Tukey’s Depth for Object Data

Xiongtao Dai* and Sara Lopez-Pintado
for the Alzheimer’s Disease Neuroimaging Initiative

*Department of Statistics, Iowa State University, Ames, IA; bDepartment of Health Sciences, Northeastern University, Boston, MA

ABSTRACT
We develop a novel exploratory tool for non-Euclidean object data based on data depth, extending celebrated Tukey’s depth for Euclidean data. The proposed metric halfspace depth, applicable to data objects in a general metric space, assigns to data points depth values that characterize the centrality of these points with respect to the distribution and provides an interpretable center-outward ranking. Desirable theoretical properties that generalize standard depth properties postulated for Euclidean data are established for the metric halfspace depth. The depth median, defined as the deepest point, is shown to have high robustness as a location descriptor both in theory and in simulation. We propose an efficient algorithm to approximate the metric halfspace depth and illustrate its ability to adapt to the intrinsic data geometry. The metric halfspace depth was applied to an Alzheimer’s disease study, revealing group differences in the brain connectivity, modeled as covariance matrices, for subjects in different stages of dementia. Based on phylogenetic trees of seven pathogenic parasites, our proposed metric halfspace depth was also used to construct a meaningful consensus estimate of the evolutionary history and to identify potential outlier trees.

Supplementary materials for this article are available online.

ARTICLE HISTORY
Received February 2021
Accepted November 2021

KEYWORDS
Data depth; Nonparametric statistics; Ranks; Robust inference

1. Introduction

1.1. Backgrounds

Complex data objects are increasingly generated across science and rapidly gaining relevance. Finite-dimensional non-Euclidean data is an important class of object data (Marron and Alonso 2014), which models, for example, directions (Mardia and Jupp 2009), covariance matrices (Pennec, Fillard, and Ayache 2006), and trees (Billera, Holmes, and Vogtmann 2001). There has been extensive development in methods and theory to address the complexity of these objects, including location measures (Fréchet 1948), statistical inference (Bhattacharya and Patrangenaru 2005), and classification (Dai and Müller 2018).

However, exploratory data analysis is a crucial paradigm that lacks development for these nonstandard data. As the basic data units become more complex and multifaceted, there is an escalated need for an agnostic exploratory data analysis. Data exploration before modeling will reveal properties of the data distribution and help identify extreme versus typical observations. In this regard, a first step is to overcome the absence of a canonical ordering for complex objects and propose principled definitions of rank, median, and order statistics.

Data depth has been proven to be a powerful exploratory and data-driven tool that can be used to rank observations and reveal features of the underlying data distribution. The notion of data depth was originally introduced for multivariate Euclidean data and provides a way of measuring how “representative” or “outlying” an observation is with respect to a probability distribution. In particular, a depth function assigns a nonnegative depth value to a given observation within a distribution, where the larger this value is the more central/deep the observation is within the distribution. Points with low depth values correspond to observations near the outskirt of the distribution and “far” from the center. These observations could be potential outliers worthwhile of investigation. Hence, a notion of depth provides a “center-outward” ordering for a sample of multivariate observations and allows generalization of ranks, order statistics, central regions (see Zuo and Serfling 2000), and robust inferential and classification methods to multivariate data (Li, Cuesta-Albertos, and Liu 2012).

For multivariate Euclidean data, Tukey’s halfspace depth (Tukey 1975) has attracted much attention. Due to its intuitive properties (Zuo and Serfling 2000) and the robustness of depth induced median (Donoho and Gasko 1992), Tukey’s depth stands out as the first and one of the most popular data summaries (Tukey 1975; Rousseeuw, Ruts, and Tukey 1999) and robust nonparametric rank tests (Liu and Singh 1993; Chenouri and Small 2012). However, Tukey’s depth relies on the Euclidean geometry and is inappropriate for non-Euclidean data objects.
Though defining depth notions for non-Euclidean data has garnered wide interest, the literature has focused on specialized spaces, such as a unit sphere (Small 1987; Liu and Singh 1992; Pandolfo, Paindaveine, and Porzio 2018), positive definite matrices (Fletcher et al. 2011; Chau et al. 2019), networks (Fraiman, Fraiman, and Fraiman 2017), data on a graph (Small 1997), and infinite-dimensional functional data (Fraiman and Muniz 2001; López-Pintado and Romo 2009). Chen, Gao, and Ren (2018) and Paindaveine and Van Bever (2018) considered halfspace depth for the scatter matrix of Euclidean data points. Fraiman, Gamboa, and Moreno (2019) proposed a spherical depth that applies to Riemannian manifold data. Targeting general settings, Carrizosa (1996) sketched a halfspace depth based on dissimilarity measures without methodological development, which differs from our proposal in general; see Section S5, supplementary material. Carrizosa (1996) also introduced an extension of the halfspace depth to a regression setting closely related to the regression depth proposed by Rousseeuw and Hubert (1999); see also Zuo (2021) for a discussion of the theoretical properties of regression depths.

