Journal of the American Statistical Association

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/uasa20

Randomness of Shapes and Statistical Inference on Shapes via the Smooth Euler Characteristic Transform

Kun Meng, Jinyu Wang, Lorin Crawford & Ani Eloyan

To cite this article: Kun Meng, Jinyu Wang, Lorin Crawford & Ani Eloyan (2025) Randomness of Shapes and Statistical Inference on Shapes via the Smooth Euler Characteristic Transform, Journal of the American Statistical Association, 120:549, 498-510, DOI: 10.1080/01621459.2024.2353947

To link to this article: https://doi.org/10.1080/01621459.2024.2353947

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2025, VOL. 120, NO. 549, 498–510: Theory and Methods
https://doi.org/10.1080/01621459.2024.2353947

Randomness of Shapes and Statistical Inference on Shapes via the Smooth Euler Characteristic Transform

Kun Meng[a], Jinyu Wang[b], Lorin Crawford[c,d], and Ani Eloyan[c]

[a]Division of Applied Mathematics, Brown University, Providence, RI; [b]Data Science Institute, Brown University, Providence, RI; [c]Department of Biostatistics, Brown University School of Public Health, Providence, RI; [d]Microsoft Research New England, Cambridge, MA

ABSTRACT

In this article, we establish the mathematical foundations for modeling the randomness of shapes and conducting statistical inference on shapes using the smooth Euler characteristic transform. Based on these foundations, we propose two Chi-squared statistic-based algorithms for testing hypotheses on random shapes. Simulation studies are presented to validate our mathematical derivations and to compare our algorithms with state-of-the-art methods to demonstrate the utility of our proposed framework. As real applications, we analyze a dataset of mandibular molars from four genera of primates and show that our algorithms have the power to detect significant shape differences that recapitulate known morphological variation across suborders. Altogether, our discussions bridge the following fields: algebraic and computational topology, probability theory and stochastic processes, Sobolev spaces and functional analysis, analysis of variance for functional data, and geometric morphometrics. Supplementary materials for this article are available online, including a standardized description of the materials available for reproducing the work.

ARTICLE HISTORY
Received January 2023
Accepted April 2024

KEYWORDS
Functional data analysis; Karhunen–Loève expansion; o-minimal structures; Persistence diagrams; Reproducing kernel Hilbert spaces

1. Introduction

The quantification of shapes has become an important research direction. It has brought advances to many fields including geometric morphometrics (Boyer et al. 2011; Gao, Kovalsky, and Daubechies 2019; Gao et al. 2019), biophysics and structural biology (Wang et al. 2021; Tang et al. 2022), and radiogenomics (Crawford et al. 2020). When shapes are considered as random variables, their corresponding quantitative summaries are also random, implying that such summaries of random shapes are statistics. The statistical inference on shapes based on these quantitative summaries has been of particular interest (Fasy et al. 2014; Roycraft, Krebs, and Polonik 2023).

In this article, we bring together mathematical and statistical approaches to make three significant contributions to shape statistics: (i) we provide mathematical foundations for the randomness of shapes encountered in applications, bridging algebraic topology (Hatcher 2002) and stochastic processes (Hairer 2009); (ii) we connect the statistical inference on shape-valued data to the well-studied analysis of variance for functional data (fdANOVA, Zhang 2013), bridging topological data analysis (TDA, Edelsbrunner and Harer 2010) and functional data analysis (FDA, Hsing and Eubank 2015); and (iii) our framework does not rely on any assumptions about diffeomorphisms or prespecified landmarks.

1.1. A Motivating Scientific Question

Through modeling the randomness of shapes, we aim to address the following statistical inference question: *Is the observed difference between two groups of shapes statistically significant?* For example, the mandibular molars in Figure 1 are from four genera of primates. A pertinent question from a morphological perspective is: *In Figure 1, do the molars from genus Tarsius exhibit significant differences from those from the other genera?*

The primary objective of this article is to propose a powerful approach for testing hypotheses on random shapes. This would help address morphology-motivated statistical inference questions like the one raised above. In achieving this objective, we lay down the mathematical foundations that justify our hypothesis testing methods. We take two key steps: In Step 1, we find the appropriate representations of shapes; and in Step 2, we test hypotheses on shapes using these representations. In Section 1.2, we provide a literature review on shape representations and introduce the topological summary employed in this article. Section 1.3 begins by presenting the main theme of our hypothesis testing approach, followed by an overview of our contributions. Since the molars in Figure 1 are diffeomorphic to the two-dimensional unit sphere, some existing diffeomorphism-related methods can be considered for representing the molars (e.g., parameterized surfaces; Kurtek et al. 2011). In contrast, we aim to propose an approach that does not rely on any diffeomorphic assumptions, allowing for a wider range of applications.

1.2. Overview of Shape and Topological Data Analysis

In classical geometric morphometrics, shapes are represented using prespecified points called landmarks (Kendall 1989). The manual landmarking of a collection of shapes requires domain

CONTACT Kun Meng kun_meng@brown.edu Division of Applied Mathematics, Brown University, 182 George Street, Providence, RI 02912.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

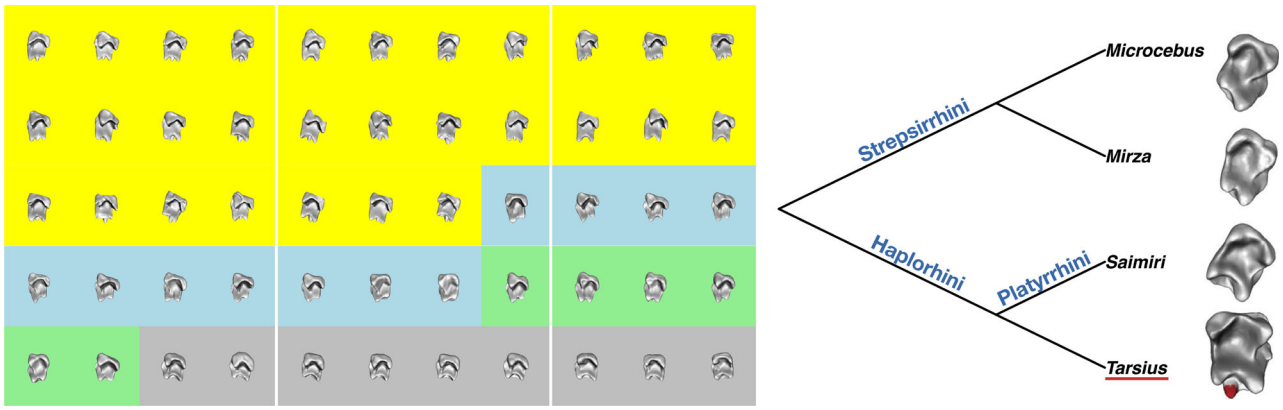

Figure 1. Left: Molars from two suborders of the primates: Haplorhini and Strepsirrhini. The Haplorhini suborder has genera *Tarsius* (yellow) and *Saimiri* (grey). The Strepsirrhini suborder has genera *Microcebus* (blue) and *Mirza* (green). Right: Relationship between the four primate genera. Tarsier molars exhibit additional high cusps (highlighted in red). A similar figure was published in Wang et al. (2021).

knowledge, can be very labor intensive, and is subject to bias (Boyer et al. 2011). Furthermore, an equal number of landmarks must be selected for each shape in a study in order to make comparisons (e.g., the Procrustes framework discussed in sec. 2.1 of Gao et al. 2019). This necessitates comprehensive information about entire collections of shapes for consistency, which can be difficult to obtain (e.g., landmarking cancer tumors, which can have very different morphology across a population of patients). Unfortunately, many datasets do not come with prespecified landmarks (e.g., Goswami 2015). Although many algorithms can automatically sample reasonable landmarks on shapes when their parameters are fine-tuned (e.g., Gao et al. 2019; Gao, Kovalsky, and Daubechies 2019), using a finite number of landmarks extracted from a continuum inevitably results in the loss of information. Diffeomorphism-based approaches (Dupuis, Grenander, and Miller 1998; Gao et al. 2019) are part of the "computational anatomy" that was historically studied by the "pattern theory school" pioneered by Ulf Grenander (Grenander and Miller 1998). They enable the comparison of (dis-)similarity between shapes with benefit of bypassing the need for landmarks. However, these approaches are based on the assumption that the shapes being compared are diffeomorphic to one another, making them unsuitable for many datasets (e.g., fruit fly wings in Miller 2015). Furthermore, parameterized curves and surfaces (PCS) provide a toolbox for assessing the heterogeneity of shapes with summary statistics that are invariant to reparameterizations (Kurtek et al. 2010, 2011, 2012). Despite their effectiveness in analyzing real data (e.g., DT-MRI brain fibers; Kurtek et al. 2012), PCS are based on assumptions about the diffeomorphism types of the shapes of interest. For example, Kurtek et al. (2011) focuses on surfaces that are diffeomorphic to the two-dimensional unit sphere.

TDA opens the door for landmark-free and diffeomorphism-free representations of shapes. Motivated by differential topology, Turner, Mukherjee, and Boyer (2014) proposed the persistent homology transform (PHT) with the capability to sufficiently encode all information within shapes (Ghrist, Levanger, and Mai 2018). To describe the PHT, we briefly provide some basics of TDA. One common statistical invariant in TDA is the persistence diagram (PD, Edelsbrunner and Harer 2010). When equipped with the Wasserstein distance, the collection of PDs, denoted as $\mathscr{D}$, is a Polish space (Mileyko, Mukherjee, and Harer 2011). Thus, probability measures can be applied, and the randomness of shapes can be represented using the probability measures on $\mathscr{D}$. The PHT takes values in $C(\mathbb{S}^{d-1};\mathscr{D}^d) = \{\text{continuous maps } F : \mathbb{S}^{d-1} \to \mathscr{D}^d\}$, where $\mathbb{S}^{d-1}$ denotes the sphere $\{x \in \mathbb{R}^d : \|x\| = 1\}$ and $\mathscr{D}^d$ is the d-fold Cartesian product of $\mathscr{D}$ (Turner, Mukherjee, and Boyer 2014, Lemma 2.1 and Definition 2.1). A single PD does not preserve all information of a shape (Crawford et al. 2020). In contrast, the PHT is injective, which means it preserves all the information of a shape. However, since $\mathscr{D}$ is not a vector space and the distances on $\mathscr{D}$ are abstract (e.g., the Wasserstein and bottleneck distances, Cohen-Steiner et al. 2007), many fundamental statistical concepts do not easily apply to summaries resulting from the PHT. For example, the definition of moments corresponding to probability measures on $\mathscr{D}$ (e.g., means) is highly nontrivial (Mileyko, Mukherjee, and Harer 2011). The difficulty in defining these concepts hinders the application of PHT-based statistical methods in $C(\mathbb{S}^{d-1};\mathscr{D}^d)$.

The smooth Euler characteristic transform (SECT, Crawford et al. 2020) offers an alternative summary statistic for shapes. The SECT not only preserves the information of shapes (Ghrist, Levanger, and Mai 2018, Corollary 1) but also represents shapes using continuous functions instead of PDs. More precisely, the values of the SECT are maps from the sphere $\mathbb{S}^{d-1}$ to a separable Banach space $\mathcal{B} \overset{\text{def}}{=} C([0,T])$, the collection of continuous functions on a compact interval $[0,T]$ (values of T will be given in (3.1)). Hence, for any shape K, its SECT, denoted as $\{\text{SECT}(K)(v)\}_{v\in\mathbb{S}^{d-1}}$, lies in $\mathcal{B}^{\mathbb{S}^{d-1}} = \{\text{maps } F : \mathbb{S}^{d-1} \to \mathcal{B}\}$. Specifically, $\text{SECT}(K)(v)$ belongs to $\mathcal{B}$ for each $v \in \mathbb{S}^{d-1}$. As a result, the randomness of shapes K is represented via the SECT by a collection of $\mathcal{B}$-valued random variables. Probability theory in separable Banach spaces is better developed than in $\mathscr{D}$ (e.g., Hairer 2009). In particular, a $\mathcal{B}$-valued random variable is a stochastic process with its sample paths in $\mathcal{B}$. As we will demonstrate in Section 3, $\mathcal{B}$ here can be replaced with a reproducing kernel Hilbert space (RKHS). The theory of stochastic processes has evolved over a century and FDA is a well-developed branch of statistics. Consequently, a myriad of tools are available to underpin both the randomness of shapes and the statistical inference on shapes.

