Visualization and Outlier Detection for Multivariate Elastic Curve Data

Weiyi Xie, Oksana Chkrebtii, Sebastian Kurtek, *Senior Member, IEEE*

Abstract—We propose a new method for the construction and visualization of geometrically-motivated boxplot displays for elastic curve data. We use a recent shape analysis framework, based on the square-root velocity function representation of curves, to extract different sources of variability from elastic curves, which include location, scale, shape, orientation and parametrization. We then focus on constructing separate displays for these various components using the Riemannian geometry of their representation spaces. This involves computation of a median, two quartiles, and two extremes based on geometric considerations. The outlyingness of an elastic curve is also defined separately based on each of the five components. We evaluate the proposed methods using multiple simulations, and then focus our attention on real data applications. In particular, we study variability in (a) 3D spirals, (b) handwritten signatures, (c) 3D fibers from diffusion tensor magnetic resonance imaging, and (d) trajectories of the Lorenz system.

Index Terms—Shape variability, Square-root velocity function, Geometric boxplots, Elastic curves

1 Introduction

Curve data objects are becoming ubiquitous in the current digital era. In particular, improvements in acquisition technology have enabled collection of large and densely-sampled curve datasets of various sorts. For example, contours in a topographic map can be considered as planar curves. Furthermore, advancement of medical imaging, computer vision and image processing technology is allowing radiologists to acquire a large number of various types of medical images. Studying the morphology of the outlines of anatomical structures is then important for disease diagnosis and monitoring, and may enable new and early treatment strategies. In fact, multiple cutting-edge application areas, including medical imaging and diagnostics, computer vision, graphics, astronomy, geology, and others, regard curves as the main data objects under study.

Curves, however, are complex data objects because they are infinite dimensional and possess different sources of variability. Thanks to recent progress in the shape analysis community [1], [2], [3], statistical methods for analyzing such data are now well established. In particular, Kurtek et al. [3] define different feature spaces that enable decomposition of variability in curve data into (a) translation, (b) scale, (c) rotation, (d) parameterization, and (e) shape. They then define statistical methods on the representation spaces of these various components. However, the number of visualization toolboxes for assessing these different types of variability in curve data is very small. Visualization is an important part of exploratory data analysis. Furthermore, effective visualization tools are necessary to communicate results of statistical analyses to experts in applied fields.

Our focus in this paper is on visualization for shape analysis of elastic curves, open and closed, which have four physical properties: location, scale, shape and orientation. Note that while parametrization is not technically an intrinsic property of the curves, it is used in this work as a way to compute optimal correspondences; this is what makes the curves and statistical analyses elastic [2]. We provide a motivating example in Fig. 1 based on the MPEG-7 dataset[1]. It is extremely difficult to extract any useful information when the original data is shown in a single plot. However, separation into the different sources of variability, i.e., location, scale, shape, orientation and parametrization (for closed curves parameterization is composed of a registration function called phase, which we make precise later, and a starting or seed point on the curve), reveals the true nature of variation hidden in the original data. Each boxplot-type display constructed using the proposed method reflects the particular variation from that component of the curve only, which facilitates intuitive interpretation of results. Additionally, one can detect componentwise outliers in the data, providing more information to the user than outlier detection using the original curves.

A key premise of the proposed visualization approach for elastic curves is that the shape boxplots are invariant to how the original data objects are parameterized. We build our visualization toolbox on the square-root velocity function framework proposed in [2], which has been shown to have such an invariance property. In [3] and [4], the authors extend this framework to include different sources of variability, in addition to shape, in the statistical analysis. An important result in these papers is that under the SRVF representation, the so-called elastic metric (a first order Sobolev metric that measures the amount of bending and stretching deformations) simplifies to the standard $\mathbb{L}^2$ metric, enabling efficient computation of statistics. The proposed approach works as follows. First, we extract the various components of variability in the original curves. Second, we use the Riemannian geometry of the respective representation spaces

- *W. Xie is with Statistical Sciences, Nutrition Research & Development, Abbott Laboratories.*
- *O. Chkrebtii and S. Kurtek are with the Department of Statistics, The Ohio State University.*

1. http://www.dabi.temple.edu/~shape/MPEG7/dataset.html

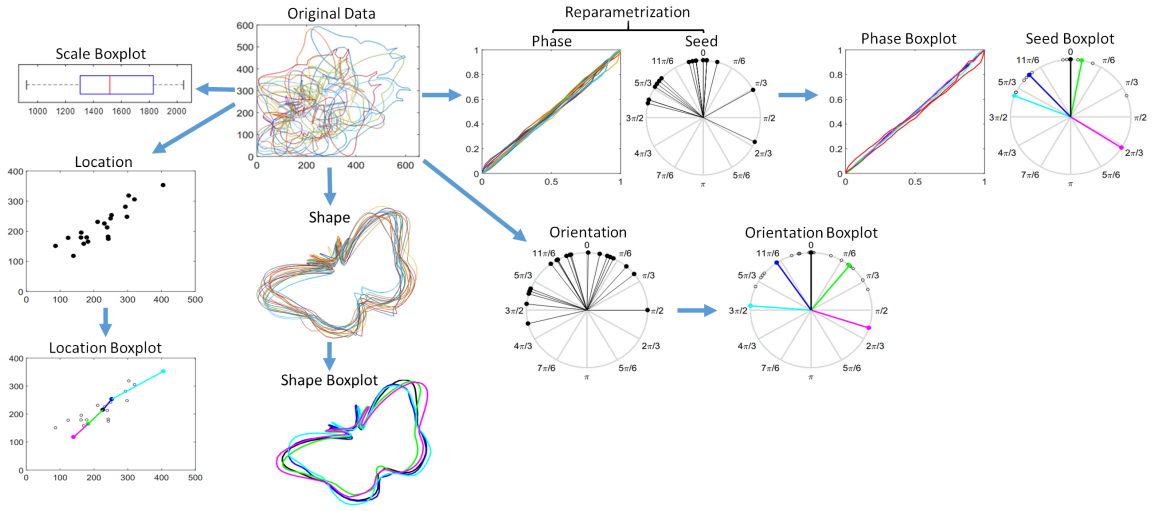

Fig. 1. Visualization of elastic curve data. The proposed method first decomposes the data into the location, scale, shape, orientation, and reparametrization components, and then generates a boxplot-type display for each one separately (black=median, blue and green=quartiles, cyan and magenta=extremes). This allows for effective visualization of variability in each component.

to compute order statistics, which are used for subsequent boxplot-type visualization. We additionally study the ability of the proposed method to detect various types of outliers, and compare our approach to a distance-based approach [4].

1.1 Related Work

The most closely related work that considers boxplot visualizations for curve data is [5]. Their approach is similar to that of [6], which was the first to construct boxplot displays for functional data. The authors in [5] first generalize the concept of functional band depth [7], [8], [9], which itself is a generalization of data depth, to multivariate curves. The depth values are used to rank the curve observations and enable outlier detection. Then, the feature curves of the boxplot are constructed according to the contiguous band swept by a percentage of the deepest ensemble members using the Constructive Solid Geometry (CSG) union operator. The main drawback of this method is inherited from the functional band depth-based methods, in that they require a pre-registration of the data. If the data has significant translation, scale, rotation and/or parameterization variability, then the structure of the boxplot (without accounting for these different sources of variability) may not be very informative. This is because parts of the boxplot are constructed in a pointwise manner, e.g., the 50% band is swept by the 50% deepest ensemble members. A second disadvantage of this approach is that it is not able to differentiate the nature of curve outlyingness.

1.2 Contributions and Paper Organization

To overcome drawbacks of previous approaches, we build on the recent method in [10]. In particular, this work provides an extension of their method from univariate functional data to the case of multivariate curves; our approach defines a comprehensive exploratory data analysis pipeline for multivariate curve data. The main contributions of this paper are as follows:

1) We define a unified approach for computing quartiles and extremes for different sources of variability in multivariate curves. We extract the translation, scale, rotation, parameterization and shape components of the data under an elastic shape analysis framework, and use the Riemannian geometry of their representation spaces to compute a median, two quartiles and two extremes.
2) We construct boxplot-type visualizations for each source of variability, thus allowing the user to clearly see the contribution of each component to the total variation in the data.
3) We provide a new definition of the interquartile range and use it to identify different types of outliers in the data, based on the five different sources of variation in elastic curves.

The rest of this paper is organized as follows. In Section 2, we review elastic shape analysis [2], [3], which allows for extraction of the different sources of variability in elastic curve data. Section 3 provides details of the construction of the shape boxplot, while Section 4 describes the construction of the orientation boxplot. Sections 5 and 6 present multiple simulations as well as results of visualizing variability in real 2D and 3D elastic curves. Finally, we close with a brief summary and some ideas for future work in Section 7. The Supplementary Material includes (a) a description of our approach to construct the location boxplot using curve centroids, (b) algorithms for computing the shape and orientation medians, and extracting the orientation and parameterization components from elastic curves, (c) detailed results for Simulations 1, 2 and 4, and (d) a detailed assessment of computational cost for all real data examples.

2 Background: Elastic Shape Analysis

The data objects in this article are elastic curves. Let $\mathcal{B} = \{\beta : D \to \mathbb{R}^d | \beta \text{ is absolutely continuous}\}$ be the space of absolutely continuous parametrized curves

in Euclidean d-dimensional space, where $d = 2$ or 3, and $D = [0,1]$ for open curves or $D = \mathbb{S}^1$ for closed curves. As most real datasets consider planar and three-dimensional curves, this is the focus of this article. Define $SO(d) = \{R \in \mathbb{R}^{d\times d} | R^T R = RR^T = I_d, \det(R) = +1\}$ as the rotation group, and $\Gamma = \{\gamma : D \to D | \gamma \text{ is an orientation-preserving diffeomorphism}\}$ as the reparameterization group. In the case of closed curves, we decompose the reparameterization into a starting point on the curve (seed) and an orientation-preserving diffeomorphism of $[0,1]$ (a process called unwrapping). For any $\beta \in \mathcal{B}$, $R \in SO(d)$ and $\gamma \in \Gamma$, $\beta \circ \gamma$ is a reparametrization of a curve β by γ and $R\beta$ is a rotation of β by R.

