

Topo-Geometric Analysis of Variability in Point Clouds Using Persistence Landscapes

James Matuk , Sebastian Kurtek , *Senior Member, IEEE*, and Karthik Bharath

Abstract—Topological data analysis provides a set of tools to uncover low-dimensional structure in noisy point clouds. Prominent amongst the tools is persistence homology, which summarizes birth-death times of homological features using data objects known as persistence diagrams. To better aid statistical analysis, a functional representation of the diagrams, known as persistence landscapes, enable use of functional data analysis and machine learning tools. Topological and geometric variabilities inherent in point clouds are confounded in both persistence diagrams and landscapes, and it is important to distinguish topological signal from noise to draw reliable conclusions on the structure of the point clouds when using persistence homology. We develop a framework for decomposing variability in persistence diagrams into topological signal and topological noise through alignment of persistence landscapes using an elastic Riemannian metric. Aligned landscapes (amplitude) isolate the topological signal. Reparameterizations used for landscape alignment (phase) are linked to a resolution parameter used to generate persistence diagrams, and capture topological noise in the form of geometric, global scaling and sampling variabilities. We illustrate the importance of decoupling topological signal and topological noise in persistence diagrams (landscapes) using several simulated examples. We also demonstrate that our approach provides novel insights in two real data studies.

Index Terms—Amplitude-phase separation, persistence landscapes, topological data analysis.

I. INTRODUCTION

IT IS difficult to draw statistical insights from datasets where each observation corresponds to an object with rich structure. Consider the data shown in Fig. 1, where panels (a) and (b) correspond to three-dimensional brain artery trees of different subjects, previously studied by [1] to understand how demographic factors are associated with brain structure; panels (c) and (d), on the other hand, display two example prostate gland biopsy

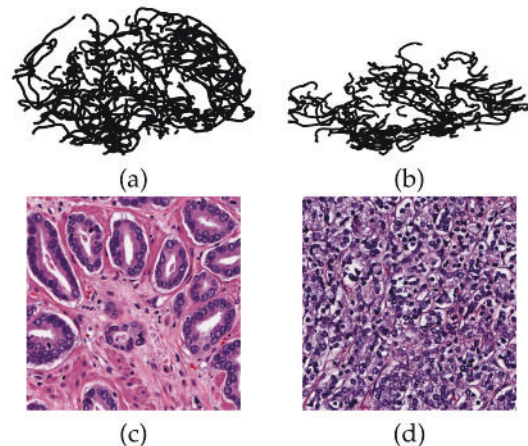


Fig. 1. Left: Brain artery trees for (a) 20 year old and (b) 79 year old subjects from [1]. Right: Images of a prostate biopsy with (c) benign and (d) malignant carcinoma from [2].

images, which were studied by [2] to aid prostate cancer prognosis. The visual differences between the observations within each of the two studies are striking. However, formally quantifying the differences between these objects, to enable statistical analysis, is a challenging and important problem. Topological Data Analysis (TDA) focuses on applying tools from (algebraic) topology to summarize and quantify the structure in complex data objects. In essence, TDA can be viewed as a general toolbox that enables discovery of topological and geometric features in complex data that can be used for subsequent analysis using existing statistical and machine learning methods.

Persistence homology is a prominent tool within TDA that provides a multi-resolution view of the topological and geometric features of data represented as point clouds in \mathbb{R}^d , e.g., samples of points on the brain artery trees or outlines of prostate glands shown in Fig. 1. As a resolution parameter changes, so do the features of the data, and these changes are recorded in persistence diagrams. Statistical analysis of samples of persistence diagrams are based on distances, such as the Wasserstein distance or bottleneck distance, which enable computation of descriptive statistics [3], [4], [5] and confidence regions [6]. Carrying out statistical analysis directly on the space of persistence diagrams is difficult since they are multisets of planar points. This motivates using functional representations (summaries) of diagrams that are more amenable for statistical analysis [2] using tools from functional data analysis [7]. In this paper, we analyze persistence landscapes [8], although it will become clear that the proposed methods can be used on other functional

Received 24 March 2023; revised 19 January 2024; accepted 22 August 2024. Date of publication 28 August 2024; date of current version 5 November 2024. The work of Sebastian Kurtek and Karthik Bharath was supported by NIH under Grant R37-CA214955, the work of Karthik Bharath was supported by NSF under Grant DMS-2015374, the work of Sebastian Kurtek was supported by NSF under Grant CCF-1740761, Grant DMS-2015226, and Grant CCF-1839252, and the work of James Matuk was supported by NIH under and Grant R01-ES028804. Recommended for acceptance by T. Tommasi. (Corresponding author: Sebastian Kurtek.)

James Matuk is with the Epidemiology Data Center, University of Pittsburgh, Pittsburgh, PA 15260 USA (e-mail: jam925@pitt.edu).

Sebastian Kurtek is with the Department of Statistics, The Ohio State University, Columbus, OH 43210 USA (e-mail: kurtek.1@stat.osu.edu).

Karthik Bharath is with the School of Mathematical Sciences, University of Nottingham, NG7 2RD Nottingham, U.K. (e-mail: Karthik.Bharath@nottingham.ac.uk).

This article has supplementary downloadable material available at <https://doi.org/10.1109/TPAMI.2024.3451328>, provided by the authors.

Digital Object Identifier 10.1109/TPAMI.2024.3451328

summaries, e.g., silhouettes [9], density estimates [10], rank functions [11], persistence entropy functions [12], persistence intensity functions and images [13], [14], with suitable modifications. Importantly, features extracted from persistence diagrams as well as the aforementioned functional summaries of diagrams have utility in modern machine learning approaches [15].

For point cloud data in \mathbb{R}^d generated from a distribution with support on a lower-dimensional manifold M , persistence diagrams are typically computed by constructing geometric simplicial complexes based on an open cover of metric balls centered at data points with respect to the Euclidean distance on \mathbb{R}^d . The construction engenders noise in persistence diagrams that is complementary to topological signal in the point cloud, since distances between data points in \mathbb{R}^d are sensitive to, mainly, three choices: (i) arbitrary (global) scaling of the point cloud; (ii) geometric configuration of the point cloud in \mathbb{R}^d with respect to the manifold M ; (iii) sampling variability, or density, of the points. The choices are linked to an implicit geometry of M : they imply an embedding $M \hookrightarrow \mathbb{R}^d$, under which a diffeomorphism of M , which preserves its topology, affects distances between points in M when measured using the Euclidean distance in the image of the embedding in \mathbb{R}^d . This results in different persistence diagrams for point clouds sampled from topologically identical manifolds M . Since the map that takes a persistence diagram to a persistence landscape is invertible [8], it is natural to query how such ‘topological noise’ manifests in a persistence landscape, and whether it is possible to exploit the structure of the space of landscapes to mitigate noise and amplify topological signal. The resulting amplification of topological signal has the potential to enhance statistical or machine learning analyses.

We refer to topological noise as any variation in the data that is complementary to topological information, although we emphasize that the terminology does not imply that geometric features resulting from the above-mentioned choices are of no use in downstream statistical tasks. The situation is similar in spirit to Kendall’s definition of landmark shape as all geometric information that remains after accounting for translation, scale and rotation variabilities [16]; in this setting, position, global scale and orientation of a set of landmark points are viewed as variation complementary to geometric shape. However, in many applications, these features of an object may be valuable descriptors [17]. Thus, the main focus in Kendall’s shape analysis lies in separating geometric shape information from the other sources of variability, and using them as complementary features of landmark configurations in downstream analyses. From this perspective, in the present setting, the only true source of nuisance variation is measurement error, which in general, is confounded with geometric and topological information.

A persistence diagram for a point cloud in \mathbb{R}^d is a multiset of points on the plane that offers a multi-resolution summary of the homology of the data, constructed using geometric filtered complexes on \mathbb{R}^d based on balls of radius $t > 0$ around each datum; the radius t acts as the resolution parameter in the sense that as it is increased, births and deaths of topological features of the point cloud are encoded in the corresponding persistence diagram. A persistence landscape is a collection of triangular functions $t \mapsto \lambda(t) \geq 0$, and is a bijective multivariate functional summary of a persistence diagram. As such, how the value of t is increased is a data analytic artefact, and should not affect the

topological signal in the point cloud. In practice, however, increasing t at different rates will result in different persistence diagrams, and hence, persistence landscapes. In particular, we focus on establishing a relationship between topological signal/noise and the two main sources of variation in a functional dataset consisting of persistence landscapes: amplitude or shape,¹ which captures y -axis variation, and phase, which tracks variation in the relative timing of shape features, e.g., extrema. We further show that phase variation in persistence landscapes is tied to the rate of increase of the resolution parameter t . In functional data analysis, the perils of not accounting for both sources of variation when computing summaries such as the mean or exploring dominant directions of variation via (functional) principal component analysis (PCA) are well-documented (see e.g., [18], [19]). Evidently, such perils plague analysis of persistence landscapes, when viewed as points in a Banach space equipped with the L^p norm: the pointwise mean of a sample of persistence landscapes can fail to be one, and this affects interpretability of the corresponding persistence diagram. Instead, computing a mean landscape using only the amplitude components of a sample of landscapes by registering, or aligning, them will better preserve shape, and mitigate effects of topological noise. In such a setting, our main contributions are as follows.

- We establish an explicit link between the rate of increase of the resolution parameter t in a simplicial filtration that generates a persistence diagram and magnitude of phase variation present in the component functions λ_k , as captured through a reparameterization $s \mapsto \gamma(s)$ of the landscape $\Lambda(s) = (\lambda_1(s), \dots, \lambda_K(s))$. Specifically, we show how γ is related to variation in persistence diagrams induced by (i) (global) scaling of the data (Fig. 3), and (ii) sampling variability of data (Fig. 4).
- We show that alignment of landscapes $\{\Lambda_i\}_{i=1}^n$ by determining optimal reparameterizations $\{\gamma_i\}_{i=1}^n$ leads to an average landscape that better preserves the structure of the sample of landscapes. This induces a separation of variability in persistence landscapes into amplitude or shape, which captures the topological information in a dataset, and phase, which captures leftover variation due to geometric, global scaling and sampling variabilities. A key consequence is the ‘denoising’ of points in the corresponding persistence diagrams by transforming them using $\{\gamma_i\}$; therefore, computing persistence diagrams for datasets $\{X_i\}$ using simplicial filtrations with balls of transformed radii $\{t \rightarrow \gamma_i(t)\}$ enhances topological information in persistence diagrams (Fig. 6).
- We demonstrate that the proposed approach for statistical analysis of landscapes offers new insight, and adds substantially, to the analyses of the brain artery tree data in [1] (Section IV) and prostate cancer data in [2] (Section V).