From an application perspective, Crawford et al. (2020) applied the SECT to magnetic resonance images taken from tumors in a cohort of glioblastoma multiforme (GBM) patients. Using summary statistics derived from the SECT as predictors within Gaussian process regression, the authors demonstrated that the SECT can predict clinical outcomes more effectively than existing tumor shape quantification approaches and common molecular assays. The relative performance of the SECT in the GBM study suggests a promising future for its utility in medical imaging and broader statistical applications related to shape analyses. Similarly, Wang et al. (2021) used derivatives of the Euler characteristic transform (ECT) as predictors in statistical models for subimage analysis. This analysis is akin to variable selection, aiming to identify physical features that are important for distinguishing between two classes of shapes. Lastly, Marsh et al. (2022) highlighted that the SECT outperforms the standard measures employed in organoid morphology.

1.3. Overview of Contributions and Article Organization

Our goal is to address the hypothesis testing question posed in Section 1.1 by employing a landmark-free and diffeomorphism-free approach, which opens up possibilities for further applications in the future. We formulate the question more generically here. Let $\mathbb{P}^{(1)}$ and $\mathbb{P}^{(2)}$ be two distributions that generate two collections of random shapes, $\{K_i^{(1)}\}_{i=1}^n$ and $\{K_i^{(2)}\}_{i=1}^n$. Detecting whether there is a significant difference between $\{K_i^{(1)}\}_{i=1}^n$ and $\{K_i^{(2)}\}_{i=1}^n$ is equivalent to rejecting the hypothesis $\mathbb{P}^{(1)} = \mathbb{P}^{(2)}$. Since each shape $K_i^{(j)}$ is random, $\mathrm{SECT}(K_i^{(j)})$ is a random variable taking values in a vector space (as discussed in Section 1.2) and can be decomposed as follows (see Theorem 5.1 for a rigorous version)

$$\mathrm{SECT}(K_i^{(j)}) = m^{(j)} + \text{random terms}, \quad \text{for } j \in \{1, 2\}, \tag{1.1}$$

where $m^{(j)}$ denotes the mean of $\mathrm{SECT}(K_i^{(j)})$ with respect to the distribution $\mathbb{P}^{(j)}$. The random terms in (1.1) can be characterized by the Karhunen–Loève (KL) expansion (Hsing and Eubank 2015, sec. 7.3). To reject $\mathbb{P}^{(1)} = \mathbb{P}^{(2)}$, it suffices to reject $m^{(1)} = m^{(2)}$. That is, the question posed in Section 1.1 can be addressed by testing for the equality of two means. The important component of the test is the variance represented by the random terms in (1.1). In Section 5, we formulate this test as a two-sample problem in the fdANOVA literature (Zhang 2013, sec. 5.2). In addition, using the KL expansion, we provide a χ^2-statistic in Section 5 to test the hypothesis. Throughout the article, our focus is on the two-sample problem. However, one may also consider employing the one-way fdANOVA to compare the means of three or more groups of shapes. The theoretical foundation and numerical experiments for this aspect are left for future research.

To develop our framework, we have to address the following mathematical foundation related questions: (i) *What underlying probability spaces allow the randomness of shapes and their corresponding SECT?* and (ii) *Are the conditions of the KL expansion satisfied in our setting?* We answer these questions in Sections 3 and 4—we model the randomness of shapes via the SECT using RKHS-valued random fields. The "theory of random sets" is a well-established framework for characterizing set-valued random variables (Molchanov 2005). However, its application to persistent homology-based statistics (e.g., the SECT) remains underexplored. In this article, we introduce a new probability space to characterize the randomness of shapes in a manner compatible with the SECT.

We first propose a collection of shapes as our sample space on which the SECT is well-defined. We then demonstrate that every shape in this collection has its SECT in $C(\mathbb{S}^{d-1}; \mathcal{H}) = \{\text{continuous maps } F : \mathbb{S}^{d-1} \to \mathcal{H}\}$, where $\mathcal{H} = H_0^1([0, T])$ is not only a Sobolev space (Brezis 2011) but also an RKHS (reasons for using $[0, T]$ instead of $(0, T)$ for $H_0^1([0, T])$ are in Appendix A.1). Importantly, $C(\mathbb{S}^{d-1}; \mathcal{H})$ is a separable Banach space (Theorem C.1) and, hence, a Polish space. It helps construct a probability space to characterize the distributions of shapes. Building on this probability space, we define the mean and covariance of the SECT. Using the Sobolev embedding, we present some properties of the mean and covariance, which pave the way for the KL expansion of the SECT.

Traditionally, the statistical inference on shapes in TDA is conducted in the persistence diagram space $\mathscr{D}$, which is unsuitable for exponential family-based distributions and requires any corresponding statistical inference to be highly nonparametric (Fasy et al. 2014; Robinson and Turner 2017). The PHT-based statistical inference in $C(\mathbb{S}^{d-1}; \mathscr{D}^d)$ is even more difficult. With the KL expansion of the SECT, we propose a χ^2-statistic for testing hypotheses on shapes. Beyond the mathematical foundations, we also provide simulation studies to illustrate the performance of our proposed hypothesis testing method. Lastly, we apply our proposed framework to answer the motivating question raised in Section 1.1.

We organize this article as follows. In Section 2, we provide the mathematical preparations. In Section 3, we define the SECT for a specific collection of shapes, highlighting its properties. In Section 4, we construct a probability space to model shape distributions. In Section 5, we propose the KL expansion of the SECT, leading to a statistic for hypothesis testing. In Section 6, we conduct simulation studies to evaluate our method. In Section 7, we apply our method to real data. In Section 8, we conclude the article. The Appendix provides the proofs of theorems, further data analysis, and future research topics.

2. Notations and Mathematical Preparations

To model the shapes discussed in our motivating question from Section 1.1, we need certain preparations regarding (i) topology and (ii) function spaces.

Topology. The first question we must address is: *What are the "shapes" in our framework?* Ghrist, Levanger, and Mai (2018) and Curry, Mukherjee, and Turner (2022) applied o-minimal structures (van den Dries 1998) to prove the injectivity of the PHT. Subsequent to this, o-minimal structures have been applied in many TDA studies to model shapes (e.g., Jiang, Kurtek, and Needham 2020; Kirveslahti and Mukherjee 2023). To stay consistent with the existing literature, we also model shapes using o-minimal structures. An o-minimal structure is a sequence $\mathcal{S} = \{\mathcal{S}_n\}_{n\geq 1}$ of subset collections $\mathcal{S}_n \subseteq 2^{\mathbb{R}^n}$ satisfying six set-theoretical axioms, where $2^{\mathbb{R}^n}$ denotes the power set of $\mathbb{R}^n$. The

precise definition of o-minimal structures is available in van den Dries (1998) and is provided in Appendix A.3 for the reader's convenience.

A typical example of o-minimal structures is the collection of *semialgebraic sets*. Specifically, a set $K \subseteq \mathbb{R}^n$ is semialgebraic if it can be expressed as a finite union of sets of the form $\{x \in \mathbb{R}^n \mid p(x) = 0, q_1(x) > 0, \ldots, q_k(x) > 0\}$, where $p, q_1, \ldots, q_k$ are polynomial functions on $\mathbb{R}^n$. If we define $\mathcal{S}_n$ as the collection of semialgebraic subsets of $\mathbb{R}^n$, then $\mathcal{S} = \{\mathcal{S}_n\}_{n\geq 1}$ is an o-minimal structure (van den Dries 1998, chap. 2). The unit sphere $\mathbb{S}^{d-1}$, open ball $B(0,R) = \{x \in \mathbb{R}^d \mid \|x\|^2 < R^2\}$ for any $R > 0$, and all polyhedra (e.g., polygon meshes in computer graphics) are semialgebraic, given that they can be represented using either the polynomial $\|x\|^2$ or affine functions. We assume the following:

Assumption 1. The o-minimal structure $\mathcal{S}$ of interest contains all semialgebraic sets.

Hereafter, a "shape" refers to a compact set $K \in \bigcup_{n\geq 1} \mathcal{S}_n$ for a prespecified o-minimal structure $\mathcal{S} = \{\mathcal{S}_n\}_{n\geq 1}$ satisfying Assumption 1. Assumption 1 incorporates many common shapes (e.g., balls and polyhedra) in our framework. More importantly, it implies the subsequent Theorem 2.1 through the "triangulation theorem" (van den Dries 1998, chap. 8). Although the definition of an o-minimal structure $\mathcal{S}$ is highly abstract (see Appendix A.3), each compact set in $\mathcal{S}$ resembles a polyhedron, which is precisely stated as follows.

Theorem 2.1. Suppose $\mathcal{S} = \{\mathcal{S}_n\}_{n\geq 1}$ is an o-minimal structure satisfying Assumption 1 and $K \in \bigcup_{n\geq 1} \mathcal{S}_n$. If K is compact, there exists a finite simplicial complex S such that the polyhedron $|S| \overset{\text{def}}{=} \bigcup_{s\in S} s$ is homeomorphic to K, where each $s \in S$ denotes a simplex.

Herein, a finite simplicial complex S is a finite collection of simplexes. Each face of a simplex $s \in S$ also belongs to S (i.e., S is a so-called "closed complex" referred to in chap. 8 of van den Dries 1998). Theorem 2.1 directly results from the "triangulation theorem" (van den Dries 1998); hence, its proof is omitted. For the dth component $\mathcal{S}_d$ of $\mathcal{S} = \{\mathcal{S}_n\}_{n\geq 1}$, Theorem 2.1 indicates that the compact sets $K \in \mathcal{S}_d$ are subsets of $\mathbb{R}^d$ that are homeomorphic to polyhedra. Theorem 2.1 also implies that the homology groups of each compact $K \in \mathcal{S}_d$ are well-defined and finitely generated; hence, the Betti numbers and Euler characteristic of K are well-defined and finite (Hatcher 2002, chap. 2).

Function Spaces. We apply the following notations throughout this article:
(i) For any normed space $\mathcal{V}$, let $\|\cdot\|_{\mathcal{V}}$ denote its norm. Denote $\|\cdot\|_{\mathbb{R}^d}$ as $\|\cdot\|$ for succinctness.
(ii) Let X be a compact metric space equipped with metric d_X, and let $\mathcal{V}$ denote a normed space. $C(X;\mathcal{V})$ is the collection of continuous maps from X to $\mathcal{V}$. Furthermore, $C(X;\mathcal{V})$ is a normed space equipped with $\|f\|_{C(X;\mathcal{V})} = \sup_{x\in X} \|f(x)\|_{\mathcal{V}}$. The Hölder space $C^{0,\frac{1}{2}}(X;\mathcal{V})$ is defined as $\left\{f \in C(X;\mathcal{V}) \,\middle|\, \sup_{x,y\in X, x\neq y} \left(\frac{\|f(x)-f(y)\|_{\mathcal{V}}}{\sqrt{d_X(x,y)}}\right) < \infty\right\}$. Here, $C^{0,\frac{1}{2}}(X;\mathcal{V})$ is a normed space equipped with $\|f\|_{C^{0,\frac{1}{2}}(X;\mathcal{V})} = \|f\|_{C(X;\mathcal{V})} + \sup_{x,y\in X, x\neq y} \left(\frac{\|f(x)-f(y)\|_{\mathcal{V}}}{\sqrt{d_X(x,y)}}\right)$. Obviously, $C^{0,\frac{1}{2}}(X;\mathcal{V}) \subseteq C(X;\mathcal{V})$. For simplicity, we denote $C(X) = C(X;\mathbb{R})$ and $C^{0,\frac{1}{2}}(X) = C^{0,\frac{1}{2}}(X;\mathbb{R})$. For a given $T > 0$ (e.g., see (3.1)), we denote $C([0,T])$ as $\mathcal{B}$.
(iii) The inner product of $\mathcal{H} = H_0^1([0,T]) = \{f \in L^2([0,1]) \mid f' \in L^2([0,T])$ and $f(0) = f(T) = 0\}$ is defined as $\langle f, g\rangle = \int_0^T f'(t)g'(t)\,dt$ (Brezis 2011, chap. 8.3, Remark 17).
(iv) Suppose (Y, d_Y) is a metric space (not necessarily compact). Both $\mathscr{B}(Y)$ and $\mathscr{B}(d_Y)$ denote the Borel algebra generated by the metric topology corresponding to d_Y.
(v) $\{F(z)\}_{z\in Z}$ denotes a function F defined on the set Z.
The following inequalities are useful for deriving many results presented in this article

$$\|f\|_{\mathcal{B}} \leq \|f\|_{C^{0,\frac{1}{2}}([0,T])} \leq \tilde{C}_T \|f\|_{\mathcal{H}}, \quad \text{for all } f \in \mathcal{H}, \tag{2.1}$$

where $\tilde{C}_T$ is a constant depending only on T. The first inequality in (2.1) results from the definition of $\|\cdot\|_{C^{0,\frac{1}{2}}([0,T])}$, while the second inequality is from Brezis (2011) (Corollary 9.14; also see Appendix L.2). Equation (2.1) implies the following Sobolev embedding

$$H_0^1([0,T]) \overset{\text{def}}{=} \mathcal{H} \subseteq C^{0,\frac{1}{2}}([0,T]) \subseteq \mathcal{B} \overset{\text{def}}{=} C([0,T]). \tag{2.2}$$

