Topological Data Analysis through alignment of Persistence Landscapes

James Matuk
Department of Statistics
The Ohio State University
Columbus, OH, USA
matuk.3@osu.edu

Sebastian Kurtek
Department of Statistics
The Ohio State University
Columbus, OH, USA
kurtek.1@stat.osu.edu

Karthik Bharath
School of Mathematical Sciences
University of Nottingham
Nottingham, UK
Karthik.Bharath@nottingham.ac.uk

Abstract

Persistence landscapes are functional summaries of persistence diagrams designed to enable analysis of the diagrams using tools from functional data analysis. They comprise a collection of scalar functions such that birth and death times of topological features in persistence diagrams map to extrema of functions and intervals where they are non-zero. As a consequence, topological information is encoded in both amplitude and phase components of persistence landscapes. Through functional data analysis of persistence landscapes under an elastic Riemannian metric, we show how meaningful statistical summaries of persistence landscapes (e.g., mean, dominant directions of variation) can be obtained by decoupling topological signal present in amplitude and phase variations. The estimated phase functions are tied to the resolution parameter that determines the filtration of simplicial complexes used to construct persistence diagrams. For a dataset obtained under scale and sampling variabilities, the phase function prescribes an optimal rate of increase of the resolution parameter for enhancing the topological signal in a persistence diagram. We demonstrate benefits of alignment through several simulation examples and a real data example concerning structure of brain artery trees represented as 3D point clouds.

1 Introduction

Persistent homology is a prominent tool within Topological Data Analysis (TDA) that provides a multi-resolution view of the topological features of data. As a resolution parameter changes, so do the features of the data, and these changes are recorded in persistence diagrams or barcodes. 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 (Mileyko et al., 2011; Turner et al., 2014; Wasserman, 2018) and confidence regions (Pasy et al., 2014). Carrying out TDA directly on the space of persistence diagrams is difficult since they are multisets of planar points. This motivates using functional representations of diagrams, or functional summaries, that are more amenable for statistical analysis (Berry et al., 2020) using tools from functional data analysis (Ramsay and Silverman, 2005). In this paper, we consider the persistence landscape (Bubenik, 2015), but other functional summaries can also be used (silhouettes (Chazal et al., 2014); density estimates (Anirudh et al., 2016); rank functions (Robins and Turner, 2016); persistence entropy functions (Atienza et al., 2020); persistence intensity functions and images (Chen et al., 2015; Adams et al., 2017)).

In general, there are two main sources of variation in a functional dataset: amplitude, which captures $y$-axis variation, and phase, which tracks variation in the relative timing of amplitude features, e.g., extrema.
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) PCA are well-documented within the statistics literature (e.g., Marron et al. 2015; Srivastava and Klassen 2016). Accordingly, there is a need to understand how topo-geometric information in persistence diagrams manifests in amplitude and phase components of the observed data. As a first step, in this paper we examine topological noise due to scaling and sampling variabilities.

Given a dataset $X$, the persistence landscape $\Lambda_X = \{\lambda_i : I \rightarrow \mathbb{R}^+; \; i = 1, \ldots, K\}$ corresponding to a degree-$p$ persistence diagram consists of $K$ triangular-like functions $\lambda_i$, defined on an interval $I$, that start and end at zero. We thus view $\Lambda_X$ as a parameterized closed curve in $\mathbb{R}^K$, and obtain curves $\Lambda_{X,i}, \; i = 1, \ldots, n$ as landscapes from $n$ datasets. Our main contributions are as follows.

1. 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 $\lambda_i$, as captured through a reparameterization $s \mapsto \gamma(s)$ of the curve $\Lambda_X(s)$. Specifically, we show how $\gamma$ is related to noise in persistence diagrams induced by (i) arbitrary (global) scaling of the data (Figure 2) and (ii) sampling variability of data (Figure 3).

2. We show how alignment of landscapes $\{\Lambda_{X,i}\}$ by determining optimal reparameterizations $\{\gamma_i\}$ leads to a mean (average) landscape that better preserves the structures of the sample of landscapes. A key consequence is the denoising of points in the corresponding persistence diagrams by transforming the points 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 the persistence diagrams (Figure 3).

3. We demonstrate the benefits of carrying out this program on simulated examples and a real dataset consisting of brain artery trees, also studied in Bendich et al. (2016) (Figure 5).

We summarize our approach through a simple example in Figure 1 with three point clouds $X_1, X_2, X_3$ consisting of brain artery trees, also studied in Bendich et al. (2016) (Figure 5).
(b, d) in the persistence diagram that has larger b and d coordinates when compared to the unscaled case, and thus shifts the location of the peak (of a component) of the persistence landscape to the right. This represents topological noise since the underlying topological signal is invariant to scaling. Matching peaks across persistence diagrams is achieved through their alignment, and the resulting phase functions $\gamma$ can be used to transform the (b, d) pairs accordingly. The second factor is how densely or sparsely are the points $x_i$ sampled. This is related to the scaling issue, since sparse sampling relates to requiring balls with larger radius $t$ to discover topological features, and hence result in a larger $b$ coordinate. In both cases, alignment removes topological noise.

To our knowledge, this is the first work in TDA literature to establish a concrete link between misalignment of persistence landscapes to noise in persistence diagrams. The only other related work we are aware of, although unrelated to use of persistence landscapes, is by Yoon and Ghrist (2020) who use a multiscale approach by allowing each ball $B_{x_i}(t_i)$ to have a possibly different radius $t_i$. In a certain sense, our approach in determining 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 $x_i$.