To the best of our knowledge, this is the first work in the literature to establish a concrete link between misalignment of persistence landscapes, and topological noise in persistence diagrams. However, in a certain sense, our approach in *determining*

¹The mathematical definitions of amplitude and shape are different in functional data analysis and shape analysis literatures. However, the two notions are very closely related for persistence landscapes, and we hence use them interchangeably.

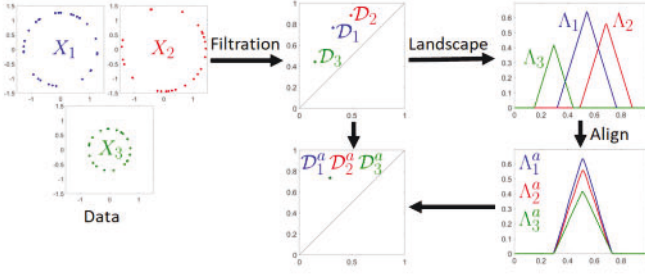


Fig. 2. Example of topological noise: Point clouds X_1, X_2, X_3 (with different sampling) from topologically identical spaces (differing only in scale) lead to different persistence diagrams $\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3$ and hence landscapes $\Lambda_1, \Lambda_2, \Lambda_3$. Our approach: Construct aligned landscapes $\Lambda_1^a, \Lambda_2^a, \Lambda_3^a$ and use alignment information to get transformed/denoised diagrams $\mathcal{D}_1^a, \mathcal{D}_2^a, \mathcal{D}_3^a$; use aligned landscapes for statistical analysis.

an optimal rate of increase of t , given by $\gamma(t)$, falls between the standard approach of fixing a t for each x_i and having t change with x_i , as with a multiscale approach that allows each ball $B_{x_i}(t_i)$ to have a possibly different radius t_i [20].

The remainder of the paper is organized as follows. In Section II, we further motivate alignment of persistence landscapes and provide technical details of our approach. In Section III, we use simulation studies to illustrate the importance of amplitude-phase separation in persistence landscapes for topological denoising. In Sections IV and V, we analyze the brain artery tree [1] and prostate cancer [2] datasets, respectively. In both cases, we focus on the scientific questions that motivated the two studies, and illustrate the benefits and novel insights gained from separate statistical analysis of the amplitude and phase components of persistence landscapes. In Section VI, we provide a short discussion. Appendices A-C in the supplement contain additional simulated examples and real data analysis results.

II. ELASTIC FUNCTIONAL DATA ANALYSIS OF PERSISTENCE LANDSCAPES

Before providing the technical details of our approach, we provide a summary of the proposed analysis pipeline in Fig. 2 using three point clouds X_1, X_2, X_3 with degree-1 (loops), one-dimensional persistence landscapes $\Lambda_1, \Lambda_2, \Lambda_3$. Topological noise is induced purely through scale (radii of circles) and sampling variability. Notice how transforming the diagrams $\{\mathcal{D}_i\}$ using $\{\gamma_i\}$ from alignment of $\{\Lambda_i\}$ collapses the three points to a single one (denoising), as it should be since spaces from which $\{X_i\}$ are sampled are topologically identical. In this setting, it is evident that the global scale and sampling variability are captured purely in the phase component of the persistence landscapes. The aligned landscapes, in turn, capture the topological information about the underlying spaces from which the data was sampled: they all contain a single maximum corresponding to a single cycle, a topological feature that arises at the same exact time across the three point clouds. Further, the aligned landscapes are identical to each other up to a uniform scaling of the function values, i.e., they have the same shape. In the remainder of this section, we review the basics of persistence diagrams and landscapes, discuss distinct sources of variability in persistence landscapes, and specify a statistical framework

to analyze these sources of variability using tools from elastic functional data analysis.

A. Persistence Diagram and Landscape

For a point cloud $X = \{x_1, \dots, x_N\}$ in \mathbb{R}^d , equipped with the standard Euclidean norm, generated from a distribution with support on a lower-dimensional manifold M , persistence homology is a tool that tracks homological features, such as connected components (degree-0), loops (degree-1), voids (degree-2), etc., of the point cloud at different resolutions [21]. Homology is computed using geometric filtered complexes constructed from the union of balls $\cup_{i=1}^N B_{x_i}(t)$ around each point, each with the same radius $t > 0$. The Čech complex, $\check{C}ech(X, t)$, consists of k -simplices whose nodes have $k+1$ many balls with a non-empty intersection. In contrast, the Vietoris-Rips complex, or just Rips complex, $Rips(X, t)$, is easier to compute and consists of k -simplices whose nodes have $k+1$ many balls with a non-empty pairwise intersection. At each fixed radius t , the homology of the simplicial complex is a snapshot of the features of the point cloud. Considering all radii, $t > 0$, provides a multi-resolution view of the features of the point cloud where features are born and die at different values of t . Persistence homology tracks the features with a persistence diagram, a function from a countable set to $\{(x, y) \in \mathbb{R}^2 | x < y\}$, consisting of birth-death pairs, (b_j, d_j) , the times at which the j^{th} feature was born and its corresponding death time. A persistence diagram is thus a multiset consisting of these points and represents a multi-resolution summary of the homology of the point cloud; see [21] for more general treatments of persistence homology and persistence diagrams.

A persistence landscape is an invertible functional representation of persistence homology computed from a persistence diagram [8]. For $X = \{x_1, \dots, x_N\}$, let $\mathcal{D}^p(X)$ denote its degree- p persistence diagram consisting of m birth-death pairs $\{(b_j, d_j)\}_{j=1}^m$. The basic units of persistence landscapes are triangular functions computed using coordinates of points in a persistence diagram, $\ell_j^p(t) = (t - b_j)\mathbb{I}_{\{b_j \leq t \leq \frac{1}{2}(b_j + d_j)\}} + (d_j - t)\mathbb{I}_{\{\frac{1}{2}(b_j + d_j) \leq t \leq d_j\}}$. For $k \in \mathbb{N}$, the k^{th} landscape function is defined as $\lambda_k^p(t) = \max_{j=1, \dots, m} \ell_j^p(t)$, which is the k^{th} maximum of the triangular functions with $\lambda_k(t) = 0$ for all $k > m$ by definition. Each function λ_k^p thus begins and ends at zero. In practice, we truncate the number of landscape functions used for data analysis to the K many that have some positive values along their domain. The degree- p persistence landscape for X is defined as the collection of landscape functions $\Lambda_X^p(t) = \{\lambda_k^p(t)\}_{k=1}^K$. In this work, we consider degree- p persistence landscapes of samples of point clouds X_1, \dots, X_n , simply denoted by $\Lambda_1(t), \dots, \Lambda_n(t)$; we explicitly specify the degree p under consideration in all simulated and real data examples.

B. Effects of Global Scaling, Sampling and Geometric Variabilities

The construction of geometric simplicial complexes on X engenders noise in a persistence diagram that is complementary to the topological signal in the point cloud, since distances between points in \mathbb{R}^d are sensitive to (i) (global) scaling, (ii) geometric configuration, and (iii) sampling variability of the

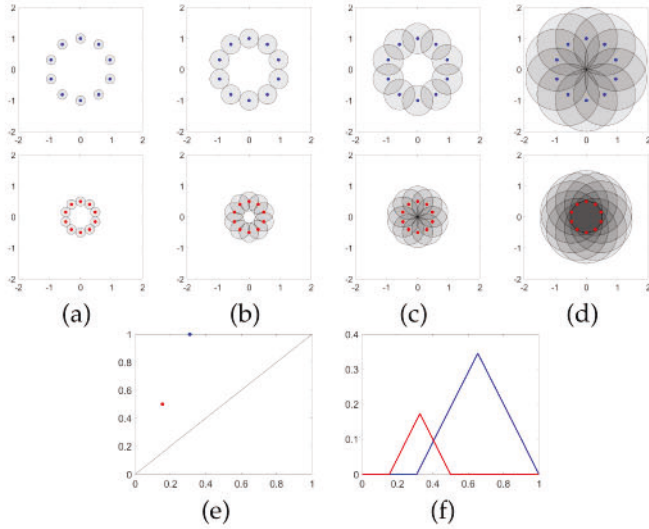


Fig. 3. Same topology with scale variability only: Construction of Rips filtration for two point clouds on circles with radii 0.5 (red) and 1 (blue) at resolutions (a) $t = .1545$, (b) $t = .3090$, (c) $t = .5$, and (d) $t = 1$. (e) Corresponding persistence diagrams and (f) landscapes.

point cloud. Our interest lies in studying how changes in (i)–(iii) result in different persistence diagrams constructed using Rips or Čech simplicial filtrations.