1.2. Our Contributions

The goal of our work is to generalize Tukey’s depth to data objects taking values on an arbitrary metric space, defining a general depth notion that shares the desirable properties of Tukey’s depth. This will make available depth-based exploratory and robust inferential toolsets to general object data.

We propose the metric halfspace depth in Section 2.1, which is a generalization of Tukey’s depth to object data on a general metric space. Metric halfspace depth incorporates the data space geometry through the distance metric $d$, a feature available on any metric space. The proposed metric halfspace depth, therefore, applies to a wide range of non-Euclidean data objects. This includes data lying on smooth Riemannian manifolds such as directional data on a sphere (Mardia and Jupp 2009); bivariate molecular torsion angles on a flat torus (Eltzner, Huckemann, and Mardia 2018); and constrained matrix-valued data, such as rotations (Bingham, Nordman, and Vardeman 2009) and covariance matrices (Dai, Lin, and Müller 2020). Nonsmooth objects with possible degeneracy lying on a geodesic space, such as phylogenetic trees (Feragen and Nye 2020), networks (Kolaczyk et al. 2020), and shapes (Dryden and Mardia 2016) can also be investigated by the proposed depth.

We establish desirable properties of the metric halfspace depth in Section 3, extending much of the properties enjoyed by Tukey’s depth (Tukey 1975) for Euclidean data. The axiomatic properties of depth notions introduced in Zuo and Serfling (2000) are satisfied to a great extent in many commonly investigated data spaces. The metric halfspace depth is invariant to a large class of transformations; if the data are symmetrically distributed around a center, then the center has the maximal metric halfspace depth; the depth values have a center-outward tendency and monotonically decrease from the deepest point to the peripheral points; and the depth vanishes as one moves away from the center. The metric halfspace depth function is upper semi-continuous, which implies that the nested deepest regions are compact. We establish a root-$n$ rate of convergence of the sample depth to the true depth function, and the consistency of the sample deepest point to the population deepest point, assuming uniqueness of the latter. Moreover, the metric halfspace depth is shown to be robust to contamination, having a high breakdown point for symmetric distributions regardless of the dimension of the data space. All proofs are included in the supplementary material.

Tukey’s depth for Euclidean data has a well-known weakness in its high computation cost even in moderate dimensions. To overcome this obstacle, we propose efficient algorithms in Section 4 to approximate the metric halfspace depth by looking into finitely many halfspaces as informed by the dataset. Our proposed approximation algorithm for calculating the depth function and the deepest point has a complexity of $O(n^3)$ with respect to the sample size $n$, independent of the dimension of the data space. The approximation algorithm is able to achieve arbitrary precision to the truth by densening the discretization of the space, which we establish in our theoretical results and demonstrate in simulation studies. The proposed depth is shown to have excellent numerical performance in terms of efficiency and robustness in Section 5. The approximate depth algorithm respects the intrinsic data geometry independent of the ambient space as demonstrated in Section S10, supplementary material.

We showcase the practical relevance of the metric halfspace depth in two applications in Section 6, which include (a) neuroconnectivity matrices from functional magnetic resonance imaging (fMRI) data of patients with dementia and healthy controls and (b) phylogenetic trees comparing the genetic materials from different species. The application to fMRI data discovered differences in the brain connectivity among groups of normal controls and patients at different dementia stages progressing to Alzheimer’s disease using depth-based rank tests. The second application considers estimating the phylogenetic history of seven Apicomplexan species, which are pathogenic parasites, in the tree space of Billera, Holmes, and Vogtmann (2001). We obtained the most representative tree for estimating a consensus evolutionary history of the Apicomplexa and also identified outliers in the individual gene trees.