2 Persistence diagrams, landscapes and scaling

2.1 Persistent Homology through diagrams and landscapes

Persistent homology is a tool that tracks homological features, such as connected components (degree-0), loops (degree-1), voids (degree-2), etc., of data at different resolutions (Wasserman, 2018). As an example of persistent homology, consider a point cloud $X = \{x_1, \ldots, x_N\}$, with each $x_i \in \mathbb{R}^d$. A ball of radius $t$ can be drawn around each of the points, $B_{x_i}(t)$, and the union of the balls, $\cup_{i=1}^N B_{x_i}(t)$, can be used to compute the homology. In practice, homology is computed from a simplicial complex, where the points in the point cloud are viewed as nodes in a graph and edges are drawn between them based on how the balls intersect. Two common simplicial complexes based on the balls are the Čech complex and Vietoris-Rips complex. The Čech complex consists of $k$-simplices whose nodes have $k + 1$ many balls with a non-empty intersection. The Vietoris-Rips complex 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 complexes is a snapshot of the features of the point cloud. However, considering all radii, $t > 0$, provides a multi-resolution view of the features of the data where features are born and die at different values of $t$. Persistent homology tracks the features with 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 data. For more general treatments of persistent homology and persistence diagrams, we refer the reader to Edelsbrunner et al. (2002) and Zomorodian and Carlsson (2005).

Persistence landscapes are functional representations of persistent homology computed from persistence diagrams (Bubenik, 2015). For data $X = \{x_1, \ldots, x_N\}$, let $D_p(X)$ denote its degree-$p$ persistence diagram consisting of $m$ birth-death pairs $\{(b_j, d_j)\}_{j=1}^m$. The basic unit of persistence landscapes are triangular functions based on points in a persistence diagram,

$$\ell_j^p(t) = (t - b_j)\mathbb{I}_{\{b_j \leq t \leq d_j\}} + (d_j - t)\mathbb{I}_{\{\frac{1}{2}(b_j + d_j) \leq t \leq d_j\}}.$$  \hspace{1cm} (1)

For $k \in \mathbb{N}$, the $k$th landscape function is defined as $\lambda_k^p(t) = k^{\text{th}} \max_{j=1,\ldots,m} \ell_j^p(t)$, which is the $k$th maximum of the triangular functions with $\lambda_k(t) = 0$ for all $k > m$ by definition. Each function $\lambda_k^p$ 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 the data $X$ is defined as the collection of landscape functions $\Lambda^p_X(t) = \{\lambda_k^p(t)\}_{k=1}^K$, which we represent as a $K$-dimensional closed Euclidean curve

$$t \mapsto \Lambda^p_X(t) := (\lambda_1^p(t), \ldots, \lambda_K^p(t)) \in \mathbb{R}^K.$$ \hspace{1cm} (2)

In this work, we will consider order $p$ persistence landscapes of samples of datasets $X_1, \ldots, X_n$, which we will denote simply by $\Lambda_1(t), \ldots, \Lambda_n(t)$, where context will clarify the degree $p$. 

3
2.2 Effect of scaling and sampling variabilities

Persistence diagrams are constructed with the Rips or the Čech simplicial filtrations based on distances between data points, which are sensitive to the scale and sampling of the data. To elaborate on intuition offered in Section 1, Figure 2 considers two datasets consisting of 10 equidistant points along circles with radii 1 (blue) and 0.5 (red), respectively. We consider degree $p = 1$ persistent homology (loops) with $K = 1$-dimensional landscapes when topological noise is entirely due to scale effects.

![Persistence Diagrams](image)

In Figure 2(a)-(d), balls of different radii $t$ are drawn around the points in the point clouds. In (a), when $t = 0.1545$, a loop forms for the red point cloud, while there is no loop present for the blue point cloud. In (b), when $t = 0.3090$, a loop forms for the blue point cloud, and the loop persists for the red point cloud. In (c), when $t = 0.5$, 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 $(0.1545, 0.5)$ and $(0.3090, 1)$ 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 this example, we see that when topological noise is due to scaling, alignment of peaks of the persistence landscapes will move points in a persistence diagram towards each other, and thus amplify the topological signal. In other words, such topological noise manifests entirely through phase variation in the landscapes. A similar phenomenon occurs when points in a point cloud have been sampled non-uniformly with sampling variability (see Section 4.2, Figure 4).

3 Elastic functional data analysis of persistence landscapes

More generally, in contrast to the above example concerning scale, noise in a persistence diagram can arise from several other factors, including sampling variability and measurement error, and these are confounded in both amplitude and phase variations in the corresponding set of landscapes. However, transforming points in
a persistence diagram using the reparameterization functions obtained by aligning or registering the landscapes will continue to mitigate noise in the persistence diagram.

Consider a sample of landscapes $\Lambda_1, \ldots, \Lambda_n$ from $n$ datasets (for some fixed degree $p$ homology). Without loss of generality, we can assume that they share a common domain $[0,1]$; in practice, there always exists an $0 < s < \infty$ such that $\Lambda_i(t) = 0, \forall t > s, i = 1, \ldots, n$; we can thus choose $[0,s]$ as a common domain and rescale by $1/s$. Under this representation, each landscape $\Lambda_i$ is thus a $K$-dimensional parameterized closed curve $[0,1] \ni t \mapsto \Lambda_i(t) \in \mathbb{R}^K$ with $\Lambda_i(0) = \Lambda_i(1)$. Alignment of $\{\Lambda_i\}$ amounts to establishing correspondence between $(K$-dimensional) points in the images $t \mapsto \Lambda_i(t)$. This is achieved by estimating optimal orientation-preserving diffeomorphisms (reparameterizations) $\{\gamma_i : [0,1] \to [0,1] : \gamma_i > 0, \gamma_i(0) = 0, \gamma_i(1) = 1\}$ ($\dot{\gamma}$ is the derivative of $\gamma$) such that the collection $\{\Lambda_i(\gamma_i)\}$ are ‘best’ aligned, where optimality is defined with respect to a metric-based matching functional on the landscapes. For each $i$, $\gamma_i$ represents common phase variation in the component functions $(\lambda_{1i}, \ldots, \lambda_{Ki})$ of $\Lambda_i$.