To define a proper metric on the shape space of elastic curves, i.e., the quotient space $\mathcal{B}/(SO(d) \times \Gamma)$, we require that these two groups act by isometries under this metric. It is well-known that the elastic family of metrics has this property [1], [11]. However, this metric is difficult to use in practice. To simplify computation, [2] defined the square-root velocity function (SRVF) transformation as follows. For a curve $\beta \in \mathcal{B}$, its SRVF $q : D \to \mathbb{R}^d$ is defined using a mapping $Q : \mathcal{B} \mapsto \mathbb{L}^2(D, \mathbb{R}^d)$ (henceforth referred to as $\mathbb{L}^2$) as $q = Q(\beta) = \dot{\beta}/\sqrt{|\dot{\beta}|}$, where $|\cdot|$ is the Euclidean norm in $\mathbb{R}^d$ and $\dot{\beta}$ is the time derivative of β. The SRVF transform simplifies the elastic metric to the $\mathbb{L}^2$ metric. The mapping $\beta \leftrightarrow (q, \beta(0))$ between $\mathcal{B}$ and $\mathbb{L}^2 \times \mathbb{R}^d$ is a bijection, and the original curve β can be reconstructed from its SRVF using $Q^{-1}(q)(t) = \beta(t) = \beta(0) + \int_0^t q(s)|q(s)|ds,\ \forall t$. Thus, before computing the SRVF for all curves in a dataset, we extract their location component defined via the centroid. For $R \in SO(d)$, $R\beta \leftrightarrow Rq$, and for $\gamma \in \Gamma$, $\beta\circ\gamma \leftrightarrow (q,\gamma) \coloneqq (q\circ\gamma)\sqrt{\dot{\gamma}}$.

Once the curves are mapped to their SRVFs, we compute their lengths: $\int_D |\dot{\beta}(t)|dt = \int_D |q(t)|^2 dt = \|q\|^2$. This comprises the scale component. Then, we define $\mathcal{C} = \{q \in \mathbb{L}^2 | \|q\|^2 = 1\}$ as the pre-shape space of open curves (i.e., SRVFs of all unit length open curves). $\mathcal{C}$ is the unit Hilbert sphere (in the case of closed curves, it is a submanifold of the unit Hilbert sphere), and is called the pre-shape space because rotation and reparametrization variabilities have not yet been extracted from the curves. Importantly, the action of the product group $\Gamma \times SO(d)$ on $\mathcal{C}$ is by isometries under the $\mathbb{L}^2$ metric: $\|R(q_1,\gamma) - R(q_2,\gamma)\| = \|q_1 - q_2\|$, for all $q_1, q_2 \in \mathcal{C}$, $\gamma \in \Gamma$ and $R \in SO(d)$. Thus, the SRVF becomes essential to our study, because it can be used to separate the rotation and parametrization variabilities from the shape variability in elastic curves.

This framework enables us to extract different sources of variability in elastic curves and analyze them individually in the following representation spaces: location in $\mathbb{R}^d$, scale in $\mathbb{R}_+$, shape in $\mathcal{C}/(\Gamma \times SO(d))$, orientation in $SO(d)$ and reparametrization in Γ. Thus, our general approach for visualization and outlier detection is to first extract the location and scale components from the curves. Next, we transform them into their SRVFs and further separate the orientation and reparametrization components from the shape component. Finally, we construct feature summary statistics based on the unique Riemannian geometry of each representation space, and provide a separate display and outlier detection for the different sources of variability.

The construction of the scale boxplot based on curve lengths is trivial as its representation space is $\mathbb{R}_+$; we use the standard Tukey boxplot [12]. The construction of the translation boxplot is similar to the shape and orientation boxplots introduced in Sections 3.2 and 4, and is included in Section 1 of the Supplementary Material. For the reparameterization boxplot, we use the same method as [10]; please refer to that paper for details. In the case of closed curves, we use the circular boxplot introduced in Section 4.1 to display the seed variability (since it is an element of $\mathbb{S}^1$).

3 Shape Space, Median and Boxplot

We first provide detailed steps to extract the rotation and reparameterization variabilities from elastic curves via the shape median, and then to construct a boxplot-type display for the shape component. Our description focuses on the space of open curves. Closed curves can be handled similarly with minor adjustments in the algorithms.

3.1 Definition of Shape Space and Shape Median

For any $q \in \mathcal{C}$, we define its orbit as $[q] = \{R(q,\gamma) | (\gamma, R) \in \Gamma \times SO(d)\}$. Shape is uniquely associated with an orbit and a distance between shapes can be viewed as a distance between the orbits of their corresponding SRVFs; the shape distance is defined as:

$$D_s([q_1],[q_2]) = \min_{\gamma\in\Gamma,\ R\in SO(d)} \cos^{-1}(\langle q_1, R(q_2,\gamma)\rangle), \quad (1)$$

where $\langle\cdot,\cdot\rangle$ is the $\mathbb{L}^2$ inner product. The orientation and parameterization of an elastic curve is usually defined relative to some template. The template we use in this work is the so-called geometric median [13], which is also used in the construction of the shape boxplot. We define the shape geometric median of a sample of SRVFs $\{q_1, \ldots, q_n\}$ as [4]:

$$[\bar{q}] = \operatorname*{argmin}_{[q]\in\mathbb{L}^2/(\Gamma\times SO(d))} \sum_{i=1}^{n} D_s([q],[q_i]), \quad (2)$$

where D_s is given in Eqn. 1. A gradient-descent algorithm to compute this median is given in Section 2 of the Supplementary Material. Solving the optimization problem in Eqn. 2 results in (a) shape median $[\bar{q}]$; (b) shape distances, $\{D_s^1, \ldots, D_s^n\}$, from the shape median to all of the SRVFs in the data; (c) optimal rotations of the data with respect to the median, $\{R_1, \ldots, R_n\}$; (d) optimal reparametrizations of the data with respect to the median, $\{\gamma_1, \ldots, \gamma_n\}$; and (e) the shape component of the data, $\{\tilde{q}_1, \ldots, \tilde{q}_n\}$, after applying the optimal rotations and reparametrizations to $\{q_1, \ldots, q_n\}$. Note that the shape median is technically defined as an entire orbit. However, in practice, one obtains a single representative element of the orbit $\bar{q} \in [\bar{q}]$. This is done using the orbit centering method [14], which guarantees that the median of $\{\gamma_1, \ldots, \gamma_n\}$ is $\gamma_{id}(t) = t$ and the median of $\{R_1, \ldots, R_n\}$ is the identity matrix I_d.

The algorithm to compute the shape median, and the procedure to compute shape boxplots, require two main geometric tools on the Hilbert sphere. Let the tangent space at any point $q \in \mathcal{C}$ be denoted by $T_q(\mathcal{C})$. Then, the exponential map, $\exp_q : T_q(\mathcal{C}) \to \mathcal{C}$, maps points from the tangent space to the representation space: $\exp_q(v) = \cos(\|v\|)q + \sin(\|v\|)\frac{v}{\|v\|}$, for $v \in T_q(\mathcal{C})$ and $q \in \mathcal{C}$; the

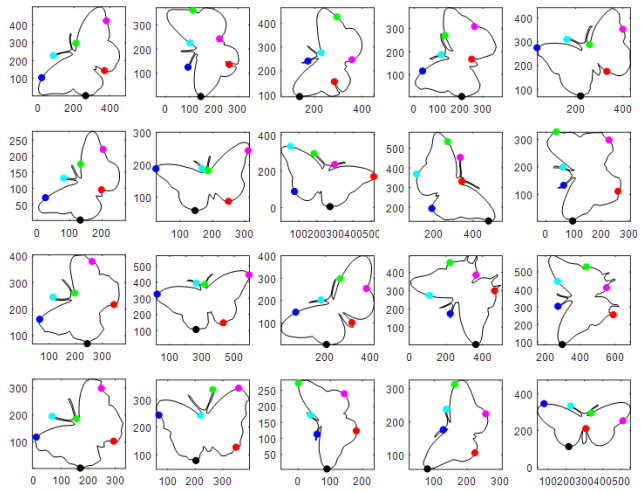

Fig. 2. 20 planar butterfly outlines with parametrization shown as color. Points of the same color correspond to the same parameter values $t = i/99,\ i = 0, 17, 33, 50, 66, 84$ corresponding to black, blue, cyan, green, magenta and red, respectively.

