that we are working with finite simplicial complexes, there is a finite discrete set of values over which the filtration changes. For each $k \geq 0$, we obtain another diagram $\mathrm{H}_k(F) \in \mathsf{Vec}^{(\mathbb{R},\leq)}$ by setting

$$\mathrm{H}_k(F)(x) := \mathrm{H}_k(F_x, \mathbb{F}),$$

where $\mathrm{H}_k : \mathsf{Simp} \to \mathsf{Vec}$ is the homology functor of order $k \geq 0$ with coefficients over a field $\mathbb{F}$. We shall also refer to the diagram given by

$$\mathrm{H}(F)(x) := \mathrm{H}(F_x, \mathbb{F}) = \bigoplus_{k \geq 0} \mathrm{H}_k(F_x, \mathbb{F})$$

using the notation $\mathrm{H}(F) \in \mathsf{Vec}^{(\mathbb{R},\leq)}$.

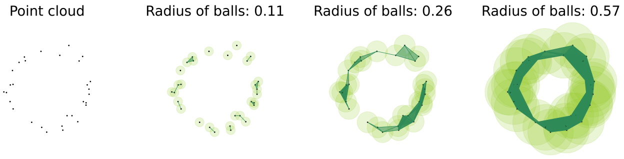


Figure 1: *Vietoris–Rips filtration of a point cloud of 30 points over a circle with Gaussian noise of scale* 0.1 *added, at 4 different filtration values (i.e., radius of the balls centered on the points). The simplicial complexes are built by adding an edge between any two points with overlapping circles (at a distance less or equal than twice the filtration value); and higher k-dimensional simplices for each* $k+1$ *subset of connected points. From the third image to the fourth, a 1-cycle (loop) appears in the filtration, encircling the hole in the shape. This is the type of feature that PH aims to capture.*

The p.f.d. persistence module $\mathrm{H}(F) \in \mathsf{Vec}^{(\mathbb{R},\leq)}$ is called the *persistent homology* module of the filtration. The previous construction is dimensionality-independent, allowing its application to both 2D and 3D data: the Vietoris–Rips filtration relies solely on pairwise distances, making it adaptable to any dimension. We demonstrate its use with 2D data in Section 4 and with 3D data in Section 5.

Observe that we obtain a spectrum of linear maps connecting the vector spaces, and thus, a natural way to study its structure is to consider the *ranks* of these maps. Let $\mathbb{R}^{2+} := \{(x,y) \in (\{-\infty\} \cup \mathbb{R}) \times (\mathbb{R} \cup \{\infty\}) : x \leq y\}$.

Definition 1 (Rank Function)**.** Given a p.f.d. persistence module $M \in \mathsf{Vec}^{(\mathbb{R},\leq)}$, its *rank function* is defined as

$$\begin{array}{rccl} \beta^M : & \mathbb{R}^{2+} & \to & \mathbb{Z} \\ & (x,y) & \mapsto & \operatorname{rank} M(x \leq y) = \dim \operatorname{Im}(M(x \leq y)). \end{array}$$

The space of rank functions will be denoted by $\mathcal{I}_1$.

Deterministic and probabilistic properties of the rank functions have been also studied under the name of *persistent Betti numbers*, first introduced by [ELZ03]. For instance, [DHS16, KP23] investigated the asymptotic normality and stabilizing properties of central limit theorems of persistent Betti numbers under the homogeneous Poisson and binomial processes and variants. This allows for implementations of the bootstrap procedure on the persistent Betti numbers [RKP23]. Further, in [BH21] the consistency and asymptotic normality of multiparameter persistent Betti numbers in large domains was determined, which is an important foundation to constructing statistical hypothesis tests. It is worth highlighting that this line of work deals mainly with *probabilistic* properties of rank functions (i.e., persistent Betti numbers), as opposed to their *statistical* performance in real data analysis, which is the focus of this work.

2.2. Persistence Diagrams and Barcodes

A *complete invariant* is a specific invariant assigned to a persistence module: it has the same value for all isomorphic persistence modules and is different for non-isomorphic ones. In single-parameter persistent homology, rank functions are equivalent to *persistence diagrams* and *barcodes*—two complete, discrete invariants obtained from the following distinct approaches.

Persistence diagrams can be traced back to the study of discontinuities of rank functions [FL01], later reinterpreted to visually capture the persistent homology of a filtered simplicial complex $H(F) \in \mathsf{Vec}^{(\mathbb{R},\leq)}$ [ELZ02]. We introduce them for general persistence modules $M \in \mathsf{Vec}^{(\mathbb{R},\leq)}$. Let $T = \{t_1,\dots,t_\ell\} \subset \mathbb{R}$ be the discrete set of values over which the module changes, and consider a sequence $\{s_0, s_1, \dots, s_\ell\}$ of real numbers interleaved with the elements of T: $s_{i-1} \leq t_i \leq s_i$. Also, set $s_{-1} = t_0 = -\infty$ and $s_{\ell+1} = t_{\ell+1} = +\infty$ and call $\overline{T} = T \cup \{-\infty, +\infty\}$. Lastly, define

$$\mu_i^j := \beta^M(s_{i-1}, s_j) - \beta^M(s_i, s_j) + \beta^M(s_i, s_{j-1}) - \beta^M(s_{i-1}, s_{j-1}). \tag{1}$$

Definition 2. The *persistence diagram* of $M \in \mathsf{Vec}^{(\mathbb{R},\leq)}$ is the multiset of points given by

$$\mathrm{Dgm}\,(M) := \{(t_i,t_j) \in \overline{T} \times \overline{T} : t_i < t_j\},$$

where each point (t_i, t_j) has multiplicity μ_i^j, union all the points in the diagonal $\partial = \{(x,y) \in \mathbb{R}^{2+} : x = y\}$ counted with infinite multiplicity. The space of persistence diagrams is denoted by $\mathcal{D}$.

Figure 2 shows persistence diagrams for 0- and 1-homology, representing components and loops, respectively, in three 3D point clouds. The sphere exhibits no persistent points far from the diagonal, indicating the absence of non-trivial loops on its surface—every loop can be deformed to a point. In contrast, the torus displays two persistent points representing its horizontal and vertical loops. The Stanford Bunny [TL94] (see `https://faculty.cc.gatech.edu/~turk/bunny/bunny.html`) exhibits a more complex interpretation, featuring one persistent 1-cycle.

The Structure Theorem due to [ZC05] (with the version in full generality due to [CB15] and [BCB20]) asserts that any p.f.d. persistence module M is isomorphic to an essentially unique (up to reordering) finite direct sum of indecomposable persistence modules

$$M \simeq M_1 \oplus \cdots \oplus M_m.$$

For single-parameter persistence, each M_j is an *interval persistence module*, i.e., there is a pair of values $b_j < d_j$, where d_j may be infinite, such that $M_j(x)$ is a copy of the field for all values $b_j \leq x < d_j$ and zero elsewhere. We denote these persistence modules by $\mathbb{I}[b_j, d_j)$.

Definition 3 ([ZC05,CdS10])**.** The *barcode* of $M \in \mathsf{Vec}^{(\mathbb{R},\leq)}$ is the list of indecomposables given by the Structure Theorem, or equivalently, the collection of intervals that define these indecomposables

$$\mathrm{Bar}\,(M) := \{[b_j, d_j) : M_j(x) \text{ is nontrivial for } b_j \leq x < d_j\}.$$

The length of a bar is called its *persistence*.

A barcode translates to a persistence diagram by plotting the left and right endpoint of each interval persistence module as an ordered pair. A persistence diagram translates to a barcode by turning each point (x,y) with $x < y$ in the persistence diagram into an interval persistence module beginning at x and ending at y. In this way, the persistence diagram is equivalent to a barcode, although the two definitions arise from different perspectives.

Rank functions are also in bijection with persistence diagrams and barcodes. We have seen above how to define persistence diagrams from an inclusion-exclusion formula (1) on rank functions. Moreover, rank functions can be seen as cumulative functions on the persistence diagrams: the value of the rank function $\beta^M(x,y)$ corresponds to the number of points (counted with multiplicities) in the region $(-\infty, x] \times [y, \infty)$ of the persistence diagram, providing the converse direction of the bijection. Figure 3 illustrates a persistence diagram and its corresponding rank function, where the equivalence between both objects becomes apparent.

2.3. Metrics and Stability in Persistent Homology

Stability results, which are the main theoretical contribution of this paper, give bounds for metrics defined on invariants of persistence modules. Given that there exist various metrics in persistent homology, the question of choosing appropriate metrics depends on the eventual goal.

Metrics on Barcodes and Persistence Diagrams. We recall the most widely used metrics to compare persistence diagrams in single-parameter persistent homology and their stability properties.

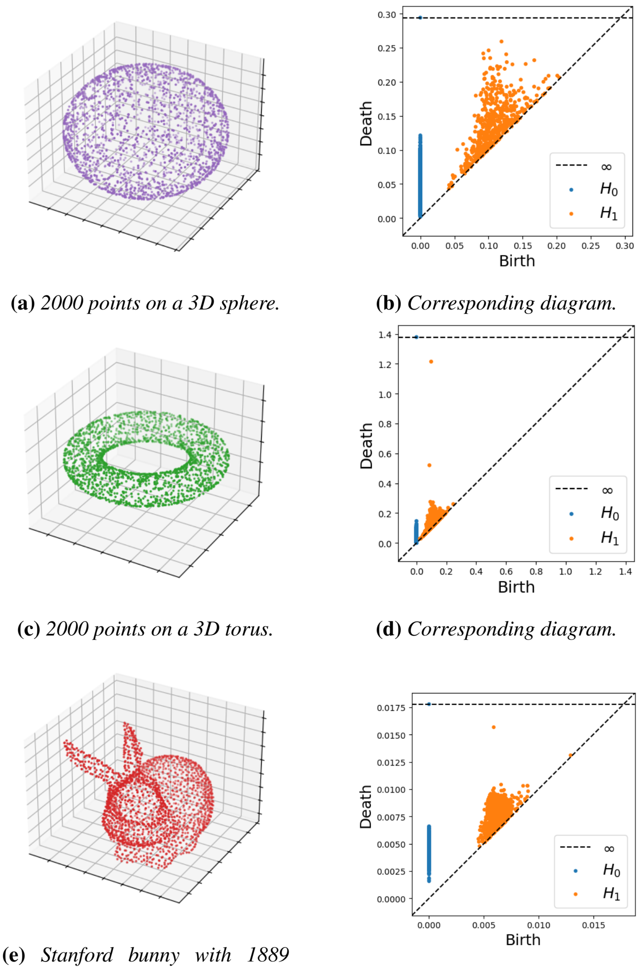
Definition 4. The *bottleneck distance* between the persistence diagrams D_1 and D_2 is defined as

$$d_B(D_1, D_2) := \inf_{\phi: D_1 \to D_2} \sup_{x \in D_1} \|x - \phi(x)\|_\infty,$$

where $\|\cdot\|_\infty$ is the infinity norm ℓ^∞ on $\mathbb{R}^2$ and ϕ ranges over all bijections between D_1 and D_2.

The first stability result in TDA concerns the bottleneck distance [CSEH07], proving that the map sending a tame function $f : X \to \mathbb{R}$ defining a filtration $F(x) := f^{-1}(-\infty, x]$ to its persistence diagram $\mathrm{Dgm}\,\mathrm{H}(F)$ is 1-Lipschitz with respect to the bottleneck distance for diagrams and the L^∞ distance for functions.

The bottleneck distance can be extended by replacing ℓ^∞ with ℓ^p norms, which give a stronger sense of proximity.



(a) *2000 points on a 3D sphere.*

(b) *Corresponding diagram.*

(c) *2000 points on a 3D torus.*

(d) *Corresponding diagram.*

(e) *Stanford bunny with 1889 points.*

(f) *Corresponding diagram.*

Figure 2: *Examples of 3D point clouds and their corresponding persistence diagrams.*

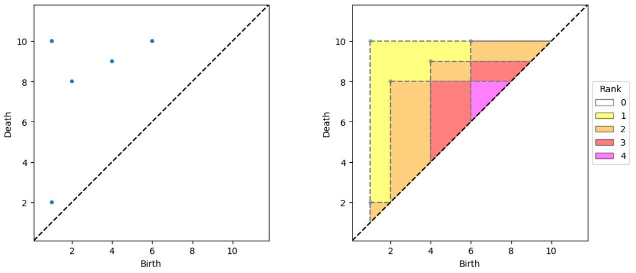


Figure 3: *Persistence diagram and its corresponding rank function.*

Definition 5. For $1 \leq p < \infty$ and $1 \leq q \leq \infty$ we define the *p,q-Wasserstein distance* between two persistence diagrams D_1 and D_2 as

$$W_{p,q}(D_1,D_2) := \inf_{\phi: D_1 \to D_2} \left[\sum_{x \in D_1} \|x - \phi(x)\|_q^p \right]^{1/p}$$

where $\|\cdot\|_q$ is the ℓ^q norm on $\mathbb{R}^2$ and ϕ ranges over all bijections between D_1 and D_2.

In our work, we will assume $p = q$ and refer to this metric simply as the p-Wasserstein distance W_p.