From the definition of a landscape, the parameter $t$ corresponds to the resolution parameter in a simplicial filtration used to compute a persistence diagram; that is, we can map a value of $t$ precisely to a birth-death $(b,d)$ point on the diagram. As a consequence, rescaling the domain of landscapes $\{\Lambda_i\}$ by $s$ corresponds to considering scaled persistence diagrams $\{\{(b_{ij}/s,d_{ij}/s)\}_{i=1,j=1}^{m}\}$, so that the birth-death pairs are now in $[0,1]^2$. Then, this correspondence implies that, for each $i = 1, \ldots, n$, the (group) action $(\Lambda_i, \gamma_i) \mapsto \Lambda_i(\gamma_i)$ induces a transformation $\{(b_{ij},d_{ij})\} \to \{(\gamma_i^{-1}(b_{ij}), \gamma_i^{-1}(d_{ij}))\}$. Since the aligned $\{\Lambda_i(\gamma_i)\}$ are such that the extrema (mainly peaks) of the component functions $(\lambda_{1i}(\gamma_i), \ldots, \lambda_{Ki}(\gamma_i))$ line up, the points $(\gamma_i^{-1}(b_{ij}), \gamma_i^{-1}(d_{ij}))$ will tend to cluster, the number of which will depend on the topology of the underlying manifolds (see Figures 3 and 4 below, and Figures 8 and 10 in the Supplementary Material (Appendix B.2)). Another way to summarize the above discussion is as follows: if Čech or Rips filtration for dataset $X_i$ is constructed with balls of radius $\gamma_i(t)$, the corresponding persistence diagram will be ‘denoised’.

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 the algorithm based on the highly successful Elastic Functional Data Analysis (EFDA) framework based on a Riemannian-geometric approach, described in detail in the book by Srivastava and Klassen (2016). We refer the reader to the book and Supplementary Material (Appendix A) for more details, but provide a terse summary below.

The framework is characterized by the square-root velocity function (SRVF) representation/transform

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

where $\dot{\Lambda}(t)$ is the component-wise derivative and $|\cdot|$ is the Euclidean norm on $\mathbb{R}^K$. The map is invertible with $\Lambda(t) = \int_0^t q(u)|q(u)|du$. The motivation for the SRVF representation comes from intractability of computing the distance between, say, $\Lambda_1$ and $\Lambda_2$ with respect to the complicated elastic metric, which possesses the desirable property of being invariant to simultaneous reparameterizations $\Lambda_1(\gamma)$ and $\Lambda_2(\gamma)$; this property is crucial for aligning curves based on reparameterizations. The elastic distance under the SRVF representation, however, reduces to the usual $L^2$ distance $\|q_1 - q_2\|_2 = [\int_0^1 |q_1(t) - q_2(t)|^2 dt]^{1/2}$, where $q_1, q_2$ are the SRVFs of $\Lambda_1, \Lambda_2$ (Section 10.3 in Srivastava and Klassen (2016)).

A template is needed in order to obtain the $\{\gamma_i\}$. In an iterative algorithm, the Karcher mean, $\hat{\mu}$, is used as a template, and optimal $\{\gamma_i\}$ are simultaneously determined with respect to the elastic metric, by using the SRVF representation (Section 10.4.4 in Srivastava and Klassen (2016)). The mean persistence landscape $\hat{\mu}$ will thus be computed using the aligned landscapes $\{\Lambda_i(\gamma_i)\}$, and will be closer in shape to a persistence landscape than one computed without alignment (see, e.g., Figure 2(g)). Once a mean $\hat{\mu}$ on the aligned space is available, it is possible to carry out functional principal component analysis (FPCA) on the space of aligned landscapes by diagonalizing the (empirical) covariance operator to obtain dominant directions of amplitude variation (along geodesics from $\hat{\mu}$ with respect to the elastic metric). Then, projections of aligned landscapes $\{\Lambda_i(\gamma_i)\}$ along a small number of PC directions result in an efficient finite-dimensional representation of $\{\Lambda_i(\gamma_i)\}$.

4 Numerical examples

In this section we present:
1. Two simulation examples which demonstrate: (i) denoising of persistence diagrams, obtained under scale and sampling variabilities in the data, through alignment of landscapes; (ii) benefits of computing 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. Versions of both examples when data is sampled with noise, and additional examples on torii and spirals, are available in Appendices B.1 and B.2 respectively of the Supplementary Material.

2. A real data example on 3D brain artery trees which: (i) through alignment of persistence landscapes for male and female groups, uncovers how apparent difference in the unaligned mean landscapes of the two groups can be attributed to a difference in scale; (ii) confirms this finding by comparing the distributions of the total artery lengths of males and females. These substantially add to the findings in Bendich et al. (2016).

For the simulated examples, we use the \texttt{ripsDiag} function to compute persistence diagrams using the Vietoris-Rips simplicial complex for point clouds, and the \texttt{landscape} function to compute landscapes from persistence diagrams; both functions are part of the TDA R package \cite{FasyEtAl2015}. In the EFDA framework, registration, mean estimation and PCA for a sample of landscapes are implemented in MATLAB; the entire procedure requires approximately 1 minute on a standard laptop for each example. We have included additional simulated examples in the Supplementary Material (Appendix B.2). Data and MATLAB code to reproduce the simulated examples are also included in the Supplementary Material (the directory entitled ‘code and data’).

4.1 Simulation 1: Mean from aligned landscapes

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 \([N(1,0.3^2)]\); (iii) generating \( M \) points uniformly on a circle with radius \( r \). Figure 3 shows three (from 20) point clouds along with the corresponding degree \( p = 1, K = 1 \)-dimensional landscapes. Figure 3(a) shows all 20 landscapes \( \{\Lambda_i\}_{i=1}^{20} \).

The amplitude and phase variations in the landscapes are related to the 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) using reparameterizations \( \{\gamma_i\} \) (e); transforming points in the persistence diagram (d) using \( \{\gamma_i\} \) results in denoising (f) by collapsing all points to a single one, since the topology of each of the 20 point clouds is the same.