Consider the simple setting where the manifold M is a circle of radius $t > 0$, embedded into \mathbb{R}^2 as $\theta \mapsto (t \cos \theta, t \sin \theta)$. Note that all circles with radius $t \geq 0$ are topologically identical. However, changing the radius t changes the metric $dt^2 + t^2 d\theta^2$ on \mathbb{R}^2 , and hence its restriction to the circle; this amounts to changing the embedding that provides coordinates for the observed points (geometric configuration), which ultimately changes the distance between points, as measured in \mathbb{R}^2 , used to construct the filtration. Fig. 3 illustrates this by considering two point clouds consisting of ten equidistant points along circles with radii 1 (blue) and 0.5 (red), respectively. We consider degree $p = 1$ persistence homology (loops) with $K = 1$ -dimensional landscapes when topological noise is entirely due to scale effects. In Fig. 3(a)–(d), balls of different radii t are drawn around the points in the point clouds. In (a), when $t = .1545$, a loop forms for the red point cloud, while there is no loop present for the blue point cloud. In (b), when $t = .3090$, a loop forms for the blue point cloud, and the loop persists for the red point cloud. In (c), when $t = .5$, the loop closes for the red point cloud, and persists for the blue point cloud. Finally, in (d), when $t = 1$, the loop closes for the blue point cloud. The loops can be summarized by the birth-death pairs $(.1545, .5)$ and $(.3090, 1)$ for the red and blue point clouds, respectively. Panel (e) shows them in the persistence diagram, while panel (f) displays the corresponding misaligned persistence landscapes.

In the persistence diagram, (b, d) coordinates of the red point are half of those for the blue, and this matches the ratio of the radii of the two circles; this implies that the persistence landscape for the red point cloud is shorter and shifted to the left by a commensurate amount as compared to the landscape for the blue point cloud. This scale-induced topological noise thus arises by

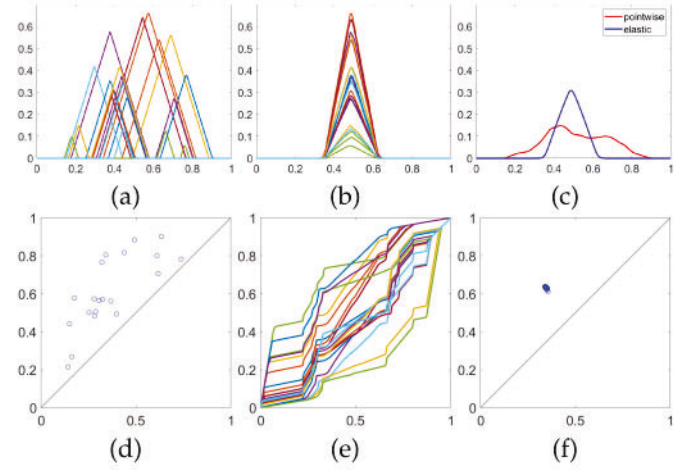


Fig. 4. Same topology with scale and sampling variabilities: (a) Persistence landscapes $\{\Lambda_i\}_{i=1}^{20}$ of 20 point clouds of the type in Fig. 2. (b) Aligned persistence landscapes $\{\Lambda_i(\gamma_i)\}_{i=1}^{20}$. (c) Mean landscape after (blue) and without (red) alignment. (d) Noisy persistence diagrams $\{(b_i, d_i)\}_{i=1}^{20}$ from 20 point clouds. (e) Estimated reparameterizations $\{\gamma_i\}_{i=1}^{20}$. (f) Denoised persistence diagrams $\{(\gamma_i^{-1}(b_i), \gamma_i^{-1}(d_i))\}_{i=1}^{20}$.

a common scaling of the triangular function ℓ_1^1 , given by

$$\ell_1^1(t) = (t - \alpha b) \mathbb{I}_{\{\alpha b \leq t \leq \frac{\alpha}{2}(b+d)\}} + (\alpha d - t) \mathbb{I}_{\{\frac{\alpha}{2}(b+d) \leq t \leq \alpha d\}}, \quad (1)$$

where $\alpha = 2$ (when blue point cloud is derived from red). If $\alpha < 1$, the triangular function will be shifted to the left along the domain and will be shorter relative to $\alpha = 1$; if $\alpha > 1$, the function will be shifted to the right and taller.

In this example based on the circle, we are able to explicitly link topological noise to a single parameter, the radius t of the circle, which governs the magnitude of both scale and geometric configuration of points. In essence, *when topological noise is due to scaling, alignment of peaks of the persistence landscapes will move points in a persistence diagram toward each other, and thus amplify the topological signal*. In other words, such topological noise manifests entirely through phase variation in the landscapes. In higher dimensions, it is not possible in general to carry out this program since the restriction of the metric on \mathbb{R}^d induced by the embedding of the manifold M is more complicated. Nevertheless, we demonstrate through numerical examples in Section III and Appendices A and B in the supplement that alignment of persistence landscapes acts as a denoising mechanism for the corresponding diagrams, even when topological noise is not only due to scaling.

From the discussion above, it is clear that changing the embedding (e.g., $x \mapsto (x, \sqrt{t - x^2})$) would have generated similar topological noise in the persistence diagram. This of course amounts to a change in the metric which ultimately results in a change in the simplicial filtration. When sampling variability is present (e.g., non-equispaced points on the circle), the situation can be viewed as one involving *local* scaling of the point cloud, and alignment of peaks of the landscapes will again induce points on the persistence diagrams to move toward each other (Section III-B, Figs. 4 and 6); similar comments apply to the situation involving measurement error (points do not lie exactly on M ; Appendix A in the supplement, Figs. 1–3).

C. Reparameterizing a Landscape and Denoising a Diagram

The discussion in the previous subsection suggests alignment of persistence landscapes, by lining up peaks and valleys of the component functions, as a viable denoising mechanism for persistence diagrams. We propose to do this through shape analysis of landscapes, viewed as parameterized curves. Our approach is based on the elastic metric to compare shapes of curves [19], based on a convenient transform of the landscape curves. Specifically, the transformation maps landscapes into a Hilbert space, where geometric computations become simplified; this stands in contrast to the Banach space setting typically used for persistence landscapes. An important consequence of this is that we are able to compute a mean landscape based on its amplitude which, in contrast to the pointwise landscape currently computed in a Banach space setting, better preserves the shape of a landscape; moreover, the Hilbert space structure provides an inner product to carry out PCA on the amplitude or shape component of landscapes, which enables one to study dominant modes of variation in samples of point clouds.

By virtue of its definition, a landscape is parameterized by the resolution (filtration) parameter t , used to construct the Rips or the Čech simplicial filtrations, which in principle can be any positive real. In order to choose a closed interval of \mathbb{R} as a parameter domain, note that, given n persistence landscapes, there always exists an $0 < s < \infty$ such that $\Lambda_i(t) = 0$, $\forall t > s$, $i = 1, \dots, n$. Given this, one can assume, without loss of generality, the parameter domain to be the unit interval obtained through rescaling by $1/s$. Then, for each $i = 1, \dots, n$, landscape Λ_i is a K -dimensional piecewise linear parameterized closed curve $[0, 1] \ni t \mapsto \Lambda_i(t) \in \mathbb{R}^K$ with $\Lambda_i(0) = \Lambda_i(1) = 0$. This in turn results in scaled persistence diagrams $\{(b_{i,j}/s, d_{i,j}/s)\}_{i=1, j=1}^{n, m_i}$, so that the birth-death pairs are in $[0, 1]^2$.

Definition II.1: A persistence landscape (diagram) Λ ($\{(b_j, d_j)\}$) obtained by rescaling in the above manner is referred to as a scaled persistence landscape (diagram).

We will simply henceforth refer to a scaled persistence landscape (diagram) as a persistence landscape (diagram), unless explicitly mentioned to the contrary.

Since a reparameterization of a landscape Λ preserves its image, its shape, modulo scale, is preserved. As the set of reparameterizations, consider

$$\Gamma = \{\gamma : [0, 1] \rightarrow [0, 1] : \dot{\gamma} > 0, \gamma(0) = 0, \gamma(1) = 1\},$$

the set of orientation-preserving diffeomorphisms of $[0, 1]$, which forms a group under composition ($\dot{\gamma}$ is the derivative of γ). The group Γ acts on the set of landscapes from the right as function composition: $(\Lambda, \gamma) \rightarrow \Lambda(\gamma)$. Alignment of landscapes $\{\Lambda_i\}$ thus amounts to establishing correspondence between (K -dimensional) points in the images $t \mapsto \Lambda_i(t)$, achieved by determining optimal $\gamma_i \in \Gamma$ such that the collection $\{\Lambda_i(\gamma_i)\}$ is ‘optimally’ aligned, where optimality is defined with respect to a metric-based matching functional. For each i , γ_i represents common phase variation in the component functions $(\lambda_{1,i}, \dots, \lambda_{K,i})$ of Λ_i .

Since the resolution/filtration parameter t is the parameter for a landscape when viewed as a closed curve, we observe the following.

Proposition 1: The map $\Lambda \mapsto \Phi(\Lambda) := \{(b_j, d_j)\}$ that takes a scaled persistence landscape Λ to a unique scaled persistence diagram $\{(b_j, d_j)\}$ is equivariant with respect to the action of Γ on the set of scaled persistence landscapes, i.e., $\Phi(\Lambda(\gamma)) = \gamma^{-1}\{(b_j, d_j)\} = \{(\gamma^{-1}(b_j), \gamma^{-1}(d_j))\}$.

A useful way to think of the induced transform $\{(\gamma^{-1}(b_j), \gamma^{-1}(d_j))\}$ on a persistence diagram is as a *nonlinear local scale change* of the multiset of points that generalizes the global scale change described in (1). This attempts to reverse the effects of (i) working with rescaled persistence landscapes, and (ii) potential topological noise induced through the geometric construction of simplicial filtrations. Indeed, this is tantamount to considering a geometric Čech or Rips filtration with parameter $\gamma(t)$: for a point cloud X , the corresponding Rips simplicial complex is defined as

$$\begin{aligned} \sigma = [x_1, \dots, x_k] \in \text{Rips}(X, \gamma(t)) &\iff |x_i - x_j| \leq \gamma(t) \\ &\iff \gamma^{-1}(|x_i - x_j|) \leq t, \quad \forall i, j \end{aligned}$$

since γ is strictly increasing. Equivalently, the equivariant action on Φ implies that the metric induced by the Euclidean norm $|\cdot|$ on \mathbb{R}^d is deformed by a diffeomorphism γ^{-1} to track local scale changes needed to preserve the integrity of topological signal in the presence of topological noise.