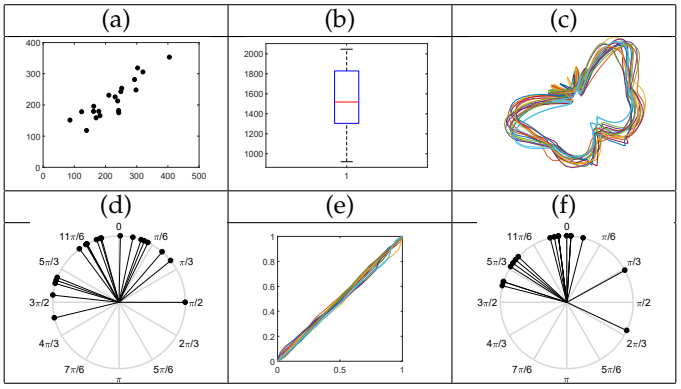

Fig. 3. Separation of (a) location, (b) scale, (c) shape, (d) orientation and reparametrization, (e) phase and (f) seed, variabilities in the butterfly outlines in Fig. 2.

inverse exponential map, $\exp_{q_1}^{-1} : \mathcal{C} \to T_{q_1}(\mathcal{C})$, maps points from the representation space to the tangent space: $\exp_{q_1}^{-1}(q_2) = \frac{\theta}{\sin(\theta)}(q_2 - \cos(\theta)q_1)$, where $q_1,\ q_2 \in \mathcal{C}$ and $\theta = \cos^{-1}(\langle q_1, q_2 \rangle)$. Intuitively, the exponential map takes a vector v in the linear tangent space, $T_q(\mathcal{C})$, and maps it along the geodesic path on $\mathcal{C}$ to a point $q_n = \exp_q(v)$. An important property of the exponential map is that the length of the vector v in the tangent space (as measured using the defined Riemannian metric) is exactly the same as the Riemannian distance on $\mathcal{C}$ between the points q and q_n. In subsequent sections, we additionally use the exponential and inverse exponential mappings to redefine certain difficult computational problems on a nonlinear representation space to equivalent simpler ones on a linear tangent space.

An example of the full separation of different sources of variability in elastic curves is given in Figs. 2 and 3. The data here are outlines of 20 butterflies from the MPEG-7 dataset mentioned earlier. When we plot the butterfly outlines separately as in Fig. 2, it is obvious that those outlines clearly differ in five main aspects: (a) location (Fig. 3(a)), (b) scale (Fig. 3(b)), (c) shape (Fig. 3(c)), (d) relative directions the butterflies are facing or orientation (Fig. 3(d)), and (e) reparametrization (relative timing of prominent geometric features in Fig. 3(e) and starting points in Fig. 3(f)). Next, we propose an innovative method to construct boxplot-type displays for the shape and orientation components, which require specialized geometric tools relevant to their representation spaces.

3.2 Construction of Shape Boxplot

The construction of the proposed shape boxplot requires the computation of the median, two quartiles and two extremes for the curve shapes in a given dataset. We have already outlined a procedure for computing the median, and now focus on the quartiles and extremes. We first order the curves according to their distances from the shape median, and select the 50% of the curves, $\{\tilde{q}_{(1)}, \ldots, \tilde{q}_{(\lceil n/2 \rceil)}\}$, that are closest to $\bar{q}$; this defines the 50% "central shape region". To identify the two shape quartiles, we prefer the two SRVFs $\tilde{q}$ within the 50% central shape region to both be far away from each other and in "opposite directions" from the shape median. This allows us to capture a lot of the shape variability in the given curves. It is easier to construct these feature summary statistics in a linear tangent space than directly on the unit sphere. Thus, we first construct the quartiles and extremes in the tangent space at the shape median, map them back to the shape space under the SRVF representation and finally to the original space of curves for visualization. To facilitate these tasks, we take advantage of the analytical expressions for the exponential and inverse exponential maps, as defined earlier.

The tangent space defined at the shape median is a linear space with the standard $\mathbb{L}^2$ metric. We use the inverse exponential map to transfer $\{\tilde{q}_1, \ldots, \tilde{q}_n\}$ to $T_{\bar{q}}(\mathcal{C})$: $v_i = \exp_{\bar{q}}^{-1}(\tilde{q}_i)$ for $i = 1, \ldots, n$. Note that $\|v_i\|$ is equal to the shape distance D_s^i that we already computed. We optimize the following expression over the 50% central shape region to define the two shape quartiles (v_{Q_1}, v_{Q_3}):

$$\underset{v_1, v_2 \in \{v_{(1)}, \ldots, v_{(\lceil n/2 \rceil)}\}}{\operatorname{argmax}} (1-\lambda)\left(\frac{\|v_1\|}{\max_i \|v_{(i)}\|} + \frac{\|v_2\|}{\max_i \|v_{(i)}\|}\right) - \\ - \lambda\left(\left\langle \frac{v_1}{\|v_1\|}, \frac{v_2}{\|v_2\|} \right\rangle + 1\right). \quad (3)$$

The first term ensures that the two quartiles are far away from each other, the second term ensures that they are in opposite directions and the parameter λ controls the weight of each term. Different choices of λ result in different types of boxplots. In all of our experiments, we use $\lambda = 0.5$ to ensure equal contribution of both terms. To display the two quartiles, we map them back to the original space of (unit length) elastic curves using $\tilde{\beta}_{Q_1} = Q^{-1}(\exp_{\bar{q}}(v_{Q_1}))$ and $\tilde{\beta}_{Q_3} = Q^{-1}(\exp_{\bar{q}}(v_{Q_3}))$, where Q^{-1} was defined in Section 2. Given the two quartiles, the shape interquartile range (IQR) is defined as the sum of the shape distances from each quartile to the geometric median: $IQR_s = \|v_{Q_1}\| + \|v_{Q_3}\|$. Then, the two shape outlier cutoffs are defined as:

$$v_{W_1} = v_{Q_1} + k_s \times IQR_s \times \frac{v_{Q_1}}{\|v_{Q_1}\|}, \text{ and}$$
$$v_{W_3} = v_{Q_3} + k_s \times IQR_s \times \frac{v_{Q_3}}{\|v_{Q_3}\|}. \quad (4)$$

The choice of k_s is not trivial in this setting since these quantities are defined on the shape space; we discuss it in more detail in Section 5. Importantly, the ability to select k_s in Eqn. 4 allows the user to tune the outlier detection procedure in a data dependent manner. Shape outliers are defined as any $\tilde{q}$ that is farther away from the shape median than both of the outlier cutoffs; that is, $\tilde{q}$ is identified as a shape outlier if $\|v\| > \max\{\|v_{W_1}\|, \|v_{W_3}\|\}$, where $v = \exp_{\bar{q}}^{-1}(\tilde{q})$. The two shape extremes are defined as the two shapes closest to each of the two shape outlier cutoffs, with the constraints that (a) they lie outside of the 50% central shape region, and (b) they are not flagged as shape outliers.

Fig. 4 shows the shape boxplot for the butterfly data in Fig. 2. Compared to Fig. 3(c) where all of the aligned shapes are plotted together in a single display, we plot them individually in Fig. 4(a). We display the full shape boxplot in Fig. 4(b) and the two outlier cutoffs in Fig. 4(c). The relative separation between any pair of the feature summary shapes in the boxplot signifies the relative similarity between the two shapes. This idea of visualization is the same as in the standard Tukey boxplot for univariate Euclidean data. From these boxplot-type displays, we are able to visualize the trend of shape variation from one end of the boxplot to the other. Specifically, we can see in the deformation from the black shape median to the green shape quartile that the antennae and the top part of the wings of the butterflies expand. The corresponding outlier cutoff (magenta) amplifies this effect, where the antennae and the top part of the wings become increasingly wider and flatter. In contrast, the deformation from the shape median to the blue shape quartile shows the antennae and the top part of the wings narrowing, which makes the middle part of the wings "push out". The corresponding outlier cutoff (cyan) again amplifies this effect: the antennae and the top part of the wings become increasingly narrow. We find no shape outliers in this butterfly dataset. The corresponding shape extremes (also cyan and magenta) are identified as the shapes closest to each of the outlier cutoffs.

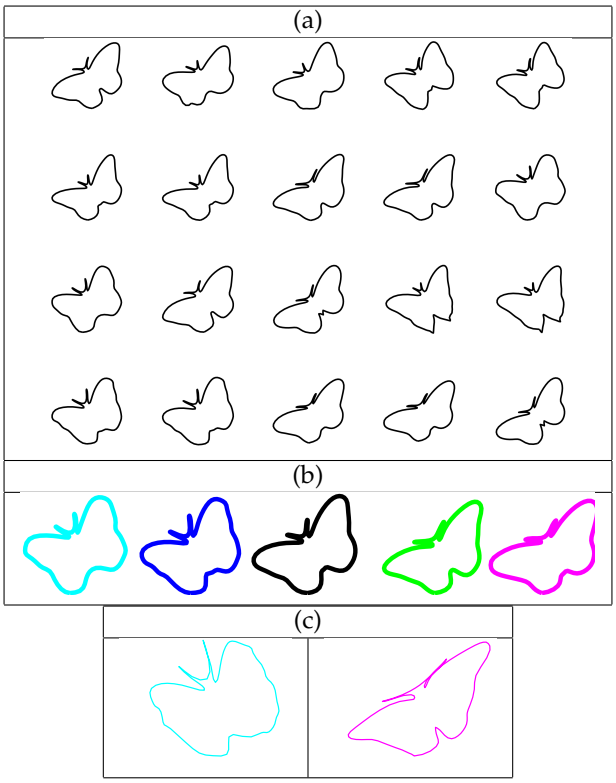

Fig. 4. Shape boxplot for the butterfly outlines in Fig. 2: (a) aligned shapes, (b) full shape boxplot with shape median (black), shape quartiles (blue and green) and shape extremes (cyan and magenta), and (c) shape outlier cutoffs.

4 Construction of Orientation Boxplot

The orientation component is a set of rotation matrices $\{R_1, \ldots, R_n\}$ obtained from aligning the data to the shape median as described in Section 3.1. Their representation space is $SO(d)$, the special orthogonal group of $d \times d$ rotation matrices, whose geometry is nonlinear. Therefore, we are going to use the same strategy as in Section 3.2: (a) map all of the rotation matrices to a tangent space whose geometry is linear, (b) construct the orientation boxplot in the tangent space, and (c) map all of the orientation summary statistics back to $SO(d)$. This procedure requires an appropriate Riemmanian geometry of $SO(d)$, which we describe next.