Wasserstein metrics are widely used in applications, especially p-Wasserstein distances with $p = 1,2$, and in some instances have been able to reveal more insight in application settings than the bottleneck distance. For example, in a protein flexibility analysis study, [BW20] show that p,∞-Wasserstein metrics provide more accurate results when comparing two conjugate point clouds obtained using atom-specific persistent homology [CW17, CMW18]. [GH10] explore the power of the Wasserstein distance in the context of statistical analysis of landmark-shape data. [Gid17] shows that the $W_{2,\infty}$ distance is able to detect premature evidence for critical transitions is financial data. [HBHM22] study barcodes endowed with Wasserstein distances for protein folding data and compare them with the Gaussian integral tuned [BBB*21] vector representation.

Despite the desirable performance of Wasserstein distances in applications, their stability properties have not been as broadly studied until recently, due to the level of additional technicality required. In [ST21], the following cellular Wasserstein Stability Theorem for p-Wasserstein metrics was established, validating many of the existing results in applications and justifying their continued use.

Theorem 6 (Cellular Wasserstein Stability Theorem, [ST21])**.** *Let $f,g: K \to \mathbb{R}$ be monotone functions on a finite CW-complex K. Then $W_p(\mathrm{Dgm}(f), \mathrm{Dgm}(g)) \leq \|f - g\|_p$.*

The existence of this result also justifies comparison with the Wasserstein metric to establish stability.

Metrics on Rank Functions. Rank functions lie in the space of functions from $\mathbb{R}^{2+}$ to $\mathbb{R}$ [RT16]. By fixing a metric on the space $\mathbb{R}^{2+}$, we can then define the L^p metric on this space of functions as follows:

$$d_{L^p}(f,g) := \left(\int_{\mathbb{R}^{2+}} (f-g)^p d\omega \right)^{1/p},$$

where ω is the measure on $\mathbb{R}^{2+}$ corresponding to the fixed metric. For notational simplicity, we write $\|f - g\|_p = d_{L^p}(f,g)$, keeping in mind that this metric comes from the L^p norm. This metric space is naturally endowed with a Hilbert structure for $p = 2$, which is a basic requirement for many FDA methodologies.

There are many choices for the metric as well as for the measure ω, however, the choice should avoid the pairwise distance between rank invariants being infinite. This happens, for example, when the metric is taken to be the Euclidean distance restricted to $\mathbb{R}^{2+}$, which implies ω is the Lebesgue measure. Here, two rank invariants have finite distance if and only if their infinite cycles have identical birth times.

Remark 7. In our work, such issues are circumvented by keeping in mind the goal of real data applications, where the posets over which we are defining our diagrams are always finite and in which the filtrations always end in a simplicial complex with trivial homology. This means that every cycle in the filtration is destroyed at some point, except for the 0-cycle representing the connected component of the space, which is always born at time 0. Thus, we can work with the Lebesgue measure without worrying about infinite distances between rank invariants: all bars in our barcodes will be finite, and therefore, every two rank functions will have finite pairwise distance, as desired.

3. L^p-Stability of Rank Functions

In this section, we present our contributions of two stability guarantees for rank functions endowed with L^p metrics with respect to the bottleneck distance and 1-Wasserstein distance for persistence diagrams. We focus on the L^p metric, as opposed to [ST21], who study the weighted version,

$$\|f-g\|_p^{\mathrm{w}} = \left(\int_{\mathbb{R}^{2+}} |f-g|^p \phi(x-y)\,dx\,dy\right)^{1/p}, \tag{2}$$

where the weight function $\phi(\cdot)$ inside the integral satisfies $\int_{\mathbb{R}} \phi(t) < \infty$. In particular, they choose $\phi(t) = e^{-t}$ and obtain the following stability result as a Corollary from Theorem 6.

Corollary 8 ([ST21])**.** *Rank functions with the L^q weighted metric (2) are 1-Lipschitz with respect to the p-Wasserstein distance between diagrams if and only if $p = q = 1$.*

The weight function in (2) ensures finite distances between rank functions [RT16, Lemma 2.1.], which allows for the definition of an inner product structure on finite sets of rank functions and thus justifies the rank FPCA method proposed in [RT16]. In our study, finiteness of L^p distances between rank functions is guaranteed as explained in Remark 7, and the Hilbert space structure follows directly. As a result, the use of such a weight is not needed in our setting, and in fact it is not helpful for our purposes: its introduction fundamentally changes the expressions in the computations involving the metric, so that the proof of Corollary 8 by [ST21] does not apply and cannot be replicated in our work. This also underlines the inherent differences between our contribution compared to [ST21].

Other Hindrances to Rank Function Stability. Under the original name of size functions [Fro92], several notions of "pseudo-stability" were established. These were achieved under *pseudometrics*, where the distance between two distinct points can be zero, which makes pseudometrics less desirable for use in real data analyses (due to difficulties in intepretability) and therefore also makes pseudo-stability less desirable as a validating property to justify the use of rank functions as topological summary statistics. Some examples are the *deformation distance*, which were adapted to persistence diagrams and is more widely known today as the 1-Wasserstein distance between persistence diagrams; the *Hausdorff pseudo-distance*, which similarly gave rise to the bottleneck distance between persistence diagrams; and the L^p *pseudo-distance*, which exhibited an unstable nature and did not appear to inspire any well-known distance in persistent homology.

In particular, it is worth noting that [dFL03] and [dFL10] renamed the Hausdorff pseudo-distance to the *matching distance* to emphasize the fact that its computation amounts to finding an optimal matching between multisets, in the same way that the bottleneck distance does, and which was used in establishing the first results of stability for persistent homology. The matching distance was used to establish stability under noisy perturbations when restricted to a subset of size functions called *reduced* size functions.

3.1. Stability Under the Bottleneck Distance

The most straightforward way to achieve stability for rank functions is to restrict away from the diagonal, which is known to complicate the metric geometry of the space of persistence diagrams [TMMH14, CM22]. To do this, we introduce a truncation of the rank function that will allow us to compare its sensitivity to noise to that of the bottleneck distance.

Definition 9. For any rank function β and any $\delta > 0$, we define the δ-*truncated rank function* as

$$\beta_\delta := \beta \cdot \mathbb{1}_{\mathbb{R}^{2+}_\delta}$$

where $\mathbb{1}_{\mathbb{R}^{2+}_\delta}$ is the indicator function of the set $\mathbb{R}^{2+}_\delta := \{(x,y) \in \mathbb{R}^{2+} : y > x + \delta\}$.

In other words, the truncated rank function is just the rank function excluding a strip of width $\delta > 0$ above the diagonal ∂ (see Definition 2). The truncated rank function locally satisfies a Hölder inequality for the L^p norm with respect to the bottleneck distance on persistence diagrams.

Proposition 10 (Bottleneck Stability for Truncated Rank Functions)**.** *Let $1 \le p < \infty$ and M be a p.f.d. persistence module with finite intervals in its barcode decomposition. For every $\delta > 0$, there exist $1 \ge \eta > 0$ and $K_{M,p} > 0$ such that any persistence module N satisfying*

$$d_B(\mathrm{Dgm}(M), \mathrm{Dgm}(N)) < \eta$$

also satisfies

$$\left\|\beta_\delta^M - \beta_\delta^N\right\|_p \le K_{M,p} \cdot d_B(\mathrm{Dgm}(M), \mathrm{Dgm}(N))^{1/p}. \tag{3}$$

In other words, the map $(\mathcal{D}, d_B) \to (\mathcal{I}_1, L^p)$ which sends each persistence diagram to its corresponding rank function is locally Hölder with exponent $1/p$.

Remark 11. The constant appearing in Proposition 10 is precisely

$$K_{M,p} = m(2R+2)^{1/p} \tag{4}$$

where m is the number of points in $\mathrm{Dgm}(M)$ and $R = \max\{|d_i - b_i| : 1 \le i \le m\}$ is the maximum persistence in such diagram.

Notice that we can always obtain a bound similar to that in (3) where the constant depends on both persistence modules M and N (see Appendix A). Proposition 10 refines this approach by obtaining a constant that only depends on the persistence module M. Nevertheless, an important limitation of Proposition 10 is that it discards the points close to the diagonal—an important component in the definition of persistence diagrams—even though it sheds light on the behavior of rank functions in discrete settings. The bounds appearing in the proof (see Appendix) will be useful in our next derivations.

3.2. Stability Under the 1-Wasserstein Distance

The previously-mentioned limitation of Proposition 10 is that it holds only for points away from the diagonal, which highlights the differences in sensitivity to noise between the L^p norms on rank functions and the bottleneck distance on persistence diagrams. This observation was already made by [LF97]; we further develop this observation with a formal study in this paper.

As we will now establish, *full* rank functions satisfy a stability property with respect to the 1-Wasserstein distance for persistence diagrams. As mentioned in Section 2.3, stability properties of the Wasserstein metric on persistence diagrams were not studied in detail until very recently, which limited their applicability as upper bounds in stability studies. We use the Cellular Wasserstein Stability Theorem (Theorem 6) [ST21] to establish stability for rank functions.

Theorem 12 (1-Wasserstein Stability for Rank Functions)**.** *Let $p = 1, 2$; and M be a p.f.d. persistence module with finite intervals in its barcode decomposition. Then there exists a constant $C_{M,p} > 0$ such that for any other p.f.d. persistence module N satisfying $W_1(\mathrm{Dgm}(M), \mathrm{Dgm}(N)) \leq 1$, we have*

$$\left\|\beta^M - \beta^N\right\|_p \leq C_{M,p} \cdot W_1(\mathrm{Dgm}(M), \mathrm{Dgm}(N))^{1/p}. \qquad (5)$$

In other words, the map $(\mathcal{D}, W_1) \to (\mathcal{I}_1, L^1)$ sending a persistence diagram to its corresponding rank function is locally Lipschitz, and the same map between the spaces $(\mathcal{D}, W_1) \to (\mathcal{I}_1, L^2)$ is locally Hölder with exponent $1/2$.

Remark 13. The constants appearing in Theorem 12 are

$$K_{M,1} = 2R + 2, \qquad (6)$$

and

$$K_{M,2} = 2 \cdot \max\left\{(2(R+1)m)^{1/2}, \frac{1}{\sqrt{2}}\right\} \qquad (7)$$

where m is the number of points in $\mathrm{Dgm}(M)$ and $R = \max\{|d_i - b_i| : 1 \leq i \leq m\}$ is the maximum persistence in such diagram.

Theorem 12 provides a stronger theoretical guarantee than Proposition 10, not only because it also covers the diagonal, but also because the constant $C_{M,p}$ in (5) is smaller than the constant $K_{M,p}$ in (3) ($p = 1, 2$). For $p = 1$, the latter depends on the number of points in the persistence diagram of M and R, its maximum persistence, whereas the former only depends on R. For $C_{M,2}$ we maintain a dependence on the number of points in the diagram of M, but it still provides a tighter bound than $K_{M,2}$, since this dependence is squared instead of linear.

4. Inference with Rank Functions

In this section, we study the performance of rank functions in machine learning tasks on real data. Specifically, we focus on *inferential* tasks—namely, classification and prediction—in the single-parameter setting.

4.1. Using Persistent Homology in Data Analysis

Persistent homology captured by persistence diagrams and barcodes fulfils the essential requirements for data analysis: interpretability (via the Structure Theorem) and stability (with various stability results available). Moreover, it is known to be a viable space for probability and statistics [MMH11, BGMP14]. Despite these desirable properties, there remain challenges in utilizing persistence diagrams and barcodes in the full scope of statistical analysis, mainly due to their complicated geometry which results in, for example, non-unique geodesics and Fréchet means [TMMH14]. As a consequence, in statistical questions there are, broadly speaking, two approaches to handling persistent homology in data analysis. One approach entails developing new data analytic methodology, such as machine learning algorithms and statistical models, to accommodate barcodes or persistence diagrams directly (e.g., [FLR*14, RHBK15, HKNU17]). The other approach entails vectorizing barcodes persistence diagrams to apply existing methods (e.g., [CFL*14, Bub15, AEK*17]).

Our approach diverges from both strategies by exploring rank functions as equivalent, alternative representations of persistent homology to leverage theory from functional data analysis (FDA). Methods for FDA, which were constructed to analyze data in the form of functions, are well-established in statistics and many of them arise as extensions of methodologies from multivariate data analysis. Rank functions equipped with the L^2 metric, as discussed above, form a metric space that admits a Hilbert structure, and thus become a viable data structure amenable to integrating FDA with persistent homology.

We emphasize here that we are not modifying the output of persistent homology, as vectorization methods do to persistence diagrams or barcodes, since rank functions are equivalent to barcodes and persistence diagrams. Equally, we are not developing new methodology to accommodate persistent homology, as the existing field of FDA is directly applicable to persistent homology captured by rank functions.

4.2. Functional Support Vector Machine on Single-Parameter Rank Functions

Our first study is the performance of rank functions in the single-parameter persistence setting in classification. Specifically, we study the clinical application of discerning heart rate variability between healthy individuals and post-stroke (acute ischemic) patients using a *functional support vector machine* (FSVM) [RV06].

Functional Data Analysis and High Dimensionality. Functional data, where datasets are collections of functions, are inherently infinite dimensional, which means that the discrete realizations of the underlying surfaces or curves are high dimensional and can cause various problems, such as overfitting.