However, some care is needed with this interpretation since $\gamma^{-1}(|\cdot|)$ fails to be a metric on \mathbb{R}^d if the function γ^{-1} is not concave, since the triangle inequality will otherwise not be satisfied. Moreover, the chain of inclusions

$$\text{Rips}(X, t') \subset \check{\text{Cech}}(X, t) \subset \text{Rips}(X, t), \text{ when}$$

$$\frac{t}{t'} \geq \sqrt{\frac{2d}{d+1}}$$

that characterize Rips and Čech complexes with parameter t [22] need not be preserved under $\gamma(t)$ for all $\gamma \in \Gamma$; since preservation would depend on the magnitude of the derivative of γ , it is difficult to provide a lower bound for $\gamma(t)/\gamma(t')$ that holds for all $\gamma \in \Gamma$. In principle, it is possible to restrict attention to a subset of diffeomorphisms γ with a concave inverse that preserve the chain of inclusions, but this may restrict how well one is able to denoise persistence diagrams by aligning persistence landscapes.

We can summarize the practical consequence of the above discussion in the following manner: on point clouds $\{X_i\}$, the equivariant action of Γ ensures that, in the optimally aligned landscapes $\{\Lambda_i(\gamma_i)\}$, the extrema (mainly peaks) of the component functions $(\lambda_{1,i}(\gamma_i), \dots, \lambda_{K,i}(\gamma_i))$ line up, and the transformed points $\{(\gamma_i^{-1}(b_{ij}), \gamma_i^{-1}(d_{ij}))\}$, consequently, will tend to cluster, the number of which will depend on the topology of the underlying manifold. As a consequence, if a Čech or Rips filtration for X_i is constructed with balls of radius $\gamma_i(t)$, the corresponding persistence diagram will be ‘denoised’.

D. Optimal Reparameterizations, Mean Amplitude Landscape and PCA

The program described above rests on determining the optimal reparameterizations $\{\gamma_i\}$ from observed landscapes $\{\Lambda_i\}$. In principle, any registration or alignment procedure for curves in \mathbb{R}^K can be used. Our choice is based on the highly successful Elastic Functional Data Analysis (EFDA) framework, a

Riemannian-geometric approach that utilizes the elastic metric for curves in \mathbb{R}^K . This framework is characterized by two important theoretical considerations for the curve alignment problem: (i) isometry, i.e., invariance to simultaneous reparameterization of curves, and (ii) invariance of optimal reparameterizations to rescaling of curves. These are addressed through the use of the elastic Riemannian metric for comparing absolutely continuous curves, which is difficult to compute in practice. For ease of exposition, we refrain from providing the definition of the metric and its properties; see Chapter 10 in [19] for details.

Let \mathcal{F} denote the space of absolutely continuous curves in \mathbb{R}^K equipped with the elastic metric. Practical use of the metric is greatly simplified through use of the square-root velocity function (SRVF) representation. For any curve $\beta \in \mathcal{F}$, its SRVF is defined as

$$\beta \mapsto Q(\beta) = q := \dot{\beta}(|\dot{\beta}|)^{-1/2},$$

where $\dot{\beta}$ is the componentwise derivative and $|\cdot|$ is the Euclidean norm on \mathbb{R}^K . The map $Q : \mathcal{F} \rightarrow \mathbb{L}^2([0, 1], \mathbb{R}^K)$ is a homeomorphism [23] with inverse $\beta(t) = \int_0^t q(u)|q(u)|du$, and effectively ‘flattens’ the complicated elastic metric: the distance $d(\beta_1, \beta_2)$ between two absolutely continuous curves with respect to the elastic metric equals $\|Q(\beta_1) - Q(\beta_2)\|_2 = \|q_1 - q_2\|_2 = [\int_0^1 |q_1(t) - q_2(t)|^2 dt]^{1/2}$, and the standard \mathbb{L}^2 metric on SRVFs of curves possesses desiderata (i) and (ii) mentioned above.

Absolute continuity of a curve in \mathbb{R}^K is defined via absolute continuity of its one-dimensional component functions. Absolutely continuous functions in one dimension have constant speed parameterization [24]. In the present setting, persistence landscapes $\Lambda : [0, 1] \rightarrow \mathbb{R}^K$ are piecewise linear curves. It is known that the set of piecewise linear curves in \mathbb{R}^K is dense in \mathcal{F} [25], and continuity of the map Q ensures that its image under Q , consisting of piecewise constant SRVFs, is dense in $\mathbb{L}^2([0, 1], \mathbb{R}^K)$.

Our definition of the amplitude (shape) of a curve and subsequent statistical analysis approach are analogous to the definitions presented in [26] for univariate functions. The group Γ acts on \mathcal{F} through composition, and since the map $Q : \mathcal{F} \rightarrow \mathbb{L}^2([0, 1], \mathbb{R}^K)$ is bijective, the action $(q, \gamma) = Q(\Lambda(\gamma)) = (q(\gamma))\sqrt{\gamma}$ of Γ can be defined on $\mathbb{L}^2([0, 1], \mathbb{R}^K)$, under which, the amplitude of a landscape Λ is its orbit $[q] := \{(q, \gamma) \mid \gamma \in \Gamma\}$. Since $\|(q, \gamma)\|_2 = \|q\|_2$ for every $q \in \mathbb{L}^2([0, 1], \mathbb{R}^K)$ and $\gamma \in \Gamma$, we note that Γ acts by isometries on $\mathbb{L}^2([0, 1], \mathbb{R}^K)$. Under this definition, two curves, Λ_1, Λ_2 , have the same amplitude if their corresponding SRVFs are in the same orbit, i.e., there exists a $\gamma \in \Gamma$ such that $q_1 = (q_2, \gamma)$. The set of all orbits forms a partition of \mathcal{Q} and is the quotient space \mathcal{Q}/Γ . Hence, \mathcal{Q}/Γ defines the amplitude space of persistence landscapes.

The amplitude distance between two landscapes $\Lambda_1, \Lambda_2 \in \mathcal{F}$ is defined as the distance between their corresponding SRVF orbits $[q_1], [q_2] \in \mathcal{Q}/\Gamma$:

$$d_a(\Lambda_1, \Lambda_2) = d([q_1], [q_2]) = \min_{\gamma \in \Gamma} \|q_1 - (q_2, \gamma)\|_2. \quad (2)$$

Key to the definition of the above distance is the invariance of the \mathbb{L}^2 metric, under the SRVF representation, to simultaneous reparameterization of curves. The function

$$\gamma^* = \operatorname{argmin}_{\gamma \in \Gamma} \|q_1 - (q_2, \gamma)\|_2$$

is then the optimal reparameterization of Λ_2 to register or align it to Λ_1 . Furthermore, $\gamma^* = \operatorname{argmin}_{\gamma \in \Gamma} \|q_1 - (q_2, \gamma)\|_2 = \operatorname{argmin}_{\gamma \in \Gamma} \|c_1 q_1 - (c_2 q_2, \gamma)\|_2$, $c_1, c_2 \in \mathbb{R}_+$, i.e., rescaling of the curves does not alter the optimal reparameterization.

Denosing persistence diagrams from point clouds X_1, \dots, X_n requires determining optimal reparameterizations $\gamma_1, \dots, \gamma_n$ that *jointly* align persistence landscapes $\Lambda_1, \dots, \Lambda_n$. This requires a template landscape to align the individual ones to. We use a data-driven template given by the mean amplitude landscape. Denote by q_1, \dots, q_n the SRVFs of $\Lambda_1, \dots, \Lambda_n$. This mean is defined as the quantity that minimizes the sum of squared amplitude distances:

$$[\hat{\mu}_q] = \operatorname{argmin}_{[q] \in \mathcal{Q}/\Gamma} \sum_{i=1}^n \min_{\gamma \in \Gamma} \|q - (q_i, \gamma)\|_2^2. \quad (3)$$

An orbit representative $[\hat{\mu}_q]$ is found by iteratively aligning q_1, \dots, q_n to the current estimate of the mean and averaging the aligned SRVFs to produce a new mean estimate; this is repeated until convergence. For identifiability, we use the center of the orbit of $[\hat{\mu}_q]$ as the representative element of the elastic mean; henceforth, we simply refer to this element of the mean orbit as $\hat{\mu}_q$. For additional algorithmic details and the orbit centering step, we refer to [26]. The corresponding mean amplitude landscape $\hat{\mu} \in \mathcal{F}$ is defined as $Q^{-1}(\hat{\mu}_q)$.

The joint alignment of $\Lambda_1, \dots, \Lambda_n$ can then be achieved via pairwise alignment of each Λ_i , $i = 1, \dots, n$ to the mean $\hat{\mu}$ using (2) by determining the optimal reparameterizations

$$\gamma_i = \operatorname{argmin}_{\gamma \in \Gamma} \|\hat{\mu}_q - (q_i, \gamma)\|_2, \quad i = 1, \dots, n,$$

which can be used to study phase variability. Details of methods for statistical analysis of reparameterization functions, including computation of a distance, averaging and PCA are available in [27], and are omitted here for brevity.

Since the aligned landscapes $\Lambda_i(\gamma_i)$, $i = 1, \dots, n$, or equivalently their SRVFs, (q_i, γ_i) , describe amplitude variability in the sample, a sample amplitude covariance function can be defined as

$$\widehat{C_q}(t, u) := \frac{1}{n-1} \sum_{i=1}^n ((q_i, \gamma_i)(t) - \hat{\mu}_q(t))((q_i, \gamma_i)(u) - \hat{\mu}_q(u))^T. \quad (4)$$

Amplitude-based PCA is carried out via eigendecomposition of $\widehat{C_q}(t, u)$,

$$\widehat{C_q}(t, u) = \sum_{b=1}^{\infty} \hat{\tau}_b \hat{\phi}_b(t) \hat{\phi}_b(u)^T, \quad (5)$$

where $\hat{\phi}_b$, $b \in \mathbb{N}$ are the primary directions of amplitude variability (amplitude PCs) and $\hat{\tau}_b$, $b \in \mathbb{N}$ are variances in the corresponding directions. Typically, one selects a finite number, B , of principal components that describe a large portion of amplitude variability. The aligned SRVFs can then be projected onto the B directions of amplitude variability with largest variance, $\beta_{i,b} := \int_0^1 \langle (q_i, \gamma_i)(t) - \hat{\mu}_q(t), \hat{\phi}_b(t) \rangle dt$, $b = 1, \dots, B$, $i = 1, \dots, n$, where $\langle \cdot, \cdot \rangle$ is the Euclidean inner product in \mathbb{R}^K . The PC scores, $\beta_i = (\beta_{i,1}, \dots, \beta_{i,B})^T$, $i = 1, \dots, n$ serve as a low dimensional Euclidean representation of the amplitude of landscapes. To visualize the primary directions of amplitude variability, we

compute $\mathcal{F} \ni Q^{-1}(\hat{\mu}_q + \nu\sqrt{\hat{\tau}_b}\hat{\phi}_b)$, i.e., a landscape that is ν standard deviations from $\hat{\mu}_q$ in the direction of $\hat{\phi}_b$.