The space $SO(d)$ is a Lie group with matrix multiplication as the group operation. Thus, the tangent space of $SO(d)$ at the identity element I_d is the Lie algebra $so(d)$, which consists of all skew-symmetric $d \times d$ matrices: $so(d) = \{S \in \mathbb{R}^{d \times d} | S + S^T = 0\}$; the tangent space at $R \in SO(d)$ is defined as $T_R(SO(d)) = \{RS | S \text{ is skew-symmetric}\}$. Given $S_1, S_2 \in so(d)$, the inner product between S_1 and S_2 is defined as: $\langle S_1, S_2 \rangle = \text{trace}(S_1^T S_2)$. Under this Riemannian metric, the Lie group exponential and inverse exponential maps are defined as follows. For $S \in so(d)$, the exponential map, $\exp : so(d) \mapsto SO(d)$, is $\exp(S) = e^S$, where e is the matrix exponential; for $R \in SO(d)$, the inverse exponential map, $\exp^{-1} : SO(d) \mapsto so(d)$, is $\exp^{-1}(R) = \log(R)$, where $\log$ is the matrix logarithm. The orientation distance between two rotations R_1 and R_2 is defined as:

$$D_o(R_1, R_2) = \| \log(R_1^T R_2) \|_F, \tag{5}$$

where $\| \cdot \|_F$ denotes the Frobenius norm.

Using these tools, we can define the geometric median of a set of rotation matrices $\{R_1, \ldots, R_n\}$ as:

$$\bar{R} = \underset{R \in SO(d)}{\text{argmin}} \sum_{i=1}^{n} D_o(R_i, R) = \underset{R \in SO(d)}{\text{argmin}} \sum_{i=1}^{n} \| \log(R_i^T R) \|_F. \tag{6}$$

The solution to this optimization problem can be found using a gradient-descent algorithm provided in Section 2 of the Supplementary Material. After the orientation median is computed, we adjust all of the rotations such that the orientation median $\bar{R}$ is equal to I_d, i.e., we perform the orbit centering step discussed before.

Next, we use the inverse exponential map to transfer all of the rotation matrices to the tangent space at the orientation median identified with I_d: $w_i = \exp^{-1}(R_i)$ for $i = 1, \ldots, n$. We continue our construction of the orientation boxplot in that tangent space. We order the rotation matrices according to their orientation distances to the orientation median $D_o^i = \| \log(R_i) \|_F = \|w_i\|_F$, $i = 1, \ldots, n$, and extract the 50% of rotation matrices that are closest to $\bar{R}$: $\{R_{(1)}, \ldots, R_{(\lceil n/2 \rceil)}\}$. These rotations define the 50% "central

orientation region". We solve the following optimization problem over the 50% central orientation region to find the two orientation quartiles (w_{Q_1}, w_{Q_3}):

$$\underset{w_1,w_2\in\{w_{(1)},\ldots,w_{(\lceil n/2\rceil)}\}}{\operatorname{argmax}} (1-\lambda)\left(\frac{\|w_1\|_F}{\max_i\|w_{(i)}\|_F}+\frac{\|w_2\|_F}{\max_i\|w_{(i)}\|_F}\right)- \\ -\lambda\left(\left\langle\frac{w_1}{\|w_1\|_F},\frac{w_2}{\|w_2\|_F}\right\rangle+1\right). \quad (7)$$

The interpretation of the two terms and the parameter λ are the same as for the shape boxplot. We again use $\lambda = 0.5$ in all of our experiments. Finally, we compute $R_{Q_1} = \exp(w_{Q_1})$ and $R_{Q_3} = \exp(w_{Q_3})$. Given the two orientation quartiles, the orientation IQR is defined as $IQR_o = \|w_{Q_1}\|_F + \|w_{Q_3}\|_F$. The two orientation outlier cutoffs are defined as:

$$w_{W_1} = w_{Q_1} + k_o \times IQR_o \times \frac{w_{Q_1}}{\|w_{Q_1}\|_F}, \text{ and}$$
$$w_{W_3} = w_{Q_3} + k_o \times IQR_o \times \frac{w_{Q_3}}{\|w_{Q_3}\|_F}. \quad (8)$$

As in the case of the shape boxplot, the choice of k_o is not trivial and we address this issue in Section 5 via simulations. A rotation R is identified as an orientation outlier if that rotation matrix is farther away from the orientation median than both of the outlier cutoffs; that is, $w = \exp^{-1}(R)$ satisfies $\|w\|_F > \max\{\|w_{W_1}\|_F, \|w_{W_3}\|_F\}$. Again, the ability to select k_o in Eqn. 8 allows tuning in a data dependent manner. The two orientation extremes are defined as those closest to each of the two orientation outlier cutoffs w_{W_1} and w_{W_3}, with the constraints that (a) they lie outside of the 50% central orientation region, and (b) they are not flagged as orientation outliers. This construction of the orientation boxplot is similar in spirit to our construction of the shape boxplot. As seen in subsequent sections, our boxplot display shows the standard frame in $\mathbb{R}^3$ rotated by the corresponding summary statistics (median, quartiles and extremes).

4.1 Construction of the Circular Boxplot

When the curves are bivariate ($d = 2$), the rotation group $SO(2)$ is one-dimensional and its elements can be represented as angles, i.e., elements of the unit circle $\mathbb{S}^1$. Specifically, for any $R \in SO(2)$, one can write R as $R = \begin{bmatrix}\cos(r) & -\sin(r)\\ \sin(r) & \cos(r)\end{bmatrix}$, where $r \in [0, 2\pi)$. Therefore, in the bivariate case, we can simplify our construction of the orientation boxplot by directly working on $\mathbb{S}^1$ rather than $SO(2)$. In this way, the boxplot construction is similar to the univariate Euclidean case, except that we have to consider the special geometry of the circle.

Given a set of angles $\{r_1, \ldots, r_n\}$, the median angle is defined as:

$$\bar{r} = \underset{r\in[0,2\pi)}{\operatorname{argmin}} \sum_{i=1}^{n} D_{o'}(r_i, r), \quad (9)$$

where $D_{o'}(r_1, r_2) = \min\{|r_1 - r_2|, 2\pi - |r_1 - r_2|\}$ is the distance between two angles, r_1 and r_2, in $\mathbb{S}^1$. The median angle (and its cut locus) splits the circle into two equal-sized semi-circles. The two quartile angles are defined as the two median angles within the range of each semi-circle. The 50% central angle region is the arc connecting the two quartile angles. The angle IQR is defined as $IQR_{o'} = D_{o'}(r_{Q_1}, \bar{r}) + D_{o'}(\bar{r}, r_{Q_2})$. The two angle outlier cutoffs are defined as:

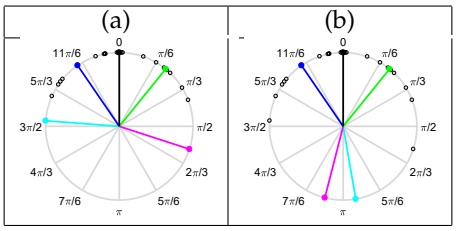

Fig. 5. Orientation boxplot for the butterfly outlines in Fig. 2: (a) median $\bar{r}$ (black), quartiles r_{Q_1} (blue) and r_{Q_3} (green) and extremes (cyan and magenta), and (b) same as (a) but instead of extremes we plot the outlier cutoffs r_{W_1} (cyan) and r_{W_3} (magenta).

$$r_{W_1} = r_{Q_1} + k_{o'} \times IQR_{o'} \times \frac{r_{Q_1} - \bar{r}}{|r_{Q_1} - \bar{r}|}, \text{ and}$$
$$r_{W_3} = r_{Q_3} + k_{o'} \times IQR_{o'} \times \frac{r_{Q_3} - \bar{r}}{|r_{Q_3} - \bar{r}|}. \quad (10)$$

These two angle outlier cutoffs have to be within the range of each semi-circle. If either of the angle outlier cutoffs is out of range, then this guarantees that there are no outliers. An angle is detected as an outlier if it lies outside of the arc connecting the median angle and the corresponding angle outlier cutoff in the same semi-circle. Finally, the two extreme angles are found in each semi-circle as the angles that are closest to each of the two angle outlier cutoffs, r_{W_1} and r_{W_3}, with the constraints that they are (a) in the same semi-circle as the corresponding outlier cutoffs, (b) outside of the 50% central region, and (c) not flagged as outliers.

Since the orientation of bivariate elastic curves is represented by a set of angles, i.e., circular data, one can use the von Mises distribution to determine the outlier cutoff constant $k_{o'}$. The von Mises distribution is a standard statistical model for directional data [15] and has two parameters: a mean angle μ and a concentration parameter κ. As κ increases, the von Mises distribution with mean μ and concentration κ converges to a Gaussian distribution with mean μ and variance $1/\kappa$ [16]. This connection motivates us to estimate the value of the outlier cutoff constant, $k_{o'}$, using the von Mises distribution in a similar manner to using the Gaussian distribution to justify the 1.5 factor for the standard outlier cutoff in Tukey's boxplot. In the standard Euclidean setting, an easy derivation shows that if the data is normally distributed (with any variance), and the outlier cutoff constant is 1.5, then the range between the two outlier cutoffs contains approximately 99.3% of the data, and the rest of the data are flagged as outliers. We want the outlier cutoff constant, $k_{o'}$, to behave similarly in the case of the von Mises distribution: approximately 99.3% of the circular data to be contained between the two outlier cutoffs. In the case of the von Mises distribution, the value of the outlier cutoff constant depends on the concentration parameter. We plug in the maximum likelihood estimate (MLE) of the concentration parameter [15] and compute the value of $k_{o'}$ satisfying the desired probability.

Fig. 5 provides an example of constructing a circular boxplot for the butterfly data in Fig. 2. Our implementation

of the circular boxplot uses the MATLAB Directional Statistics Toolbox [17]. As before, after the orientation median is computed, we adjust all of the rotations such that the orientation median is $\bar{r} = 0$ (black). The full circular boxplot is displayed in panel (a); it provides a nice summary of the given data. The two angle outlier cutoffs (cyan and magenta) can be seen in panel (b). We notice that both of the outlier cutoffs fall outside of their corresponding semicircles resulting in no orientation outliers.