However, FDA methodologies are generally insensitive to dimensionality; they circumvent the problem of high dimensionality in broadly two ways. One way is via dimensionality reduction, where projecting the data onto a smaller collection of orthogonal bases produces lower dimensional vectors that are robust to discretizations. The choice of basis function depends on the underlying functions; e.g., the Fourier basis functions can be used to approximate functions that exhibit cyclical properties, while wavelet basis functions can be used to approximate functions that exhibit fluctuations. Alternatively, a data driven approach may be adopted, with one of the most commonly used techniques being *functional* PCA (FPCA) [DPR82], which is a descriptive technique (rather than inferential, which is the focus of our work), and has previously been implemented on rank functions by [RT16]. FPCA works on the principle of finding orthonormal basis functions and projecting onto a finite subset of them with the greatest variation.

Another alternative is to ensure within the construction of the methodology that it is invariant to the choice of grid points for evaluation, such that as the number of grid points increases, convergence to an appropriate result is guaranteed by construction. This is known as the *refinement invariance principle* [CL08] and will not be our focus here.

Functional Support Vector Machines. Classical SVMs are supervised binary classification methods that seek to find the optimal boundary in the feature space which distinguishes between observations of the two categories in a way such that the distance to the boundary from any data point is maximized. For our FSVM application on rank functions, let $\{f_1, f_2, \dots, f_N\}$ be a collection of centered, discretized rank functions with corresponding labels $(y_i)_{i=1,\dots,N} \in \{-1, 1\}$ identifying the two groups. We adopt the *soft margin* approach to determine the boundary for conventional reasons, as it is used in most computational packages. Here, "soft" refers to certain deviations from the boundary being allowed for classification in the approach. The boundary can be defined for some function $\psi \in \mathcal{H}$, $\mathcal{H}$ being a Hilbert space, and scalar $b \in \mathbb{R}$ as $\langle \psi, f_i \rangle + b = 0$, such that $y_i(\langle \psi, f_i \rangle + b) \geq 1 - \zeta_i, \forall\, i = 1, \dots, n$, where ζ_is are slack variables in the soft margin approach providing trade-offs between accuracy and overfitting. The optimal boundary is one which maximizes the margin given by $\frac{2}{\|\psi\|}$ and the optimization problem can be solved more easily through its dual formulation, by sequential minimal optimization [Pla98].

Practically, however, not all data are linearly separable, in which case, a workaround is to project the data to higher dimensions where a clearer division between the two classes then becomes observable. This technique is referred to as the *kernel trick* [BGV92]. Let ϕ be the projection. Then the inner product from the previous optimization problem is replaced by a kernel function, i.e., $\langle \phi(f), \phi(g) \rangle$. Examples include the *polynomial* kernel (of order d), $K(f,g) = (\langle f, g \rangle + 1)^d$, and the *Gaussian radial basis function* (GRBF) kernel, $K(f,g) = \exp\left(-\gamma \|f - g\|^2\right)$ [RV06]. Their usage can be seen in a wide range of biomedical applications, for example, in the identification of PTSD patients based on resting state functional magnetic resonance imaging (fMRI) [SRS*22] and also in the classification of brain functions on electroencephalographic signals [XWZZ08]. In these two examples, the GRBF kernel outperforms other kernels and classifiers.

Data Description. The dataset we study consists of 86 sequences of 512 beat-to-beat time intervals (RR series) extracted from electrocardiograms from a clinical study between two groups of people in a similar age category: one group of 46 healthy individuals used as control and one group of 40 patients who have recently experienced stroke episodes [GGR*21, NGGP21]. Stroke patients generally show reduced heart rate variability compared to healthy individuals [LSKS*18]. We aim to discern differences in heart rate variability between the two groups using persistent homology rank functions. For computations, we first linearly interpolated between the points in the RR series to construct continuous functions over time and then we constructed a sublevel set filtration based on the height function in the positive y-direction. We compute the zero-dimensional persistent homology rank function for each individual's RR series in the dataset. An example of steps in this process is visualized in Figure 4.

Training the FSVM Classifier and Evaluating Performance. On the set of rank functions computed from the data, we train FSVM classifiers using the linear kernel, GRBF kernel, and polynomial kernels of three different degrees (2, 3, and 5). Since we are working with discretized functions, we consider both the rank functions as computed from the data, and transformed versions using dimension reduction. We experiment with both a set of data-driven basis functions obtained from FPCA and a set of standard basis functions—the Haar wavelets. Haar wavelets, in particular, have also been used in other inferential tasks in persistent homology; [HBM23] use them in persistent homology density estimation.

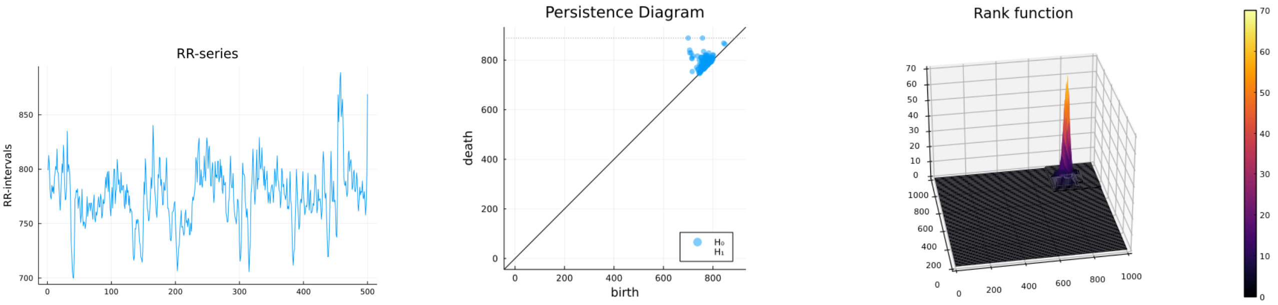
We evaluate the performance of the binary classifiers using two metrics: the accuracy and the area under the curve of the Receiver Operator Curve (AUC–ROC). The evaluation is carried out and averaged over ten iterations of five-fold cross-validation.

Results. The performance results of the FSVM classifier are summarized in Table 1. Overall, the FSVM classifiers with degree two and three polynomial kernels produce highest accuracy, $> 80\%$, compared to the performance of other kernels implemented and gave on average AUC–ROC values of over 0.8, indicating excellent discrimination between the two categories. We also include the runtimes for these computations, including: (a) the computation of the PH rank functions; (b) the training of the corresponding SVM; and (c) the computation of 10 iterations of the accuracy and AUC–ROC over five-fold cross-validation, and the corresponding average. Although we observed a higher runtime for the linear SVM on PCA-projected data as well as convergence limitations within the specified maximum number of iterations, the computations are nevertheless manageable in terms of runtime.

Table 1: *Average accuracy, AUC–ROC and runtimes of classifiers constructed on rank functions and projected rank functions with linear, GRBF, and polynomial kernels over ten iterations of five-fold cross-validation.*

Kernel	Discretized rank functions + SVM		
	Accuracy	AUC-ROC	Runtime (s)
Linear	39.5	0.408	10.5
GRBF	75.8	0.762	10.2
Polynomial (d=2)	82.6	0.829	8.51
Polynomial (d=3)	81.6	0.829	9.06
Polynomial (d=5)	76.0	0.775	8.28
	Projection on PCA basis + SVM		
Linear	66.8	0.681	53.2
GRBF	50.3	0.502	3.14
Polynomial (d=2)	84.2	0.842	3.17
Polynomial (d=3)	82.0	0.829	3.14
Polynomial (d=5)	75.5	0.774	3.17
	Projection on Wavelet basis + SVM		
Linear	39.7	0.404	11.5
GRBF	75.0	0.756	12.0
Polynomial (d=2)	84.0	0.842	9.99
Polynomial (d=3)	80.3	0.816	10.4
Polynomial (d=5)	76.6	0.786	10.0

In general, linear kernels perform less well in discriminating be-



(a) **(b)** **(c)**

Figure 4: *An example showing (a) the RR series for a healthy individual, (b) its respective persistence diagram and (c) its rank function.*

tween functions of the two categories, except when the dimensionality of the rank functions is reduced by projecting onto its principal component functions. In doing so, we considered only the first 30 principal component functions with the largest eigenvalues, which explains 95% of the variation. For the two transformations, working with lower dimensional vectors and polynomial kernels, we see similar, and slight improvements, in accuracy and AUC–ROC of the classifiers than the original rank function.

Following [GGP*21], we take the SVM classifier quadruples of clinical indices related to the RR series and quadruples of features derived from persistence diagrams as our input data; we now discuss this analysis in further detail. The AUC–ROCs of the optimal models for both approaches can be found in Table 2 as reported in [GGP*21]. The optimized performance of the classifier using rank functions was, on average, better than the performance of standard heart rate variability indices on the frequency and time domain, which only achieved an average AUC–ROC of 0.79 and 0.75 respectively [GGP*21]. For the persistence-based approach in [GGP*21], a wide range of topological indices were extracted from the persistent diagrams. Some indices were more typical, such as the total number of intervals, the sum of the lengths of all the persistent intervals, various mean and standard deviations; some were less conventional, such as the *persistent entropy* (due to [AGDST20] and given by $h(r) = \sum_{i=1}^{n} -\frac{\ell_i}{L}\log_2\frac{\ell_i}{L}$, where ℓ_i is the length of an interval and L is the sum of the length of all intervals), the *frac 5%, 100, 200* (the number of intervals whose lengths are shorter than the threshold of 5% the length of the longest interval, 100ms, and 200ms), or the *signal-to-noise ratio* (which is the ratio of the sum of "signal" intervals over the sum of "noise" intervals, where the signal is considered to be the intervals longer than the threshold of 5% the length of the longest interval and those not passing this threshold are considered as "noise"). They introduce new geometric measures called the *topological triangle indices*, which are based on the triangular interpolation on the RR interval histograms classically used in heart rate variability analysis and work by constructing a triangle on the persistence diagram with one side lying on the diagonal, enclosing a set of points with a small percentage of outliers and such that the triangle is as compact as possible. With these topological indices, it was shown that optimized combinations of parameters were able to achieve up to 0.84 AUC performance. However, indeed, computing these indices is an involved process.

Table 2: *AUC–ROC of linear SVM conducted on quadruples of persistence and non-persistence based features as reported in [GGP*21] using three-fold cross-validation and a standard scaler on the input data.*

Model parameters	AUC–ROC
Optimal model for frequency parameters	0.75
Optimal model for time domain parameters	0.79
Triangle height, location, no. of intervals, length sum	0.83
Triangle location, misalignment, no. of intervals, frac 5%	0.83
Triangle location, no. of intervals, length sum, signal to noise	0.84
Triangle width, no. of intervals, length sum, frac 5%	0.84

For an additional comparison to a typical persistence-based approach, we also trained SVM classifiers on persistence images [AEK*17] and persistence landscapes [Bub15]—stable vectorizations computed from the barcodes. For persistence images, SVM with a linear kernel achieved optimal performance, with an accuracy of 68.5% and an AUC–ROC of 0.793, amongst SVM with alternative kernels, sparse SVM, and SVM applied after dimensionality reduction with PCA. For persistence landscapes, SVM with a GRBF kernel achieved optimal performance, with an accuracy of 81.0% and an AUC–ROC of 0.903. Full results are shown in Appendix C.

In conclusion, the performance we find using rank functions, as a direct and equivalent representation of persistent homology (as opposed to vectorized and manipulated) has better performance over classification with persistence images and is on par and slightly improved over the much more involved approach of computing topological indices from [GGP*21] and the one using persistence landscapes.

5. Rank Functions in Multiparameter Persistent Homology

In this section, we explore the use of *multiparameter* persistent homology [CZ07] in real data applications. This is currently an active area of research in TDA, due to interpretive and computational difficulties.

5.1. Multiparameter Persistent Homology: The Struggle to Generalize Barcodes

There is an important distinction between *single-parameter* persistence modules, i.e., diagrams such as those we have considered so far $M \in \mathsf{Vec}^{(\mathbb{R},\leq)}$, and *multiparameter* persistence modules [CZ09], i.e., diagrams which allow indexing over the poset $(\mathbb{R}^n, \preceq)$. Here, $\preceq$ is the product order inherited from the total order in the reals, namely $(x_1,\dots,x_n) \preceq (y_1,\dots,y_n)$ if $x_i \leq y_i$ for all $i = 1,\dots,n$. The construction discussed in Section 2.1 giving rise to persistent homology can be replicated for these posets to obtain *multifiltrations* and *multiparameter persistent homology*.

The Structure Theorem in Section 2.2 can be extended to general p.f.d. persistence modules indexed over a small category [BCB20], which includes the case of multiparameter persistence. As mentioned before, for single-parameter persistence, the only possible indecomposable modules are interval modules, i.e., modules $\mathbb{I}[b,d)$ supported on intervals $[b,d) \subset \mathbb{R}$, allowing for the definition of barcodes as multisets of intervals, as well as the interpretability of births and deaths of topological features corresponding to the intervals. Although there is a natural extension of the concept of an interval for general posets, the representation type of indecomposable modules over these posets is wider than those supported on intervals, so that no direct, parallel definition of barcode exists. Moreover, it has been shown that there is in fact no hope for a complete, discrete invariant in multiparameter persistence [CZ07].