III. SIMULATION STUDIES

In this section, we present simulation examples which demonstrate (i) denoising of persistence diagrams, obtained under scale and sampling variabilities in point clouds, through alignment of landscapes, and (ii) benefits of computing the mean landscape and PC directions on the set of aligned landscapes as opposed to computing a pointwise mean with unaligned ones, as currently done in practice. The supplement contains additional examples that (i) consider data sampled with additive noise (Appendix A), and (ii) illustrate mean estimation (Appendix B).

For these examples, we use the `ripsDiag` function to compute persistence diagrams using the Vietoris-Rips simplicial complex for point clouds, and the `landscape` function to compute landscapes from persistence diagrams; both functions are part of the TDA R package [28]. In the EFDA framework, registration, mean estimation and PCA for a sample of landscapes are implemented in MATLAB. Code and data to reproduce simulated examples are available here: <https://github.com/jamesmatuk/EFDA-of-Persistence-Landscapes>.

A. Examples 1 and 2: Mean From Aligned Landscapes

Example 1: We consider 20 point clouds, where each point cloud is generated by (i) sampling M from a Discrete-Uniform(10, 30), (ii) sampling r from $|N(1, 0.3^2)|$, and (iii) generating M points uniformly on a circle with radius r . Fig. 2 shows three (from 20) point clouds along with the corresponding degree $p = 1$, $K = 1$ -dimensional landscapes. Fig. 4(a) shows all 20 landscapes $\{\Lambda_i\}_{i=1}^{20}$. The amplitude and phase variations in the landscapes are related to variability in the radii, sample size and dispersion. Panels (b)-(f) demonstrate the benefit of alignment of landscapes $\{\Lambda_i\}$: a visually better mean (c) is obtained by using the aligned landscapes $\{\Lambda_i(\gamma_i)\}$ (b); transforming points in the persistence diagrams (d) using reparameterizations $\{\gamma_i\}$ (e) results in denoising (f) by collapsing all points to a single one, since the topology of the 20 point clouds is the same.

Example 2: In Fig. 5, we consider mean estimation based on degree $p = 0$, $K = 1$ -dimensional persistence landscapes computed from 20 point clouds that consist of 2000 points uniformly sampled along two interwoven spirals. The tightness of the spirals is random, so that the spirals complete Uniform(2, 5) revolutions. Panels (a) and (c) show two examples of point clouds generated in such a manner with the corresponding landscapes shown in panels (b) and (d). The tighter spirals in (a) have points closer together, and the resulting landscape is smaller and shifted to the left as compared to the spirals in (c). When computing landscapes, we only considered the point in persistence diagrams that corresponded to the death time that coincided with the intersection of the two spirals present in each point cloud. Panels (e)-(g) show landscapes for all 20 point clouds, their alignment, and a comparison of the mean before (red) and after (blue) alignment. The mean based on aligned landscapes appears to have sharper features that are consistent with the observed landscapes. Based on the denoised persistence diagrams in (j), in contrast to the noisy persistence diagrams in (h), it is evident

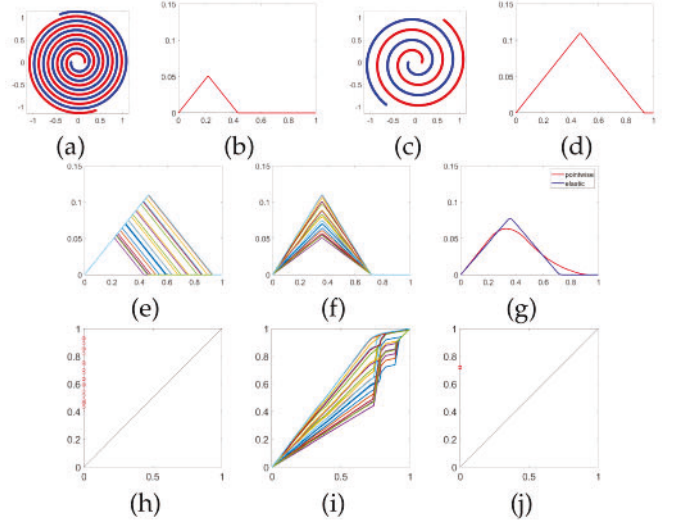


Fig. 5. Same topology with scale variability: (a) and (c) Two examples, from 20, of randomly generated point clouds. (b) and (d) Corresponding persistence landscapes. (e) Persistence landscapes $\{\Lambda_i\}_{i=1}^{20}$ of 20 point clouds. (f) Aligned persistence landscapes $\{\Lambda_i(\gamma_i)\}_{i=1}^{20}$. (g) Mean landscape after (blue) and without (red) alignment. (h) Noisy persistence diagrams $\{(b_i, d_i)\}_{i=1}^{20}$ from 20 point clouds. (i) Estimated reparameterizations $\{\gamma_i\}_{i=1}^{20}$. (j) Denoised persistence diagrams $\{(\gamma_i^{-1}(b_i), \gamma_i^{-1}(d_i))\}_{i=1}^{20}$.

that reparameterization of landscapes completely accounts for the scale variability associated with the tightness of the spirals, i.e., all points collapse to a single point in the denoised diagrams.

B. Example 3: PCA on Aligned Landscapes

We consider a more involved setting involving 20 point clouds from two topologically different spaces: (i) one circle, and (ii) two connected circles. Point clouds from (i) are drawn in the same manner as in Example 1, but for the fact that the sample size M is drawn from a Discrete-Uniform(20, 30). For point clouds from (ii), the radius of the larger circle is drawn from a $|N(1, 0.3^2)|$, while the radius of the smaller circle is a random proportion of the larger circle, drawn from a Beta(10, 10). Panels (a) and (c) in Fig. 6 show one point cloud each from (i) and (ii). We consider degree $p = 1$, $K = 2$ -dimensional landscapes $\{\Lambda_i = (\lambda_{i1}, \lambda_{i2})\}_{i=1}^{20}$. For point clouds from (i), λ_{i1} will have one peak and $\lambda_{i2} = 0$ for all t ; for point clouds from (ii), λ_{i1} will have two peaks and λ_{i2} will have a single peak.

The top and bottom rows in Fig. 6(e)-(g) show the primary PC direction of variability within one standard deviation of the mean following alignment and without alignment, respectively. Panel (h) highlights the benefits of alignment of landscapes through better separation of the two settings, (i) and (ii), when projected along the first two PC directions. Specifically, in the top row, when PCA is carried out on aligned landscapes, all of the point clouds that have two loops have a negative first PC score, while all of the point clouds with only one loop have a positive first PC score. There is no such clear separation of the two groups when PCA is performed on unaligned landscapes, as seen in the bottom row of (h). The separation in PC scores after alignment is directly related to the improved interpretability of the primary PC direction shown in the top row in (e)-(g): the landscape in (e) corresponding to -1 standard deviation from the mean in

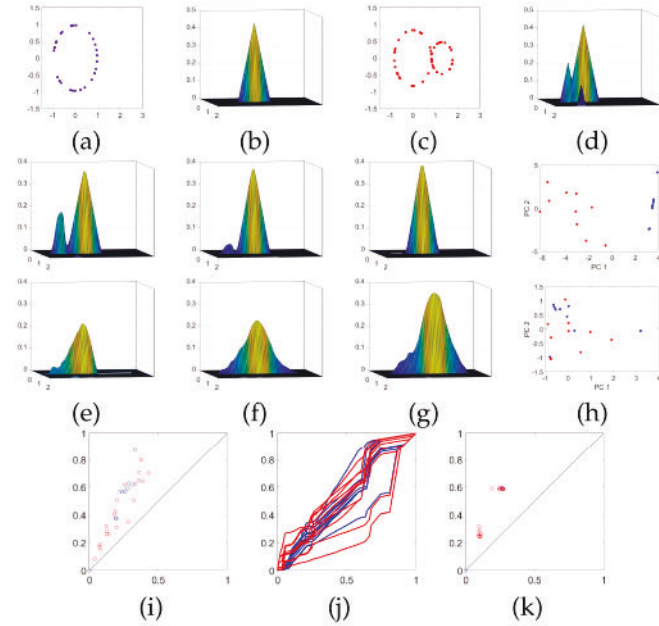


Fig. 6. Different topology with scale and sampling variabilities: (a) and (c) Two examples, from 20, of randomly generated point clouds from topologically different spaces (blue and red, respectively, in all relevant panels). (b) and (d) Corresponding degree $p = 1$, $K = 2$ -dimensional persistence landscapes. (e) -1 , (f) 0 , (g) $+1$ standard deviation from the mean landscape in the first PC direction, and (h) projection of landscapes onto the first two PC directions: following alignment (top) and without alignment (bottom). (i) Noisy and (k) denoised persistence diagrams. (j) Estimated reparameterizations.

(f) exhibits features of a landscape for a point cloud with two loops, while the landscape in (g) corresponding to $+1$ standard deviation exhibits features of a landscape for a point cloud with one loop.