5 Simulations

We begin with several simulations to guide the appropriate choice of the values of the outlier cutoffs for the shape and orientation boxplots. We are interested in the distribution of two quantities: p_c, the percentage of correctly detected outliers (number of correctly detected outliers divided by the total number of outliers), and p_f, the percentage of falsely detected outliers (number of falsely detected outliers divided by the total number of non-outliers). For each simulation, we generate 100 replicates and report the estimated values $\hat{p}_c$ and $\hat{p}_f$ (as a figure for Simulations 1, 2 and 4, and a table for Simulation 3). More detailed tabulated results for Simulations 1, 2 and 4 are provided in Section 3 of the Supplementary Material. For shape outliers, we also apply the methods of [4], which are distance-based. They first compute the geodesic distances between the estimated median shape and each of the shapes in the data $\{D_s^1, \ldots, D_s^n\}$. The quartiles, Q_1 and Q_3, of the distances are used to compute the IQR. Then, the observations corresponding to distances that are greater than $Q_3 + 1.5IQR$ are labeled as outliers. When computing the distances, one has the choice of either accounting for (referred to as elastic) or not accounting for (referred to as arc-length) parametrization variability.

Simulation 1: We generate 20 ellipses; 15 have both semi-major and semi-minor axes (independently) following a uniform distribution on $(0.9, 1.1)$, $U(0.9, 1.1)$, and five have semi-minor axes following a $U(0.9, 1.1)$ and semi-major axes following a $U(a_1, a_2)$. We try three different settings for a_1 and a_2: $U(1.3, 1.5)$, $U(1.4, 1.6)$ and $U(1.5, 2.5)$. Thus, we introduce exactly 25% of shape outliers into the dataset for each setting. One example simulated dataset for each setting is displayed in Figs. 6(a)-(c). We focus on outlier detection for the shape component in this simulation.

The results of this simulation are reported in Fig. 7(a). Mild shape outliers were created with the outlying semi-major axis between $(1.3, 1.5)$ (blue). The performance of true positive shape outlier detection improves substantially when the outlying semi-major axis length increases by only 0.1 of a unit to $(1.4, 1.6)$ (red). Finally, with semi-major axis length being between $(1.5, 2.5)$ (green), the shape outlier detection performance is very stable in terms of the choice of k_s. The elastic method of [4] provides average true detection rates of $54.0\%(23.9\%)$, $63.2\%(21.0\%)$ and $87.8\%(13.6\%)$, respectively, over the three simulated datasets (the standard deviations are reported in the parentheses). The nonelastic method of [4] reports average true detection rates of $55.8\%(23.8\%)$, $69.4\%(20.4\%)$ and $88.8\%(13.4\%)$, respectively. Both methods in [4] report false detection rates of $0\%(0\%)$ for all cases. The performance of the proposed method is favorable compared to the methods in [4] for shape outlier detection.

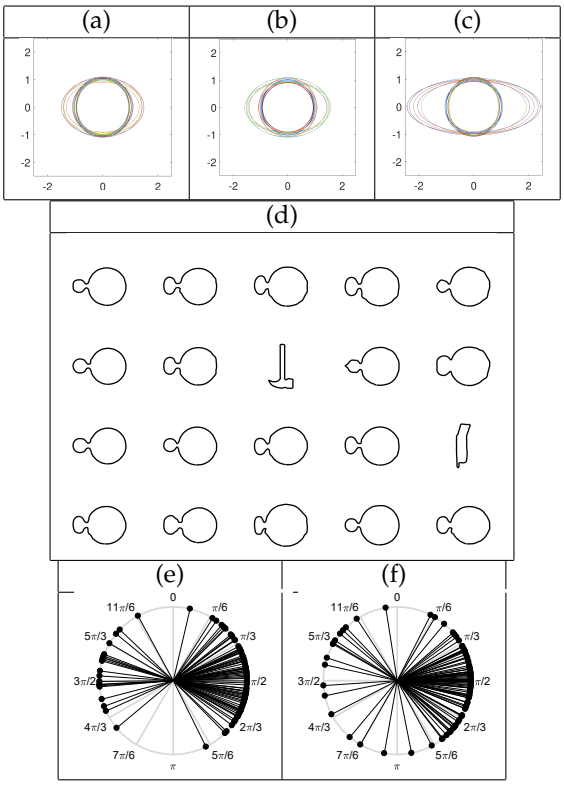

Fig. 6. Sample datasets for Simulation 1 (a) $U(1.3, 1.5)$, (b) $U(1.4, 1.6)$, (c) $U(1.5, 2.5)$; (d) Simulation 2; and Simulation 3 with (κ_1, κ_2) equal to (e) $(5, 5)$, and (f) $(5, 2)$.

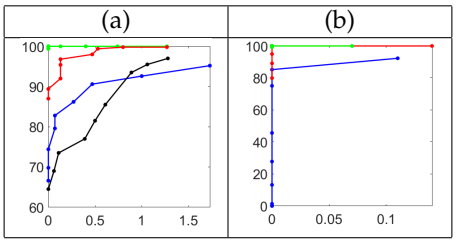

Fig. 7. Plots of average false (x-axis) and true (y-axis) positive outlier detection rates in (%) for (a) Simulation 1 with settings $U(1.3, 1.5)$ (blue), $U(1.4, 1.6)$ (red) and $U(1.5, 2.5)$ (green), and Simulation 2 (black); and (b) Simulation 4 with settings $[\pi/6, \pi/3]$ (blue), $[\pi/3, \pi/2]$ (red) and $[\pi/2, 2\pi/3]$ (green). The marks in (a) correspond to $k_s = 1.5, 1.4, 1.3, 1.2, 1.1, 1.0, 0.8, 0.7, 0.6$, and in (b) to $k_o = 1.4, 1.3, 1.2, 1.0, 0.9, 0.8, 0.6, 0.5, 0.4$.

Simulation 2: Next, we consider 20 sets of closed curves from the MPEG-7 dataset, with each set containing exactly 20 shapes. For each replication, we randomly select one set from the 20 sets of outlines as the non-outlying class and randomly replace two outlines within the non-outlying class by two outlines randomly chosen from the other sets to introduce 10% of outliers. We again focus on outlier detection for the shape component. One example simulated dataset is displayed in Fig. 6(d). The results, shown in Fig. 7(a) in black, demonstrate that our method performs well in this setting. The elastic method of [4] reports an

TABLE 1
Average true and false positive outlier detection rates (in % with standard deviations in parentheses) for data in Simulation 3.

(κ_1, κ_2)	(5,5)	(5,4)	(5,3)	(5,2)
$\hat{p}_c$	98.7 (3.2)	97.6 (4.2)	93.8 (6.3)	86.8 (8.8)
$\hat{p}_f$	0.1 (0.3)	0.2 (0.5)	0.2 (0.6)	0.2 (0.6)

average true detection rate of $91.5\%(24.7\%)$ and an average false detection rate of $1.6\%(3.5\%)$, while the nonelastic one gives an average true detection rate of $73.5\%(37.2\%)$ and an average false detection rate of $1.2\%(3.2\%)$. Again, the proposed method performs well compared to [4].

Combining our observations from Simulations 1 and 2, we advise the following general settings for the shape outlier cutoff constant k_s: mild shape outliers detected with $k_s \in [0.7, 1.1)$, regular shape outliers detected with $k_s \in [1.1, 1.4)$ and severe shape outliers detected with $k_s \geq 1.4$. This multiscale approach to outlier detection allows for better exploration of complex elastic curve datasets.

Simulation 3: We use this simulation to test the proposed outlier detection method for circular data based on cutoffs motivated by the von Mises distribution. Specifically, we generate 100 angles using a mixture of two von Mises distributions: each angle is generated with a probability of 0.8 from a von Mises distribution with mean $\frac{\pi}{2}$ and concentration parameter κ_1, and with a probability of 0.2 it is generated from a von Mises distribution with mean $\frac{3}{2}\pi$ and concentration parameter κ_2. Thus, we introduce approximately 20% of outliers into the dataset. We use the following pairs of (κ_1, κ_2): (5,5), (5,4), (5,3) and (5,2). Two example simulated datasets are displayed in Figs. 6(e)-(f). The results of this simulation, shown in Table 1, confirm that the proposed outlier detection method, assuming the von Mises distribution for the circular data, is effective and robust. To the best of our knowledge, there are no methods in the current literature that can detect 2D orientation outliers in elastic curve data.

Simulation 4: We generate 3D rotation matrices via matrix multiplication of three basic rotation matrices, $R_x(\alpha)$, $R_y(\beta)$, and $R_z(\gamma)$, which rotate a shape about the x, y and z-axes, respectively, using $R = R_x(\alpha)R_y(\beta)R_z(\gamma)$. Here, α, β and γ denote the angles of rotation and take values in $[0, 2\pi)$. For each replication, we generate 100 rotation matrices, such that with probability 0.8 α, β, $\gamma \in U(0, \pi/6)$ and with probability 0.2 α, β, $\gamma \in U(\theta_1, \theta_2)$. We choose the following three settings for this simulation: (a) $\theta_1 = \pi/6$, $\theta_2 = \pi/3$, (b) $\theta_1 = \pi/3$, $\theta_2 = \pi/2$, and (c) $\theta_1 = \pi/2$, $\theta_2 = 2\pi/3$; the last setting generates rotations farthest from identity. As a result, we introduce approximately 20% of 3D orientation outliers into the dataset.