![Figure 3: Same topology with both scale and sampling variabilities](image)

Figure 3: Same topology with both scale and sampling variabilities: (a) Persistence landscapes \( \{\Lambda_i\}_{i=1}^{20} \) of 20 point clouds of the type in Figure 1 (b) Aligned persistence landscapes \( \{\Lambda_i(\gamma_i)\}_{i=1}^{20} \). (c) Mean landscape after (blue) and without (red) alignment. (d) Noisy (rescaled) persistence diagram \( \{(b_{ij},d_{ij})\}_{i=1}^{20} \) from 20 point clouds. (e) Estimated phase functions \( \{\gamma_i\}_{i=1}^{20} \). (f) Denoised/transformed persistence diagram \( \{\gamma_i^{-1}(b_{ij}),\gamma_i^{-1}(d_{ij})\}_{i=1}^{20} \).

4.2 Simulation 2: Principal component analysis 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 Simulation 1, but for the fact that 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 \([N(1,0.3^2)]\), 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) of Figure 4 show one point cloud each from (i) and (ii).
Figure 4: Different topology with both scale and sampling variabilities: (a) & (c) Two examples, from 20, of randomly generated point clouds from topologically different spaces (blue and red, respectively, in all relevant panels). (b) & (d) Corresponding degree $p = 1, K = 2$-dimensional persistence landscapes. (e)-(g) -1, 0, 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/transformed persistence diagrams. (j) Estimated reparameterizations.

We consider degree $p = 1$ homology and $K = 2$-dimensional persistence landscapes $\{\Lambda_i = (\lambda_{i1}, \lambda_{i2})\}_{i=1}^{20}$. For point clouds from (i), $\lambda_{i1}$ will have one peak and $\lambda_{i2} = 0$; on the other hand, for point clouds from (ii), $\lambda_{i1}$ will have two peaks and $\lambda_{i2}$ will have a single peak.

Top and bottom rows of panels (e)-(g) in Figure 4 show mean landscapes and PC directions following alignment and without alignment, respectively. Panel (h) clearly highlights the benefits of alignment of landscapes through better separation of the two settings, (i) and (ii), when projected along 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 panel (h).

Following results from Simulation 1, we expect to see two clear clusters in the denoised persistence diagram, corresponding to two distinct topological features, using reparameterizations $\{\gamma_i\}$, shown in panel (j), and this is indeed the case as seen in panel (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 point clouds. 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 feature (larger circle). It is very difficult to discern such topological information from the noisy diagram in panel (i).
4.3 Real data example: 3D brain artery trees

We now demonstrate the utility of the proposed approach on real data from 3D point clouds representing brain artery trees discussed in Bendich et al. (2016). For a description of the experiment and data generation see Bullitt et al. (2005). Information regarding human subjects in the experiment is available at http://insight-journal.org/midas/community/view/21. As per the terms of use listed on the above URL, we do not have explicit permission to share these data. We are thus unable to include the code and data used in this section in the Supplementary Material.

Bendich et al. (2016) computed persistence diagrams from trees which represented arteries in 98 healthy human subjects, and used persistence diagrams to compute the differences \( \{d_{i,j} - b_{i,j}\}_{i=1,j=1}^{98,100} \) for subjects based on the 100 largest differences sorted from largest to smallest. 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 topological noise (Fasy et al., 2014).

This dataset is apt to demonstrate our approach for two reasons, one a priori and the other 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, Bendich et al. (2016) subsampled 3000 points from each point cloud, thus creating large sampling variability; (ii) we uncover a significant scale effect between two groups of subjects. We note that this finding is purely exploratory and serves merely as a proof of concept for the proposed methodology.

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 landscapes from degree \( p = 1 \) homology and \( K = 100 \)-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, computed from degree \( p = 1 \) homology persistence diagrams, grouped by sex. One major finding of Bendich et al. (2016) is the existence of sex differences in their mean degree \( p = 1 \) feature vectors \( \{b_{ij} - d_{ij}\} \). For convenience, we denote the mean landscape for the male (female) group following alignment (within each group) by \( \hat{\mu}_m^a (\hat{\mu}_f^a) \), and pointwise mean computed without alignment for the males (females) by \( \hat{\mu}_m (\hat{\mu}_f) \).

The means \( \hat{\mu}_m^a \) and \( \hat{\mu}_f^a \) are shown in the top of Figure 5(a)-(b), respectively. The difference between the means is obtained by first aligning the group means to the common pooled mean and then taking their difference, where all operations are carried out under the SRVF representation; this difference is shown in the
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 (Bullitt et al., 2005). For each group, we estimated a cumulative distribution function (CDF) of total artery length using a kernel density estimate using the \texttt{ksdensity} function in MATLAB (default bandwidth). From the estimated CDFs shown in Figure 5(d), it appears that total artery length is stochastically ordered by sex, with females having stochastically longer brain artery trees.

Given this global scale disparity and sampling variability, we would expect \( \hat{\mu}_a^m \) to be shifted to the right relative to \( \hat{\mu}_a^f \). This behavior can be extracted from the phase difference between \( \hat{\mu}_a^m \) and \( \hat{\mu}_a^f \) to the mean computed following alignment of the pooled sample, as shown in Figure 5(e). The blue reparameterization shifts \( \hat{\mu}_a^f \) to the left while the red shifts \( \hat{\mu}_a^m \) to the right. The misalignment caused by differences in global scale and sampling variabilities 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 is due to global scale 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 thereby confounding of, amplitude and phase variation in landscapes.