Following results from previous simulations, we expect to see two clear clusters in the denoised persistence diagrams, using reparameterizations $\{\gamma_i\}$ shown in (j), corresponding to two distinct topological features; this is indeed the case as seen in (k). This is explained as follows: points concentrated around $(b, d) \approx (0.25, 0.6)$ correspond to the single circle in the blue point clouds and the large circle in the red ones. This is consistent with the data generating process where the large circles across the two groups correspond to each other. The points associated with the second feature for the red point clouds are concentrated around $(b, d) \approx (0.1, 0.25)$ and correspond to the additional significant homological feature (smaller circle) that generally has smaller persistence than the larger circle. It is very difficult to discern such topological information from the noisy diagrams in (i).

IV. ANALYSIS OF BRAIN ARTERY TREES

We now demonstrate the utility of the proposed approach on 3D point clouds representing brain artery trees. These data were collected to understand population attributes of brain arteries and how these attributes vary with demographic covariates. For a description of the experiment and data generation, see [29]. Information regarding human subjects in the experiment is available at <http://insight-journal.org/midas/community/view/21>. Chapter 10.1 in [30] provides an overview and comparison of past approaches used to analyze this data.

The approach of Bendich et al. [1] computed persistence diagrams from artery trees for 98 healthy human subjects, and used persistence diagrams to extract the 100 largest birth-death differences, $\{d_{i,j} - b_{i,j}\}_{i=1,j=1}^{98,100}$, for each subject. Restricting focus to the largest differences serves as a denoising step, since points close to the line $b = d$ in a persistence diagram can be thought of as noise [6].

This dataset is apt to demonstrate our approach for three reasons, one *a priori* and the other two *a posteriori*: (i) since each point cloud contains a large number (order of 10^5) of points, in order to be able to compute the diagrams, [1] subsampled 3000 points from each point cloud, thus creating large sampling variability; (ii) we uncover a significant scale effect between the two sex groups of subjects (males versus females); and (iii) we confirm the finding of [1] that there exists a significant correlation between the topological structure of the brain artery trees (as captured by PCs) and age. We note that these findings are exploratory and serve as a proof of concept for the proposed approach.

A. Exploration of Sex Differences Among Subjects

The starting point for our analysis are the persistence diagrams available at <https://marron.web.unc.edu/brain-artery-tree-data/>, and not the original 3D point clouds. From these, we compute degree $p = 1$, $K = 100$ -dimensional persistence landscapes. Information on the sex of each subject is also available along with the tree data. We investigate differences between mean persistence landscapes grouped by sex. One major finding of [1] is the existence of sex differences in their mean degree $p = 1$ feature vectors $\{d_{i,j} - b_{i,j}\}_{i=1,j=1}^{98,100}$. For convenience, we denote the mean landscape for the male (female) group following alignment (within each group) by $\hat{\mu}_a^m$ ($\hat{\mu}_a^f$), and pointwise mean computed without alignment for the males (females) by $\hat{\mu}^m$ ($\hat{\mu}^f$).

The means $\hat{\mu}_a^m$ and $\hat{\mu}_a^f$ are shown in the top row of Fig. 7(a) and (b), respectively. The difference between the means is obtained by first aligning the group means to the common pooled mean (computed with alignment) and then taking their difference, where all operations are carried out under the SRVF representation; this difference is shown in the top row of panel (c). The bottom row of Fig. 7(a)–(c) shows the pointwise means $\hat{\mu}^m$ and $\hat{\mu}^f$, and the corresponding $(\hat{\mu}^m - \hat{\mu}^f)$, when no alignment is carried out. The difference between the pointwise means has very large features indicating topological and geometric structural differences across males and females. However, these features are essentially non-existent in the difference of the aligned means. *This indicates that the large difference in the pointwise means is potentially due to misalignment, and can be construed as topological noise.*

We surmise that global scale differences and sampling variability in observed data between the male and female groups may be responsible for this phenomenon. To confirm this, we use the total artery length for each subject as a measure of global scale, which is also available as part of the tree data [31]. For each group, we estimate a cumulative distribution function (CDF) of total artery length using a kernel density estimate using the `ksdensity` function in MATLAB (default bandwidth). From the estimated CDFs shown in Fig. 7(d), it appears that total artery length is stochastically ordered by sex, with females having

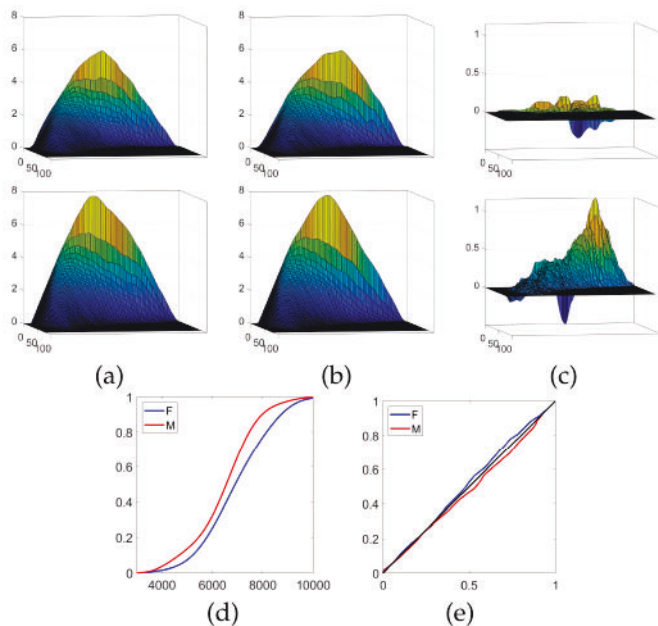


Fig. 7. Top. Mean persistence landscapes and their differences for males and females in the brain artery example with (top row) and without (bottom row) alignment: (a) Male mean, (b) female mean, and (c) difference between (a) and (b). Bottom. Relationship between groupwise total artery length and relative phase of groupwise means following alignment, across sexes: (d) Estimated groupwise CDFs of total artery length, and (e) reparameterizations that align groupwise aligned means to a mean computed from alignment of the pooled data. The identity parameterization (black) is shown for reference.

stochastically longer brain artery trees. Given this global scale disparity and sampling variability, we would expect $\hat{\mu}_a^f$ to be shifted to the right relative to $\hat{\mu}_a^m$. This behavior can be extracted from the phase difference between $\hat{\mu}_a^m$ and $\hat{\mu}_a^f$ after alignment to the pooled sample mean, as shown in Fig. 7(e). The blue reparameterization (female) shifts $\hat{\mu}_a^f$ to the left while the red (male) shifts $\hat{\mu}_a^m$ to the right. Thus, the misalignment caused by differences in global scale and sampling between the two groups appear to explain the reason behind the large difference between the groupwise pointwise means. In summary, the above analysis suggests that the sex effect detected via pointwise analysis without alignment of the landscapes, corresponding to some of the results presented in Table 10.1 in [30], is due to global scale and/or sampling differences of the observed data rather than differences in homology, and thus makes a compelling case study of the perils in ignoring the distinction between, and confounding of, amplitude and phase in landscapes.

B. Correlation Between Age and Topological Structure of Brain Artery Trees

Bendich et al. [1] show that age is significantly correlated with the dominant PCs estimated using degree $p = 0$ persistence diagrams. One potential confounding variable for this relationship is total artery length, since this quantity is also significantly correlated with age (correlation of -0.63 ; see Fig. 6(a) in Appendix C in supplement). To account for this, they first rescaled each subject's persistence diagram by their total artery length and then measured correlation between age and the dominant

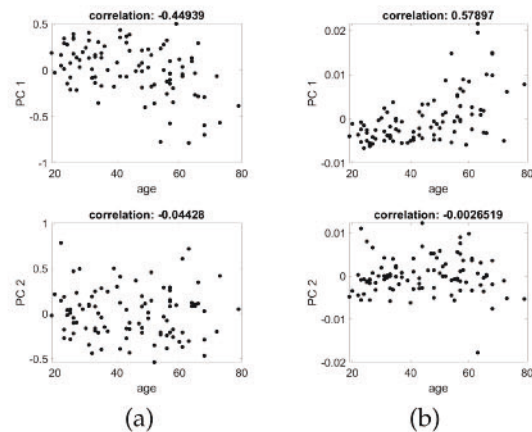


Fig. 8. Correlations and scatterplots of PC 1 (top) and PC 2 (bottom) estimated using (a) aligned and (b) unaligned landscapes, computed from rescaled persistence diagrams, versus age.

PCs estimated using rescaled diagrams. Even after rescaling, the relationship holds.

Here, we reanalyze the data using $K = 100$ -dimensional persistence landscapes from rescaled degree $p = 0$ persistence diagrams. We estimate PCs using aligned and unaligned landscapes, and project onto the first two PC directions to explore whether there is a correlation between either of the first two PCs and age. Fig. 8(a) and (b) shows the relationship between age and the first two PCs for aligned and unaligned landscapes, respectively. In both panels, age appears to be strongly associated with the first PC. Thus, our findings are consistent with [1] and the results presented in Table 10.1 in [30]. We further note that alignment of the persistence landscapes computed using rescaled persistence diagrams decreases the correlation slightly from 0.58 to 0.45 (note that the sign of the correlation coefficient is irrelevant here due to lack of directionality in the PCs); this is likely due to residual scale effects (and sampling variability) after accounting for the total artery length. Correlations computed using PCs from both unaligned and aligned persistence landscapes are similar to the correlation of 0.53 reported under TDA H_0 in Table 10.1 in [30]. Appendix C in the supplement reports results of this analysis when original persistence diagrams (without rescaling by total artery length) are used to derive persistence landscapes; there, we show that scale is confounded with topo-geometric structure when unaligned landscapes are used to carry out PCA.