2. Metric Halfspace Depth

2.1. General Definition

We consider extending the concept of data depth to data objects taking values on a general metric space. Let $M$ be a metric space equipped with distance $d$, and $X$ be an $M$-valued random object defined on probability space $(\Omega, F, P)$ and measurable with respect to the Borel $\sigma$-algebra $B(M)$. To define a halfspace depth, the key lies in suitably generalizing the notion of halfspaces. For two points $x_1, x_2 \in M$ on the metric space, we denote the metric halfspace, or halfspace for brevity, as

$$H_{x_1, x_2} = \{ y \in M \mid d(y, x_1) \leq d(y, x_2) \},$$

which is said to be anchored at $(x_1, x_2)$. Halfspace $H_{x_1, x_2}$ contains all points of $M$ that lie no further away from $x_1$ than from $x_2$. Let $H = \{ H_{x_1, x_2} \mid x_1 \neq x_2 \in M \}$ be the collection of all halfspaces and $H_x = \{ H_{x_1, x_2} \in H \mid x \in H_{x_1, x_2} \}$ the set of halfspaces containing $x$, understanding that the same halfspace
may arise from different pairs of anchors. The proposed metric halfspace depth (MHD) at \( x \in M \) w.r.t. the probability measure \( P_x \) induced by \( X \) is defined as

\[
D(x) = D(x; P_x) = \inf_{H \in \mathcal{H}_x} P_x(H)
\]

\[
= \inf_{x_1, x_2 \in M} \inf_{d(x_1, x_2) \leq d(x, x_2)} P(d(X, x_1) \leq d(X, x_2)).
\]

Depth \( D(x) \) is the least probability measure of the halfspaces containing \( x \), which is well-defined since the halfspaces are closed and thus, measurable. Analogously, given iid observations \( X_1, \ldots, X_n \in M \), the sample metric halfspace depth at \( x \in M \) w.r.t. the empirical distribution \( P_n \) is

\[
D_n(x) = D(x; P_n)
\]

\[
= \inf_{H \in \mathcal{H}_x} P_n(H)
\]

\[
= \inf_{x_1, x_2 \in M} \frac{1}{n} \sum_{i=1}^{n} I(d(X_i, x_1) \leq d(X_i, x_2)),
\]

where \( I(\cdot) \) is the indicator function.

It is immediately seen that if \( M \) is a Euclidean space, then each halfspace is a closed Euclidean halfspace of the form \( \{ x \in \mathbb{R}^m \mid x^T v \leq c \} \) for some vector \( v \in \mathbb{R}^m \) and \( c \in \mathbb{R} \), and the metric halfspace depth coincides with Tukey’s halfspace depth (Tukey 1975). The metric halfspace depth specializes to the metrical halfspace depth (Tukey 1975). The metric halfspace depth coincides with Tukey’s halfspace depth at all points. The distance between two points (solid dots) is given by the length of the segment of a great circle connecting them. Lower right: The space of 3-spider consisting of three Euclidean positive axes issuing from the origin. This is a geodesic space but not a Riemannian manifold due to the singularity at the origin. A geodesic connecting two points on different branches is highlighted, and the distance between points is the length of the geodesic.

2.2. Preliminaries on Metric Spaces

A map \( \gamma \) from a closed interval \( I \subset \mathbb{R} \) to \( M \) is said to be a geodesic if there exists a constant \( \lambda \) such that \( d(\gamma(t), \gamma(t')) = \lambda |t - t'| \) for all \( t, t' \in I \); if further \( \lambda = 1 \), then \( \gamma \) is said to be a unit speed geodesic. We say that a geodesic \( \gamma \) joins \( x \in M \) to \( y \in M \) if \( I = [0, l] \), \( \gamma(0) = x \), and \( \gamma(l) = y \) for some constant \( l \). Now, \((M, d)\) is said to be a geodesic space if any two points \( x, y \in M \) are joined by a geodesic. Riemannian manifolds are smooth submanifolds embedded in an ambient Euclidean space. The definitions of the manifolds and additional geometrical quantities, such as the tangent space \( T_x M \) and exponential map \( \exp_x \), are reviewed in Section S1, supplementary material. The distance between two points \( x, y \) on a Riemannian manifold \( M \) is the length of the shortest path on \( M \) connecting them. Riemannian manifolds are geodesic spaces by the Hopf–Rinow theorem (Lee 2018).