3. Smooth Euler Characteristic Transform

In this section, we give the background on the SECT and propose corresponding mathematical foundations. Notably, we specify the "sample space"—a collection of shapes on which the SECT is well-defined. The SECT of the shapes in this sample space has properties that are suitable for the probability theory developed in Section 4. The molars in the motivating question from Section 1.1 will be modeled as elements of the sample space.

Suppose an o-minimal structure $\mathcal{S} = \{\mathcal{S}_n\}_{n\geq 1}$ satisfying Assumption 1 is given, and we focus on shapes in d-dimensional space $\mathbb{R}^d$. We assume the shape $K \in \mathcal{S}_d$ is compact and $K \subseteq B(0,R) = \{x \in \mathbb{R}^d : \|x\| < R\}$, for example, the $K \subseteq \mathbb{R}^2$ in Figure 2 or the surfaces of the mandibular molars in $\mathbb{R}^3$ as presented by Figure 1. For each direction $v \in \mathbb{S}^{d-1}$, we define a filtration $\{K_t^v\}_{t\in[0,T]}$ of sublevel sets by the following (see Figure 2 for an illustration)

$$K_t^v \overset{\text{def}}{=} \{x \in K \mid x \cdot v \leq t - R\}, \quad \text{for all } t \in [0,T], \quad \text{where } T \overset{\text{def}}{=} 2R. \tag{3.1}$$

We then have the following Euler characteristic curve (ECC, denoted as χ_t^v) in direction v

$$\chi_t^v(K) \overset{\text{def}}{=} \text{the Euler characteristic of } K_t^v = \chi(K_t^v) = \sum_{k=0}^{d-1} (-1)^k \cdot \beta_k(K_t^v), \tag{3.2}$$

for $t \in [0,T]$, where $\beta_k(K_t^v)$ is the kth Betti number of K_t^v. The sum in (3.2) ends at $d-1$ because higher homology groups are trivial (Curry, Mukherjee, and Turner 2022, sec. 4). If K_t^v

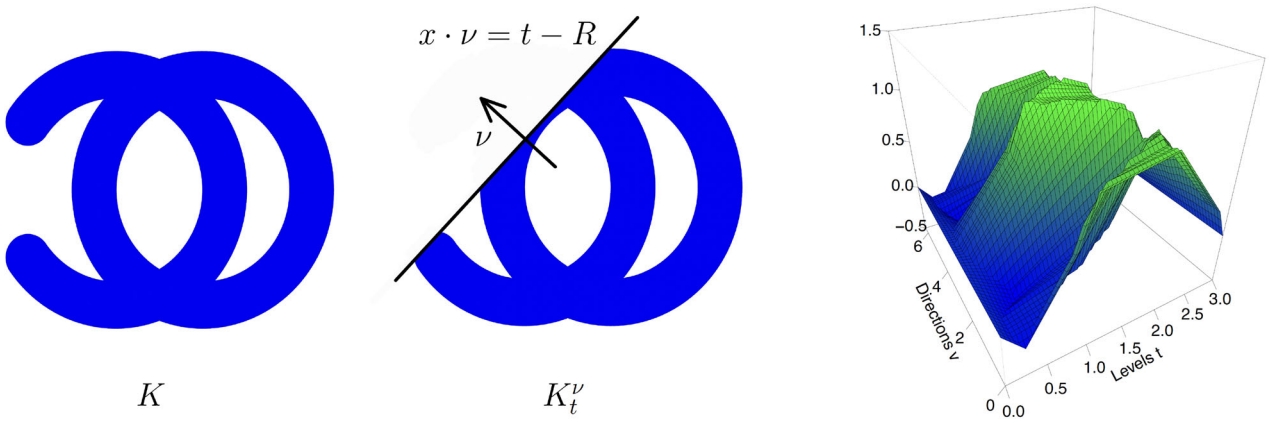

Figure 2. Consider the two-dimensional shape $K \in \mathcal{S}_2$ in the left panel. For each pair of ν and t, the equation $x \cdot \nu = t - R$ represents a straight line (or a hyperplane in a high-dimensional space). The subset K_t^ν denotes the region below this line. Let $\phi_\nu(x) = x \cdot \nu + R$, then $K_t^\nu = \{x \in K \,|\, \phi_\nu(x) \le t\}$. The right panel presents the function $(\nu, t) \mapsto \mathrm{SECT}(K)(\nu, t)$, where $\nu \in \mathbb{S}^1$ is identified by $\theta \in [0, 2\pi]$ through $\nu = (\cos\theta, \sin\theta)$. Procedures for generating the shape K and the right panel are given in Appendix D.1.

is a triangle mesh, $\chi(K_t^\nu) = \#V - \#E + \#F$, where $\#V$, $\#E$, and $\#F$ denote the number of vertices, edges, and faces of the mesh, respectively. Due to Theorem 2.1, the compactness of K guarantees that the Betti numbers in (3.2) are well-defined and finite.

The Euler characteristic transform (ECT) defined as $\mathrm{ECT}(K) : \mathbb{S}^{d-1} \to \mathbb{Z}^{[0,T]}$, $\nu \mapsto \{\chi_t^\nu(K)\}_{t\in[0,T]}$ was proposed by Turner, Mukherjee, and Boyer (2014) as an alternative to the PHT. Based on the ECT, Crawford et al. (2020) further proposed the SECT as follows

$$\begin{aligned} &\mathrm{SECT}(K): \ \mathbb{S}^{d-1} \to \mathbb{R}^{[0,T]}, \\ &\nu \mapsto \mathrm{SECT}(K)(\nu) = \{\mathrm{SECT}(K)(\nu, t)\}_{t\in[0,T]}, \\ &\text{where } \ \mathrm{SECT}(K)(\nu, t) \overset{\text{def}}{=} \int_0^t \chi_\tau^\nu(K)\, d\tau - \frac{t}{T}\int_0^T \chi_\tau^\nu(K)\, d\tau. \end{aligned} \tag{3.3}$$

A visualization of the function $(\nu, t) \mapsto \mathrm{SECT}(K)(\nu, t)$ is presented in Figure 2. The following lemma implies that the Lebesgue integrals in (3.3) are well-defined.

Lemma 3.1. For any fixed $K \in \mathcal{S}_d$ and $\nu \in \mathbb{S}^{d-1}$, the function $t \mapsto \chi(K_t^\nu)$ is piecewise constant with only finitely many discontinuities.

Through the "cell decomposition theorem" (van den Dries 1998, chap. 3), Lemma 3.1 directly follows from either Lemma 3.4 of Curry, Mukherjee, and Turner (2022) or "(2.10) Proposition" in Chapter 4 of van den Dries (1998). Hence, the proof of Lemma 3.1 is omitted.

To investigate the distribution of $\mathrm{SECT}(K)$ over different shapes K, we introduce the following condition to restrict our attention to a subset of $\mathcal{S}_d$.

Condition 3.1. Let $K \in \mathcal{S}_d$. The condition is that K satisfies the following inequality

$$\sup_{k\in\{0,\ldots,d-1\}} \left[\sup_{\nu\in\mathbb{S}^{d-1}} \left(\#\left\{ \xi \in \mathrm{Dgm}_k(K;\phi_\nu) \,\middle|\, \mathrm{pers}(\xi) > 0 \right\} \right) \right] \le \frac{M}{d}, \tag{3.4}$$

where $\mathrm{Dgm}_k(K;\phi_\nu)$ is the PD of K associated with the function $\phi_\nu(x) = x \cdot \nu + R$ (also see Figure 2), $\mathrm{pers}(\xi)$ is the persistence of the homology feature ξ, $\#\{\cdot\}$ denotes the cardinality of a multiset, and $M > 0$ is a sufficiently large prespecified number.

Condition 3.1 involves technicalities from computational topology (Edelsbrunner and Harer 2010). To maintain the flow of the article, we relegate the details of this condition, as well as the definitions of $\mathrm{Dgm}_k(K;\phi_\nu)$ and $\mathrm{pers}(\xi)$, to Appendix B. Heuristically, Condition 3.1 implies the existence of a uniform upper bound on the number of nontrivial homology features of K across all directions ν. Hereafter, we focus on shapes in the following collection

$$\mathcal{S}_{R,d}^M \overset{\text{def}}{=} \left\{ K \in \mathcal{S}_d \,\middle|\, K \subseteq B(0,R) \text{ is compact and satisfies Condition 3.1 with fixed } M > 0 \right\}.$$

Our proposed collection $\mathcal{S}_{R,d}^M$ is suitable for modeling shapes in many applications. For example, the surfaces of the molars in Figure 1 are compact subsets of $\mathbb{R}^3$, bounded by a common open ball, and can be approximately represented by triangle meshes (hence, modeled by an o-minimal structure satisfying Assumption 1). In addition, the four genera of primates in Figure 1 share a phylogentic relationship which implies that their molars have common baseline features and satisfy Condition 3.1 with a shared upper bound M. In each application, the dimension d and radius R of the ball $B(0,R)$ can easily be determined based on observed shapes. Although our mathematical framework requires the existence of such an M in (3.4), the value of M is not needed for our statistical methodology (see Section 5). Thus, Condition 3.1 does not hinder our proposed statistical methodology.

Lemma 3.1 implies that the function $\{\chi_t^\nu(K)\}_{t\in[0,T]}$ of t belongs to $L^1([0,T])$. Therefore, the function $\mathrm{SECT}(K)(\nu) = \{\mathrm{SECT}(K)(\nu,t)\}_{t\in[0,T]}$ of t is absolutely continuous on $[0,T]$. Furthermore, we have the following regularity result of the Sobolev type.

Lemma 3.2. For any $K \in \mathcal{S}_{R,d}^M$ and $\nu \in \mathbb{S}^{d-1}$, the function $\mathrm{SECT}(K)(\nu)$ belongs to $\mathcal{H}$.

Lemma 3.2 is a special case of Lemma C.3. It indicates $\mathrm{SECT}(\mathcal{S}^M_{R,d}) \subseteq \mathcal{H}^{\mathbb{S}^{d-1}} = \{\text{all maps } F : \mathbb{S}^{d-1} \to \mathcal{H}\}$, which is enhanced by the following result.