Mild 3D orientation outliers were created with α, β, $\gamma \in U(\frac{\pi}{6}, \frac{\pi}{3})$. From the results shown in Fig. 7(b), we can see that the true detection rates drop very fast as k_o increases. In contrast, with α, β, $\gamma \in U(\frac{\pi}{3}, \frac{\pi}{2})$ and with α, β, $\gamma \in U(\frac{\pi}{2}, \frac{2\pi}{3})$, the performance of both the true and false detection rates is very good across almost all of the k_o values in the table. We advise using the following scale for k_o: mild 3D orientation outliers are detected with $k_o \in [0.5, 0.9)$, regular 3D orientation outliers are detected with $k_o \in [0.9, 1.3)$ and severe 3D orientation outliers are detected with $k_o \geq 1.3$.

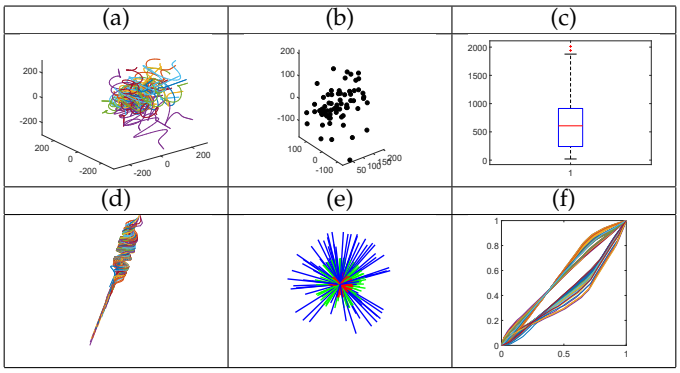

Fig. 8. Separation of (b) location, (c) scale, (d) shape, (e) orientation, and (f) phase variabilities in the (a) 3D spiral data.

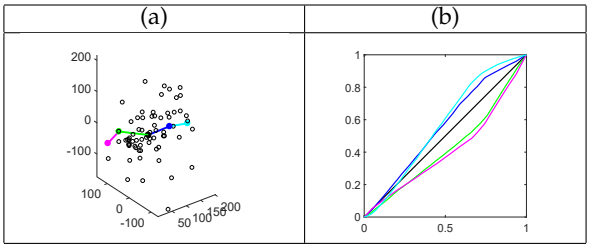

Fig. 9. (a) Location, and (b) phase boxplots for the spiral data.

6 Applications to Real Data

Next, we assess our method on four elastic curve datasets. The first example considers artificially generated 3D helical curves. The second example visualizes variability in signatures, which are planar open curves. In the third example, we visualize variability in a set of three-dimensional brain fibers extracted from a diffusion tensor magnetic resonance image (DT-MRI). Finally, in the last example, we study variability in 3D curves generated by the Lorenz system.

Example 1: This study considers variability in 70 spiral curves. We plot the original data in Fig. 8(a). The variability of the spirals is quite complex and involves differences in shape, orientation, length and location. As a byproduct of computing the shape median, we additionally extract the parameterization variability. While this component is often regarded as a nuisance variable in shape analysis, it may be informative in some multivariate curve data settings. We display the separation of different sources of variability in Figs. 8(b)-(f). Based on the length of each spiral, we discover two scale outliers that are substantially longer than other spirals (Fig. 8(c)). The plot of the 3D orientations in Fig. 8(e) shows that this component has a lot of variability.

The location boxplot is displayed in Fig. 9(a); we find no location outliers. The two clusters of reparameterizations in Fig. 8(f) indicate that in order to match the timing of the features on the median shape, some of the parameterization functions must go faster in the beginning and then slower, while others go slower first and then faster; few go at a similar pace as the median. The phase boxplot in Fig. 9(b) summarizes these features: the distance from either of the two phase quartiles to its corresponding extreme is

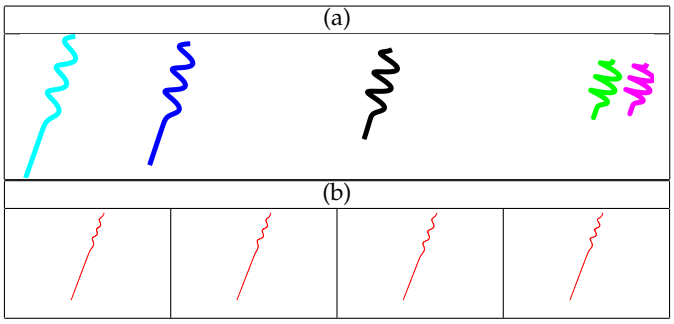

Fig. 10. (a) Shape boxplot, and (b) mild shape outliers for the spiral data.

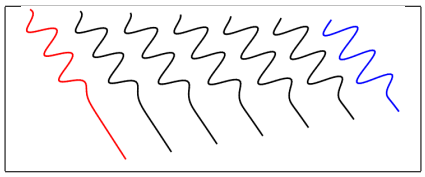

Fig. 11. Deformation between shape centroids of two clusters formed according to the phase functions in Fig. 8(f).

much less than to the median, which implies that the phase functions are spread out toward the two phase extremes where the clusters are formed. Note that this clustering is closely related to the variability in shape of the spirals, which we discuss next.

Fig. 10 summarizes the shape component variability in the spirals. The shape boxplot is displayed in Fig. 10(a). The shapes in this dataset vary in mostly two ways. From the shape median to the blue shape quartile and corresponding cyan extreme, the spiral stretches, becomes thinner and the straight segment extends; from the shape median to the green shape quartile and corresponding magenta extreme, the spiral contracts, becomes thicker and the straight segment shortens. We can also visualize the relative similarities between each pair of the feature summary shapes: the green quartile and magenta extreme are much closer to each other than the blue and cyan pair. This is consistent visually as the green and magenta spirals are more similar to each other in terms of shape than the blue and cyan spirals. We detect four mild shape outliers (Fig. 10(b)), all of which have very elongated, thin spirals and long straight segments. In comparison, the elastic method of [4] flags six outliers.

The variability in the shape boxplot can be further related to the observed phase variability. On the one hand, compared to the shape median, the straight segments of the blue quartile and cyan extreme are longer, and the spirals are thinner; on the other hand, the straight segments of the green quartile and magenta extreme are shorter, and the spirals are thicker. In order to match the features across these different summaries, the phase functions corresponding to the shapes similar to the blue quartile and cyan extreme have to traverse slower along the spiral and then faster along the straight segment; in contrast, the phase functions corresponding to the green quartile and magenta extreme have to traverse along the spiral faster and then slower along the straight segment. This explains the two

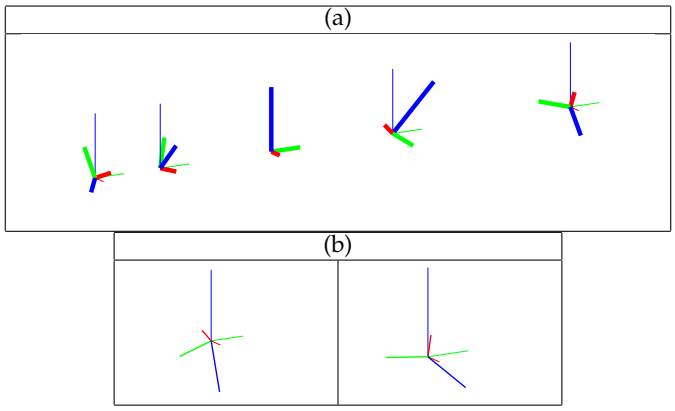

Fig. 12. (a) Orientation boxplot, and (b) mild orientation outliers for the spiral data.

clusters evident in the extracted phase functions. Fig. 11 displays the geodesic (minimal deformation) path between the shapes of the centroids of spirals clustered according to phase sampled at five equally spaced points (black shapes along the path). The first cluster corresponds to the phase functions in Fig. 8(f) concentrated below the 45 degree line (i.e., identity reparameterization). Most shapes in this cluster look like the blue quartile and cyan extreme in Fig. 10(a), and the resulting shape mean in this cluster is the red curve in the geodesic path (thin spiral with long straight segment). The second phase cluster corresponds to functions in Fig. 8(f) concentrated above identity reparameterization. Most shapes in this cluster look like the green quartile and magenta extreme in Fig. 10(a), and the resulting shape mean in this cluster is the blue curve in the geodesic path (thick spiral with short straight segment).

The three-dimensional orientations extracted from the spiral data are displayed in Fig. 12 as rotations of the standard frame in $\mathbb{R}^3$. In each plot, the thin red, green and blue axes represent the original x, y and z-axes, respectively, corresponding to an identity rotation. The thick red, green and blue axes represent the x, y, and z axes, respectively, after applying the desired rotation. Fig. 12(a) displays the orientation boxplot. Since the orientation component of the spirals appears to be quite random, there does not seem to be an intuitive pattern in this display. As in the shape boxplot, the relative separation between any pair of the feature summary rotations is based on the orientation distance between them; this improves visualization as it is challenging to observe in the original plot in Fig. 8(e). Two rotations are flagged as mild orientation outliers and are displayed in Fig. 12(b).

Example 2: This real data study considers a dataset of 40 handwritten signatures "Lau" from the SVC 2004 dataset [18]. The signatures are shown in Fig. 13(a). The reparametrization variability in this example can possibly represent different speed of writing.