The left panel in Figure 1 illustrates the relationship between different types of complete and connected metric spaces and highlights four common examples. The unit sphere \( S^2 \) in \( \mathbb{R}^3 \) is a Riemannian manifold where a geodesic is a segment of a great circle (upper right, Figure 1), and the 3-spider that models trees with three leaves (lower right, Figure 1) is an example of a geodesic space that is not a Riemannian manifold, since the origin is degenerate and does not have a neighborhood resembling a real interval.

If \( M \) is an unbounded Riemannian manifold such as the space of symmetric positive definite matrices and the hyperbolic space, depth notions could alternatively be developed through mapping data on \( M \) to a tangent space \( T_x M \) through the inverse exponential map \( \exp^{-1}_x : M \to T_x M \), and then employing Euclidean depth notions such as Tukey’s depth on the linear tangent space. However, this tangent space approach has limited applicability since it relies on the exponential map being injective, which is violated on bounded spaces such as the unit sphere \( S^k \) and the rotational group \( SO(k) \); even if this approach can be applied, it is in general not possible to fully preserve the data geometry reflected by the distance metric while working on the linear tangent space \( T_x M \); and the base point \( x \) must be chosen. In contrast, our metric halfspace depth is well-defined and geometry preserving on any metric space.

2.3. Examples: Metric Halfspace Depth in Common Spaces

In what follows, we provide examples of commonly investigated data spaces and illustrate metric halfspace depth in these spaces. Examples 1–3 concern Riemannian manifolds, and Example 4 considers the tree space as a geodesic space that is not a Riemannian manifold. Depth values in the examples are calculated using the approximation algorithm described in Section 4.1.
More details about the setups can be found in Section S3.3, supplementary material.

**Example 1 (Euclidean space).** When data lie in the Euclidean space \( \mathcal{M} = \mathbb{R}^m \), the proposed metric halfspace depth coincides with Tukey’s depth. With norm \( \|x\| = (x^T x)^{1/2} \), distance \( d(x_1, x_2) = \|x_1 - x_2\| \), and Euclidean halfspace \( H_{x_0, v} = \{y \in \mathbb{R}^m | (y - x_0)^T v \leq 0\} \), classical Tukey’s depth is,

\[
D_{\text{Tukey}}(x) = \inf P_X(H'), \quad x \in \mathbb{R}^m,
\]

where the infimum is taken over all Euclidean halfspaces \( H' \) containing \( x \). Any metric halfspace \( H = H_{x_1, x_2} \in \mathcal{H} \) coincides with a Euclidean halfspace \( H'_{(x_0, v)} = \{y \in \mathbb{R}^m | (y - x_0)^T v \leq 0\} \) with \( x_0 = (x_1 + x_2)/2 \) and \( v = (x_2 - x_1)/\|x_2 - x_1\| \) if \( x_1 \neq x_2 \); vice versa, each Euclidean halfspace can be expressed as a metric halfspace. Thus, the metric halfspace depth coincides with Tukey’s depth because the infimums are taken over an identical set, noting that \( P_X(H_{x_1, x_2}) = 1 \) if \( x_1 = x_2 \) which does not influence the infimum for the metric halfspace depth.

Tukey’s depth in the Euclidean space satisfies all four axiomatic properties of a depth function introduced in Zuo and Serfling (2000), is a continuous function of the depth location (Massé 2004) and can be consistently estimated by its sample version (Massé 2004); moreover, the deepest point, that is, Tukey’s median, has a high breakdown point (Donoho and Gasko 1992; Liu, Zuo, and Wang 2017) and can also be consistently estimated (Bai and He 1999; Chen, Gao, and Ren 2018; Zuo 2020). We will show in Section 3 that many of these properties generalize on geodesic spaces.