Given the lack of complete invariants for multiparameter persistent homology, a central research interest has been the development of incomplete, interpretable, and computable invariants. Some strategies to define incomplete invariants include viewing n-dimensional persistence modules as n-graded modules over polynomials [CZ07] and capitalizing the invariants already existing for such objects, such as minimal presentations and multigraded betti numbers [LW15,LW22] or multigraded associated primes and local cohomology [HOST19]. Several other proposals bypass the Structure Theorem entirely. Patel [Pat18] generalizes the Möbius inversion in single parameter persistence connecting rank functions and persistence diagrams to define generalized persistence diagrams. Kim and Mémoli [KM21] introduced generalized rank invariants, proving they are the courterpart to generalized persistence diagrams in the Möbius inversion by Patel [Pat18] in the multiparameter setting. Developing a theory of modules over posets, Miller [Mil20] defined QR codes for n-dimensional modules. Lastly, using resolutions and rank-exact structures, Botnan et al. [BOO22] defined signed decompositions and signed barcodes, extending single parameter barcodes and including the generalized persistence diagrams by [KM21]. As in single-parameter persistence, a third approach entails vectorizing the output of persistent homology by embedding the modules in a Hilbert space. Some of these vectorizations are known to result in a loss of information, however. Popular vectorization methods in multiparameter persistence include persistence landscapes [Vip20], images [CB20], and kernels [CFK*19], among others.

Remarkably, rank functions can be extended to multiparameter persistence quite naturally. Let $\mathbb{R}^{2n+} := \left\{(x,y) \in (\{-\infty\} \cup \mathbb{R})^n \times (\mathbb{R} \cup \{\infty\})^n : x \preceq y\right\}$.

Definition 14 (Rank Invariant)**.** Given a p.f.d. multiparameter persistence module $M \in \mathsf{Vec}^{(\mathbb{R}^n,\preceq)}$, its *rank invariant* is defined as

$$\begin{array}{rccl} \beta^M: & \mathbb{R}^{2n+} & \to & \mathbb{Z} \\ & (x,y) & \mapsto & \operatorname{rank} M(x \preceq y) = \dim \operatorname{Im}\left(M(x \preceq y)\right). \end{array}$$

The space of rank invariants for n-dimensional persistence modules will be denoted by $\mathcal{I}_n$.

In any of these approaches, including in the case of rank invariants, applications to real data are in their infancy. A main obstacle is the lack of efficient software to compute the invariants; the current technology also being restricted to two parameters. Rank invariants for biparameter persistence modules can be computed using RIVET [The20,LW15,LW22]—currently the standard software for most strategies in defining multiparameter invariants. Developing efficient algorithms and optimizing existing software remains an active research area [KN19, SIDFL20, FKR23]

The question of metrics for rank invariants is equally important as for rank functions. The well-established *matching distance* for rank invariants restricts multiparameter persistence modules to lines [dFL03, dFL06, dFL10]. The matching distance is known to be stable for rank invariants for filtrations obtained as sublevel sets of a function $f : X \to \mathbb{R}^n$ on X a triangulable space and with respect to the L^∞ distance between two filter functions [CFF*13]. In a more general setting, the matching distance is also known to be stable with respect to the *interleaving distance* [Les15,Lan18]; and it is computable in polynomial time [KLO19, KR21].

A significant challenge, however, in using the matching distance in applications—especially in inferential tasks—despite its computability is that it does not induce a Hilbert structure on the space of rank invariants, which is often a condition needed in order to adapt FDA methods (e.g., [CMC*20]). Thus, in our real data application, our focus remains on the L^2 distance on rank invariants, which is also efficiently computed over a discretized grid, providing the necessary Hilbert structure to integrate with FDA methods.

5.2. Application of Biparameter Rank Functions in Lung Tumor Classification

We now demonstrate the inferential ability of the biparameter rank functions using nonparametric supervised learning methods on real data. The application focus is to predict lung tumor malignancies from computed tomography (CT) images, which has been studied previously by [VMS*23] using single-parameter topological summary statistics. Here, we aim to show that using biparameter persistent homology captures additional distinguishing features of the tumor morphology, both on a local and global scale, which, together with the rank functions, leads to improved classification.

Data Description. We study images from the Lung Image Database Consortium (LIDC), which is freely available from The Cancer Imaging Archive (TCIA) [AIMB*11, AIMB*15]. From the LIDC data, we extract a subgroup of 70 chest CT scans, complete with annotations and masks, consisting of those with primary tumors that have either been diagnosed as benign (29) or malignant (41).

Following the approach in [VMS*23], we convert the collection of CT scan images and masks into 3D point clouds of landmarks on the tumor surfaces by sampling, as shown in Figure 5. On the resulting point clouds, we compute the biparameter rank invariants using two types of bifiltrations, both of which are extensions of the Vietoris–Rips filtration—namely, the *degree–Rips* filtration and the *height–Rips* filtration. The degree refers to the degree of connectivity measured on each vertex of the 1-skeleton, while the height is measured along the z coordinate, in the direction of stacking of the tumor slices. Using the bifiltration captures prominent features as they develop on the tumor surface along both filtration functions.

Classification. We utilize the following two supervised classification methods:

- ***k*-Nearest Neighbors** [CH67]: This algorithm is a fundamental classification technique for both multivariate and functional data, where the decision for a new datum is made based on the majority vote of its k-closest neighbors. The method is adaptable to general metric spaces since the proximity can be measured using various metrics; the method has been studied in persistent homology by [MMM17, CLM24]. Here, we work with the rank invariants in L^2.
- **Functional Maximum Depth** [LPR09]: This method uses an extended notion of *depth* on functional data to classify curves and surfaces. For a collection of rank invariants, $f_1(x), \dots, f_n(x)$, $x \in \mathcal{X}$, with $\mathcal{X}$ its the domain, we define a *band* as the region or hyperspace bounded by an upper and lower function as follows:

$$B(f_1, \dots, f_n) = \{(x,y) : x \in \mathcal{X}, \min_{i=1,\dots,n} f_i(x) \le y \le \max_{i=1,\dots,n} f_i(x)\}.$$

The *band depth BD* is the total number of times that f lies within the band formed by a subcollection of the functions $BD_{n,J}(f) := \sum_{j=2}^{J} BD_n^{(j)}(f)$ for a fixed value J, where $2 \le J \le n$ and

$$BD_n^{(j)}(f) = \binom{n}{j}^{-1} \sum_{1 \le i_1 < i_2 < \cdots < i_j \le n} \mathbb{1}_{Bd(f)}. \tag{8}$$

Here, $\mathbb{1}_{Bd(f)}$ is the indicator function of the set $Bd(f) := \{G(f) \subseteq B(f_{i_1}, f_{i_2}, \dots, f_{i_j})\}$ for $G(f) = \{(x, f(x)) : x \in \mathcal{X}\}$. For our application, we use a modified band depth *MBD* where instead of using a strict indicator function in (8), we consider the proportion of the hyperspace for which f lies within the band:

$$MBD_n^{(j)}(f) = \binom{n}{j}^{-1} \sum_{1 \le i_1 < \cdots < i_j \le n} \omega(A(f; f_{i_1}, \dots, f_{i_j}))/\omega(\mathcal{X}), \tag{9}$$

where ω is a Lebesgue measure on $\mathcal{X}$ and $A(f; f_{i_1}, f_{i_2}, \dots, f_{i_j}) \equiv \{x \in \mathcal{X} : \min_{r=i_1,\dots,i_j} f_r(x) \le f(x) \le \max_{r=i_1,\dots,i_j} f_r(x)\}$. Hence, for any new invariant f, it will be assigned to the class in which the modified band depth (9) is maximized.

In the dataset we study, each of the tumor images is classified as either benign or primary malignant. Our task is to use topological summaries of the images as predictors to determine whether a primary tumor is benign or malignant. We train the classifiers on the biparameter rank invariants computed from the whole dataset. Taking a 75/25 split of the data for training and testing and averaging over 50 iterations, we obtain the results in Table 3. Furthermore, 24 of the 29 CT scans of benign tumors and 17 of the 41 CT scans of malignant tumors were taken with added contrast material. Refining to this smaller set, we see further improvements in the predictive accuracies reported in Table 4.

Results. Without added contrast, our results show that by training a modified maximum depth (*MBD*) classifier, we attain an optimal accuracy and AUC–ROC of 70.8 and 72.0 with the degree–Rips filtration. Overall the performances of MBD classifiers trained on the different bifiltrations are better than the performance of k-NN classifiers and also the optimized model in [VMS*23] which achieved an AUC–ROC of 67.7 on this dataset.

Table 3: *Accuracy and AUC–ROC of classification between benign and malignant primary tumors in the LIDC dataset.*

Primary Benign vs Malignant		k-NN		MBD	
		Accuracy	AUC–ROC	Accuracy	AUC–ROC
	Height–Rips	61.4	0.599	68.8	0.691
	Degree–Rips	63.3	0.618	70.8	0.720

Table 4: *Accuracy and AUC–ROC of classification between benign and malignant primary tumors in LIDC dataset with added contrast material.*

Primary Benign vs Malignant (with contrast)		k-NN		MBD	
		Accuracy	AUC–ROC	Accuracy	AUC–ROC
	Height–Rips	83.8	0.830	76.9	0.768
	Degree–Rips	80.0	0.791	72.5	0.727

Moreover, comparing the performance on the subset of data with added contrast material, we find that the k-NN classifiers achieved better AUC–ROC with both filtrations than the optimal model in [VMS*23] which had an AUC–ROC of 78.0 on average. In fact, the average AUC–ROC for the best k-NN classifier based on height–Rips filtration was 83.0. Therefore, indeed we find that the additional information captured by the bifiltration leads to better predictions.

6. Discussion

In this paper, we revisited persistent homology—which provides a geometric representation of data and can be used as a tool for point cloud processing—represented by rank functions in inferential, nonparametric FDA settings. In order to be able to validate our findings from the data analyses, we derived stability conditions on rank functions endowed with an appropriate metric over function space for FDA implementation: namely, the L^2 distance, which provides a Hilbert structure on the space of rank functions. Stability of rank functions, alternatively known as persistent Betti numbers, was well established with respect to the matching distance [CFF*13], while, to the best of our knowledge, a thorough understanding of the stability behavior of rank functions endowed with the L^p metric was previously missing in the literature. We fill this gap with Proposition 10, showing that we can compare to the bottleneck distance for barcodes only when restricting to points away from the diagonal; and Theorem 12, where we are also able to find bounds for rank functions with respect to the 1-Wasserstein

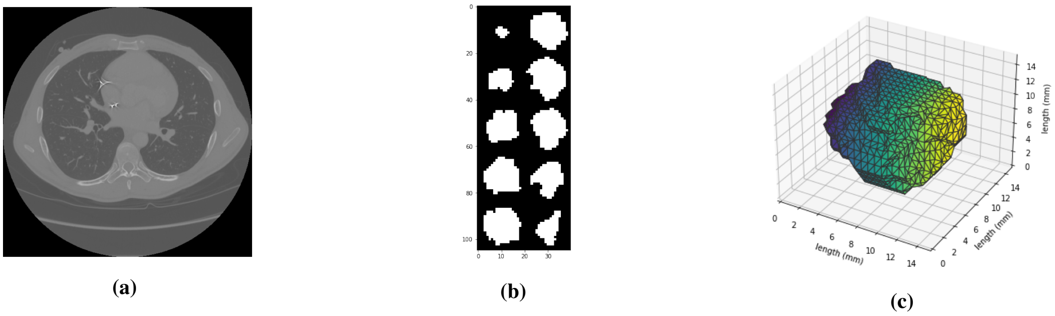


Figure 5: *An example showing the data extraction process. Annotated CT images (a) from the LIDC combined with masks (b) are converted into a 3D surface (c) from which we can sample a point cloud.*

distance. We also evaluated the performance of the topological representation of data as rank functions in two real-world applications and found that incorporating topological information outperforms previous non-topological methods, as well as other persistence-inspired approaches that use complicated constructions rather than equivalent representations of persistence diagrams. In addition to performing less well, these topological constructions based on persistent homology are more difficult to interpret and relate back to the original data. A particularly important contribution in this work that we highlight is in the second application where we used biparameter rank invariants (i.e., rank functions adapted to multiparameter persistent homology). The adaptation of multiparameter persistent homology to real data is still in its infancy and far from as widespread as in the single-parameter case, because, given the lack of direct extensions for the barcodes and persistence diagrams to higher dimensions, much of the work in the recent years has been foundational and devoted to finding alternative invariants that capture as much information as possible from computing persistent homology. We have found that using rank invariants directly in our real data analysis and machine learning task of classification provides excellent results, encouraging the use of this invariant in multiparameter persistent homology.

This naturally inspires several directions for future research. The first would be extending the theoretical stability results in Section 3 to multiparameter persistent homology represented by rank invariants. As in this work, this would be an important theoretical result needed to validate the experimental findings in this paper as well as justify its continued use in applying multiparameter persistent homology in real data applications. An additionally important direction to study would be comparative: given the multitude of invariants proposed in the literature on multiparameter persistent homology, understanding the performance of rank invariants in comparison to that of other existing variants would provide a basis and guideline for invariant usage in real data applications. Specifically, we would like to know whether rank invariants are able to capture more information, as a direct invariant obtained from persistent homology, than other functional vectorizations which embed modules into Hilbert spaces. In the single-parameter setting, this is true, as we explored in this work.