Theorem 3.2. For each $K \in \mathcal{S}^M_{R,d}$, we have: (i) There exists a constant $C^*_{M,R,d}$ depending only on M, R, and d such that the following inequality holds for any directions $\nu_1, \nu_2 \in \mathbb{S}^{d-1}$,

$$\begin{aligned} &\| \mathrm{SECT}(K)(\nu_1) - \mathrm{SECT}(K)(\nu_2) \|_{\mathcal{H}} \\ &\quad \le C^*_{M,R,d} \cdot \sqrt{\|\nu_1 - \nu_2\| + \|\nu_1 - \nu_2\|^2}; \end{aligned} \tag{3.5}$$

and (ii) $\mathrm{SECT}(K) \in C^{0,\frac{1}{2}}(\mathbb{S}^{d-1}; \mathcal{H})$, where $\mathbb{S}^{d-1}$ is equipped with the geodesic distance $d_{\mathbb{S}^{d-1}}$.

Results complementary to Theorem 3.2 can be found in Theorem C.3, which imply that the function $(\nu, t) \mapsto \mathrm{SECT}(K)(\nu, t)$ belongs to $C^{0,\frac{1}{2}}(\mathbb{S}^{d-1} \times [0, T])$. Theorem 3.2(i) is an analog of Lemma 2.1 in Turner, Mukherjee, and Boyer (2014). Theorem 3.2(ii) implies $\mathrm{SECT}(\mathcal{S}^M_{R,d}) \subseteq C^{0,\frac{1}{2}}(\mathbb{S}^{d-1}; \mathcal{H}) \subseteq C(\mathbb{S}^{d-1}; \mathcal{H}) \subseteq \mathcal{H}^{\mathbb{S}^{d-1}}$. As a result, (3.3) defines the following map

$$\mathrm{SECT} : \mathcal{S}^M_{R,d} \to C(\mathbb{S}^{d-1}; \mathcal{H}), \quad K \mapsto \mathrm{SECT}(K). \tag{3.6}$$

In Appendix D.1, we provide detailed proof-of-concept examples (similar to Figure 2) to visually illustrate the SECT and support the regularity results in Theorems 3.2 and C.3.

Corollary 1 of Ghrist, Levanger, and Mai (2018) implies the following result, which shows that the SECT preserves all the information of shapes $K \in \mathcal{S}^M_{R,d}$.

Theorem 3.3. The map SECT defined in (3.6) is injective for all dimensions d.

The map $\mathrm{SECT} : \mathcal{S}^M_{R,d} \to C(\mathbb{S}^{d-1}; \mathcal{H})$ is injective, but not surjective. Specifically, Theorem 3.2 suggests that the image of SECT does not lie outside of $C^{0,\frac{1}{2}}(\mathbb{S}^{d-1}; \mathcal{H})$. An explicit characterization of the image $\mathrm{SECT}(\mathcal{S}^M_{R,d})$ remains a topic for future research.

Inspired by Theorem 3.3, one may consider reconstructing a shape K from either the $\mathrm{SECT}(K)$ or the $\mathrm{ECT}(K)$. From a theoretical standpoint, a shape K can be reconstructed using the "Schapira's inversion formula" (Schapira 1995). Further details are available in Ghrist, Levanger, and Mai (2018). From an algorithmic perspective, the proof of Theorem 3.1 in Turner, Mukherjee, and Boyer (2014) offers an algorithm to reconstruct low-dimensional meshes from their ECT. Nevertheless, effective algorithmic approaches to reconstructing shapes are still underdeveloped. Challenges in reconstructing shapes are extensively discussed in Fasy et al. (2018). A comprehensive exploration of the reconstruction using SECT is also left for future research.

Together with (3.6), Theorem 3.3 allows us to represent each $K \in \mathcal{S}^M_{R,d}$ by $\mathrm{SECT}(K) \in C(\mathbb{S}^{d-1}; \mathcal{H})$. This perspective aids us in modeling the randomness of shapes using probability measures on the separable Banach space $C(\mathbb{S}^{d-1}; \mathcal{H})$. Here, we prefer $C(\mathbb{S}^{d-1}; \mathcal{H})$ over $\frac{1}{2}$-Hölder space $C^{0,\frac{1}{2}}(\mathbb{S}^{d-1}; \mathcal{H})$. This is because $\frac{1}{2}$-Hölder spaces are typically not separable (Hairer 2009, Remark 4.21). The separability condition is essential for probability measures on Banach spaces to exhibit nonpathological behavior (Hairer 2009, sec. 4).

4. Probabilistic Distributions over the SECT

To address the motivating question outlined in Section 1.1 using hypothesis testing, we need to view the observed shapes (e.g., the molars in Figure 1) as shape-valued random variables. In this section, we construct a probability space to model the randomness of shapes and make the SECT a random variable (in the measurable sense) taking values in $C(\mathbb{S}^{d-1}; \mathcal{H})$. More importantly, this probability space helps justify the KL expansion of the SECT, which lays the foundations for our hypothesis testing method in Section 5.

Probability Space. Suppose $\mathcal{S}^M_{R,d}$ is equipped with a σ-algebra $\mathscr{F}$. A distribution of shapes K across $\mathcal{S}^M_{R,d}$ is represented by a probability measure $\mathbb{P} = \mathbb{P}(dK)$ on $\mathscr{F}$. Then, $(\mathcal{S}^M_{R,d}, \mathscr{F}, \mathbb{P})$ is a probability space. For each fixed (ν, t), the integer-valued map $\chi^\nu_t : K \mapsto \chi^\nu_t(K)$ is defined on $\mathcal{S}^M_{R,d}$. Hereafter, we assume the following:

Assumption 2. For each fixed $(\nu, t) \in \mathbb{S}^{d-1} \times [0, T]$, the map $\chi^\nu_t : (\mathcal{S}^M_{R,d}, \mathscr{F}) \to (\mathbb{R}, \mathscr{B}(\mathbb{R}))$ is a measurable function and, hence, a real-valued random variable.

A σ-algebra $\mathscr{F}$ satisfying Assumption 2 exists. Here, we construct a metric ρ on $\mathcal{S}^M_{R,d}$ and show that the Borel algebra $\mathscr{B}(\rho)$ induced by ρ satisfies Assumption 2. We define

$$\rho(K_1, K_2) \overset{\mathrm{def}}{=} \sup_{\nu \in \mathbb{S}^{d-1}} \left\{ \left(\int_0^T \left| \chi^\nu_\tau(K_1) - \chi^\nu_\tau(K_2) \right|^2 d\tau \right)^{1/2} \right\}, \quad \text{for all } K_1, K_2 \in \mathcal{S}^M_{R,d}. \tag{4.1}$$

Theorem 4.1. The map ρ is a metric on $\mathcal{S}^M_{R,d}$. Assumption 2 is satisfied if $\mathscr{F} = \mathscr{B}(\rho)$.

Under Assumption 2, the ECC $\{\chi^\nu_t\}_{t \in [0,T]}$, for each $\nu \in \mathbb{S}^{d-1}$, is a stochastic process defined on the probability space $(\mathcal{S}^M_{R,d}, \mathscr{F}, \mathbb{P})$. Since each sample path $\{\chi^\nu_t(K)\}_{t \in [0,T]}$ has finitely many discontinuities (Lemma 3.1), $\int_0^t \chi^\nu_\tau(K)\, d\tau$ for each $t \in [0, T]$ is a Riemann integral, which is equal to the limit of Riemann sum $\int_0^t \chi^\nu_\tau(K)\, d\tau = \lim_{n \to \infty} \left\{ \frac{t}{n} \sum_{l=1}^n \chi^\nu_{\frac{tl}{n}}(K) \right\}$. Given that each $\chi^\nu_{\frac{tl}{n}}$ is a random variable under Assumption 2, the limit of the Riemann sum for each $t \in [0, T]$ is a random variable as well. Therefore, for each $\nu \in \mathbb{S}^{d-1}$, $\{\int_0^t \chi^\nu_\tau d\tau\}_{t \in [0,T]}$ with $\int_0^t \chi^\nu_\tau d\tau : K \mapsto \int_0^t \chi^\nu_\tau(K) d\tau$ is a stochastic process. Then, under Assumption 2, (3.3) defines the following stochastic process on $(\mathcal{S}^M_{R,d}, \mathscr{F}, \mathbb{P})$ for each $\nu \in \mathbb{S}^{d-1}$

$$\mathrm{SECT}(\nu) \overset{\mathrm{def}}{=} \left\{ \int_0^t \chi^\nu_\tau d\tau - \frac{t}{T} \int_0^T \chi^\nu_\tau d\tau \overset{\mathrm{def}}{=} \mathrm{SECT}(\nu, t) \right\}_{t \in [0,T]}. \tag{4.2}$$

Precisely, for each fixed ν, we have the stochastic process $\mathrm{SECT}(\nu) : K \mapsto \mathrm{SECT}(K)(\nu) = \{\mathrm{SECT}(K)(\nu, t)\}_{t \in [0,T]}$ defined on $(\mathcal{S}^M_{R,d}, \mathscr{F}, \mathbb{P})$; and, for each fixed (ν, t), we have the real-valued random variable $\mathrm{SECT}(\nu, t) : K \mapsto \mathrm{SECT}(K)(\nu, t)$ defined on $(\mathcal{S}^M_{R,d}, \mathscr{F}, \mathbb{P})$.

Since $\mathcal{H}$ is an RKHS (Appendix A.1), Lemma 3.2 and Theorem 3.2, together with Theorem 7.1.2 of Hsing and Eubank (2015), imply the following. Its proof is omitted.

Theorem 4.2. (i) For each $\nu \in \mathbb{S}^{d-1}$, $\mathrm{SECT}(\nu)$ is a real-valued stochastic process with sample paths in $\mathcal{H}$. Equivalently, $\mathrm{SECT}(\nu)$ is a random variable taking values in $(\mathcal{H}, \mathscr{B}(\mathcal{H}))$. (ii) The map SECT defined in (3.6) is a random variable taking values in $C(\mathbb{S}^{d-1}; \mathcal{H})$.

Using Theorem 4.2 in conjunction with Theorem 3.3, we can represent random shapes (which model the surfaces of the mandibular molars in Figure 1) as $C(\mathbb{S}^{d-1}, \mathcal{H})$-valued random variables. This representation through the SECT has no loss of information.

In Appendix D.2, we provide proof-of-concept examples to illustrate random shapes and their SECT representations visually. These examples relate the SECT to Fréchet regression (Petersen and Müller 2019), Wasserstein regression (Chen et al. 2023), and manifold learning (Dunson and Wu 2021; Meng and Eloyan 2021; Li, Mukhopadhyay, and Dunson 2022).

Mean and Covariance of the SECT. For deriving the KL expansion in Section 5, we define the mean and covariance of the SECT. To do so, we need the following lemma.

Lemma 4.1. For any probability measure $\mathbb{P}$ defined on the measurable space $(\mathcal{S}^M_{R,d}, \mathscr{F})$, we have $\mathbb{E}\left\{\sup_{\nu\in\mathbb{S}^{d-1}} \|\mathrm{SECT}(\nu)\|^2_{\mathcal{H}}\right\} = \int_{\mathcal{S}^M_{R,d}} \left\{\sup_{\nu\in\mathbb{S}^{d-1}} \|\mathrm{SECT}(K)(\nu)\|^2_{\mathcal{H}}\right\} \mathbb{P}(dK) < \infty$.

Lemma 4.1, together with (2.1), implies that $\mathbb{E}|\mathrm{SECT}(\nu,t)|^2 \le \tilde{C}^2_T \cdot \mathbb{E}\|\mathrm{SECT}(\nu)\|^2_{\mathcal{H}} < \infty$ for all $(\nu,t) \in \mathbb{S}^{d-1} \times [0,T]$. Then, we define the mean and covariance functions as follows

$$m_\nu(t) = \mathbb{E}\{\mathrm{SECT}(\nu,t)\} = \int_{\mathcal{S}^M_{R,d}} \mathrm{SECT}(K)(\nu,t)\,\mathbb{P}(dK),$$
$$\Xi_\nu(s,t) = \mathrm{cov}\Big(\mathrm{SECT}(\nu,s), \mathrm{SECT}(\nu,t)\Big), \quad \text{for } s,t \in [0,T] \text{ and } \nu \in \mathbb{S}^{d-1}. \tag{4.3}$$

Lemma C.4 provides several properties of the mean m_ν and covariance Ξ_ν that validate our KL expansion of $\mathrm{SECT}(\nu)$ in Section 5. Additionally, Lemma C.4 demonstrates that the mean $\boldsymbol{m} \overset{\mathrm{def}}{=} \{m_\nu\}_{\nu\in\mathbb{S}^{d-1}}$ of SECT belongs to $C(\mathbb{S}^{d-1}; \mathcal{H})$. A tentative discussion on the "pseudo-inverse" $\mathrm{SECT}^{-1}(\boldsymbol{m})$ is provided after Lemma C.4 in Appendix C.