**Example 2 (Spheres).** The \( m \)-dimensional unit sphere \( S^m = \{x \in \mathbb{R}^{m+1} | x^T x = 1\} \subset \mathbb{R}^{m+1} \) is a Riemannian manifold. The distance between \( x, y \in S^m \) is the great arc distance \( d(x, y) = \arccos(x^T y) \).

The metric halfspace depth specializes to angular Tukey’s depth for spherical data considered by Small (1987); Liu and Singh (1992), where the latter is defined as the least probability measure of any hemisphere covering \( x \). This is because a metric halfspace \( H_{x_1, x_2} = \{x \in S^m | x^T (x_2 - x_1) \leq 0\} \) is a closed hemisphere in this context.

An example of the metric halfspace depth applied to data on \( S^2 \) is shown in the upper left panel of Figure 2, where data were generated according to the wrapped normal distribution with isotropic variance 1/2; the setup is described in Section S3.3, supplementary material. The depth values follow a center-outward pattern, monotonically decreasing from the deepest point near the center of symmetry.

The deepest point meaningfully characterizes a representative point well-encompassed by the point cloud, and the points with the lowest depth all lie on the peripheral.

**Example 3 (Symmetric positive definite matrices).** Let \( \mathcal{M} = \text{SPD}(k) \) be the manifold of \( k \times k \) symmetric positive definite (SPD) matrices. This matrix manifold has seen wide application in modeling brain connectivity matrices (Dai, Lin, and Müller 2020) and diffusion tensors (Pennec, Fillard, and Ayache 2006). Endowed with the affine-invariant geometry (Pennec, Fillard, and Ayache 2006), the geodesic distance on \( \mathcal{M} \) is defined as

\[
d(P, Q) = \| \logm(P^{-1/2} Q P^{-1/2}) \|_F \quad \text{for } P, Q \in \mathcal{M},
\]

where \( \| \cdot \|_F \) is the Frobenius norm, \( \logm \) is the matrix logarithm, and \( P^{-1/2} \) is the inverse of the symmetric positive definite square root \( P^{1/2} \) of \( P \). The geometry is invariant under affine transformations in the sense that \( d(APA^T, AQQA^T) = d(P, Q) \) for any invertible matrix \( A \) and thus, have been widely adopted in applications. As the data space is non-Euclidean with a complex geometry, the halfspaces in general have rather complex shapes. An example of a halfspace \( H_{x_1, x_2} \) is shown in Figure S2, supplementary material. Here we evaluate the depth of an SPD matrix with respect to a sample of SPD matrices as the data units, which is different from the scenario considered for scatter depth (Chen, Gao, and Ren 2018; Paindaveine and Van Bever 2018) where the depth of an SPD matrix is evaluated with respect to Euclidean data units for estimating the covariance matrix.

Illustrated for \( \mathcal{M} = \text{SPD}(2) \), the lower panel of Figure 2 displays nonisotropic log-normal matrix data points that are colored according to the proposed metric halfspace depth. Each point represents the lower diagonal values of an SPD matrix \( (x; y; z) \). The proposed depth produces reasonable results by showing a center-outward profile analogous to Tukey’s depth in the Euclidean space. The deepest point in red is tightly surrounded by data points with gradually decreasing depth values, and the deepest point is not heavily drawn by data points with large values in the diagonal elements \( x \) and \( z \). The peripheral points all have the least depth.

**Example 4 (BHV space of phylogenetic trees).** We model phylogenetic trees in the Billera–Holmes–Vogtmann (BHV) tree space (Billera, Holmes, and Vogtmann 2001), a widely investigated
geodesic space with nice geometry. Let $\mathcal{M} = \mathbb{T}^k$ denote the space of rooted phylogenetic trees with $k$ leaves endowed with the BHV geometry (Billera, Holmes, and Vogtmann 2001), where a brief summary for the BHV geometry and the associated metric halfspaces is included in Section S2.4, supplementary material.