Software & Data Availability

All code and data to reproduce results in this paper are available from `https://github.com/Qiquan-Wang/rank_function_stability.git`. The heart rate variability data were collected and studied by [GGR*21] and [NGGP21]. The lung imaging LIDC data were obtained from TCIA [AIMB*11].

Acknowledgements

We wish to thank Yueqi Cao, Lorin Crawford, Herbert Edelsbrunner, Claudia Landi, Amit Patel, and Nicole Solomon for helpful discussions. Q.W. is grateful to Matthew Williams for his support and guidance. G.H.P. and A.M. also wish to acknowledge the Max-Planck Institute for Mathematics in the Sciences in Leipzig, Germany for hosting their visit in February 2018, where this work began.

I.G.R. is funded by the UK Engineering and Physical Sciences Research Council [EP/S021590/1] Centre for Doctoral Training in Geometry and Number Theory (The London School of Geometry and Number Theory), University College London. Q.W. is funded by a CRUK–Imperial College London Convergence Science PhD studentship at the UK EPSRC Centre for Doctoral Training in Modern Statistics and Machine Learning (2021 cohort, PIs Monod/Williams), which is supported by Cancer Research UK under grant reference [CANTAC7210021]. A.M. is supported by the Engineering and Physical Sciences Research Council under grant reference [EP/Y028872/1], a London Mathematical Society Emmy Noether Fellowship [EN-2223-01], and an Imperial College London Elsie Widdowson Fellowship.

References

[AEK*17] Adams H., Emerson T., Kirby M., Neville R., Peterson C., Shipman P., Chepushtanova S., Hanson E., Motta F., Ziegelmeier L.: Persistence images: A stable vector representation of persistent homology. *The Journal of Machine Learning Research 18*, 1 (2017), 218–252. 6, 8

[AGDST20] Atienza N., Gonzalez-Díaz R., Soriano-Trigueros M.: On the stability of persistent entropy and new summary functions for topological data analysis. *Pattern Recognition 107* (2020), 107509. 8

[AIMB*11] Armato III S. G., McLennan G., Bidaut L., McNitt-Gray M. F., Meyer C. R., Reeves A. P., Zhao B., Aberle D. R., Henschke C. I., Hoffman E. A., Kazerooni E. A., MacMahon H., van Beek E. J. R., Yankelevitz D., Biancardi A. M., Bland P. H., Brown M. S., Engelmann R. M., Laderach G. E., Max D., Pais R. C., Qing D. P.-Y., Roberts R. Y., Smith A. R., Starkey A., Batra P., Caligiuri P., Farooqi A., Gladish G. W., Jude C. M., Munden R. F., Petkovska I., Quint L. E., Schwartz L. H., Sundaram B., Dodd L. E., Fenimore C., Gur D., Petrick N., Freymann J., Kirby J., Hughes B., Vande Casteele A., Gupte S., Sallam M., Heath M. D., Kuhn M. H., Dharaiya E., Burns R., Fryd D. S., Salganicoff M., Anand V., Shreter U., Vastagh S., Croft B. Y., Clarke L. P.: The lung image database consortium (lidc) and image database resource initiative (idri): A completed reference database of lung nodules on ct scans. *Medical Physics 38*, 2 (2011), 915–931. URL: https://aapm.onlinelibrary.wiley.com/doi/abs/10.1118/1.3528204, arXiv:https://aapm.onlinelibrary.wiley.com/doi/pdf/10.1118/1.3528204, doi:https://doi.org/10.1118/1.3528204. 9, 11

[AIMB*15] Armato III S. G., McLennan G., Bidaut L., McNitt-Gray M. F., Meyer C. R., Reeves A. P., Zhao B., Aberle D. R., Henschke C. I., Hoffman E. A., Kazerooni E. A., MacMahon H., Van Beek E. J. R., Yankelevitz D., Biancardi A. M., Bland P. H., Brown M. S., Engelmann R. M., Laderach G. E., Max D., Pais R. C., Qing D. P. Y., Roberts R. Y., Smith A. R., Starkey A., Batra P., Caligiuri P., Farooqi A., Gladish G. W., Jude C. M., Munden R. F., Petkovska I., Quint L. E., Schwartz L. H., Sundaram B., Dodd L. E., Fenimore C., Gur D., Petrick N., Freymann J., Kirby J., Hughes B., Casteele A. V., Gupte S., Sallam M., Heath M. D., Kuhn M. H., Dharaiya E., Burns R., Fryd D. S., Salganicoff M., Anand V., Shreter U., Vastagh S., Croft B. Y., Clarke L. P.: Data from lidc-idri., 2015. URL: https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=1966254. 9

[BBB*21] Burley S. K., Bhikadiya C., Bi C., Bittrich S., Chen L., Crichlow G. V., Christie C. H., Dalenberg K., Di Costanzo L., Duarte J. M., Dutta S., Feng Z., Ganesan S., Goodsell D. S., Ghosh S., Green R. K., Guranović V., Guzenko D., Hudson B. P., Lawson C., Liang Y., Lowe R., Namkoong H., Peisach E., Persikova I., Randle C., Rose A., Rose Y., Sali A., Segura J., Sekharan M., Shao C., Tao Y.-P., Voigt M., Westbrook J., Young J. Y., Zardecki C., Zhuravleva M.: RCSB Protein Data Bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. *Nucleic Acids Research 49*, D1 (Jan. 2021), D437–D451. doi:10.1093/nar/gkaa1038. 4

[BCB20] Botnan M., Crawley-Boevey W.: Decomposition of persistence modules. *Proceedings of the American Mathematical Society 148*, 11 (Aug. 2020), 4581–4596. URL: https://www.ams.org/proc/2020-148-11/S0002-9939-2020-14790-9/, doi:10.1090/proc/14790. 3, 9

[BCF*08] Biasotti S., Cerri A., Frosini P., Giorgi D., Landi C.: Multidimensional Size Functions for Shape Comparison. *Journal of Mathematical Imaging and Vision 32*, 2 (2008), 161–179. URL: https://doi.org/10.1007/s10851-008-0096-z, doi:10.1007/s10851-008-0096-z. 1

[BDFF*08] Biasotti S., De Floriani L., Falcidieno B., Frosini P., Giorgi D., Landi C., Papaleo L., Spagnuolo M.: Describing shapes by geometrical-topological properties of real functions. *ACM Comput. Surv. 40*, 4 (2008), 12:1–12:87. URL: http://doi.acm.org/10.1145/1391729.1391731, doi:10.1145/1391729.1391731. 1

[BdSS15] Bubenik P., de Silva V., Scott J.: Metrics for generalized persistence modules. *Foundations of Computational Mathematics 15*, 6 (Dec. 2015), 1501–1531. arXiv:1312.3829 [cs, math]. doi:10.1007/s10208-014-9229-5. 2

[BGK15] Bhattacharya S., Ghrist R., Kumar V.: Persistent Homology for Path Planning in Uncertain Environments. *IEEE Transactions on Robotics 31*, 3 (June 2015), 578–590. URL: http://ieeexplore.ieee.org/document/7078886/, doi:10.1109/TRO.2015.2412051. 1

[BGMP14] Blumberg A. J., Gal I., Mandell M. A., Pancia M.: Robust Statistics, Hypothesis Testing, and Confidence Intervals for Persistent Homology on Metric Measure Spaces. *Foundations of Computational Mathematics 14*, 4 (Aug 2014), 745–789. URL: https://doi.org/10.1007/s10208-014-9201-4, doi:10.1007/s10208-014-9201-4. 6

[BGV92] Boser B. E., Guyon I. M., Vapnik V. N.: A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory* (New York, NY, USA, 1992), COLT '92, Association for Computing Machinery, p. 144–152. URL: https://doi.org/10.1145/130385.130401, doi:10.1145/130385.130401. 7

[BH21] Botnan M. B., Hirsch C.: On the consistency and asymptotic normality of multiparameter persistent betti numbers, 2021. arXiv:2109.05513. 3

[BOO22] Botnan M. B., Oppermann S., Oudot S.: Signed Barcodes for Multi-Parameter Persistence via Rank Decompositions. In *38th International Symposium on Computational Geometry (SoCG 2022)* (Dagstuhl, Germany, 2022), Goaoc X., Kerber M., (Eds.), vol. 224 of *Leibniz International Proceedings in Informatics (LIPIcs)*, Schloss Dagstuhl – Leibniz-Zentrum für Informatik, pp. 19:1–19:18. URL: https://drops.dagstuhl.de/opus/volltexte/2022/16027, doi:10.4230/LIPIcs.SoCG.2022.19. 9

[BRI*17] Biwer C., Rothberg A., IglayReger H., Derksen H., Burant C. F., Najarian K.: Windowed persistent homology: A topological signal processing algorithm applied to clinical obesity data. *PLOS ONE 12*, 5 (May 2017), e0177696. Publisher: Public Library of Science. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0177696, doi:10.1371/journal.pone.0177696. 1

[BS14] Bubenik P., Scott J. A.: Categorification of persistent homology. *Discrete & Computational Geometry 51*, 3 (Apr. 2014), 600–627. doi:10.1007/s00454-014-9573-x. 2

[Bub15] Bubenik P.: Statistical topological data analysis using persistence landscapes. *The Journal of Machine Learning Research 16*, 1 (2015), 77–102. 6, 8, 17

[BW20] Bramer D., Wei G.-W.: Atom-specific persistent homology and its application to protein flexibility analysis. *Computational and Mathematical Biophysics 8*, 1 (Jan. 2020), 1–35. Publisher: De Gruyter Open Access. doi:10.1515/cmb-2020-0001. 4

[CB15] Crawley-Boevey W.: Decomposition of pointwise finite-dimensional persistence modules. *Journal of Algebra and Its Applications 14*, 05 (June 2015), 1550066. Publisher: World Scientific Publishing Co. doi:10.1142/S0219498815500668. 3

[CB20] Carrière M., Blumberg A.: Multiparameter Persistence Image for Topological Machine Learning. In *Advances in Neural Information Processing Systems* (2020), vol. 33, Curran Associates, Inc., pp. 22432–22444. URL: https://papers.nips.cc/paper/2020/hash/fdff71fcab656abfbefaabecab1a7f6d-Abstract.html. 9

[CdS10] Carlsson G., de Silva V.: Zigzag persistence. *Foundations of computational mathematics 10*, 4 (2010), 367–405. 3

[CFF*13] Cerri A., Fabio B. D., Ferri M., Frosini P., Landi C.: Betti numbers in multidimensional persistent homology are stable functions. *Mathematical Methods in the Applied Sciences 36*, 12 (2013), 1543–1557. doi:10.1002/mma.2704. 9, 10

[CFK*19] Corbet R., Fugacci U., Kerber M., Landi C., WANG B.: A kernel for multi-parameter persistent homology. *Computers & Graphics: X 2* (Dec. 2019), 100005. URL: https://www.sciencedirect.com/science/article/pii/S2590148619300056, doi:10.1016/j.cagx.2019.100005. 9

[CFL*14] CHAZAL F., FASY B. T., LECCI F., RINALDO A., WASSERMAN L.: Stochastic convergence of persistence landscapes and silhouettes. In *Proceedings of the thirtieth annual symposium on Computational geometry* (2014), ACM, p. 474. 6

[CH67] COVER T., HART P.: Nearest neighbor pattern classification. *IEEE Transactions on Information Theory 13*, 1 (1967), 21–27. doi:10.1109/TIT.1967.1053964. 10

[CL08] COX D. D., LEE J. S.: Pointwise testing with functional data using the westfall–young randomization method. *Biometrika 95* (2008), 621–634. 7

[CLM24] CAO Y., LEUNG P., MONOD A.: *k*-Means Clustering for Persistent Homology. *Advances in Data Analysis and Classification* (2024), 1–25. 10

[CM22] CAO Y., MONOD A.: A Geometric Condition for Uniqueness of Fréchet Means of Persistence Diagrams. *arXiv preprint arXiv:2207.03943* (2022). 5

[CMC*20] CRAWFORD L., MONOD A., CHEN A. X., MUKHERJEE S., RABADÁN R.: Predicting clinical outcomes in glioblastoma: An application of topological and functional data analysis. *Journal of the American Statistical Association 115*, 531 (2020), 1139–1150. URL: https://doi.org/10.1080/01621459.2019.1671198, arXiv: https://doi.org/10.1080/01621459.2019.1671198, doi:10.1080/01621459.2019.1671198. 1, 2, 9

[CMW18] CANG Z., MU L., WEI G.-W.: Representability of algebraic topology for biomolecules in machine learning based scoring and virtual screening. *PLOS Computational Biology 14*, 1 (Jan. 2018), e1005929. Publisher: Public Library of Science. URL: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005929, doi:10.1371/journal.pcbi.1005929. 1, 4

[CSEH07] COHEN-STEINER D., EDELSBRUNNER H., HARER J.: Stability of Persistence Diagrams. *Discrete & Computational Geometry 37*, 1 (Jan. 2007), 103–120. doi:10.1007/s00454-006-1276-5. 3

[CW17] CANG Z., WEI G.-W.: Analysis and prediction of protein folding energy changes upon mutation by element specific persistent homology. *Bioinformatics 33*, 22 (Nov. 2017), 3549–3557. 1, 4