In most shape analysis studies, data are preprocessed by alignment. In Appendix E, we introduce the "ECT alignment" as a preprocessing step before any statistical inference. Throughout the article, we assume that the data have been aligned using this method. The ECT alignment exploits rigid motions, does not rely on landmarks, and is equivalent to the alignment approach outlined in Wang et al. (2021) (Supplementary Section 4). The primary objective of the ECT alignment is to minimize the differences between two shapes that arise from rigid motions. For instance, the molars in Figure 1 were aligned using the ECT alignment. Furthermore, the ECT alignment is compatible with our SECT framework. Appendix E demonstrates that the ECT alignment does not alter the qualitative properties of SECT (e.g., the measurability, Sobolev-regularity, and $\frac{1}{2}$-Hölder continuity).

In applications, it is infeasible to sample infinitely many directions $\nu \in \mathbb{S}^{d-1}$ and levels $t \in [0,T]$. For given shapes K, we compute $\mathrm{SECT}(K)(\nu,t)$ for finitely many directions $\{\nu_1,\ldots,\nu_\Gamma\} \subseteq \mathbb{S}^{d-1}$ and levels $\{t_1,\ldots,t_\Delta\} \subseteq [0,T]$. To retain information about shapes K, one needs to properly set the numbers of directions and levels (i.e., Γ and Δ). From a theoretical viewpoint, Curry, Mukherjee, and Turner (2022) comprehensively discussed the number Γ of directions needed to recover shapes K from $\mathrm{ECT}(K)$ when K are "piecewise linearly embedded shapes with plausible geometric bounds." From the numerical perspective, we note the following: (i) Wang et al. (2021) provided detailed simulation studies on the choices of Γ and Δ in sub-image analysis, and a general guideline for setting Γ and Δ in practice was presented in Supplementary Table 1 therein; and (ii) in our Appendix K, we provide detailed numerical experiments on the tradeoffs between the choices of Γ and Δ, the statistical power of our proposed algorithms (Algorithms 1 and 2), and computational cost.

5. Testing Hypotheses on Shapes

In this section, we apply the probabilistic formulation from Section 4 and Lemma C.4 to test hypotheses on shapes. Suppose $\mathbb{P}^{(1)}$ and $\mathbb{P}^{(2)}$ are two distributions on the measurable space $(\mathcal{S}^M_{R,d}, \mathscr{F})$. Let $\mathbb{P}^{(1)} \otimes \mathbb{P}^{(2)}$ be the product probability measure defined on the product σ-algebra $\mathscr{F} \otimes \mathscr{F}$, satisfying $\mathbb{P}^{(1)} \otimes \mathbb{P}^{(2)}(A \times B) = \mathbb{P}^{(1)}(A) \cdot \mathbb{P}^{(2)}(B)$ for all $A, B \in \mathscr{F}$. To address the motivating question from Section 1.1, we test the following hypotheses

$$H_0^*: \ \mathbb{P}^{(1)} = \mathbb{P}^{(2)}, \quad \textit{vs.} \quad H_1^*: \ \mathbb{P}^{(1)} \neq \mathbb{P}^{(2)}, \tag{5.1}$$

for example, suppose $\mathbb{P}^{(1)}$ and $\mathbb{P}^{(2)}$ model the distributions of molars from two genera of primates (Figure 1). Rejecting the H_0^* in (5.1) helps distinguish the two genera of primates.

Define $m^{(j)}_\nu(t) = \int_{\mathcal{S}^M_{R,d}} \mathrm{SECT}(K)(\nu,t)\,\mathbb{P}^{(j)}(dK)$ for $j \in \{1,2\}$ as the mean functions corresponding to $\mathbb{P}^{(1)}$ and $\mathbb{P}^{(2)}$. To reject the null H_0^* in (5.1) (equivalently, distinguish two collections of shapes), it suffices to reject the null hypothesis H_0 in the following

$$\begin{aligned} &H_0: m^{(1)}_\nu(t) = m^{(2)}_\nu(t) \text{ for all } (\nu,t), \quad \text{versus} \\ &H_1: m^{(1)}_\nu(t) \neq m^{(2)}_\nu(t) \text{ for some } (\nu,t). \end{aligned} \tag{5.2}$$

Analysis of Variance for Functional Data (fdANOVA). Considering the hypotheses in (5.2) for all directions $\nu \in \mathbb{S}^{d-1}$ results in simultaneous multiple-comparisons and inflation of the Type I error. To address this issue, we focus on a specific direction, motivated by the observation that the null hypothesis H_0 in (5.2) is equivalent to $\sup_{\nu\in\mathbb{S}^{d-1}}\{\|m^{(1)}_\nu - m^{(2)}_\nu\|_{\mathcal{B}}\} = 0$. Hence, the direction of interest is defined as

$$\nu^* \overset{\mathrm{def}}{=} \arg\max_{\nu\in\mathbb{S}^{d-1}} \left\{\|m^{(1)}_\nu - m^{(2)}_\nu\|_{\mathcal{B}}\right\}. \tag{5.3}$$

Lemma C.4 and (2.2) imply $\{m^{(j)}_\nu\}^2_{j=1} \subseteq \mathcal{B}$ for all ν. Lemma C.4, together with (2.1), confirms the existence of a maximizer in (5.3). The maximizer in (5.3) may not be unique. If there are multiple maximizers, we arbitrarily choose one, as this choice

does not influence our framework. The null hypothesis H_0 in (5.2) is then equivalent to $\|m^{(1)}_{\nu^*} - m^{(2)}_{\nu^*}\|_{\mathcal{B}} = 0$, where the ν^* defined in (5.3) is called a *distinguishing direction*. Hereafter, we investigate the distribution of SECT(ν^*).

Based on the discussion above, testing the hypotheses in (5.2) is equivalent to testing $m^{(1)}_{\nu^*}(t) = m^{(2)}_{\nu^*}(t)$ for $t \in [0, T]$ using SECT(ν^*), which is a fdANOVA problem that has been well-studied in the literature (e.g., Zhang 2013, sec. 5.2). However, many state-of-the-art fdANOVA approaches are incompatible with SECT(ν^*). For example, the Gaussianity of SECT(ν^*) is not guaranteed (Remark C.1), and the "two-sample problem assumptions" in Section 5.2 of Zhang (2013) may not be satisfied. Besides, the L^2-norm-based test (Zhang and Chen 2007) and F-type test (Shen and Faraway 2004) are not preferred when the functional data are not Gaussian (Zhang 2013, chap. 5). Additionally, many fdANOVA methods are time-consuming. For example, tests based on random projections (TRP, Cuesta-Albertos and Febrero-Bande 2010) require the computation of (at least 30) L^2-projections for each observed function, followed by the application of appropriate ANOVA tests to these projections. To address the Gaussianity issue and achieve computational efficiency, we propose a method for fdANOVA using the KL expansion. Our test has a foundation that aligns with the probabilistic framework of SECT(ν^*) in Section 4; it is comparable with the existing methods in terms of size and power (see Appendix J); and it is also computationally efficient, allowing for the permutation test used with our method.

Karhunen–Loève Expansion. Let $\Xi^{(j)}_{\nu^*}(s, t)$ be the covariance function of the stochastic process SECT(ν^*) corresponding to $\mathbb{P}^{(j)}$, for $j \in \{1, 2\}$ (see (4.3)). Hereafter, we assume the following, which is true under the null hypothesis $H_0^* : \mathbb{P}^{(1)} = \mathbb{P}^{(2)}$ in (5.1).

Assumption 3 (Homoscedasticity). $\Xi_{\nu^*} \overset{\text{def}}{=} \Xi^{(1)}_{\nu^*} = \Xi^{(2)}_{\nu^*}$, where ν^* is defined in (5.3).

This is a standard assumption in the fdANOVA literature (e.g., Zhang 2013, sec. 5.2) and can be tested using the methods proposed by Jia Guo and Zhang (2019).

We define an integral operator on $L^2([0, T]^2)$ as $f \mapsto \int_0^T f(s) \cdot \Xi_{\nu^*}(s, \cdot)\, ds$. This operator is compact and self-adjoint (Hsing and Eubank 2015, Theorems 4.6.2 and Example 3.3.4). Moreover, the Hilbert-Schmidt theorem (Reed and Simon 1972, Theorem VI.16) suggests that there is a complete orthonormal basis $\{\phi_l\}_{l=1}^{\infty}$ for $L^2([0, T])$ so that (i) each ϕ_l is an eigenfunction with eigenvalue λ_l, (ii) $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$, and (iii) $\lim_{l\to\infty} \lambda_l = 0$. Lemma C.4 and Theorem 7.3.5 of Hsing and Eubank (2015) imply the following KL expansion:

Theorem 5.1 (Karhunen–Loève expansion). (i) For each fixed $j \in \{1, 2\}$, we have

$$\lim_{L\to\infty} \sup_{t\in[0,T]} \mathbb{E}^{(j)} \left[\text{SECT}(\nu^*, t) - m^{(j)}_{\nu^*}(t) - \sum_{l=1}^{L} \sqrt{\lambda_l} \cdot Z_l^{(j)} \cdot \phi_l(t) \right]^2 = 0, \tag{5.4}$$

where $Z_l^{(j)}(K) \overset{\text{def}}{=} \frac{1}{\sqrt{\lambda_l}} \int_0^T \{\text{SECT}(K)(\nu^*, t) - m^{(j)}_{\nu^*}(t)\} \cdot \phi_l(t)\, dt$ for $l = 1, 2, \ldots$, and $\mathbb{E}^{(j)}$ is the expectation associated with $\mathbb{P}^{(j)}$. For each $j \in \{1, 2\}$, the random variables $\{Z_l^{(j)}\}_{l=1}^{\infty}$ are defined on the probability space $(\mathcal{S}^M_{R,d}, \mathscr{F}, \mathbb{P}^{(j)})$, are mutually uncorrelated, and have mean 0 and variance 1. (ii) There exists $\mathcal{N} \in \mathscr{F} \otimes \mathscr{F}$ so that $\mathbb{P}^{(1)} \otimes \mathbb{P}^{(2)}(\mathcal{N}) = 0$ and

$$\begin{aligned} &\delta_l\left(K^{(1)}, K^{(2)}\right) \\ &\overset{\text{def}}{=} \frac{1}{\sqrt{2\lambda_l}} \int_0^T \left\{\text{SECT}(K^{(1)})(\nu^*, t) - \text{SECT}(K^{(2)})(\nu^*, t)\right\} \cdot \phi_l(t)\, dt \\ &\qquad = \theta_l + \left(\frac{Z_l^{(1)}(K^{(1)}) - Z_l^{(2)}(K^{(2)})}{\sqrt{2}} \right), \\ &\text{where } \theta_l \overset{\text{def}}{=} \frac{1}{\sqrt{2\lambda_l}} \int_0^T \left\{ m^{(1)}_{\nu^*}(t) - m^{(2)}_{\nu^*}(t) \right\} \cdot \phi_l(t)\, dt, \end{aligned} \tag{5.5}$$

for any $(K^{(1)}, K^{(2)}) \notin \mathcal{N}$ and each fixed $l = 1, 2, \ldots$. The null set $\mathcal{N}$ is allowed to be empty.

Using the KL expansion in (5.4), the random sampling of shapes may be considered, which is discussed in Appendix M.1 and left for future research.