### 5 Discussion

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 \( \phi \) 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 \( \phi \) since the metric is accordingly transformed; that is, for a fixed \( x \in e(M) \), \( \{ y \in e(M) : \| x - y \| \phi < t \} \) will transform to \( \phi(x) \in \phi \circ e(M) \), \( \{ \phi(y) \in \phi \circ e(M) : \| \phi(x) - \phi(y) \| \phi < t \} \). In the special case of the setting considered in the paper, where \( \phi \) corresponds to a (constant) scale change, the radius \( t \) changes nonlinearly as \( t \rightarrow \gamma(t) \), for a reparameterization \( \gamma \), 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 Yoon and Ghrist (2020). Much remains to be done in this direction.

The limitations of this work inspire directions for future work. As noted in Section 3, there are many factors that influence amplitude and phase variability in landscapes, including signal-to-noise ratio. Notwithstanding the promising results for the noisy setting on the circle (see Supplementary Material, Appendix B.1), robustness of the alignment-based approach to observation noise will strongly depend on the geometry of the manifold \( M \) and magnitude of noise in observed data 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 (Matůk et al., 2021).

While the focus of this paper is on persistence landscapes (largely due to space constraints), we expect our approach to be fruitful with silhouettes (Chazal et al., 2014), since similar triangular functions are used

---

1. We emphasize that this is purely an exploratory/descriptive finding without statistical significance. We do not attribute associative or causative meaning to the finding and therefore none should be construed.
in their definition. However, feasibility of the alignment method for other functional summaries mentioned in Section 1 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 $\gamma$ 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 \cite{Grenander2007}.

Acknowledgements

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. This research was partially funded by NSF DMS-1613054, NIH R37-CA214955 (to SK and KB), EPSRC EP/V048104/1, and NSF DMS-2015374 (to KB), and NSF CCF1740761, NSF DMS-2015226, and NSF CCF-1839252 (to SK).

References

Adams, H., T. Emerson, M. Kirby, R. Neville, C. Peterson, P. Shipman, S. Chepushtanova, E. Hanson, F. Motta, and L. Ziegelmeier (2017). Persistence images: A stable vector representation of persistent homology. *Journal of Machine Learning Research* 18(8), 1–35.

Anirudh, R., V. Venkataraman, K. Ramamurthy, and P. Turaga (2016). A Riemannian framework for statistical analysis of topological persistence diagrams. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1023–1031.

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

Bendich, P., J. S. Marron, E. Miller, A. Pieloch, and S. Skwerer (2016). Persistent homology analysis of brain artery trees. *The Annals of Applied Statistics* 10(1), 198 – 218.

Berry, E., Y.-C. Chen, J. Cisewski, and B. Fasy (2020). Functional summaries of persistence diagrams. *Journal of Applied and Computational Topology* 4, 211–262.

Bubenik, P. (2015). Statistical topological data analysis using persistence landscapes. *Journal of Machine Learning Research* 16(3), 77–102.

Bullitt, E., K. Muller, I. Jung, W. Lin, and S. Aylward (2005). Analyzing attributes of vessel populations. *Medical Image Analysis* 9, 39–49.

Bullitt, E., D. Zeng, G. Gerig, S. Aylward, S. Joshi, J. Smith, W. Lin, and M. Ewend (2005). Vessel tortuosity and brain tumor malignancy. *Academic Radiology* 12, 1232–40.

Chazal, F., B. T. Fasy, F. Lecci, A. Rinaldo, and L. Wasserman (2014). Stochastic convergence of persistence landscapes and silhouettes. In *ACM Symposium on Computational Geometry*, pp. 474–483. ACM.

Chen, Y.-C., D. Wang, A. Rinaldo, and L. Wasserman (2015). Statistical analysis of persistence intensity functions. *arXiv:1510.02502v1*.

Edelsbrunner, H., D. Letscher, and A. Zomorodian (2002). Topological persistence and simplification. *Discrete & Computational Geometry* 28(4), 511–533.

Fasy, B. T., J. Kim, F. Lecci, and C. Maria (2015). Introduction to the R package TDA. *arXiv:1411.1830*.

Fasy, B. T., F. Lecci, A. Rinaldo, L. Wasserman, S. Balakrishnan, and A. Singh (2014). Confidence sets for persistence diagrams. *The Annals of Statistics* 42(6), 2301 – 2339.
Grenander, U. and M. Miller (2007). *Pattern Theory: From Representation to Inference*. Oxford University Press.

Marron, J. S., J. O. Ramsay, L. M. Sangalli, and A. Srivastava (2015). Functional data analysis of amplitude and phase variation. *Statistical Science* 30(4), 468–484.

Matuk, J., K. Bharath, O. Chkrebtii, and S. Kurtek (2021). Bayesian framework for simultaneous registration and estimation of noisy, sparse, and fragmented functional data. *Journal of the American Statistical Association*, In Press.

Mileyko, Y., S. Mukherjee, and J. Harer (2011). Probability measures on the space of persistence diagrams. *Inverse Problems* 27(12), 124007.

Ramsay, J. O. and B. W. Silverman (2005). *Functional Data Analysis*. Springer.

Robins, V. and K. Turner (2016). Principal component analysis of persistent homology rank functions with case studies of spatial point patterns, sphere packing and colloids. *Physica D: Nonlinear Phenomena* 334, 99–117.

Srivastava, A. and E. Klassen (2016). *Functional and Shape Data Analysis*. Springer.

Srivastava, A., W. Wu, S. Kurtek, E. Klassen, and J. S. Marron (2011). Registration of functional data using Fisher-Rao metric. *arXiv* 1103.3817.

Tucker, J. D., W. Wu, and A. Srivastava (2013). Generative models for functional data using phase and amplitude separation. *Computational Statistics & Data Analysis* 61, 50–66.

Turner, K., Y. Mileyko, S. Mukherjee, and J. Harer (2014). Fréchet means for distributions of persistence diagrams. *Discrete & Computational Geometry* 52(1), 44–70.

Wasserman, L. (2018). Topological data analysis. *Annual Review of Statistics and Its Application* 5(1), 501–532.