[CZ07] CARLSSON G., ZOMORODIAN A.: The theory of multidimensional persistence. *Discrete and Computational Geometry 42* (06 2007), 71–93. doi:10.1007/s00454-009-9176-0. 1, 8, 9

[CZ09] CARLSSON G., ZOMORODIAN A.: The Theory of Multidimensional Persistence. *Discrete & Computational Geometry 42*, 1 (July 2009), 71–93. doi:10.1007/s00454-009-9176-0. 9

[dFL03] D'AMICO M., FROSINI P., LANDI C.: Optimal matching between reduced size functions. *DISMI, Universit'a di Modena e Reggio Emilia 35* (2003). 5, 9

[dFL06] D'AMICO M., FROSINI P., LANDI C.: Using matching distance in size theory: A survey. *International Journal of Imaging Systems and Technology 16*, 5 (2006), 154–161. doi:10.1002/ima.20076. 9

[dFL10] D'AMICO M., FROSINI P., LANDI C.: Natural Pseudo-Distances and Optimal Matching between Reduced Size Functions. *Acta Applicandae Mathematicae 109*, 2 (2010), 527–554. URL: https://doi.org/10.1007/s10440-008-9332-1, doi:10.1007/s10440-008-9332-1. 5, 9

[DFLM09] DI FABIO B., LANDI C., MEDRI F.: Recognition of Occluded Shapes Using Size Functions. In *Image Analysis and Processing – ICIAP 2009* (Berlin, Heidelberg, 2009), Foggia P., Sansone C., Vento M., (Eds.), Springer Berlin Heidelberg, pp. 642–651. 1

[DHS16] DUY T. K., HIRAOKA Y., SHIRAI T.: Limit theorems for persistence diagrams, 2016. arXiv:1612.08371. 3

[DPR82] DAUXOIS J., POUSSE A., ROMAIN Y.: Asymptotic theory for the principal component analysis of a vector random function: some applications to statistical inference. *Journal of multivariate analysis 12*, 1 (1982), 136–154. 6

[dSG07] DE SILVA V., GHRIST R.: Coverage in sensor networks via persistent homology. *Algebraic & Geometric Topology 7*, 1 (Apr. 2007), 339–358. URL: http://www.msp.org/agt/2007/7-1/p16.xhtml, doi:10.2140/agt.2007.7.339. 1

[ELZ02] EDELSBRUNNER, LETSCHER, ZOMORODIAN: Topological Persistence and Simplification. *Discrete & Computational Geometry 28*, 4 (Nov. 2002), 511–533. doi:10.1007/s00454-002-2885-2. 2, 3

[ELZ03] EDELSBRUNNER H., LETSCHER D., ZOMORODIAN A.: Topological persistence and simplification. 3

[ESR16] EMMETT K., SCHWEINHART B., RABADAN R.: Multiscale Topology of Chromatin Folding. In *Proceedings of the 9th EAI International Conference on Bio-inspired Information and Communications Technologies (formerly BIONETICS)* (Brussels, BEL, May 2016), BICT'15, ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), pp. 177–180. URL: https://dl.acm.org/doi/10.4108/eai.3-12-2015.2262453, doi:10.4108/eai.3-12-2015.2262453. 1

[FKR23] FUGACCI U., KERBER M., ROLLE A.: Compression for 2-parameter persistent homology. *Computational Geometry 109* (2023), 101940. 9

[FL99] FROSINI P., LANDI C.: Size theory as a topological tool for computer vision. *Pattern Recognition and Image Analysis 9*, 4 (1999), 596–603. 1

[FL01] FROSINI P., LANDI C.: Size Functions and Formal Series. *Applicable Algebra in Engineering, Communication and Computing 12*, 4 (Aug. 2001), 327–349. doi:10.1007/s002000100078. 1, 3

[FLR*14] FASY B. T., LECCI F., RINALDO A., WASSERMAN L., BALAKRISHNAN S., SINGH A.: Confidence sets for persistence diagrams. *Ann. Statist. 42*, 6 (12 2014), 2301–2339. URL: https://doi.org/10.1214/14-AOS1252, doi:10.1214/14-AOS1252. 6

[Fro92] FROSINI P.: Measuring shapes by size functions. In *Intelligent Robots and Computer Vision X: Algorithms and Techniques* (1992), vol. 1607, International Society for Optics and Photonics, pp. 122–134. 1, 5

[GGP*21] GRAFF G., GRAFF B., PILARCZYK P., JABŁOŃSKI G., GĄSECKI D., NARKIEWICZ K.: Persistent homology as a new method of the assessment of heart rate variability. *Plos one 16*, 7 (2021), e0253851. 8

[GGR*21] GĄSECKI D., GRAFF B., ROJEK A., NARKIEWICZ K., GRAFF G., PILARCZYK P.: The database of normal rr-intervals of length up to 512 of 41 patients at rest hospitalized due to the episode of acute ischemic stroke, 2021. URL: https://mostwiedzy.pl/en/open-research-data/the-database-of-normal-rr-intervals-of-length-up-to-512-of-41-patients-at-rest-hospitalized-due-to-t,62108070060958-0, doi:10.34808/xs7m-3552. 7, 11

[GH10] GAMBLE J., HEO G.: Exploring uses of persistent homology for statistical analysis of landmark-based shape data. *Journal of Multivariate Analysis 101*, 9 (Oct. 2010), 2184–2199. doi:10.1016/j.jmva.2010.04.016. 4

[Gid17] GIDEA M.: Topological Data Analysis of Critical Transitions in Financial Networks. In *3rd International Winter School and Conference on Network Science* (Cham, 2017), Shmueli E., Barzel B., Puzis R., (Eds.), Springer Proceedings in Complexity, Springer International Publishing, pp. 47–59. doi:10.1007/978-3-319-55471-6_5. 1, 4

[HBHM22] HAMILTON W., BORGERT J. E., HAMELRYCK T., MARRON J. S.: Persistent Topology of Protein Space. In *Research in Computational Topology 2*, Gasparovic E., Robins V., Turner K., (Eds.), Association for Women in Mathematics Series. Springer International

Publishing, Cham, 2022, pp. 223–244. doi:10.1007/978-3-030-95519-9_10. 4

[HBM23] Häberle K., Bravi B., Monod A.: Wavelet-Based Density Estimation for Persistent Homology. *arXiv preprint arXiv:2305.08999* (2023). 7

[HKNU17] Hofer C., Kwitt R., Niethammer M., Uhl A.: Deep learning with topological signatures. In *Advances in Neural Information Processing Systems 30*, Guyon I., Luxburg U. V., Bengio S., Wallach H., Fergus R., Vishwanathan S., Garnett R., (Eds.). Curran Associates, Inc., 2017, pp. 1634–1644. URL: http://papers.nips.cc/paper/6761-deep-learning-with-topological-signatures.pdf. 6

[HOST19] Harrington H. A., Otter N., Schenck H., Tillmann U.: Stratifying Multiparameter Persistent Homology. *SIAM Journal on Applied Algebra and Geometry 3*, 3 (Jan. 2019), 439–471. Publisher: Society for Industrial and Applied Mathematics. doi:10.1137/18M1224350. 9

[KLO19] Kerber M., Lesnick M., Oudot S.: Exact Computation of the Matching Distance on 2-Parameter Persistence Modules. 15 pages. Artwork Size: 15 pages Medium: application/pdf Publisher: Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik GmbH, Wadern/Saarbruecken, Germany Version Number: 1.0. URL: http://drops.dagstuhl.de/opus/volltexte/2019/10450/, doi:10.4230/LIPICS.SOCG.2019.46. 9

[KM21] Kim W., Mémoli F.: Generalized persistence diagrams for persistence modules over posets. *Journal of Applied and Computational Topology 5*, 4 (Dec. 2021), 533–581. doi:10.1007/s41468-021-00075-1. 2, 9

[KN19] Kerber M., Nigmetov A.: Efficient approximation of the matching distance for 2-parameter persistence. *arXiv preprint arXiv:1912.05826* (2019). 9

[KP23] Krebs J., Polonik W.: On the asymptotic normality of persistent betti numbers, 2023. arXiv:1903.03280. 3

[KR21] Kerber M., Rolle A.: Fast Minimal Presentations of Bigraded Persistence Modules. In *2021 Proceedings of the Symposium on Algorithm Engineering and Experiments (ALENEX)*, Proceedings. Society for Industrial and Applied Mathematics, Jan. 2021, pp. 207–220. URL: https://epubs.siam.org/doi/10.1137/1.9781611976472.16, doi:10.1137/1.9781611976472.16. 9

[Lan18] Landi C.: *The Rank Invariant Stability via Interleavings*. Springer International Publishing, Cham, 2018, pp. 1–10. doi:10.1007/978-3-319-89593-2_1. 9

[Les15] Lesnick M.: The Theory of the Interleaving Distance on Multidimensional Persistence Modules. *Foundations of Computational Mathematics 15*, 3 (June 2015), 613–650. doi:10.1007/s10208-015-9255-y. 9

[LF97] Landi C., Frosini P.: New pseudodistances for the size function space. In *Vision Geometry VI* (1997), vol. 3168, International Society for Optics and Photonics, pp. 52–61. 5

[LF02] Landi C., Frosini P.: Size functions as complete invariants for image recognition. In *Vision Geometry XI* (2002), vol. 4794, International Society for Optics and Photonics, pp. 101–110. 1

[LPR09] López-Pintado S., Romo J.: On the concept of depth for functional data. *Journal of the American Statistical Association 104*, 486 (2009), 718–734. URL: https://doi.org/10.1198/jasa.2009.0108, arXiv:https://doi.org/10.1198/jasa.2009.0108, doi:10.1198/jasa.2009.0108. 10

[LSKS*18] Lees T., Shad-Kaneez F., Simpson A. M., Nassif N. T., Lin Y., Lal S.: Heart rate variability as a biomarker for predicting stroke, post-stroke complications and functionality. *Biomarker Insights 13* (2018). 7

[LW15] Lesnick M., Wright M.: Interactive Visualization of 2-D Persistence Modules, Dec. 2015. arXiv:1512.00180 [cs, math] version: 1. URL: http://arxiv.org/abs/1512.00180. 9

[LW22] Lesnick M., Wright M.: Computing Minimal Presentations and Bigraded Betti Numbers of 2-Parameter Persistent Homology. *SIAM Journal on Applied Algebra and Geometry 6*, 2 (June 2022), 267–298. Publisher: Society for Industrial and Applied Mathematics. URL: https://epubs.siam.org/doi/10.1137/20M1388425, doi:10.1137/20M1388425. 9

[Mil20] Miller E.: Data structures for real multiparameter persistence modules, Aug. 2020. arXiv:1709.08155 [math]. doi:10.48550/arXiv.1709.08155. 9

[MMH11] Mileyko Y., Mukherjee S., Harer J.: Probability measures on the space of persistence diagrams. *Inverse Problems 27*, 12 (2011), 124007. URL: http://stacks.iop.org/0266-5611/27/i=12/a=124007. 6

[MMM17] Marchese A., Maroulas V., Mike J.: K-means clustering on the space of persistence diagrams. In *Wavelets and Sparsity XVII* (2017), vol. 10394, SPIE, pp. 218–227. 10

[NGGP21] Narkiewicz K., Graff B., Graff G., Pilarczyk P.: The database of normal rr-intervals of length up to 512 of 46 healthy subjects at rest, 2021. URL: https://mostwiedzy.pl/en/open-research-data/the-database-of-normal-rr-intervals-of-length-up-to-512-of-46-healthy-subjects-at-rest,621020624624111-0, doi:10.34808/4k51-7n26. 7, 11

[Pat18] Patel A.: Generalized Persistence Diagrams. *Journal of Applied and Computational Topology 1*, 3-4 (June 2018), 397–419. doi:10.1007/s41468-018-0012-6. 9

[Pla98] Platt J.: Sequential minimal optimization: A fast algorithm for training support vector machines. *Advances in Kernel Methods-Support Vector Learning 208* (07 1998). 7

[Ram02] Ramsay J. O. J. O.: *Applied functional data analysis : methods and case studies*. Springer, New York, New York, 2002. 1

[Ram05] Ramsay J. O. J. O.: *Functional data analysis*, 2nd ed. ed. Springer series in statistics. Springer, New York, 2005. 1

[RHBK15] Reininghaus J., Huber S., Bauer U., Kwitt R.: A stable multi-scale kernel for topological machine learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2015), pp. 4741–4748. 6

[RKP23] Roycraft B., Krebs J., Polonik W.: Bootstrapping persistent betti numbers and other stabilizing statistics. *The Annals of Statistics 51*, 4 (Aug. 2023). URL: http://dx.doi.org/10.1214/23-AOS2277, doi:10.1214/23-aos2277. 3

[RT16] Robins V., Turner K.: Principal Component Analysis of Persistent Homology Rank Functions with case studies of Spatial Point Patterns, Sphere Packing and Colloids. *Physica D: Nonlinear Phenomena 334* (Nov. 2016), 99–117. doi:10.1016/j.physd.2016.03.007. 2, 4, 5, 6