We illustrate here the geometry of the simplest tree space $\mathbb{T}^3$ with three leaves and one interior edges. The topology of the tree is the way leaves and interior nodes are connected. There are three bifurcating tree topologies, respectively, corresponding to which of leaf A, B, and C branches out first, and a star tree topology with a degenerate interior edge. Tree space $\mathbb{T}^3$ is represented by the 3-spider $(\mathbb{R}^3 \times \{1, 2, 3\})/\sim$, formed by three rays identified at the origin $o$. Coordinate $(a, j)$ represents a point (tree) lying on the $j$th leg of the 3-spider at a distance $a$ from the origin; we refer to this representation of the trees as the (radius, branch)-coordinate. The equivalence relationship $\sim$ is defined by $(a_1, j_1) \sim (a_2, j_2)$ if and only if $(a_1, j_1) = (a_2, j_2)$ for $a_1 > 0$ and for $a_1 = a_2 = 0$. The three legs of the spider correspond to three different bifurcating tree topologies, and the position of a point on a leg corresponds to the length of the branches consists of the line segments connecting each to the two points so that the lines segment connecting $a$ center-outward ordering, Zuo and Serfling (2000) postulated that the depth of a point is invariant to affine transformations; (c) the depth of symmetry, in the data, then the depth achieves its maximum at this center; and (d) Center-outward monotonicity, that is, depth values gradually decrease as one moves away from the deepest point. These properties are satisfied by classical Tukey’s depth (Tukey 1975).

We will show that the four depth properties are satisfied to a great extent by the proposed metric halfspace depth under regularity conditions as detailed in the next theorem. To state these properties on a general metric space, due to the lack of a vector space structure, we need to address the lack of affine data transformation and introduce an invariance property, a notion of data symmetry, and monotonicity.

For an invariance property, let $f$ be a transformation from $(\mathcal{M}, d)$ to another metric space $(\mathcal{N}, e)$. For any $y \in \mathcal{N}$, let $H_{y,e} = \{H_{y,z,e} \mid z \in \mathcal{N}\}$ be the collection of halfspaces $H_{y,z,e} = \{x \in \mathcal{N} \mid e(x, y) \leq e(x, z)\} \subset \mathcal{N}$ containing $y$. We say that $f$ is halfspace preserving at $x \in \mathcal{M}$ with respect to $(\mathcal{M}, d)$ and $(\mathcal{N}, e)$, or simply halfspace preserving at $x$ if $H_{f(x),e} = f(H_x) := \{f(H) \mid H \in H_x\}$, in which case the collection of halfspaces containing $x$ is preserved by $f$. We say that $X$ is halfspace symmetric about $\theta \in \mathcal{M}$ if $P(X \in H) \geq 1/2$ for all halfspace $H$ containing $\theta$, extending the same notion defined in the Euclidean space by Zuo and Serfling (2000). To define monotonicity on a metric space, we restrict attention to geodesic spaces, where monotonicity of the depth function can be investigated along geodesics leaving from the deepest point.

For theory development, we require $(\mathcal{M}, d)$ to be a connected complete separable metric space. For a subset $S \subset \mathcal{M}$, let $\overline{S}$, $\partial S$ denote the interior, boundary, and complement of $S$, respectively. Proofs for the theoretical results and additional analytical properties of the halfspaces are included in Sections S7 and S2.5, supplementary material, respectively.

**Theorem 1.** The metric halfspace depth $D(\cdot)$ satisfies the following properties.

(a) *(Transformation invariance)* Let $f : \mathcal{M} \rightarrow \mathcal{N}$ be a bijection measurable map between metric spaces $(\mathcal{M}, d)$ and $(\mathcal{N}, e)$, and $D_f(y) = \inf_{H \in H_{y,e}} P(H_f(X))$ denote the depth at $y \in \mathcal{N}$ with respect to the pushforward measure $P_{f(X)} = P_X \circ f^{-1}$ on $\mathcal{N}$. If $f$ is halfspace preserving at $x \in \mathcal{M}$, then $D(f(x)) = D_f(f(x))$.

(b) *(Vanishing at infinity)* Let $o \in \mathcal{M}$ be an arbitrary point. Then $\sup_{x,d(o,x) > L} D(x) \rightarrow 0$ as $L \rightarrow \infty$, taking the convention here that the supremum over an empty set is 0.

(c) *(Maximality at the symmetry center)* If $X$ is halfspace symmetric about a unique center $\theta$, then $\theta$ is the unique deepest point, that is, $\theta = \arg\max_{x \in \mathcal{M}} D(x)$.