Yoon, H. R. and R. Ghrist (2020). Persistence by parts: Multiscale feature detection via distributed persistent homology. *arXiv*: 2001.01623.

Zomorodian, A. and G. Carlsson (2005). Computing persistent homology. *Discrete and Computational Geometry* 33, 249–274.
Appendix

This appendix contains three sections for the paper entitled *Topological Data Analysis through alignment of Persistence Landscapes*. In Appendix A, we provide additional background material for statistical analysis of curves under the Elastic Functional Data Analysis paradigm. In Appendix B, we discuss additional simulated examples. In Appendix C, we report additional results for the brain artery tree data considered in Section 4.3 in the main article.

A Brief review of Elastic Functional Data Analysis

In this section, we provide several key definitions that enable alignment and statistical analysis of curves under the Elastic Functional Data Analysis paradigm, and refer to [Srivastava and Klassen (2016)] for more details.

Let \( \Lambda : [0,1] \rightarrow \mathbb{R}^K \in \mathcal{F} \), where \( \mathcal{F} \) is the space of all absolutely continuous curves in \( \mathbb{R}^K \), denote a \( K \)-dimensional curve, e.g., a persistence landscape. Further, define a mapping \( (Q : \mathcal{F} \rightarrow \mathcal{Q}) \) of a curve \( \Lambda \in \mathcal{F} \) to its corresponding SRVF as \( Q(\Lambda) = q = \frac{\Lambda}{\lambda} \), where \( \Lambda \) is the derivative of \( \Lambda \) and \( | \cdot | \) is the Euclidean norm in \( \mathbb{R}^K \). Our definition of amplitude and subsequent statistical analysis approach are analogous to the definitions presented in [Srivastava et al. (2011)] for univariate functions. The amplitude of a curve \( \Lambda \) is an equivalence class defined through its SRVF \( q \), as the quantity that minimizes the sum of squared amplitude distances:

\[
[q] := \{(q, \gamma) | \gamma \in \Gamma\},
\]

where \( (q, \gamma) = Q(\Lambda(\gamma)) = (q(\gamma))\sqrt{\gamma} \) (recall that \( \Gamma \) is the set of orientation-preserving diffeomorphisms or reparameterizations of \( [0,1] \)). Under this definition, two curves, \( \Lambda_1, \Lambda_2 \), have the same amplitude if their corresponding SRVFs are in the same equivalence class, i.e., there exists a \( \gamma \in \Gamma \) such that \( q_1 = (q_2, \gamma) \). The set of all equivalence classes forms a partition of \( \mathcal{Q} \) and is the quotient space \( \mathcal{Q}/\Gamma \). Hence, \( \mathcal{Q}/\Gamma \) defines the amplitude space.

The amplitude distance between two curves \( \Lambda_1, \Lambda_2 \in \mathcal{F} \) is defined as the distance between their corresponding SRVF equivalence classes \([q_1], [q_2] \in \mathcal{Q}/\Gamma\):

\[
d_{\text{amp}}(\Lambda_1, \Lambda_2) = d([q_1], [q_2]) = \min_{\gamma \in \Gamma} \|q_1 - (q_2, \gamma)\|_2,
\]

where \( q_1 \in [q_1] \) and \( q_2 \in [q_2] \). Here, the invariance of the \( L^2 \) metric, under the SRVF representation, to simultaneous reparameterization of curves is key to the definition of the distance. The argmin of the right hand side of Equation 5 gives the optimal reparameterization of \( \Lambda_2 \) to register or align it to \( \Lambda_1 \).

For a sample of curves \( \Lambda_1, \ldots, \Lambda_n \), with corresponding SRVFs \( q_1, \ldots, q_n \), the amplitude mean is defined as the quantity that minimizes the sum of squared amplitude distances:

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

An element of \( [\hat{\mu}_q] \) is found by iteratively aligning \( q_1, \ldots, q_n \) 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 the interested reader to [Srivastava et al. (2011)]. The corresponding mean curve \( \hat{\mu} \in \mathcal{F} \) is defined through the inverse SRVF mapping:

\[
\hat{\mu}(t) = Q^{-1}(\hat{\mu}_q)(t) = \int_0^t \hat{\mu}_q(s)|\hat{\mu}_q(s)|ds.
\]

The joint alignment of \( \Lambda_1, \ldots, \Lambda_n \) can be achieved via pairwise alignment of each \( \Lambda_i \), \( i = 1, \ldots, n \) to the mean \( \hat{\mu} \) via Equation 5. The resulting optimal reparameterizations, \( \gamma_i = \arg\min_{\gamma \in \Gamma} \|\hat{\mu}_q - (q_i, \gamma)\|_2, i = 1, \ldots, n \), can be used to study phase variability. Methods for statistical analysis of reparameterization functions are detailed in [Tucker et al. (2013)] and omitted here for brevity. The aligned curves \( \Lambda_i(\gamma_i), i = 1, \ldots, n \), or equivalently their SRVFs, \( (q_i, \gamma_i), i = 1, \ldots, n \), describe amplitude variability in the sample. Based on the aligned curves, a sample amplitude covariance function can be defined as

\[
C_q(t, u) := \frac{1}{n-1} \sum_{i=1}^{n} ((q_i, \gamma_i)(t) - \hat{\mu}_q(t))((q_i, \gamma_i)(u) - \hat{\mu}_q(u))^\top.
\]
Amplitude principal components analysis is carried out via eigendecomposition of \( \hat{C}_q(t, u) \),

\[
\hat{C}_q(t, u) = \sum_{b=1}^{\infty} \hat{\tau}_b \hat{\phi}_b(t) \hat{\phi}_b(u)^\top,
\]

where \( \hat{\phi}_b, \quad b \in \mathbb{N} \) are the primary directions of amplitude variability (amplitude principal components) and \( \hat{\tau}_b, \quad b \in \mathbb{N} \) 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:

\[
\beta_{i,b} := \int_{0}^{1} \langle (q_i, \gamma)(t) - \hat{\mu}_q(t), \hat{\phi}_b(t) \rangle dt \quad b = 1, \ldots, B, \quad i = 1, \ldots, n,
\]

where \( \langle \cdot, \cdot \rangle \) is the Euclidean inner product in \( \mathbb{R}^K \). The PC scores, \( \beta_i = \beta_{i,1}, \ldots, \beta_{i,B}, \quad i = 1, \ldots, n, \) serve as a low-dimensional Euclidean representation of the amplitude of curves. To visualize the primary directions of amplitude variability, we compute \( \mathcal{F} \ni Q^{-1}(\hat{\mu}_q + \nu \sqrt{\hat{\tau}_b} \hat{\phi}_b) \), which is a curve that is \( \nu \) standard deviations from \( \hat{\mu}_q \) in the direction of \( \hat{\phi}_b \) (see Figure 4 in the main article and Figure 7 in Appendix B.1 for examples).

## B Additional numerical examples

In this section, we (1) study the potential impacts of observation noise (away from the manifold of interest) through simulation examples in Appendix B.1 and (2) provide three additional mean estimation examples for point clouds sampled from different topological spaces in Appendix B.2.

### B.1 Simulations 1 and 2 with added noise

We repeat Simulation 1, as discussed in Section 4.1 in the main article, with additive pointwise noise, i.e., for each point cloud, we independently generate additive noise from a zero-mean bivariate Gaussian distribution with covariance \( r(0.1)^2 I_2 \), where \( r \) is the radius of the circle that underlies the “noiseless” point cloud. Two examples of point clouds and their corresponding degree \( p = 1 \) \( K = 1 \)-dimensional landscape functions are shown in panels (a)&(c) and (b)&(d) of Figure 6, respectively. The degree \( p = 1 \) \( K = 1 \)-dimensional landscape functions for all 20 point clouds are shown in panel (e). Most of the landscapes generated from the noisy point clouds are similar in shape to those displayed in Figure 3 in the main article. However, sometimes there is an extremely small peak, corresponding to noise in some of the landscape functions. This is most visually apparent in panel (f), where there is an extremely small peak at \( t \approx 0.25 \) prior to the major peak in each landscape. Due to the small magnitude of this noise-induced feature, the addition of additive noise appears to have little effect on landscape alignment and mean computation. The mean based on aligned landscapes, visualized in panel (g), is consistent with that of a circle. As highlighted in the main article, there is considerable variance reduction in the denoised/transformed persistence diagram in panel (j), via reparameterizations shown in panel (i). Importantly, points corresponding to the main feature of the data are collapsed to a single point, while points near the diagonal, corresponding to features created by the additive noise, remain near the diagonal. Such clarity is absent in the noisy persistence diagram in panel (h).

Next, we also repeat Simulation 2, discussed in Section 4.2 in the main article, with additive noise. For point clouds in the blue group (single circle), we independently generate additive noise from a zero-mean bivariate Gaussian distribution with covariance \( r(0.1)^2 I_2 \), where \( r \) is the radius of the circle that underlies the “noiseless” point cloud. We repeat this procedure for point clouds in the red group, but noise is generated such that the covariance depends on the radius of one of the two circles that it belongs to. A single example of a point cloud in the blue and red group are shown in Figure 7(a)&(c); the corresponding degree \( p = 1 \) \( K = 2 \)-dimensional landscapes are shown in panels (b)&(d). The shapes of the landscapes in each group are similar to those displayed in Figure 4 in the main article. However, here, we notice some differences in PCA carried out on aligned landscapes. The first direction of variability, viewed in the top row of panels (e)-(g) appears to be associated with scale variability in the point cloud data. On the other hand, as confirmed in the top row of panel (h), the second direction of variability appears to be associated with the homology of the point clouds, i.e., most of the red point clouds, generated from two connected circles, have a positive second
Figure 6: Same topology with scale and sampling variabilities as well as additive noise: (a)&(c): Two examples, from 20, of randomly generated point clouds on circles with randomly chosen radii, random sample sizes and additive noise. (b) & (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) (Rescaled) Noisy persistence diagram \( \{(b_{ij}, d_{ij}) \}_{i=1}^{20} \) from 20 point clouds. (i) Estimated reparameterizations \( \{\gamma_i \}_{i=1}^{20} \). (j) Denoised/transformed persistence diagram \( \{\gamma_i^{-1}(b_{ij}), \gamma_i^{-1}(d_{ij}) \}_{i=1}^{20} \).

PC score, while most of the blue point clouds, generated from a single circle, have a negative second PC score. Since noise can distort topological features, there does appear to be some overlap between the two classes based on the first two PC scores. This observation is in contrast to that in Section 4.2 in the main article, where the two classes are clearly separated based on the first PC score alone. The corresponding PC scores computed based on unaligned landscapes, shown in the bottom of panel (h), provide no such distinction between the two classes. A comparison of the denoised/transformed persistence diagram presented in Figure 7(k) to its noisy counterpart in panel (h) shows the benefits of our approach. While the clustering of features in panel (k) is not as clear as in Figure 4 in the main article, due to the additive noise, one can still extract useful homological information from the denoised/transformed persistence diagram. On the other hand, this is not possible based on the noisy diagram in panel (i).