V. CLASSIFICATION OF GLEASON DATA

As described in [2], the Gleason grading system is a prognostic tool to help understand severity of prostate cancer. Grading groups are assigned based on features of a prostate gland biopsy. In more benign biopsies, carcinoma walls are well-defined as seen in Fig. 1(a). On the other hand, malignant carcinoma lose their structure and have very irregular shapes as shown in Fig. 1(d). Using a variety of functional summaries of persistence diagrams, [2] classified simulated point clouds representing four different Gleason grade groups that ranged from benign to unhealthy. A representative from each of the four classes used in their study is shown in the left of Fig. 9(a)–(d), from benign to most severe. The middle and right columns show the corresponding degree $p = 1$ persistence diagrams and $K = 5$

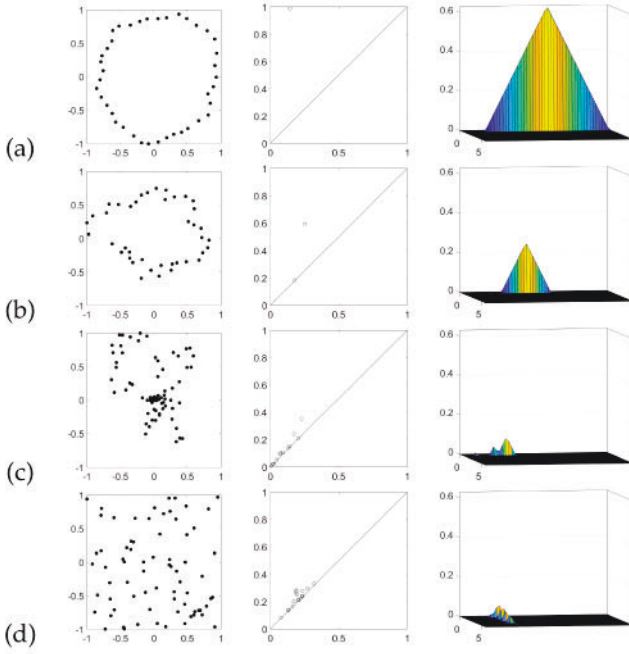


Fig. 9. Point clouds (left), persistence diagrams (middle) and landscapes (right) from each of the four Gleason grading scales: (a)–(d) benign to most severe.

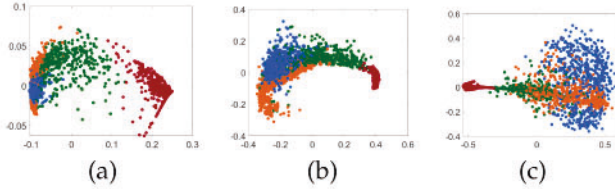


Fig. 10. Multidimensional scaling plots for the training dataset: (a) \mathbb{L}^2 distance, (b) phase distance, and (c) amplitude distance. Each point is colored according to class membership with benign = red, grade 2 = green, grade 3 = orange, and grade 4 = blue.

-dimensional landscapes, respectively. As the prognostic grade worsens, the cycle in the point clouds, corresponding to the carcinoma outline, becomes less pronounced. This is accompanied by significant changes in geometry and scale. Thus, the benign class is characterized by a persistence landscape with a large maximum in the first component function, and negligible maxima in subsequent component functions. On the other hand, the most severe class is characterized by a landscape with multiple small maxima along several component functions. Clearly, the landscapes contain significant amplitude and phase variation, and our aim is to classify the point clouds (landscapes) into the four Gleason grades based on these two components. The data consists of 2400 point clouds with 600 in each of the four Gleason grades.

In particular, using this data, we study k -nearest neighbor (KNN) classification accuracy using the amplitude and phase features of the landscapes. We also compare to one of the approaches taken in [2], which applied KNN classification to the landscapes without alignment. KNN classification is a distance-based approach, and we compare performance based on the following three distances:

(i) \mathbb{L}^2 distance: $d_{\mathbb{L}^2}(\Lambda_1, \Lambda_2) = \|\Lambda_1 - \Lambda_2\|_2$; (ii) amplitude distance: $d_a(\Lambda_1, \Lambda_2) = \min_{\gamma \in \Gamma} \|q_1 - (q_2, \gamma)\|_2$; (iii) phase distance: $d_p(\Lambda_1, \Lambda_2) = \arccos(\int_0^1 \sqrt{\gamma^*(s)} ds)$, where $\gamma^* = \operatorname{argmin}_{\gamma \in \Gamma} \|q_1 - (q_2, \gamma)\|_2$. The KNN classification procedure is implemented as follows. For a test observation (unknown class), we first compute its distance from each observation in the training data (known class). Then, we find the k nearest neighbors in the training set to the test observation, and predict its class as the one that is most frequent among the k nearest training neighbors. In case of a tie, we use the class of the nearest training neighbor. The classification accuracy is then computed as the percentage of correctly predicted classes in the test set. Using the same training and testing split as [2], data are split into 83% training (2000 landscapes, 500 in each of the four severity classes) and 17% testing (400 landscapes, 100 in each of the four severity classes).

We begin by displaying the 2D multidimensional scaling (MDS) plots, computed using the three different distances, for the training dataset. In short, MDS uses pairwise distances to compute lower-dimensional Euclidean coordinates of the data such that interpoint Euclidean distances based on these coordinates are as similar as possible to the original distances. It is evident that there is good separation between the benign (red) and Gleason grade 2 (green) classes for each of the three distances. The \mathbb{L}^2 and phase distances also provide good separation between the grade 2, grade 3 (orange) and grade 4 (blue) classes, with phase appearing more discriminative between the grade 3 and grade 4 classes. However, the amplitude distance is ineffective at separating the grade 2, grade 3 and grade 4 classes. This result is not surprising. The amplitude (shape) of persistence landscapes is effective in capturing whether and how many cycles exist in the point clouds. Thus, while it is very effective in discriminating between the benign and severe classes, it does not provide effective finer classification into the four Gleason grades. On the other hand, the signal captured in the phase component is related to the size and geometry of the homological features. Visually inspecting the four point clouds in Fig. 9, it is clear that these are the most discriminative features in the data. Finally, the \mathbb{L}^2 distance uses both amplitude and phase information of the landscapes without explicit control of the contribution of each component. We expect that these observations will lead to very good KNN classification rate based on the phase distance. On the other hand, the amplitude distance will only be effective at classifying benign versus severe classes.

A key question that has not yet been addressed is the choice of the number of nearest neighbors k . While we could fix this number *a priori* to some small number of neighbors, say 1 or 3, this approach will not result in optimal classification performance. Instead, we will learn an optimal k , for each of the three distances, based on training data and then apply the KNN classifier with the optimal k to testing data. We allowed values of $k = 1, \dots, 20$. Based on KNN classification applied to the training data, we determined the optimal values of k to be 11 for the \mathbb{L}^2 distance, 19 for the amplitude distance, and 9 for the phase distance, using leave-one-out cross-validation.

We report the overall classification rate on the testing data in Table I(a). The phase-based KNN classifier provides highest classification accuracy with the \mathbb{L}^2 distance-based approach in

TABLE I

(A) TEST CLASSIFICATION ACCURACY, BASED ON THE KNN CLASSIFIER, USING THE \mathbb{L}^2 ($k = 11$), PHASE ($k = 9$) AND AMPLITUDE ($k = 19$) DISTANCES. (B) CORRESPONDING CONFUSION MATRICES

(a)		L²				Phase				Amplitude			
		<i>Overall Accuracy</i>											
		91%				92.75%				70%			
(b)		<i>Confusion Tables</i>											
		True				True				True			
		1	2	3	4	1	2	3	4	1	2	3	4
Predicted	1	100	2	0	0	100	3	0	0	100	4	0	0
	2	0	87	9	5	0	89	5	10	0	91	23	2
	3	0	8	86	4	0	4	94	2	0	5	77	86
	4	0	3	5	91	0	4	1	88	0	0	0	12

close second; the phase distance results in correct classification of 7 more cases than the \mathbb{L}^2 distance. Finally, the amplitude distance provides the lowest classification accuracy. We further report confusion matrices in Table I(b). Overall, the phase distance is more effective than the other two distances in discriminating between neighboring classes, e.g., class 2 versus class 3. The amplitude distance is only effective in discriminating between the benign and severe classes. These observations are very similar to those reported earlier based on the MDS plots.

VI. DISCUSSION

In the analysis of the brain artery tree data, alignment of persistence landscapes adds substantially to the findings of [1], [30] by uncovering that the apparent differences in the unaligned mean landscapes of the two sex groups can be partially attributed to a difference in scale and sampling variability, and confirms this finding by comparing the distributions of the total artery lengths of males and females. In the analysis of the Gleason dataset [2], we show that the amplitude of landscapes (topological information) is most effective in discriminating between benign and severe cancer, while the phase (geometry and scale) is very effective in discriminating between all four grades. In particular, phase-based classification outperforms the standard \mathbb{L}^2 -based approach. In both settings, we demonstrate the need to consider amplitude and phase variability in persistence landscapes to address the scientific questions of interest.

The novel approach presented in this paper can be viewed as a first step toward understanding how geometry of the manifold on which point clouds are sampled influences TDA. To see this, suppose $e : M \hookrightarrow \mathbb{R}^D$ is an equivariant embedding of a d -dimensional manifold M into \mathbb{R}^D , $D \geq d$. Then, a diffeomorphism $\phi : \mathbb{R}^D \rightarrow \mathbb{R}^D$ acts on the embedding as $\phi \circ e(M)$. The map ϕ does not change the topology of M , but constructing simplicial filtrations for point clouds under the embedding in \mathbb{R}^D using balls will transform according to ϕ since the metric is accordingly transformed; that is, for a fixed $x \in e(M)$, $\{y \in e(M) : \|x - y\|_{\mathbb{R}^D} < t\}$ will transform to $\phi(x) \in \phi \circ e(M)$, $\{\phi(y) \in \phi \circ e(M) : \|\phi(x) - \phi(y)\|_{\mathbb{R}^D} < t\}$. In the special case where ϕ corresponds to a (constant) scale change, the radius t changes nonlinearly as $t \mapsto \gamma(t)$, for a reparameterization γ , since t is forced to lie within $[0, 1]$. This phenomenon also relates to when points are sampled with variability on M , since by judiciously changing the metric depending on the locations of points, balls of different (or differently changing) radii can be used to construct

the simplicial filtration, not dissimilar to the multiscale approach considered by [20]. Much remains to be done in this direction.