(d) *(Center-outward monotonicity)* Suppose $\mathcal{M}$ is a geodesic space. Let $\theta \in \mathcal{M}$ be a deepest point, $x \in \mathcal{M}$, and $\gamma : [0, 1] \rightarrow \mathcal{M}$ a geodesic joining $\theta$ to $x$. If any halfspace $H_{\gamma(s),e}$ of $\mathcal{M}$ that has a nonempty intersect with $\gamma([0, 1])$ contains at least one of $x$ and $\theta$, then $D(\gamma(t)) \leq D(x)$ holds for $t \in [0, 1]$.

**Theorem 1**(a) states that the metric halfspace depth is invariant to transformation $f$ that preserves halfspaces. It is immediate that affine transformations and rotations are halfspace preserving, respectively, between Euclidean spaces and between spheres of the same dimension at all $x \in \mathcal{M}$. Thus, this result implies the transformation invariance properties of Tukey’s depth (Donoho and Gasko 1992) and angular Tukey’s...
depth (Liu and Singh 1992). More generally, a map \( f \) is halfspace preserving at \( x \) if it preserves the order of distances at \( x \), that is, for \( x, x_1, x_2 \in M, d(x_1, x) \leq d(x_2, x) \) if and only if \( e(f(x_1), f(x)) \leq e(f(x_2), f(x)) \). This is clearly satisfied if \( f \) is an isometry, that is, \( d(x, y) = e(f(x), f(y)) \) for all \( x, y \in M \).

The depth follows a center-outward tendency. In a space where “infinite” is well-defined, Theorem 1(b) states that the depth of a point vanishes as the point moves toward infinity. Therefore, the peripheral data points will have a small depth. Theorem 1(c) states that if the data distribution is halfspace symmetric about a unique center \( t \), then the halfspace depth is maximized at this center \( t \). We consider halfspace symmetry to define data symmetry on a general metric space, which does not require the space \( M \) itself to be symmetric, thereby generalizing beyond the Euclidean space and spheres (Liu and Singh 1992). In the Euclidean space, Zuo and Serfling (2000) showed that halfspace symmetry is weaker than alternative symmetry notions such as centrally symmetric, that is, \( X - \theta \) and \( X \) equal in distribution, and angularly symmetric, which requires \( (X - \theta) / ||X - \theta|| \) to be centrally symmetric.

Between the deepest \( \theta \) and an arbitrary location \( x \), Theorem 1(d) states that the metric halfspace depth is nonincreasing along geodesics leaving from \( \theta \) if the metric space satisfies a geometric condition. The geometric condition requires that the halfspaces in \( M \) are not overly rich so they will not single out points on the geodesic connecting \( \theta \) and \( x \) while excluding the endpoints. This condition is satisfied by the model spaces, namely the Euclidean space, sphere, and hyperbolic space, as stated in Proposition 1.

**Proposition 1.** Let \( M \) be one of the \( m \)-dimensional model spaces, namely, the Euclidean space \( \mathbb{R}^m \), unit sphere \( S^m \), or hyperbolic space \( \mathbb{H}^m \), and \( \gamma : [0, 1] \rightarrow M \) be a geodesic joining \( \theta \) to \( x \). Then any halfspace \( H \subset M \) with a nonempty intersect with \( \gamma([0, 1]) \) contains at least one of \( \theta \) and \( x \).

We next show the upper semi-continuity of the depth function \( D(\cdot) \) and the compactness and nestedness of the depth regions \( D^\alpha := \{ x \in M | D(x) \geq \alpha \}, \alpha > 0 \). Define \( P_H : M \times M \rightarrow \mathbb{R} \) as \( P_H(x_1, x_2) = P_X(H(x_1, x_2)) \) and let \( E_{x_1, x_2} = \{ x \in M | d(x, x_1) = d(x, x_2) \} \) be the equidistance set anchored at \( x_1, x_2 \in M \). A metric space is locally compact if every point has a compact neighborhood. All finite-dimensional manifolds and BHV tree spaces are locally compact.