B.2 Additional mean estimation examples

In Figure 8, we consider mean estimation based on degree \( p = 1 \) \( K = 2 \)-dimensional persistence landscapes computed from 20 point clouds that consist of uniformly sampled points along two circles with different radii. The point clouds in this example are generated in the same exact way as the data in the red group in Section 4.2 in the main article. Panels (a) and (c) show two examples of randomly generated point clouds with their corresponding degree \( p = 1 \) \( K = 2 \)-dimensional landscapes in panels (b) and (d). Panels (e)-(g) show the \( K = 1 \) (top) and \( K = 2 \) landscape functions, their alignment, and a comparison of the mean before (red) and with (blue) alignment, respectively. The proposed alignment procedure results in a mean landscape that better preserves the major features along both landscape components. On the other hand, the unaligned mean landscape destroys the prominent two peak structure in the first component and results. In panel (j), the points in the denoised/transformed persistence diagram, corresponding to the two loops in the point clouds, form two separate clusters making the presence of these features in the data much clearer; the noisy
Figure 7: Different topology with scale and sampling variabilities as well as additive noise: (a) & (c) Two examples, from 20, of randomly generated point clouds from topological different spaces (blue and red, respectively, in all relevant panels). (b) & (d) Corresponding degree $p = 1, K = 2$-dimensional persistence landscapes. (e)-(g) -1, 0, 1, standard deviation from the mean landscape in the first PC direction, and (h) projection of landscapes onto the first two PC direction: following alignment (top) and without alignment (bottom). Noisy (i) and denoised/transformed (k) persistence diagrams. (j) Estimated reparameterizations. Diagram shown in panel (h) does not provide such a distinction.

In Figure 9, 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 their corresponding degree $p = 0 K = 1$-dimensional landscapes in panels (b) and (d). The tighter spirals in panel (a) have points closer together, and the resulting landscape is smaller and shifted to the left as compared to the spirals in panel (c). When computing landscape functions, 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 with (blue) alignment. The mean based on aligned landscapes appears to have sharper features that are consistent with the observed landscapes. Based on the denoised/transformed persistence diagram in panel (j), in contrast to the noisy persistence diagram in panel (h), it is evident that warping completely accounts for the scale variability associated with the tightness of the spirals, i.e., all points collapsed to a single point in the denoised/transformed persistence diagram.

In Figure 10, we consider mean estimation based on degree $p = 1 K = 2$-dimensional persistence landscapes computed from 20 point clouds that consist of 1000 points sampled uniformly on a ringed torus. The major radius of each torus is sampled from a $|\mathcal{N}(2, .3^2)|$, while the minor radius is a proportion $\text{Beta}(10, 10)$ of the major radius. Two example point clouds are shown in panels (a) and (c). We preprocess the persistence diagrams used to compute the degree $p = 1 K = 2$-dimensional persistence landscapes by disregarding all
points in the persistence diagram except for the two points with longest persistence, as these points correspond to the two loops formed by the torii that underlie the point clouds. The landscapes corresponding to point clouds in (a) and (c) are shown in (b) and (d), respectively. Clearly, the estimated landscapes can vary widely depending on the relationship between the major and minor radii. Panels (e)-(g) show the first (top) and second (bottom) landscape functions, their alignment, and a comparison of the mean without (red) and after (blue) alignment. The first landscape function is automatically weighted higher during alignment due to the relatively large magnitude of the peak as compared to the second landscape function. In panel (j), the points in the denoised/transformed persistence diagram, using the reparameterizations shown in (i), concentrate to make the presence of the features more clear; the two detected features describe the two loops. In comparison, the noisy diagram shown in panel (h) does not provide a clear distinction.
Figure 9: Same topology with scale variability: (a)&(c) Two examples, from 20, of randomly generated point clouds. (b)&(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) (Rescaled) Noisy persistence diagram $\{(b_{ij},d_{ij})\}_{i=1}^{20}$ from 20 point clouds. (i) Estimated reparameterizations $\{\gamma_i\}_{i=1}^{20}$. (j) Denoised/transformed persistence diagram $\{(\gamma_i^{-1}(b_{ij}),\gamma_i^{-1}(d_{ij}))\}_{i=1}^{20}$.

C Persistence diagrams for brain artery trees

In this section, we display three randomly chosen examples of persistence diagrams, for females and males, that were used to generate persistence landscapes used in the real data brain artery tree example presented in Section 4.3 in the main article. We also show the estimated reparameterizations and transformed/denoised diagrams.

We estimated three different sets of reparameterizations during the alignment process: $\{\gamma_{i,F}\}_{i=1}^{47}$ and $\{\gamma_{i,M}\}_{i=1}^{49}$ were used to align subjects within sex to group specific means, and $\gamma_F$, $\gamma_M$ were used to align the group specific means to a mean estimated from pooled data. In order to compare reparameterizations for subjects in different groups, we further computed the alignment of each subject to the pooled mean as $\{\gamma_{i,F} = \gamma_{i,F} \circ \gamma_F\}_{i=1}^{47}$ for females and $\{\gamma_{i,M} = \gamma_{i,M} \circ \gamma_M\}_{i=1}^{49}$ for males. The corresponding results, for three randomly selected females and three randomly selected males, are shown in Figures 11 and 12 respectively; each row corresponds to a different subject. In panels (a)-(c) of each figure, we show the noisy persistence diagram, the estimated reparameterization and the denoised/transformed diagram, respectively. Regions where reparameterizations are above the identity (diagonal line $\gamma(t) = t$) correspond to regions where the inverse reparameterizations are below the identity, and vice versa. Consequently, the effect of transforming persistence diagrams for female subjects results in points in denoised/transformed diagrams that are generally smaller in magnitude than the noisy diagrams, as seen in panels (a) and (c) in Figure 11. Conversely, the effect of transforming persistence diagrams for male subjects results in points in transformed diagrams that are generally larger in magnitude than the noisy diagrams, as seen in panels (a) and (c) in Figure 11.
effects of denoising are not as clear in this real data example as in the simulations. This is due to the large sampling and geometric variability across different subjects.
Figure 11: Alignment results for three random subjects in the female group (each row corresponds to a different subject): (a) Noisy and (c) denoised/transformed persistence diagrams. (b) Estimated reparameterizations. The identity reparameterization is shown in black for reference.
Figure 12: Alignment results for three random subjects in the male group (each row corresponds to a different subject): (a) Noisy and (c) denoised/transformed persistence diagrams. (b) Estimated reparameterizations. The identity reparameterization is shown in black for reference.