Our Approach. Consider two independent collections of random shapes $\{K_i^{(j)}\}_{i=1}^n \overset{\text{iid}}{\sim} \mathbb{P}^{(j)}$, for $j \in \{1, 2\}$ (i.e., $\{(K_i^{(1)}, K_i^{(2)})\}_{i=1}^n \overset{\text{iid}}{\sim} \mathbb{P}^{(1)} \otimes \mathbb{P}^{(2)}$). The pairing in $(K_i^{(1)}, K_i^{(2)})$ is arbitrary for the following reasons: (i) pairs $(K_i^{(1)}, K_i^{(2)})$ and $(K_i^{(1)}, K_{i'}^{(2)})$ with $i \neq i'$ have the same distribution $\mathbb{P}^{(1)} \otimes \mathbb{P}^{(2)}$, and (ii) numerical experiments in Sections 6 and 7 demonstrate that the performance of our proposed algorithms is numerically invariant to shuffling the index i within each collection $\{K_i^{(j)}\}_{i=1}^n$. Without loss of generality, we assume that all the shapes have been aligned using the "ECT alignment" (Appendix E). Here, we present the theoretical foundation for employing $\{(K_i^{(1)}, K_i^{(2)})\}_{i=1}^n$ to test the hypotheses in (5.2). This foundation helps address the motivating question from Section 1.1.

Without loss of generality, we assume $(K_i^{(1)}, K_i^{(2)}) \notin \mathcal{N}$, for all $i = 1, 2, \ldots, n$, where $\mathcal{N}$ is the null set in Theorem 5.1 satisfying $\mathbb{P}^{(1)} \otimes \mathbb{P}^{(2)}(\mathcal{N}) = 0$. Then, we have

$$\xi_{l,i} \overset{\text{def}}{=} \delta_l\left(K_i^{(1)}, K_i^{(2)}\right) = \theta_l + \left(\frac{Z_l^{(1)}(K_i^{(1)}) - Z_l^{(2)}(K_i^{(2)})}{\sqrt{2}} \right), \tag{5.6}$$

where δ_l and θ_l are defined in (5.5). Theorem 5.1 implies that, for each fixed l, the random variables $\{\xi_{l,i}\}_{i=1}^n$ are iid across $i = 1, \ldots, n$ with mean θ_l and variance 1; for each fixed i, the random variables $\{\xi_{l,i}\}_{l=1}^{\infty}$ are mutually uncorrelated across $l = 1, 2, 3, \ldots$. The following lemma represents the null H_0 in (5.2) using the means $\{\theta_l\}_{l=1}^{\infty}$.

Lemma 5.1. The null H_0 in (5.2) is equivalent to $\theta_l = 0$ for all positive integers l.

Recall that $\lim_{l\to\infty} \lambda_l = 0$. When eigenvalues λ_l in the denominator of (5.5) are close to zero for large l, the estimated θ_l becomes unstable. Specifically, even if $m^{(1)}_{\nu^*}(t) \approx m^{(2)}_{\nu^*}(t)$, an extremely small λ_l can move the corresponding estimated θ_l far away from zero. Using the standard approach in principal

Table 1. Rejection rates (from 1000 experiments) for different indices ε (significance $\alpha = 0.05$).

Indices ε	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.08	0.10
Algorithm 1	0.118	0.161	0.315	0.519	0.785	0.910	0.975	0.990	1.000
Algorithm 2	0.046	0.054	0.162	0.343	0.612	0.789	0.931	0.994	1.000
Algorithm 3	0.050	0.050	0.111	0.185	0.335	0.535	0.739	0.983	0.999
FP	0.136	0.153	0.308	0.539	0.810	0.924	0.986	0.997	1.000
TRP-WTPS	0.075	0.091	0.261	0.515	0.790	0.929	0.980	0.997	1.000

NOTE: Appendix J provides a comparison of Algorithms 1, 2, and 3 to other existing fdANOVA methods.

component analysis, we focus on $\{\theta_l\}_{l=1}^{L}$ with

$$L \overset{\text{def}}{=} \max\{1, \tilde{L}\}, \quad \text{where } \tilde{L} \overset{\text{def}}{=} \min\left\{ l \in \mathbb{N} \,\middle|\, \frac{\sum_{l'=1}^{l} \lambda_{l'}}{\sum_{l''=1}^{\infty} \lambda_{l''}} > 0.95 \right\}. \tag{5.7}$$

Hence, to test the hypotheses in (5.2) via Lemma 5.1, we test the following

$$\widehat{H}_0 : \theta_1 = \cdots = \theta_L = 0, \quad \text{versus} \quad \widehat{H}_1 : \text{ there exists } l' \in \{1, \ldots, L\} \text{ such that } \theta_{l'} \neq 0. \tag{5.8}$$

Under the null $\widehat{H}_0$ in (5.8), for each $l \in \{1, \ldots, L\}$, the central limit theorem indicates that $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \xi_{l,i}$ is asymptotically $N(0,1)$ when n is large. The mutual uncorrelation in Theorem 5.1 and the asymptotic normality of $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \xi_{l,i}$ provide the asymptotic independence of $\{\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \xi_{l,i}\}_{l=1}^{L}$ across $l = 1, \ldots, L$. Then, $\sum_{l=1}^{L} (\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \xi_{l,i})^2$ is asymptotically χ_L^2 under the $\widehat{H}_0$ in (5.8). At the asymptotic significance $\alpha \in (0,1)$, we reject the $\widehat{H}_0$ if

$$\sum_{l=1}^{L} \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \xi_{l,i} \right)^2 > \chi^2_{L,1-\alpha} = \text{ the } 1-\alpha \text{ lower quantile of the } \chi_L^2 \text{ distribution.} \tag{5.9}$$

In applications, neither the mean $m_v^{(j)}(t)$ nor the covariance $\Xi_v(s,t)$ is known. Hence, the KL expansions in (5.5) cannot be directly used and must be estimated. In Appendix F, we propose a numerical foundation for conducting the asymptotic χ^2-test in (5.9) and encapsulate the numerical procedures for the test in Algorithm 1. In all our analyses in Sections 6 and 7, the numerical estimates $\widehat{L}$ (see (F.4) in Appendix F) of the L in (5.7) are smaller than 10. When the $\widehat{L}$ values are large (e.g., several hundred), one may also consider applying the adaptive Neyman test proposed by Fan (1996).

In the simulation studies presented in Tables 1 and J.1, our Algorithm 1 has comparable performance with more than ten existing state-of-the-art fdANOVA methods. Nonetheless, both Algorithm 1 and the existing methods exhibit Type I error inflation (e.g., the rejection rate of Algorithm 1 is 0.118, whereas the significance is 0.05). To mitigate this inflation, we may consider applying the permutation test using one of these methods that is computationally efficient. For example, Górecki and Smaga (2015) proposed a permutation test based on an F-type statistic (FP). Specifically, Górecki and Smaga (2015) approximated each observed function by basis functions via information criteria, and the F-type statistic was approximated by a form conducive to efficiently computing permutation-based p-values. However, the FP also exhibits Type I error inflation (see Tables 1 and J.1). Motivated by the FP, we apply the permutation test to the χ^2-statistic defined in (5.9) in the following way: we first apply Algorithm 1 to our original shapes $K_i^{(j)}$ and then repeatedly reapply Algorithm 1 to the shapes with shuffled group labels j. The χ^2-test statistic derived from the original data is then compared to that from the shuffled data. A detailed description of our permutation-based approach is presented in Algorithm 2 in Appendix F. Simulations in Section 6 demonstrate that our permutation-based approach eliminates the Type I error inflation encountered by Algorithm 1. The permutation nature of Algorithm 2 is also advantageous for small sample sizes. Note, however, that the power of Algorithm 2 under the alternative is moderately weaker than that of Algorithm 1. Lastly, the runtimes of Algorithms 1 and 2, when applied to simulations, are studied in Appendix K. We present the runtimes when applying the algorithms to real data in Table 2.

Table 2. P-values of Algorithms 1, 2, and 4 for the dataset of mandibular molars.

	Algorithm 1	Algorithm 2	Algorithm 4
Tarsius vs. *Microcebus*	$< 10^{-3}$	$< 10^{-3}$	$< 10^{-3}$
Tarsius vs. *Mirza*	$< 10^{-3}$	$< 10^{-3}$	0.001
Tarsius vs. *Saimiri*	$< 10^{-3}$	$< 10^{-3}$	$< 10^{-3}$
Microcebus vs. *Mirza*	$< 10^{-3}$	0.009	0.004
Microcebus vs. *Saimiri*	$< 10^{-3}$	$< 10^{-3}$	$< 10^{-3}$
Mirza vs. *Saimiri*	$< 10^{-3}$	$< 10^{-3}$	$< 10^{-3}$
Tarsius vs. *Tarsius*	0.196 (0.220)	0.496 (0.294)	0.527 (0.273)
Overall runtimes (in hours)	≈ 3	≈ 3	≈ 20

NOTE: In the last row, we present the overall runtime for conducting all hypothesis testing tasks using each algorithm.

6. Experiments Using Simulations

We present simulations showing the performance of our Algorithms 1 and 2. In addition, we compare our algorithms with the "randomization-style null hypothesis significance test (NHST)" (Robinson and Turner 2017), the TRP using Wald-type permutation statistic (TRP-WTPS, Cuesta-Albertos and Febrero-Bande 2010; Pauly, Brunner, and Konietschke 2014), and the FP. Details of the randomization-style NHST are given in Appendix G and referred to as Algorithm 3. The application of the FP and TRP to the SECT is described in Section 5. We implement the FP and TRP-WTPS using the R package `fdANOVA` with its default parameters as recommended by Górecki and Smaga (2019). Additional simulations comparing our proposed algorithms and other existing fdANOVA methods are presented in Appendix J.

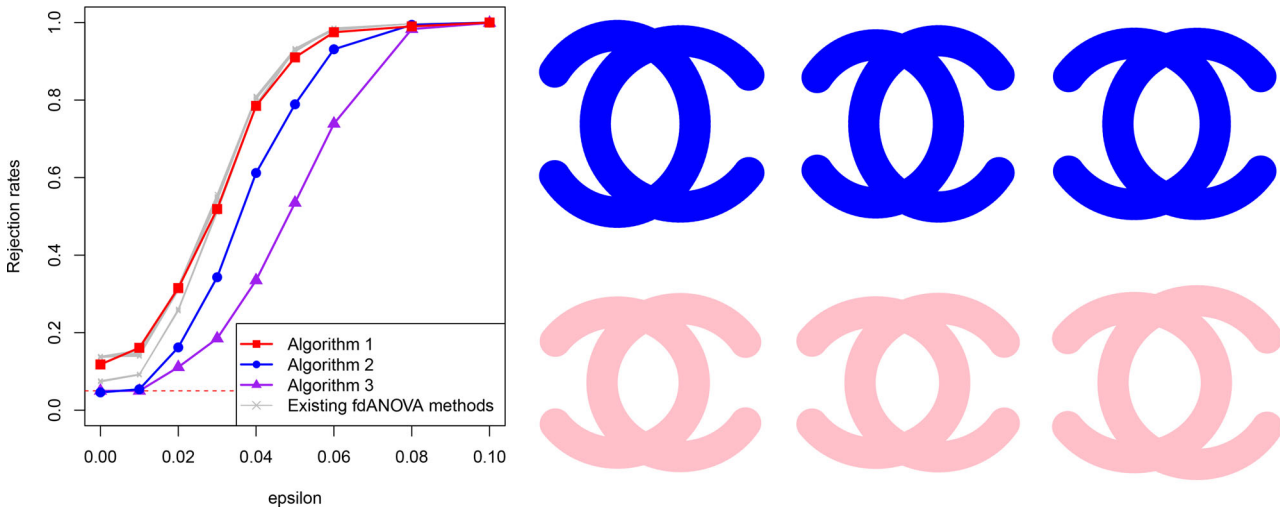

Figure 3. (Left panel) The relationship between ε and the rejection rates computed via Algorithms 1, 2, 3 (see Table 1), and 12 existing fdANOVA methods (see Table J.1 in Appendix J for details on the existing fdANOVA methods). The (red) dashed line presents the significance level $\alpha = 0.05$. (Right panel) The shapes in the first row are from $\mathbb{P}^{(0)}$, and the shapes in the second row are from $\mathbb{P}^{(0.08)}$.