[RV06] Rossi F., Villa N.: Support vector machine for functional data classification. *Neurocomputing 69*, 7–9 (Mar. 2006), 730–742. URL: http://dx.doi.org/10.1016/j.neucom.2005.12.010, doi:10.1016/j.neucom.2005.12.010. 6, 7

[SIDFL20] Scaramuccia S., Iuricich F., De Floriani L., Landi C.: Computing multiparameter persistent homology through a discrete morse-based approach. *Computational Geometry 89* (2020), 101623. 9

[SRS*22] Saba T., Rehman A., Shahzad M., Latif R., Bahaj S., Alyami J.: Machine learning for post-traumatic stress disorder identification utilizing resting-state functional magnetic resonance imaging. *Microscopy Research and Technique 85* (01 2022). doi:10.1002/jemt.24065. 7

[ST21] Skraba P., Turner K.: Wasserstein Stability for Persistence Diagrams, Mar. 2021. arXiv:2006.16824 [math]. doi:10.48550/arXiv.2006.16824. 4, 5, 6, 16, 17

[The20] The RIVET Developers: Rivet, 2020. URL: https://github.com/rivetTDA/rivet/. 9

[TL94] Turk G., Levoy M.: Zippered polygon meshes from range images. In *Proceedings of the 21st Annual Conference on Computer Graphics and Interactive Techniques* (New York, NY, USA, 1994), SIGGRAPH '94, Association for Computing Machinery, p. 311–318. URL: https://doi.org/10.1145/192161.192241, doi:10.1145/192161.192241. 3

[TMMH14] Turner K., Mileyko Y., Mukherjee S., Harer J.: Fréchet Means for Distributions of Persistence Diagrams. *Discrete & Computational Geometry 52*, 1 (Jul 2014), 44–70. URL: https://doi.org/10.1007/s00454-014-9604-7, doi:10.1007/s00454-014-9604-7. 5, 6

[VAB13] Vasudevan R., Ames A., Bajcsy R.: Persistent homology for automatic determination of human-data based cost of bipedal walking. *Nonlinear Analysis: Hybrid Systems 7*, 1 (Feb. 2013), 101–115. URL: https://linkinghub.elsevier.com/retrieve/pii/S1751570X1200026X, doi:10.1016/j.nahs.2012.07.006. 1

[Vie27] Vietoris L.: Über den höheren Zusammenhang kompakter Räume und eine Klasse von zusammenhangstreuen Abbildungen. *Mathematische Annalen 97*, 1 (Dec. 1927), 454–472. URL: https://doi.org/10.1007/BF01447877, doi:10.1007/BF01447877. 2

[Vip20] Vipond O.: Multiparameter persistence landscapes. *Journal of Machine Learning Research 21*, 61 (2020), 1–38. URL: http://jmlr.org/papers/v21/19-054.html. 9

[VMS*23] Vandaele R., Mukherjee P., Selby H. M., Shah R. P., Gevaert O.: Topological data analysis of thoracic radiographic images shows improved radiomics-based lung tumor histology prediction. *Patterns 4*, 1 (2023), 100657. URL: https://www.sciencedirect.com/science/article/pii/S2666389922002975, doi:https://doi.org/10.1016/j.patter.2022.100657. 9, 10

[VUFF93] Verri A., Uras C., Frosini P., Ferri M.: On the use of size functions for shape analysis. *Biological Cybernetics 70*, 2 (1993), 99–107. URL: https://doi.org/10.1007/BF00200823, doi:10.1007/BF00200823. 1

[XWZZ08] Xie S.-Y., Wang P.-W., Zhang H.-J., Zhao H.-T.: Research on the classification of brain function based on svm. In *2008 2nd International Conference on Bioinformatics and Biomedical Engineering* (2008), pp. 1931–1934. doi:10.1109/ICBBE.2008.812. 7

[ZC05] Zomorodian A., Carlsson G.: Computing Persistent Homology. *Discrete & Computational Geometry 33*, 2 (Feb. 2005), 249–274. doi:10.1007/s00454-004-1146-y. 3

[ZRTH03] Zhu J., Rosset S., Tibshirani R., Hastie T.: 1-norm support vector machines. In *Advances in Neural Information Processing Systems* (2003), Thrun S., Saul L., Schölkopf B., (Eds.), vol. 16, MIT Press. URL: https://proceedings.neurips.cc/paper_files/paper/2003/file/49d4b2faeb4b7b9e745775793141e2b2-Paper.pdf. 18

Appendix A: Proofs

We now give the proofs of the key theoretical results presented in Section 3.

Proof [Proof of Proposition 10]. Let $1 \leq p < \infty$ and M a p.f.d. persistence module with barcode $\mathrm{Bar}(M) = \{[b_i, d_i) : 1 \leq i \leq m\}$. Call $\delta_i := d_i - b_i < \infty$ for all $1 \leq i \leq m$, fix $\delta > 0$, and define $\eta := \min\{\delta/2, 1, \delta_i/2 : 1 \leq i \leq m\}$.

Let N be a p.f.d. persistence module such that $d_B(\mathrm{Dgm}(M), \mathrm{Dgm}(N)) < \eta$, with barcode $\mathrm{Bar}(N) = \{[\tilde{b}_j, \tilde{d_j}) : 1 \leq j \leq n\}$. By the definition of η, the optimal matching ϕ between points in $\mathrm{Dgm}(M)$ and $\mathrm{Dgm}(N)$ defined by the bottleneck distance matches all points outside of the diagonal in the diagram $\mathrm{Dgm}(M)$ to points in $\mathrm{Dgm}(N)$ outside of the diagonal. In addition, all the remaining points in $\mathrm{Dgm}(N)$ matched to the diagonal are at an ℓ^∞-distance of their orthogonal projection to the diagonal of less than $\delta/2$, which means that $\beta_\delta^N = 0$ for all of them.

With these two facts and the additivity of rank functions, we obtain

$$\begin{aligned}
\left\|\beta_\delta^M - \beta_\delta^N\right\|_p &= \left\|\sum_{i=1}^m \beta_\delta^{\mathbb{I}[b_i,d_i)} - \sum_{j=1}^n \beta_\delta^{\mathbb{I}[\tilde{b}_j,\tilde{d}_i)}\right\|_p \\
&= \left\|\sum_{i=1}^m \left(\beta_\delta^{\mathbb{I}[b_i,d_i)} - \beta_\delta^{\mathbb{I}[b_i',d_i')}\right) - \sum_{j\in J} \beta_\delta^{\mathbb{I}[\tilde{b}_j,\tilde{d}_i)}\right\|_p \\
&\leq \sum_{i=1}^m \left\|\beta_\delta^{\mathbb{I}[b_i,d_i)} - \beta_\delta^{\mathbb{I}[b_i',d_i')}\right\|_p \qquad (10)
\end{aligned}$$

where $\phi(b_i, d_i) = (b_i', d_i')$ for all $1 \leq i \leq m$ and $J \subset \{1, \ldots, n\}$ is the subset of indices of points in $\mathrm{Dgm}(N)$ matched to the diagonal. We now obtain a bound for (10). For $i \in \{1, \ldots, m\}$, define the sets

$$\begin{aligned}
A_i &:= \{(x,y) \in \mathbb{R}^{2+} : \ b_i \leq x \leq y \leq d_i\}, \\
B_i &:= \{(x,y) \in \mathbb{R}^{2+} : \ b_i' \leq x \leq y \leq d_i'\}, \\
D_i &:= A_i \Delta B_i = (A_i \cup B_i) \setminus (A_i \cap B_i), \qquad (11)
\end{aligned}$$

where Δ denotes the symmetric difference (see Figure 6 for some illustrative examples of D_i).

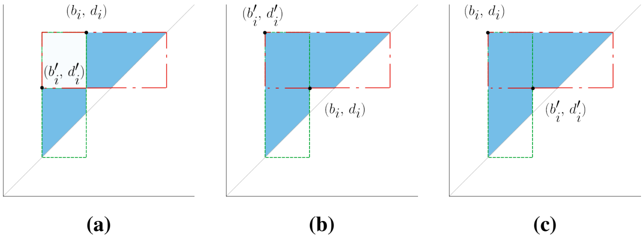


Figure 6: *Example of domains D_i (shaded in blue) on which rank functions differ and the sketched rectangles (delineated by red dashed lines and green dotted lines) indicate the bound of $\omega(D_i)$ when $M = \mathbb{I}[b_i, d_i)$ and $N = \mathbb{I}[b_i', d_i')$.*

Notice that for $(x,y) \in D_i^c$, the truncated rank functions coincide, i.e., $\beta^{\mathbb{I}[b_i,d_i)}(x,y) = \beta^{\mathbb{I}[b_i',d_i')}(x,y)$; also, these rank functions differ by one for $(x,y) \in D_j$. This implies

$$\left\|\beta_\delta^{\mathbb{I}[b_i,d_i)} - \beta_\delta^{\mathbb{I}[b_i',d_i')}\right\|_p = \omega(D_i)^{1/p} \qquad (12)$$

where ω denotes the Lebesgue measure in $\mathbb{R}^2$.
The rectangles depicted in red dashed lines and green dotted lines in Figure 6 each have one side of length

$$\max(d_i', d_i) - \min(b_i, b_i') =$$
$$\begin{cases} d_i - b_i' \leq d_i - b_i + \left\|(b_i,d_i) - (b_i',d_i')\right\|_\infty & \text{(Figure 6(a))},\\ d_i' - b_i \leq d_i - b_i + \left\|(b_i,d_i) - (b_i',d_i')\right\|_\infty & \text{(Figure 6(a))},\\ d_i' - b_i' \leq d_i - b_i + 2\left\|(b_i,d_i) - (b_i',d_i')\right\|_\infty & \text{(Figure 6(b))},\\ d_i - b_i & \text{(Figure 6(c))}. \end{cases}$$

The other side of both rectangles is bounded by $\left\|(b_i,d_i) - (b_i',d_i')\right\|_\infty$. Observe that this is also a bound for the lengths of the sides of the rectangle in the intersection.
In case 6(c), by adding the Lebesgue measure of the rectangles, we get

$$\omega(D_i) \leq 2\,(d_i - b_i)\,\left\|(b_i,d_i) - (b_i',d_i')\right\|_\infty$$

where $\omega(\cdot)$ denotes the Lebesgue measure. In cases 6(a) and 6(b), adding the Lebesgue measure of the rectangles and triangles that decompose the figures, we obtain

$$\begin{aligned}\omega(D_i) &\leq 2\,(d_i - b_i)\,\left\|(b_i,d_i) - (b_i',d_i')\right\|_\infty + 2\left\|(b_i,d_i) - (b_i',d_i')\right\|_\infty^2\\ &\leq 2\,(d_i - b_i + 1)\,\left\|(b_i,d_i) - (b_i',d_i')\right\|_\infty\end{aligned}$$

where we have used that $\left\|(b_i,d_i) - (b_i',d_i')\right\|_\infty \leq 1$ in the last inequality. Setting $R := \max\{d_i - b_i : 1 \leq i \leq m\}$, we conclude

$$\omega(D_i) \leq 2(R+1)\, d_B(\mathrm{Dgm}\,(M), \mathrm{Dgm}\,(N)). \tag{13}$$

Returning to (10), we obtain the bound

$$\left\|\beta_\delta^M - \beta_\delta^N\right\|_p \leq m\,(2R+2)^{1/p}\, d_B(\mathrm{Dgm}\,(M), \mathrm{Dgm}\,(N))^{1/p},$$

where m is the number of points in the persistence diagram of M and R is the maximum persistence (see Definition 3), so the constant only depends on M, as desired. □

As previously mentioned in Section 3, we can always obtain a bounding constant that depends on both modules M and N as follows: let R' be the maximum between the lifetimes of the bars in the barcodes of M and N, so that $\omega(D_j) \leq 2R' \cdot d_B(\mathrm{Dgm}\,(M), \mathrm{Dgm}\,(N))$ (see (11)); and then use this bound afterwards (10). In the proof of Proposition 10 above, we refine this strategy by bounding with a constant that only depends on the module M.

A natural follow-up question is whether it is possible to achieve a bound such as that in (13), but with the following dependency with respect to the bottleneck distance:

$$\omega(D_j) \leq C(b_j,d_j) \cdot d_B\left(\mathrm{Dgm}\,(M), \mathrm{Dgm}\,(N)\right)^p$$

for some values $p \geq 2$. This would imply a Lipschitz stability condition (notice that for $p = 1$, Proposition 10 is actually a Lipschitz condition). In a similar vein to Corollary 8 by [ST21], the answer to this question is negative, and the key to this fact is the following counterexample.

Proposition 15. *Take $\mathbb{I}[b,d)$ and $C(b,d) = 2(d-b+1) > 0$ such that*

$$\omega(D) \leq C(b,d) \cdot \left\|(b,d) - (b',d')\right\|_\infty$$

for some other interval module $\mathbb{I}[b',d')$, with $\left\|(b,d) - (b',d')\right\|_\infty < 1$ and $D \subset \mathbb{R}^{2+}$ defined as

$$\begin{aligned} A &:= \{(x,y) \in \mathbb{R}^{2+} :\ b \leq x \leq y \leq d\},\\ B &:= \{(x,y) \in \mathbb{R}^{2+} :\ b' \leq x \leq y \leq d'\},\\ D &:= A \Delta B = (A \cup B) \setminus (A \cap B).\end{aligned}$$

Then

$$\omega(D) > C(b,d) \cdot \left\|(b,d) - (b',d')\right\|_\infty^p$$

for every $p \geq 2$.