The limitations of this work inspire directions for future work. First, using scaled persistence diagrams by rescaling to $[0, 1]^2$ is also a source of topological noise, but is entirely driven by practical considerations. In principle, we could instead consider the group of diffeomorphisms of $[0, \infty)$ to align the persistence landscapes, although there would be very little phase variation for parameter values exceeding the maximum across different point clouds. A compromise would be to consider the subgroup of diffeomorphisms of $[0, \infty)$ based on scaling and translating diffeomorphisms of $[0, 1]$, considered in [32], to perform alignment that better reflects phase variability in the landscapes. Second, notwithstanding the promising results for the noisy simulations presented in Appendix A in the supplement, robustness of the alignment-based approach to measurement error will strongly depend on the geometry of the manifold M , sampling density and magnitude of noise in observed point clouds on \mathbb{R}^d , especially if data have been sampled from a distribution with support only on M . One possible approach would constitute of first estimating M (and its dimension) using a manifold fitting method, and using this information to construct tailored simplicial filtrations; however, additional noise induced by the fitting procedure would have to be accounted for in downstream tasks. Another option for the large noise setting is to use explicit statistical models to align persistence landscapes that account for all sources of uncertainty. For example, Bayesian models based on shape constraints to infer the pattern and number of extrema in landscapes may be profitably used [33].

While the focus of this paper is on persistence landscapes, we expect our approach to be fruitful with silhouettes [9], since similar triangular functions are used in their definition. However, feasibility of the alignment method for other functional summaries mentioned in Section I is not clear, and is worthy of further investigation. Finally, for denoising a persistence diagram directly without using landscapes, it is possible to consider generalizations of the one-dimensional transforms γ of points on diagrams to the group of diffeomorphisms of \mathbb{R}^2 , along the lines of what is done in the Large Deformation Diffeomorphic Metric Mapping (LDDMM) framework [34].

ACKNOWLEDGMENT

The authors would like to thank the two anonymous reviewers for providing constructive comments that have improved this manuscript.

The MR brain images from healthy volunteers used in this paper were collected and made available by the CASILab at the University of North Carolina at Chapel Hill and were distributed by the MIDAS Data Server at Kitware, Inc. The authors would like to thank Jessi Cisewski Kehe for sharing the Gleason dataset.

REFERENCES

- [1] P. Bendich, J. S. Marron, E. Miller, A. Pieloch, and S. Skwerer, "Persistent homology analysis of brain artery trees," *Ann. Appl. Statist.*, vol. 10, no. 1, pp. 198–218, 2016.
- [2] E. Berry, Y.-C. Chen, J. Cisewski, and B. Fasy, "Functional summaries of persistence diagrams," *J. Appl. Comput. Topol.*, vol. 4, pp. 211–262, 2020.

- [3] Y. Mileyko, S. Mukherjee, and J. Harer, "Probability measures on the space of persistence diagrams," *Inverse Problems*, vol. 27, no. 12, 2011, Art. no. 124007.
- [4] K. Turner, Y. Mileyko, S. Mukherjee, and J. Harer, "Fréchet means for distributions of persistence diagrams," *Discrete Comput. Geometry*, vol. 52, no. 1, pp. 44–70, 2014.
- [5] L. Wasserman, "Topological data analysis," *Annu. Rev. Statist. Appl.*, vol. 5, no. 1, pp. 501–532, 2018.
- [6] B. T. Fasy, F. Lecci, A. Rinaldo, L. Wasserman, S. Balakrishnan, and A. Singh, "Confidence sets for persistence diagrams," *Ann. Statist.*, vol. 42, no. 6, pp. 2301–2339, 2014.
- [7] J. O. Ramsay and B. W. Silverman, *Functional Data Analysis*. Berlin, Germany: Springer, 2005.
- [8] P. Bubenik, "Statistical topological data analysis using persistence landscapes," *J. Mach. Learn. Res.*, vol. 16, no. 3, pp. 77–102, 2015.
- [9] F. Chazal, B. T. Fasy, F. Lecci, A. Rinaldo, and L. Wasserman, "Stochastic convergence of persistence landscapes and silhouettes," in *Proc. ACM Symp. Comput. Geometry*, 2014, pp. 474–483.
- [10] R. Anirudh, V. Venkataraman, K. Ramamurthy, and P. Turaga, "A Riemannian framework for statistical analysis of topological persistence diagrams," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2016, pp. 1023–1031.
- [11] V. Robins and K. Turner, "Principal component analysis of persistent homology rank functions with case studies of spatial point patterns, sphere packing and colloids," *Physica D: Nonlinear Phenomena*, vol. 334, pp. 99–117, 2016.
- [12] N. Atienza, R. Gonzalez-Díaz, and M. Soriano-Trigueros, "On the stability of persistent entropy and new summary functions for topological data analysis," *Pattern Recognit.*, vol. 107, 2020, Art. no. 107509.
- [13] Y.-C. Chen, D. Wang, A. Rinaldo, and L. Wasserman, "Statistical analysis of persistence intensity functions," 2015, *arXiv:1510.02502v1*.
- [14] H. Adams et al., "Persistence images: A stable vector representation of persistent homology," *J. Mach. Learn. Res.*, vol. 18, no. 8, pp. 1–35, 2017.
- [15] F. Hensel, M. Moor, and B. Rieck, "A survey of topological machine learning methods," *Front. Artif. Intell.*, vol. 4, 2021, Art. no. 681108.
- [16] D. G. Kendall, "Shape manifolds, Procrustean metrics, and complex projective spaces," *Bull. London Math. Soc.*, vol. 16, pp. 81–121, 1984.
- [17] S. Kurtek, A. Srivastava, E. Klassen, and Z. Ding, "Statistical modeling of curves using shapes and related features," *J. Amer. Stat. Assoc.*, vol. 107, no. 499, pp. 1152–1165, 2012.
- [18] J. S. Marron, J. O. Ramsay, L. M. Sangalli, and A. Srivastava, "Functional data analysis of amplitude and phase variation," *Stat. Sci.*, vol. 30, no. 4, pp. 468–484, 2015.
- [19] A. Srivastava and E. Klassen, *Functional and Shape Data Analysis*. Berlin, Germany: Springer, 2016.
- [20] H. R. Yoon and R. Ghrist, "Persistence by parts: Multiscale feature detection via distributed persistent homology," 2020, *arXiv:2001.01623*.
- [21] H. Edelsbrunner, D. Letscher, and A. Zomorodian, "Topological persistence and simplification," *Discrete Comput. Geometry*, vol. 28, no. 4, pp. 511–533, 2002.
- [22] V. Silva and R. Ghrist, "Coverage in sensor networks via persistent homology," *Algebr. Geometric Topol.*, vol. 7, no. 1, pp. 339–358, 2007.
- [23] M. Bruveris, "Optimal reparametrizations in the square root velocity framework," *SIAM J. Math. Anal.*, vol. 48, no. 6, pp. 4335–4354, 2015.
- [24] E. M. Stein and R. Shakarchi, *Real Analysis: Measure Theory, Integration, and Hilbert Spaces*. Princeton, NJ, USA: Princeton Univ. Press, 2005.
- [25] S. Lahiri, E. Klassen, and D. Robinson, "Precise matching of PL curves in \mathbb{R}^n in the square root velocity framework," *Geometry Imag. Comput.*, vol. 2, no. 3, pp. 133–186, 2015.
- [26] A. Srivastava, W. Wu, S. Kurtek, E. Klassen, and J. S. Marron, "Registration of functional data using Fisher-Rao metric," 2011, *arXiv:1103.3817*.
- [27] J. D. Tucker, W. Wu, and A. Srivastava, "Generative models for functional data using phase and amplitude separation," *Comput. Statist. Data Anal.*, vol. 61, pp. 50–66, 2013.
- [28] B. T. Fasy, J. Kim, F. Lecci, and C. Maria, "Introduction to the R package TDA," 2015, *arXiv:1411.1830*.
- [29] E. Bullitt et al., "Vessel tortuosity and brain tumor malignancy," *Academic Radiol.*, vol. 12, pp. 1232–40, 2005.
- [30] J. S. Marron and I. L. Dryden, *Object Oriented Data Analysis*. Boca Raton, FL, USA: CRC Press, 2021.
- [31] E. Bullitt, K. Muller, I. Jung, W. Lin, and S. Aylward, "Analyzing attributes of vessel populations," *Med. Image Anal.*, vol. 9, pp. 39–49, 2005.
- [32] D. Bryner and A. Srivastava, "Shape analysis of functional data with elastic partial matching," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 9589–9602, Dec. 2022.
- [33] J. Matuk, K. Bharath, O. Chkrebti, and S. Kurtek, "Bayesian framework for simultaneous registration and estimation of noisy, sparse, and fragmented functional data," *J. Amer. Stat. Assoc.*, vol. 117, no. 540, pp. 1964–1980, 2022.
- [34] U. Grenander and M. Miller, *Pattern Theory: From Representation to Inference*. London, U.K.: Oxford Univ. Press, 2007.



James Matuk received the PhD degree in statistics from the Ohio State University in 2021. He is a data scientist with the Epidemiology Data Center, the University of Pittsburgh. His research interests are Bayesian statistics, functional data analysis, and latent variable models.



Sebastian Kurtek (Senior Member, IEEE) received the MS and PhD degrees in biostatistics from Florida State University, in 2009 and 2012, respectively. He is a professor of statistics with the Ohio State University. His research interests include functional data analysis, statistical shape analysis and statistics on manifolds. His research has been supported by the National Science Foundation and the National Institutes of Health.



Karthik Bharath received the PhD degree in statistics from the University of Connecticut in 2012. He is a professor of statistics with the University of Nottingham. His research interests are in geometric statistics, functional and shape data analysis, stochastic processes and quantum information. His research has been supported by the National Science Foundation, National Institutes of Health, and the Engineering and Physical Sciences Research Council, U.K.