We extract the different components of variability in the data and show them in Figs. 13(b)-(f). We also compute the length of each signature and provide the standard boxplot in Fig. 13(c); no scale outliers were found in this dataset. The small variability in the orientation component (Fig. 13(e)) is consistent with a visual inspection of the 40

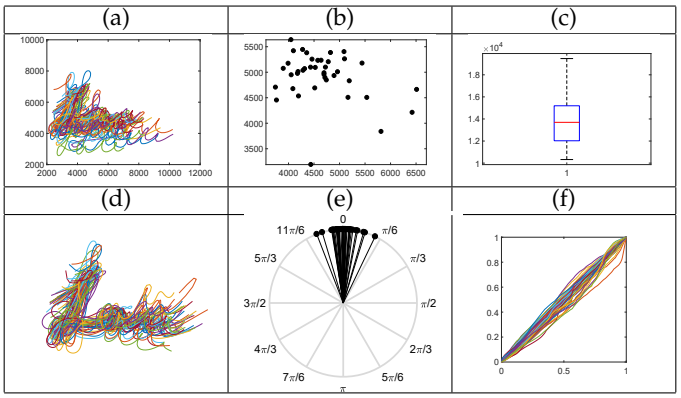

Fig. 13. Separation of (b) location, (c) scale, (d) shape, (e) orientation, and (f) phase variabilities in the (a) signature data.

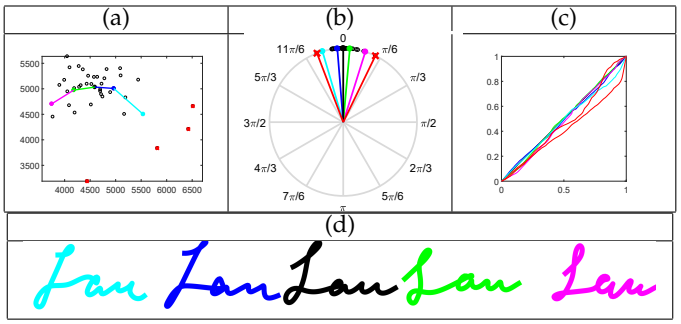

Fig. 14. (a) Location, (b) orientation, (c) phase, and (d) shape boxplots for the signature data.

signatures. Nonetheless, we still flag two orientation outliers as seen in the boxplot in Fig. 14(b). We display the location boxplot in Fig. 14(a) and detect four location outliers. We also detect two phase outliers in this signature data (red phase functions in Fig. 14(c)), which may indicate that those two signatures were written at significantly different speeds relative to the other signatures.

Finally, we assess the shape variability in this signature dataset. We display the shape boxplot in Fig. 14(d) and see that the five feature summary shapes are approximately equally separated. From the shape median to the green shape quartile and magenta extreme, the last letter "u" in the signature tends to look more like a "w" and the horizontal segment in the first letter "L" becomes tilted. On the other hand, from the shape median to the blue shape quartile and cyan extreme, the horizontal segment in the "L" shifts higher. Our method and the elastic method of [4] do not detect any shape outliers in this dataset.

Example 3: The data in this study considers fiber tracts in the human brain extracted from a DT-MRI and consists of 176 3D open curves. The data is presented in Fig. 15(a). Based on an initial observation, the fibers can be clustered into two groups (this result was reported in [3]). Additionally, we know that some of the fiber curves are replicated.

We extract the different components of variability in this data and display them in Figs. 15(b)-(f). As seen in Fig. 15(c), we flag several fibers as scale outliers based on their length. Additionally, based on the plots of the shape and orientation components in Figs. 15(d) and (e), there seem to

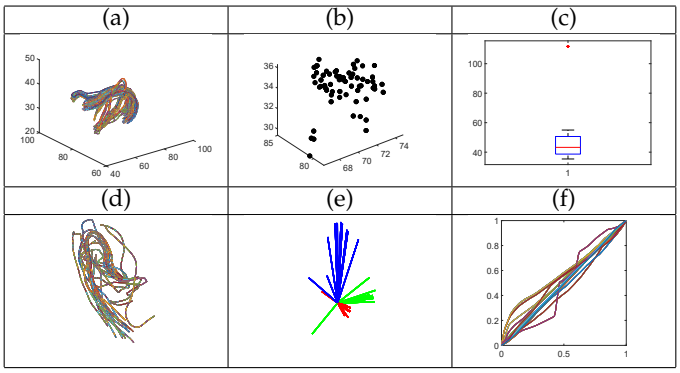

Fig. 15. Separation of (b) location, (c) scale, (d) shape, (e) orientation, and (f) phase variabilities in the (a) 3D fiber data.

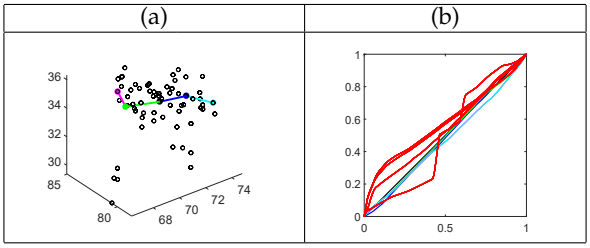

Fig. 16. (a) Location, and (b) phase boxplots for the fiber data.

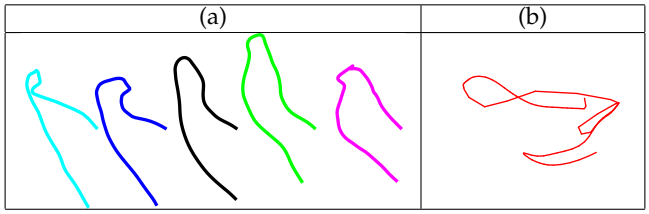

Fig. 17. (a) Shape boxplot, and (b) severe shape outliers for the fiber data.

be fewer than 176 observations in the original dataset; this is consistent with our prior knowledge that there are multiple replicates in this data. The rotations in Fig. 15(e) indicate some clustering in this component. The location boxplot is displayed in Fig. 16(a) and no location outliers are detected. The phase boxplot is shown in Fig. 16(b) and is considered a nuisance variable in this application.

Next, we investigate the variability in the shape component. We observe from Fig. 17(a) that the magenta extreme shape and the green quartile shape are separated by a larger distance than their blue and cyan counterparts. Also, from the shape median to the blue shape quartile and cyan extreme, the left part of the fiber straightens and flattens, and the right part of the fiber tends to extend rightward; from the shape median to the green shape quartile and magenta extreme, the left part of the fiber curves inward, and the right part of the fiber straightens. We detect 11 severe shape outliers. Interestingly, all of them are exactly the same, as shown in Fig. 17(b). This confirms our prior knowledge of the existence of duplicates, which were perhaps introduced during initial data pre-processing. The elastic method of [4] also flags the same 11 shape outliers.

Finally, we display the orientation boxplot in Fig. 18.

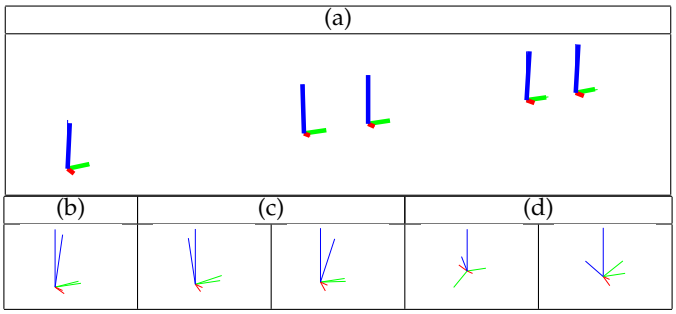

Fig. 18. (a) Orientation boxplot, (b) mild, (c) regular, and (d) severe orientation outliers for the fiber data.

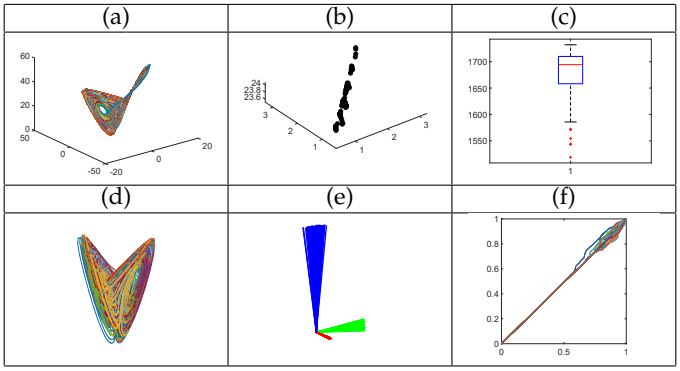

Fig. 19. Separation of (b) location, (c) scale, (d) shape, (e) orientation, and (f) phase variabilities in the (a) Lorenz trajectories.

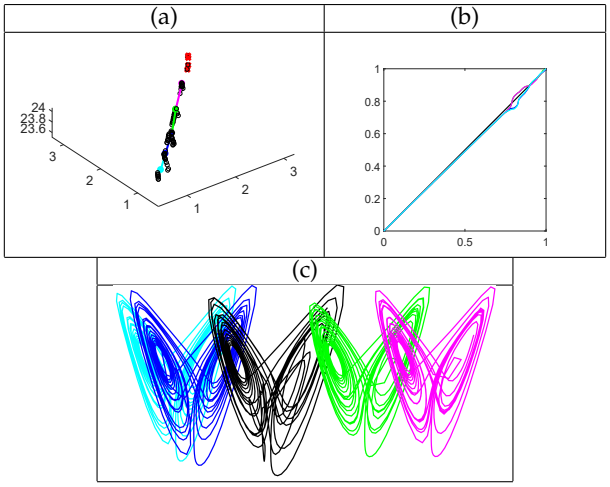

Fig. 20. (a) Location, (b) phase, and (c) shape boxplots for the Lorenz trajectories.

There appears to be little variability in this component, and this is consistent with our initial observation from Fig. 15. We visualize the distribution of orientations by examining the relative separation between pairs of boxplot features. Somewhat unexpectedly, we detect 55 orientation outliers that can be separated into five groups, where each group contains 11 rotations that are exactly the same as seen in Figs. 18(b)-(f); this is again due to duplicates in the data. Within the five different rotation outliers, one of them is flagged as a mild orientation outlier, two as regular outliers and two as severe outliers.