We focus on a family of distributions $\{\mathbb{P}^{(\varepsilon)}\}_{0\le\varepsilon\le 0.1}$ with shapes $\{K_i^{(\varepsilon)}\}_{i=1}^n \overset{\text{iid}}{\sim} \mathbb{P}^{(\varepsilon)}$ via

$$K_i^{(\varepsilon)} \overset{\text{def}}{=} \left\{ x \in \mathbb{R}^2 \,\middle|\, \inf_{y\in S_i^{(\varepsilon)}} \|x-y\| \le 0.2 \right\}, \quad \text{where} \tag{6.1}$$

$$S_i^{(\varepsilon)} \overset{\text{def}}{=} \left\{ \left(\tfrac{2}{5} + a_{1,i}\cdot\cos t,\ b_{1,i}\cdot\sin t\right) \,\middle|\, \tfrac{1-\varepsilon}{5}\pi \le t \le \tfrac{9+\varepsilon}{5}\pi \right\} \bigcup \left\{ \left(-\tfrac{2}{5} + a_{2,i}\cdot\cos t,\ b_{2,i}\cdot\sin t\right) \,\middle|\, \tfrac{6\pi}{5} \le t \le \tfrac{14\pi}{5} \right\} \quad \text{and}$$

$\{a_{1,i}, a_{2,i}, b_{1,i}, b_{2,i}\}_{i=1}^n \overset{\text{iid}}{\sim} N(1, 0.05^2)$. The ε denotes the dissimilarity between $\mathbb{P}^{(\varepsilon)}$ and $\mathbb{P}^{(0)}$. For each $\varepsilon \in [0, 0.1]$, through the discussion in Section 5, we test the following hypotheses via fdANOVA methods (i.e., FP, TRP-WTPS, Algorithms 1, and 2)

$$H_0 : m_\nu^{(0)}(t) = m_\nu^{(\varepsilon)}(t) \text{ for all } (\nu, t) \in \mathbb{S}^{d-1}\times[0,T] \quad \text{versus}$$
$$H_1 : m_\nu^{(0)}(t) \ne m_\nu^{(\varepsilon)}(t) \text{ for some } (\nu, t),$$

where the mean $m_\nu^{(\varepsilon)}(t) \overset{\text{def}}{=} \int_{\mathcal{S}_{R,d}^M} \text{SECT}(K)(\nu,t)\,\mathbb{P}^{(\varepsilon)}(dK)$, and the null hypothesis H_0 is true when $\varepsilon = 0$. We also test $H_0^* : \mathbb{P}^{(0)} = \mathbb{P}^{(\epsilon)}$ *vs.* $\mathbb{P}^{(0)} \ne \mathbb{P}^{(\epsilon)}$ using Algorithm 3.

We set $T = 3$, directions $\nu_p = (\cos\frac{p-1}{4}\pi, \sin\frac{p-1}{4}\pi)^\intercal$ for $p \in \{1,2,3,4\}$, levels $t_q = \frac{T}{50}q$ for $q \in \{1,\dots,50\}$ (i.e., $\Gamma = 4$ and $\Delta = 50$ in Algorithms 1, 2, and 3), the confidence level 95% (i.e., $\alpha = 0.05$), and the number of permutations $\Pi = 1000$. For each $\varepsilon \in \{0, 0.01, 0.02, 0.03, 0.04, 0.05, 0.06, 0.08, 0.1\}$, we independently generate two collections $\{K_i^{(0)}\}_{i=1}^n \overset{\text{iid}}{\sim} \mathbb{P}^{(0)}$ and $\{K_i^{(\varepsilon)}\}_{i=1}^n \overset{\text{iid}}{\sim} \mathbb{P}^{(\varepsilon)}$ through (6.1) with the number of shape pairs set to $n = 100$, and we compute the SECT of each generated shape in directions $\{\nu_p\}_{p=1}^4$ and at levels $\{t_q\}_{q=1}^{50}$. We then implement the fdANOVA methods and Algorithm 3 to these computed SECT statistics and get the corresponding `Accept/Reject` outputs. We repeat this procedure 1000 times and report the rejection rates across all 1000 replicates for each ε in Table 1. The rejection rates are also visually presented in Figure 3. We choose $\Gamma = 4$ as the number of directions in our simulations based on the following observation: in Appendix K, we experiment with all combinations of $\Gamma \in \{2,4,8\}$, $\Delta \in \{25,50,100\}$, and $n \in \{25,50,100\}$. When $\Delta = 50$ and $n = 100$, the number $\Gamma = 4$ is sufficiently large for our Algorithms 1 and 2 to distinguish $\mathbb{P}^{(0)}$ from $\mathbb{P}^{(\varepsilon)}$ with $\varepsilon > 0$ using the significance level $\alpha = 0.05$. Moreover, this choice allows us to demonstrate that even a relatively small number of directions (e.g., $\Gamma = 4$) is sufficient for implementing our Algorithms 1 and 2.

The results in Table 1 and Figure 3 demonstrate that our proposed algorithms are effective at detecting the difference between $\mathbb{P}^{(\varepsilon)}$ and $\mathbb{P}^{(0)}$ in terms of distinguishing their mean functions. Notably, our algorithms (especially Algorithm 2) tend to avoid falsely detecting differences between shape-generating distributions under the null hypothesis (i.e., $\varepsilon = 0$). As ε increases, $\mathbb{P}^{(\varepsilon)}$ deviates from $\mathbb{P}^{(0)}$, and the power of our algorithms in detecting the deviation increases. When $\varepsilon \ge 0.08$, the power of Algorithms 1 and 2 exceeds 0.99. For all the ε, it is difficult to see the deviation of $\mathbb{P}^{(\varepsilon)}$ from $\mathbb{P}^{(0)}$ visually. For instance, by merely observing the shapes in Figure 3, one might find it hard to differentiate between the shape collections generated by $\mathbb{P}^{(0)}$ (blue) and $\mathbb{P}^{(0.08)}$ (pink). However, in more than 99% of the simulations, our algorithms detect the difference between the two distributions. We also randomly shuffle the index i within each collection $\{K_i^{(\varepsilon)}\}_{i=1}^n$ and apply Algorithms 1 and 2 to the shuffled collections. The results obtained from the unshuffled and shuffled shape collections, respectively, are nearly identical. Algorithm 3 performs well in detecting the discrepancy between $\mathbb{P}^{(0)}$ and $\mathbb{P}^{(\varepsilon)}$. However, its power under the alternative hypotheses (i.e., $\varepsilon > 0$) is weaker than that of our Algorithms 1 and 2. Moreover, Algorithms 1 and 2 exhibit performance comparable to twelve existing state-of-the-art fdANOVA methods (see Table 1, Figure 3, and Table J.1 in Appendix J).

7. Applications

We first apply our proposed Algorithms 1 and 2 to the MPEG-7 shape silhouette database (Sikora 2001) as a toy example. Details

of this are provided in Appendix I. This analysis shows that our proposed algorithms can distinguish between shape classes in the silhouette database and do not falsely identify signals when there are no differences between groups.

In this section, we apply our algorithms to address the motivating question in Section 1.1. Specifically, we use Algorithms 1 and 2 to distinguish between the four categories of mandibular molars in Figure 1 that are from four genera of primates. The shapes in Figure 1 come from two suborders of primates: Haplorhini and Strepsirrhini (see Figure 1). In the haplorhine suborder collection, 29 molars came from the genus *Tarsius* (yellow panels in Figure 1), and 9 molars came from the genus *Saimiri* (grey panels in Figure 1). In the strepsirrhine collection, 11 molars came from the genus *Microcebus* (blue panels in Figure 1), and 6 molars came from the genus *Mirza* (green panels in Figure 1).

Before applying Algorithms 1 and 2, we preprocess the raw triangle mesh data of the surfaces of the molars by aligning them through the ECT alignment approach detailed in Appendix E. The aligned molars are presented in Figure 1. We apply our Algorithms 1 and 2 to the preprocessed molars. For each aligned molar, we compute its SECT for 2918 directions; in each direction, we use 200 sublevel sets. To compare any pair of molar groups, as a proof of concept, we select the smaller size of the two groups as the sample size input n in our algorithms. For example, when comparing the *Tarsius* and *Microcebus* groups, we choose $n = 11$; that is, we compare the first 11 molars of the *Tarsius* group to all the molars in the *Microcebus* group. We apply our algorithms to the four groups of molars and present the results in Table 2. The p-values in Table 2 are either χ^2-test p-values (Algorithm 1) or permutation-test p-values (Algorithm 2 with 1000 permutations). The small p-values ($P < 0.05$) in Table 2 show that our proposed algorithms can distinguish the four different genera of primates. Since the genera *Microcebus* and *Mirza* belong to the same suborder Strepsirrhini (see Figure 1), the p-value from Algorithm 2 is comparatively large when comparing molars from these two groups. In comparison, although the *Tarsius* and *Saimiri* both belong to the suborder Haplorhini, the molars of the two genera look different. Specifically, the paraconids (i.e., the cusp highlighted in red in Figure 1) are only retained by the genus *Tarsius* and, thus, are a key reason for the small p-values ($P < 10^{-3}$) when comparing with molars from the *Saimiri*. Other small p-values ($P < 10^{-3}$) in our analyses are a result of the corresponding genera belonging to different suborders.

In addition to testing the difference between genera, we apply our algorithms within the genus *Tarsius*. Specifically, we focus on the first 28 molars in the *Tarsius* group. We randomly split the 28 molars into two halves and apply Algorithms 1 and 2 to test the difference between the two halves. We repeat the random splitting procedure 100 times and present the corresponding p-values in Table 2. The results are summarized by their mean and standard deviation (in parentheses). These p-values show that our proposed Algorithm 2 tends to avoid the Type I error for the molars from the genus *Tarsius*.

Landmark methods are widely used in geometric morphometrics. One state-of-the-art approach is the "Gaussian process landmarking (GPL)" algorithm (Gao et al. 2019; Gao, Kovalsky, and Daubechies 2019) which can automatically sample landmarks on the surfaces of the molars in Figure 1. Gao et al. (2019) showed that these sampled landmarks could induce a continuous Procrustes distance to measure the dissimilarity between molars. A permutation test can be derived using the Procrustes distance induced by the GPL algorithm. This test is detailed in Appendix H and is encapsulated by Algorithm 4. We use the GPL-based Algorithm 4 to differentiate the four collections of molars. For this, we use the `MATLAB` code from the `GitHub` repository provided by Gao et al. (2019) to compute the Procrustes distance. Performance of Algorithm 4 is in Table 2, which shows that the GPL-based method and our Algorithm 2 have comparable performance. However, repeatedly computing the Procrustes distance is time-consuming. Hence, Algorithm 2 is more computationally efficient than Algorithm 4 while achieving similar performance (see the last row of Table 2).

We want to note that, in addition to the GPL algorithm, many other existing methods can be applied to measure dissimilarity between molars, including parameterized surfaces (Kurtek et al. 2010, 2011) and the approaches from computational anatomy (Grenander and Miller 1998). Similarly, the parameterized curves (Kurtek et al. 2012) can also be used to analyze the silhouette database in Appendix I. An even more comprehensive comparison of our algorithms with the entire edifice of existing methods is left for future research.

8. Conclusions and Discussions

In this article, we established the mathematical foundations for the randomness of shapes via the SECT. Specifically, (i) $(\mathcal{S}^M_{R,d}, \mathscr{B}(\rho), \mathbb{P})$ was constructed as the underlying probability space; (ii) the SECT was modeled as a $C(\mathbb{S}^{d-1}; \mathcal{H})$-valued random variable. We further demonstrated several properties of the SECT ensuring its KL expansion, which led to a χ^2-statistic for testing hypotheses on random shapes. We bridged the fdANOVA and TDA. Simulation studies corroborated our mathematical derivations and showed the performance of our hypothesis testing algorithms. Our approach was shown to be powerful in detecting the difference between two shape-generating distributions. We applied our proposed algorithms to silhouette and primate molar datasets. Importantly, our simulations when $\varepsilon = 0$, together with the applications to the molars and the silhouette database, indicate that our algorithms tend to avoid falsely detecting differences between shape-generating distributions when there are none. Using the molars in Figure 1, we compared the performance of our algorithms to a permutation test based on a state-of-the-art landmarking algorithm (Gao et al. 2019; Gao, Kovalsky, and Daubechies 2019), underscoring the efficiency of our algorithms. We enumerate potential future research areas in Appendix M, for example, the fdANOVA methods can be used for brain connectivity (Chen et al. 2024; Meng and Eloyan 2024) via topological summaries.