Proof Let $\epsilon > 0$ and consider $\mathbb{I}_\epsilon := \mathbb{I}[b-\epsilon, d+\epsilon)$ so that $\left\|(b,d) - (b-\epsilon, d+\epsilon)\right\|_\infty = \epsilon$. In this case we are in a situation similar to the one depicted in Figure 6(b). Define

$$D_\epsilon := \left\{(x,y) \in \mathbb{R}^{2+} : b \leq x \leq y \leq d\right\} \Delta$$
$$\left\{(x,y) \in \mathbb{R}^{2+} : b - \epsilon \leq x \leq y \leq d + \epsilon\right\}.$$

For this domain, we exactly have $\omega(D_\epsilon) = 2(d-b)\epsilon + 2\epsilon^2$ and taking $0 < \epsilon < 1$, we get

$$\omega(D_\epsilon) = 2(d-b)\epsilon + 2\epsilon^2 > 2(d-b+1) \cdot \epsilon^p = C(b,d) \cdot \epsilon^p$$

for every $p \geq 2$.
Now, for every other interval module $\mathbb{I}[b',d')$ with $\left\|(b,d) - (b',d')\right\|_\infty < 1$, we can just consider the previous case with $\epsilon = \min\left(\left(|b'-b|, |d'-d|\right) < 1\right.$, so that $D_\epsilon \subset D$ and thus $\omega(D_\epsilon) \leq \omega(D)$, concluding the proof. □

From this counterexample and the previous proof, it is not possible to obtain a Lipschitz stability condition for $p \geq 2$.

Proof [Proof of Theorem 12.] Let M be a persistence module with barcode $\mathrm{Bar}\,(M) = \{[\tilde{b}_i, \tilde{d}_i) : 1 \leq i \leq m\}$ and N a p.f.d. persistence module over $\mathbb{R}$ such that $W_1(\mathrm{Dgm}\,(M), \mathrm{Dgm}\,(N)) \leq 1$, with barcode $\mathrm{Bar}\,(N) = \{[b_j', d_j') : 1 \leq j \leq n\}$.
Let ϕ be the optimal matching between $\mathrm{Dgm}\,(M)$ and $\mathrm{Dgm}\,(N)$ induced by the 1-Wasserstein distance. Let $J \subset \{1, \ldots, n\}$ be the set of subindices corresponding to points in $\mathrm{Dgm}\,(N)$ matched to points outside of the diagonal in $\mathrm{Dgm}\,(M)$. Call $(b_j, d_j) = \phi^{-1}(b_j', d_j')$, for $j \in J$, the corresponding matched points in $\mathrm{Dgm}\,(M)$. Note that some of the points $(b_j', d_j') \in \mathrm{Dgm}\,(N)$ with $j \in J$ might be on the diagonal. For $j \in J$, define the set D_j as in (11).
From the additivity of the rank function, we obtain the following sequence on inequalities

$$\begin{aligned}\left\|\beta^M - \beta^N\right\|_p &= \left\|\sum_{i=1}^m \beta^{\mathbb{I}[\tilde{b}_i, \tilde{d}_i)} - \sum_{j=1}^n \beta^{\mathbb{I}[b_j', d_i')}\right\|_p\\ &= \left\|\sum_{j \in J}\left(\beta^{\mathbb{I}[b_i,d_i)} - \beta^{\mathbb{I}[b_i',d_i')}\right) - \sum_{j \in [m]\setminus J} \beta^{\mathbb{I}[b_j',d_j')}\right\|_p\\ &\leq \sum_{j \in J}\left\|\beta^{\mathbb{I}[b_j,d_j)} - \beta^{\mathbb{I}[b_j',d_j')}\right\|_p + \sum_{j \in [m]\setminus J}\left\|\beta^{\mathbb{I}[b_j',d_j')}\right\|_p\end{aligned} \tag{14}$$

where $[m] = \{1,\dots,m\}$.
For the first term of the sum (14), the bound in (12) still applies. Notice also that the L^p norm of $\beta^{\mathbb{I}[b'_j,d'_j)}$ for $j \in [m] \setminus J$ equals the pth root of the area of the triangle with vertices at (b'_j,d'_j), (b'_j,b'_j), and (d'_j,d'_j), which is equal to $\frac{1}{2}|d'_j - b'_j|^2$.
For $p = 1$, we can obtain the Lipschitz condition as follows:

$$\begin{aligned}
\left\|\beta^M - \beta^N\right\|_1 &\leq \sum_{j\in J} \omega(D_j) + \sum_{j\in J'\setminus J} \frac{1}{2}|d'_j - b'_j|^2 \\
&\leq \sum_{j\in J} \omega(D_j) + \sum_{j\in J'\setminus J} |d'_j - b'_j| && (15)\\
&\leq 2(R+1)\sum_{j\in J} \left\|(b'_j,d'_j) - (b_j,d_j)\right\|_\infty + \sum_{j\in J'\setminus J} |d'_j - b'_j| && (16)\\
&\leq 2(R+1)\sum_{j\in J} \left(|b_j - b'_j| + |d_j - d'_j|\right) + \sum_{j\in J'\setminus J} |d'_j - b'_j| \\
&\leq 2(R+2)\cdot W_1(\mathrm{Dgm}(M), \mathrm{Dgm}(N)),
\end{aligned}$$

where in (15), we use that $W_1(\mathrm{Dgm}(M), \mathrm{Dgm}(N)) \leq 1$ implies that $\frac{|d'_j - b'_j|}{2} \leq 1$ for $j \in J' \setminus J$ and in (16) we have used the bound in (13).
For the case where $p = 2$, applying the Cauchy–Schwarz inequality we get

$$\sum_{j\in J} \omega(D_j)^{1/2} \leq \left(|J| \sum_{j\in J} \omega(D_j)\right)^{1/2},$$

thus, we obtain

$$\begin{aligned}
\left\|\beta^M - \beta^N\right\|_2 &\leq (2(R+1)|J|)^{1/2} \left(\sum_{j\in J} \left\|(b'_j,d'_j) - (b_j,d_j)\right\|_\infty\right)^{1/2} \\
&\quad + \frac{1}{\sqrt{2}} \sum_{j\in J'\setminus J} |d'_j - b'_j| && (17)\\
&\leq \max\left\{(2(R+1)|J|)^{1/2}, \frac{1}{\sqrt{2}}\right\}\cdot \\
&\quad \left(W_1(\mathrm{Dgm}(M), \mathrm{Dgm}(N))^{1/2} + W_1(\mathrm{Dgm}(M), \mathrm{Dgm}(N))\right) \\
&\leq 2\cdot \max\left\{(2(R+1)|J|)^{1/2}, \frac{1}{\sqrt{2}}\right\}\cdot \\
&\qquad W_1(\mathrm{Dgm}(M), \mathrm{Dgm}(N))^{1/2} && (18)
\end{aligned}$$

where in (17) we take the root and use the bound in (13); and in (18), we use that $W_1(\mathrm{Dgm}(M), \mathrm{Dgm}(N)) \leq 1$. □

Appendix B: Comparison with Persistence Landscapes

One of the most popular topological vectorization methods in TDA is the *persistence landscape*, proposed by [Bub15] as follows. Let $M \in \mathsf{Vec}^{(\mathbb{R},\leq)}$ be a single-parameter persistence module and β^M the corresponding rank function.

Definition 16 (Persistence landscape, p-landscape distance [Bub15]). For $k \in \mathbb{N}$ kth-*persistence landscape* is a function $\lambda_k : \mathbb{R} \to \mathbb{R} \cup \{-\infty, +\infty\}$ defined as

$$\lambda_k^M(t) := \sup\{m \geq 0 : \beta^M(t-m, t+m) \geq \}).$$

Given two different persistence modules, $M, N \in \mathsf{Vec}^{(\mathbb{R},\leq)}$, the *p-landscape distance* between them is defined as

$$d_{\lambda,p}(M,N) = \sum_{k=1}^{\infty} \left\|\lambda_k^M - \lambda_k^N\right\|_p^p.$$

In [Bub15], several stability results are established for landscapes endowed with this metric. Although landscapes and rank functions are inherently different in nature—where the former is a vectorization of persistence diagrams and barcodes (building from the latter), while the latter is a direct and equivalent representation of diagrams and barcodes—both have been used in real-data applications: a main contribution of this work is the performance assessment of rank functions in inferential machine learning tasks. This then raises the question of comparison between the stability results associated with landscapes versus those established in this work.

A first observation is that the p-landscape metric, introduced in [Bub15], involves an infinite sum over the L^p distances of these landscapes, which is a first distinction from the direct L^p metrics that we consider over rank functions. Using the ∞-landscape distance, stability is then achieved with respect to the bottleneck distance between diagrams, which surpasses our Proposition 10 [Bub15, Theorem 13]. However, this is expected, since the persistence landscape is an *incomplete* invariant and thus sacrifices some information encompassed in the persistence diagram for improved stability in the L^∞ metric, while rank functions, as mentioned previously, are exactly equivalent to persistence diagrams and therefore comprise all topological information of the data captured by persistent homology.

Up until recently, such a stability bound was the best possible, since stability of PH was only rigorously established for the bottleneck distance. However, thanks to new stability results for the p-Wasserstein distances established by [ST21], stability is now possible with respect to these metrics. This is what we achieve in Theorem 12. Comparing this result to the p-landscape stability theorem [Bub15, Theorem 16] is challenging due to different settings and metrics. [Bub15, Theorem 16] considers filtrations over triangulable, compact metric spaces—a restriction we do not impose. In this setting, the p-landscape metric is compared to the L^∞ distance between filtering functions in sublevel-set filtrations. Our work extends beyond sublevel-set filtrations, and our L^p metrics over rank functions are thus not easily comparable to the p-landscape distances.

Appendix C: HRV Classification Results using Persistence Images and Persistence Landscapes

We include here the results of the SVM classification using the vectorization techniques of persistence images and persistence landscapes on HRV data. Table 5 shows the average accuracy, AUC–ROC and runtimes (in seconds) of the SVM classifier using persistence images under various kernels with and without dimensionality reduction using PCA. Table 6 shows the same data for the

5 first persistence landscapes λ_k, $1 \leq k \leq 5$. In these tables, the runtime includes: (a) the computation of the PH barcodes for all data, and from them, the computation of the corresponding vectorizations; (b) the training of the corresponding SVM; and (c) the computation of the accuracy and AUC-ROC over five-fold cross-validation. Experiments were run in a processor 11th Gen Intel Core i5-1135G7, with 16GB RAM. Table 7 further shows the average accuracy and AUC–ROC of linear support vector classification (LSVC) and sparse LSVC on the data. Recall that where standard LSVC adopts the L^2 penalty in the loss function, sparse LSVC adopts the L^1 norm, effectively reducing the dimensionality of the feature space [ZRTH03].

Table 5: *Average accuracy, AUC–ROC and runtimes of SVM classifiers constructed on persistence images (and persistence images reduced by PCA) computed on HRV data with linear, GRBF, polynomial and sigmoid kernels over ten iterations of five-fold cross-validation.*

Kernel	SVM		
	Accuracy	AUC-ROC	Runtime (s)
Linear	65.2	0.771	4.31
GRBF	64.9	0.756	4.64
Polynomial (d=2)	48.2	0.695	5.56
Polynomial (d=3)	44.9	0.680	4.05
Polynomial (d=5)	43.0	0.678	4.90
Sigmoid	60.1	0.724	4.56
Kernel	PCA + SVM		
	Accuracy	AUC-ROC	Runtime (s)
Linear	65.24	0.771	1.58
GRBF	65.36	0.756	1.56
Polynomial (d=2)	44.18	0.718	1.33
Polynomial (d=3)	43.02	0.686	2.01
Polynomial (d=5)	42.69	0.676	1.68
Sigmoid	61.13	0.726	2.44

Table 6: *Average accuracy, AUC–ROC and runtimes of SVM classifiers constructed on persistence landscapes (and persistence landscapes reduced by PCA) computed on HRV data with linear, GRBF, polynomial and sigmoid kernels over ten iterations of five-fold cross-validation.*

Kernel	SVM		
	Accuracy	AUC-ROC	Runtime (s)
Linear	77.0	0.843	0.421
GRBF	81.0	0.903	0.314
Polynomial (d=2)	68.2	0.866	0.427
Polynomial (d=3)	62.6	0.852	0.390
Polynomial (d=5)	58.5	0.820	0.369
Sigmoid	63.2	0.700	0.632
Kernel	PCA + SVM		
	Accuracy	AUC-ROC	Runtime (s)
Linear	77.02	0.843	0.524
GRBF	80.87	0.900	0.498
Polynomial (d=2)	60.48	0.797	0.479
Polynomial (d=3)	55.70	0.868	0.342
Polynomial (d=5)	46.52	0.828	0.461
Sigmoid	67.47	0.765	0.417

Table 7: *Average accuracy and AUC–ROC of linear SVM and sparse linear SVM classifiers on persistence images computed on HRV data over ten iterations of five-fold cross-validation.*

	Accuracy	AUC–ROC
Linear SVM	68.5	0.793
Sparse LSVM	65.7	0.682