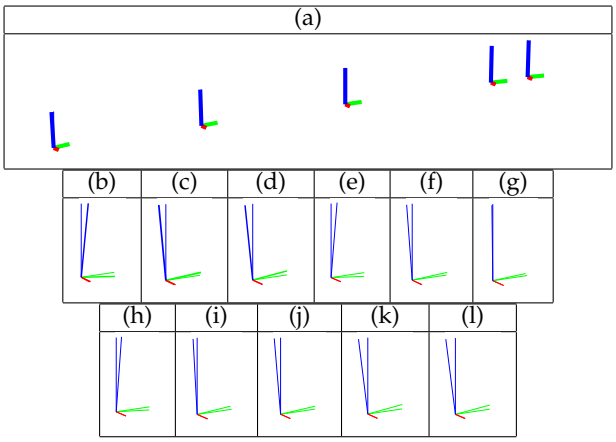

Fig. 21. (a) Orientation boxplot, (b)-(j) mild orientation outliers, and (k)&(l) regular orientation outliers for the Lorenz trajectories.

Example 4: Lastly, we examine three-dimensional trajectories of the Lorenz system [19], a set of ordinary differential equations that were originally used to describe the simplified dynamics of convection in the atmosphere. They have since also been used to model other oscillating systems, including lasers and dynamos. An important characteristic of the Lorenz system is its sensitivity to small perturbations for certain parameter regimes. In fact, small deviations over time can result in exponentially fast divergence of the system trajectory, a phenomenon called chaos. Because this system is restricted to lie on a lower-dimensional manifold, called the Lorenz attractor, individual trajectories share similar qualitative features such as time spent in each lobe of the attractor. This variability also changes over time, as small perturbations typically require some time to propagate into large differences in trajectories. We examine a posterior sample (of size 100) over the solution trajectory for the Lorenz system with fixed initial states and parameter values in the chaotic regime [20]. Because the Lorenz system does not have a closed form solution, discretization uncertainty characterized by this posterior sample qualitatively resembles a system with small perturbations across the domain. Initially, trajectories are very similar, but begin to diverge over the last third of the domain. The proposed approach provides a useful visualization tool to study the distribution of these complex nonlinear paths.

The dataset is displayed in Fig. 19(a). We separate the original data into various components of variability shown in Figs. 19(b)-(f). The scale boxplot in Fig. 19(c) detects six scale outliers, all of which are significantly shorter in length than other Lorenz attractor curves. We also find that the Lorenz attractor curves appear very similar in shape (Fig. 19(d)). The extracted orientation components also have very small variability (Fig. 19(e)). We construct the location boxplot in Fig. 20(a) and flag four location outliers. The phase component (Figs. 19(f) and 20(b)) of the Lorenz attractor curves is more interesting: there is almost no variability until the last one-third of the interval. This is consistent with the chaotic nature of this system; small initial difference in the seemingly coincident Lorenz attractor curves leads to massive divergence in the path of the curves at later stages, popularized as "the butterfly effect".

The shape boxplot is shown in Fig. 20(c). Although the shapes appear similar, we note relative differences based on the separation of the feature summaries: the blue quartile and cyan extreme are much closer to each other than the green quartile and magenta extreme. That is, the blue and cyan curves are more similar to each other in shape than the green and magenta ones, though we cannot see much of the subtle differences. Both our method and the elastic method of [4] detect no shape outliers.

Finally, we construct the orientation boxplot in Figure 21(a). As in the case of the shape component, it is difficult to see the difference in the derived rotations, since they vary in a very small range of angles. However, using the proposed tools, we detect a total of 11 orientation outliers, among which nine are mild outliers and two are regular outliers.

Computational Cost: The computational cost of the proposed approach for Examples 1-4 is 32, 27, 68 and 895 seconds, respectively. A more detailed assessment of algorithm convergence and computational cost is provided in Section 4 of the Supplementary Material.

7 Summary and Future Work

In this article, we extended the idea of Tukey's boxplot from univariate Euclidean data to elastic curve data. The goal of this paper is to define the median, the two quartiles, the two extremes and detect outliers for various sources of variability in curve datasets. To achieve this goal, we introduced a set of procedures to construct boxplot-type displays for visualization of the location, shape and orientation components often present in elastic curve observations (the reparameterization and scale components are assessed using previously introduced methods). Different from traditional approaches to visualizing curve data based on a single boxplot display, often constructed via depth-based methods, the proposed approach is metric-based and considers the different sources of variability separately; the boxplots are generated via geometric tools defined on each of their representation spaces. Our method thus allows for independent outlier detection for each component. In the future, we will focus on extending this concept of component-wise boxplot displays to more complex functional data, including images and shapes of surfaces. In both cases, the Riemannian geometry of the phase component is much more complex and requires the development of novel, computationally efficient tools.

Acknowledgments

We thank Prof. Zhaohua Ding for providing the DT-MRI fiber dataset. Sebastian Kurtek's research was partially supported by NSF DMS 1613054, NSF CCF 1740761, NSF CCF 1839252 and NIH R37 CA214955.

References

[1] L. Younes, "Computable elastic distance between shapes," *SIAM Journal of Applied Mathematics*, vol. 58, no. 2, pp. 565–586, 1998.

[2] A. Srivastava, E. Klassen, S. H. Joshi, and I. H. Jermyn, "Shape analysis of elastic curves in Euclidean spaces," *IEEE T. on Pattern Analysis and Machine Intelligence*, vol. 33, no. 7, pp. 1415–1428, 2011.

[3] S. Kurtek, A. Srivastava, E. Klassen, and Z. Ding, "Statistical modeling of curves using shapes and related features," *Journal of the American Statistical Association*, vol. 107, no. 499, pp. 1152–1165, 2012.

[4] S. Kurtek, J. Su, C. Grimm, M. Vaughan, R. Sowell, and A. Srivastava, "Statistical analysis of manual segmentations of structures in medical images," *Computer Vision and Image Understanding*, vol. 117, pp. 1036–1050, 2013.

[5] R. T. Whitaker, M. Mirzargar, and R. M. Kirby, "Curve boxplot: Generalization of boxplot for ensembles of curves," *IEEE T. on Visualization and Computer Graphics*, vol. 20, no. 12, pp. 2654–2663, 2014.

[6] Y. Sun and M. G. Genton, "Functional boxplots," *Journal of Computational and Graphical Statistics*, vol. 20, no. 2, pp. 316–334, 2011.

[7] S. López-Pintado and J. Romo, "On the concept of depth for functional data," *Journal of the American Statistical Association*, vol. 104, no. 486, pp. 718–734, 2009.

[8] J. Kuelbs and J. Zinn, "Concerns with functional depth," *Latin American Journal of Probability and Mathematical Statistics*, vol. 10, no. 2, pp. 815–839, 2013.

[9] A. Chakraborty and P. Chaudhuri, "On data depth in infinite dimensional spaces," *Annals of the Institute of Statistical Mathematics*, vol. 66, no. 2, pp. 303–324, 2014.

[10] W. Xie, S. Kurtek, K. Bharath, and Y. Sun, "A geometric approach to visualization of variability in functional data," *Journal of the American Statistical Association*, vol. 112, no. 519, pp. 979–993, 2017.

[11] W. Mio, A. Srivastava, and S. H. Joshi, "On shape of plane elastic curves," *International Journal of Computer Vision*, vol. 73, no. 3, pp. 307–324, 2007.

[12] J. W. Tukey, *Exploratory Data Analysis*. Addison-Wesley, 1977.

[13] P. T. Fletcher, S. Venkatasubramanian, and S. Joshi, "The geometric median on Riemannian manifolds with application to robust atlas estimation," *Neuroimage*, vol. 45, no. 1, pp. S143–S152, 2009.

[14] A. Srivastava, W. Wu, S. Kurtek, E. Klassen, and J. S. Marron, "Registration of functional data using Fisher–Rao metric," *arXiv:1103.3817v2*, 2011.

[15] K. V. Mardia and P. E. Jupp, *Directional Statistics*. John Wiley & Sons, 2009.

[16] G. Hughes, "Multivariate and time series models for circular data with applications to protein conformational angles," Ph.D. dissertation, Department of Statistics, University of Leeds, 2007.

[17] P. Berens, "CircStat: A MATLAB toolbox for circular statistics," *Journal of Statistical Software*, vol. 31, no. 10, pp. 1–21, 2009.

[18] D.-Y. Yeung, H. Chang, Y. Xiong, S. George, R. Kashi, T. Matsumoto, and G. Rigoll, "SVC2004: First international signature verification competition," in *Proceedings of the International Conference on Biometric Authentication (ICBA)*, 2004, pp. 16–22.

[19] E. N. Lorenz, "Deterministic nonperiodic flow," *Journal of the Atmospheric Sciences*, vol. 20, pp. 130–141, 1963.

[20] O. Chkrebtii, D. Campbell, B. Calderhead, and M. Girolami, "Bayesian solution uncertainty quantification for differential equations," *Bayesian Analysis*, vol. 11, no. 4, pp. 1239–1267, 2016.

Weiyi Xie Weiyi Xie works for Abbott Laboratories as a Senior Biostatistician. He received his MS and PhD in Statistics from The Ohio State University. He has won multiple conference awards, including JSM and M&M. He is interested in applying functional data analysis into solving critical healthcare problems. He is also passionate about developing innovative visualization tools for driving new insights.

Oksana Chkrebtii Oksana Chkrebtii is an Assistant Professor at The Ohio State University, specializing in Bayesian analysis and uncertainty quantification. She received her BMath and MSc degrees in Statistics from Carleton University in 2006 and 2008, respectively. She received her PhD in Statistics from Simon Fraser University in 2014.

Sebastian Kurtek Sebastian Kurtek is an Associate Professor of Statistics at The Ohio State University. He received his BS in Mathematics from Tulane University, and his MS and PhD in Biostatistics from Florida State University. His main research interests include statistical shape analysis, functional data analysis, and statistics on manifolds.