Supplementary Materials

The supplementary materials provide the proofs of theorems, further data analysis, and future research topics.

Acknowledgments

We are grateful to the Editor, Associate Editor, and three Referees for their thorough review of our article and the insightful suggestions that have tremendously improved its quality. We want to thank Dr. Matthew T. Harrison from the Division of Applied Mathematics at Brown University for useful comments and suggestions. KM wants to thank Mattie Ji from the Department of Mathematics at Brown University for her insightful comments. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Data Availability Statement

The source code for implementing the simulation studies and applications is publicly available online at *https://github.com/JinyuWang123/TDA.git*.

Disclosure Statement

The authors report there are no competing interests to declare.

Funding

LC would like to acknowledge the support of a David & Lucile Packard Fellowship for Science and Engineering. Research reported in this publication was partially supported by the National Institute On Aging of the National Institutes of Health under Award Number R01AG075511.

ORCID

Kun Meng http://orcid.org/0000-0001-9366-1737
Lorin Crawford http://orcid.org/0000-0003-0178-8242

References

Boyer, D. M., Lipman, Y., Clair, E. S., Puente, J., Patel, B. A., Funkhouser, T., Jernvall, J., and Daubechies, I. (2011), "Algorithms to Automatically Quantify the Geometric Similarity of Anatomical Surfaces," *Proceedings of the National Academy of Sciences*, 108, 18221–18226. DOI:10.1073/pnas.1112822108. Available at *https://www.pnas.org/doi/abs/10.1073/pnas.1112822108*. [498,499]

Brezis, H. (2011), *Functional Analysis, Sobolev Spaces and Partial Differential Equations* (Vol. 2), New York: Springer. [500,501]

Chen, Y., Lin, Z., and Müller, H.-G. (2023), "Wasserstein Regression," *Journal of the American Statistical Association*, 118, 869–882. [504]

Chen, Y., Lin, S.-C., Zhou, Y., Carmichael, O., Müller, H.-G., Wang, J.-L., and Alzheimer's Disease Neuroimaging Initiative. (2024), "Gradient Synchronization for Multivariate Functional Data, with Application to Brain Connectivity," *Journal of the Royal Statistical Society*, Series B. DOI:10.1093/jrsssb/qkad140. [508]

Cohen-Steiner, D., Edelsbrunner, H., and Harer, J. (2007), "Stability of Persistence Diagrams," *Discrete & Computational Geometry*, 37, 103–120. [499]

Crawford, L., Monod, A., Chen, A. X., Mukherjee, S., and Rabadán, R. (2020), "Predicting Clinical Outcomes in Glioblastoma: An Application of Topological and Functional Data Analysis," *Journal of the American Statistical Association*, 115, 1139–1150. [498,499,500,502]

Cuesta-Albertos, J., and Febrero-Bande, M. (2010), "A Simple Multiway Anova for Functional Data," *Test*, 19, 537–557. [505,506]

Curry, J., Mukherjee, S., and Turner, K. (2022), "How Many Directions Determine a Shape and Other Sufficiency Results for Two Topological Transforms," *Transactions of the American Mathematical Society*, Series B, 9, 1006–1043. [500,501,502,504]

Dunson, D. B., and Wu, N. (2021), "Inferring Manifolds from Noisy Data Using Gaussian Processes," arXiv preprint arXiv:2110.07478. [504]

Dupuis, P., Grenander, U., and Miller, M. I. (1998), "Variational Problems on Flows of Diffeomorphisms for Image Matching," *Quarterly of Applied Mathematics*, LVI, 587–600. [499]

Edelsbrunner, H., and Harer, J. (2010), *Computational Topology: An Introduction*, Providence, RI: American Mathematical Society. [498,499,502]

Fan, J. (1996), "Test of Significance based on Wavelet Thresholding and Neyman's Truncation," *Journal of the American Statistical Association*, 91, 674–688. [506]

Fasy, B. T., Lecci, F., Rinaldo, A., Wasserman, L., Balakrishnan, S., and Singh, A. (2014), "Confidence Sets for Persistence Diagrams," *The Annals of Statistics*, 42, 2301–2339. [498,500]

Fasy, B. T., Micka, S., Millman, D. L., Schenfisch, A., and Williams, L. (2018), "Challenges in Reconstructing Shapes from Euler Characteristic Curves," arXiv preprint arXiv:1811.11337. [503]

Gao, T., Kovalsky, S. Z., Boyer, D. M., and Daubechies, I. (2019), "Gaussian Process Landmarking for Three-Dimensional Geometric Morphometrics," *SIAM Journal on Mathematics of Data Science*, 1, 237–267. [498,499,508]

Gao, T., Kovalsky, S. Z., and Daubechies, I. (2019), "Gaussian Process Landmarking on Manifolds," *SIAM Journal on Mathematics of Data Science*, 1, 208–236. [498,499,508]

Ghrist, R., Levanger, R., and Mai, H. (2018), "Persistent Homology and Euler Integral Transforms," *Journal of Applied and Computational Topology*, 2, 55–60. [499,500,503]

Górecki, T., and Smaga, Ł. (2015), "A Comparison of Tests for the One-Way Anova Problem for Functional Data," *Computational Statistics*, 30, 987–1010. [506]

——— (2019), "fdanova: An r Software Package for Analysis of Variance for Univariate and Multivariate Functional Data," *Computational Statistics*, 34, 571–597. [506]

Goswami, A. (2015), "Phenome10k: A Free Online Repository for 3-D Scans of Biological and Palaeontological Specimens," Google Scholar. [499]

Grenander, U., and Miller, M. I. (1998), "Computational Anatomy: An Emerging Discipline," *Quarterly of Applied Mathematics*, 56, 617–694. [499,508]

Hairer, M. (2009), "An Introduction to Stochastic PDEs," arXiv preprint arXiv:0907.4178. [498,499,503]

Hatcher, A. (2002), *Algebraic Topology*, New York: Cambridge University Press. [498,501]

Hsing, T., and Eubank, R. (2015), *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators* (Vol. 997), Chichester: Wiley. [498,500,503,505]

Jia Guo, B. Z., and Zhang, J.-T. (2019), "New Tests for Equality of Several Covariance Functions for Functional Data," *Journal of the American Statistical Association*, 114, 1251–1263. DOI:10.1080/01621459.2018.1483827. [505]

Jiang, Q., Kurtek, S., and Needham, T. (2020), "The Weighted Euler Curve Transform for Shape and Image Analysis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 844–845. [500]

Kendall, D. G. (1989), "A Survey of the Statistical Theory of Shape," *Statistical Science*, 4, 87–99. [498]

Kirveslahti, H., and Mukherjee, S. (2023), "Representing Fields Without Correspondences: The Lifted Euler Characteristic Transform," *Journal of Applied and Computational Topology*, 8, 1–34. [500]

Kurtek, S., Klassen, E., Ding, Z., Jacobson, S. W., Jacobson, J. L., Avison, M. J., and Srivastava, A. (2010), "Parameterization-Invariant Shape Comparisons of Anatomical Surfaces," *IEEE Transactions on Medical Imaging*, 30, 849–858. [499,508]

Kurtek, S., Klassen, E., Gore, J. C., Ding, Z., and Srivastava, A. (2011), "Elastic Geodesic Paths in Shape Space of Parameterized Surfaces," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34, 1717–1730. [498,499,508]

Kurtek, S., Srivastava, A., Klassen, E., and Ding, Z. (2012), "Statistical Modeling of Curves Using Shapes and Related Features," *Journal of the American Statistical Association*, 107, 1152–1165. [499,508]

Li, D., Mukhopadhyay, M., and Dunson, D. B. (2022), "Efficient Manifold Approximation with Spherelets," *Journal of the Royal Statistical Society*, Series B, 84, 1129–1149. [504]

Marsh, L., Zhou, F. Y., Quin, X., Lu, X., Byrne, H. M., and Harrington, H. A. (2022), "Detecting Temporal Shape Changes with the Euler Characteristic Transform," arXiv preprint arXiv:2212.10883. [500]

Meng, K., and Eloyan, A. (2021), "Principal Manifold Estimation Via Model Complexity Selection," *Journal of the Royal Statistical Society*, Series B, 83, 369–394. DOI:10.1111/rssb.12416. [504]

Meng, K., and Eloyan, A. (2024), "Population-Level Task-Evoked Functional Connectivity via Fourier Analysis," *Journal of the Royal Statistical Society*, Series C. DOI:10.1093/jrsssc/qlae015. [508]

Mileyko, Y., Mukherjee, S., and Harer, J. (2011), "Probability Measures on the Space of Persistence Diagrams," *Inverse Problems*, 27, 124007. [499]

Miller, E. (2015), "Fruit Flies and Moduli: Interactions between Biology and Mathematics," *Notices of the AMS*, 62, 1178–1184. [499]

Molchanov, I. (2005), *Theory of Random Sets* (Vol. 19), Dordrecht: Springer. [500]

Pauly, M., Brunner, E., and Konietschke, F. (2014), "Asymptotic Permutation Tests in General Factorial Designs," *Journal of the Royal Statistical Society*, Series B, 77, 461–473. DOI:10.1111/rssb.12073. [506]

Petersen, A., and Müller, H.-G. (2019), "Fréchet Regression for Random Objects with Euclidean Predictors," *The Annals of Statistics*, 47, 691–719. [504]

Reed, M., and Simon, B. (1972), *Methods of Modern Mathematical Physics: Functional Analysis*, New York: Academic Press. [505]

Robinson, A., and Turner, K. (2017), "Hypothesis Testing for Topological Data Analysis," *Journal of Applied and Computational Topology*, 1, 241–261. [500,506]

Roycraft, B., Krebs, J., and Polonik, W. (2023), "Bootstrapping Persistent Betti Numbers and Other Stabilizing Statistics," *The Annals of Statistics*, 51, 1484–1509. [498]

Schapira, P. (1995), "Tomography of Constructible Functions," in *International Symposium on Applied Algebra, Algebraic Algorithms, and Error-Correcting Codes*, pp. 427–435, Springer. [503]

Shen, Q., and Faraway, J. (2004), "An f Test for Linear Models with Functional Responses," *Statistica Sinica*, 14, 1239–1257. [505]

Sikora, T. (2001), "The mpeg-7 Visual Standard for Content Description-an Overview," *IEEE Transactions on Circuits and Systems for Video Technology*, 11, 696–702. [507]

Tang, W. S., da Silva, G. M., Kirveslahti, H., Skeens, E., Feng, B., Sudijono, T., Yang, K. K., Mukherjee, S., Rubenstein, B., and Crawford, L. (2022), "A Topological Data Analytic Approach for Discovering Biophysical Signatures in Protein Dynamics," *PLoS Computational Biology*, 18, e1010045. [498]

Turner, K., Mukherjee, S., and Boyer, D. M. (2014), "Persistent Homology Transform for Modeling Shapes and Surfaces," *Information and Inference: A Journal of the IMA*, 3, 310–344. [499,502,503]

van den Dries, L. (1998), *Tame Topology and o-Minimal Structures* (Vol. 248), Cambridge: Cambridge University Press. [500,501,502]

Wang, B., Sudijono, T., Kirveslahti, H., Gao, T., Boyer, D. M., Mukherjee, S., and Crawford, L. (2021), "A Statistical Pipeline for Identifying Physical Features that Differentiate Classes of 3D Shapes," *The Annals of Applied Statistics*, 15, 638–661. [498,499,500,504]

Zhang, J.-T. (2013), *Analysis of Variance for Functional Data*, Boca Raton, FL: CRC Press. [498,500,505]

Zhang, J.-T., and Chen, J. (2007), "Statistical Inferences for Functional Data," *The Annals of Statistics*, 35, 1052–1079. [505]