**Proposition 2.** Suppose that \( M \) is a complete and locally compact geodesic space.

(a) \( P_H(\cdot, \cdot) \) is upper-semi continuous. If further \( P_X(E_{x_1, x_2}) = 0 \) for all \( x_1 \neq x_2 \in M \), then \( P_H(\cdot, \cdot) \) is continuous.

(b) \( D(x) \) is upper semi-continuous.

(c) \( D^\alpha \) is nested, that is, \( D^{\alpha_1} \subset D^{\alpha_2} \) for \( \alpha_1 \geq \alpha_2 \), and \( D^\alpha \) is compact for \( \alpha > 0 \).

The additional condition in Proposition 2(a) is satisfied if \( M \) is a Riemannian manifold and \( X \) has a density w.r.t. the Riemannian volume measure (Lee 2018); for example, this is satisfied if \( M \) is the unit sphere and \( X \) follows a warped normal distribution.

### 3.2. Convergence of the Depth Function and Deepest Point

Next, we show that the metric halfspace depth can be estimated consistently by its sample version uniformly over all locations by making use of empirical process theory. Let \( L_2(Q) \) be the \( L_2 \)-norm of measurable functions with respect to probability measure \( Q \) on the sigma-algebra of \( M \), so \( L_2(Q)(f) = \int f(x)^2 dQ(x) \) \( 1/2 \).

For a set of measurable functions \( F \), the covering number \( N(\varepsilon, F, L_2(Q)) \) is the minimal number of balls in \( L_2(Q) \) with radius \( \varepsilon \) required to cover \( F \).

The bracketing number \( N_l(\varepsilon, F, L_2(Q)) \) is the minimal number of \( \varepsilon \)-brackets required to cover \( F \). An \( \varepsilon \)-bracket \([ l, u] \) is the set of functions \( f \) with \( l \leq f \leq u \), given two functions \( l \) and \( u \) with \( ||u - l||_{L_2(Q)} < \varepsilon \). The covering and bracket numbers for a collection of measurable sets are by convention those of the corresponding collection of indicator functions. Either one of the following conditions is needed for the convergence results.

\[
\begin{align*}
(\text{N1}) & \quad \sup_Q \int_0^\infty \log(N(\varepsilon, H, L_2(Q))^{1/2} d\varepsilon < \infty, \\
(\text{N2}) & \quad \int_0^\infty \log(N_l(\varepsilon, H, L_2(P_X))^{1/2} d\varepsilon < \infty.
\end{align*}
\]

**Theorem 2.** Given iid observations \( X_1, \ldots, X_n \) from \( P_X \), if either (N1) or (N2) holds,

\[
E \sup_{x \in M} |D_n(x) - D(x)| = O(n^{-1/2}).
\]

Condition (N1) and (N2) are common entropy/bracketing integral conditions imposed on the complexity of the collection of halfspaces in order to guarantee convergence of the empirical process. If (N1) holds, then the statement of Theorem 2 is uniform not only in \( x \) but also over the underlying distribution \( P_X \). Condition (N1) holds if the Vapnik–Chervonenkis (VC) dimension of \( H \) is finite (Theorem 2.6.4, van der Vaart and Wellner 1996). Let \( C \) be a collection of subsets of \( M \). We say that \( C \) shatters a finite subset \( F = \{ x_1, \ldots, x_n \} \subset M \) if \( C \cap F := \{ C \cap F | C \in C \} \) is the collection of all subsets of \( F \). The Vapnik–Chervonenkis (VC) dimension of \( C \) is the smallest \( n \) for which no set of size \( n \) is shattered by \( C \), formally defined by

\[
\text{VC}(C) = \inf \{ n | \max_{x_1, \ldots, x_n} \Delta_n(C, x_1, \ldots, x_n) < 2^n \},
\]

where \( \Delta_n(C, x_1, \ldots, x_n) := || \{ C \cap \{ x_1, \ldots, x_n \} | c \in C \} || \) is the number of subsets of \( \{ x_1, \ldots, x_n \} \) picked out by \( C \). It is well known that the VC dimension of halfspaces in the Euclidean space \( \mathbb{R}^m \) is \( m + 2 \) (Wenocur and Dudley 1981). Theory on the VC dimensions of subsets of a Riemannian manifold (Narayanan and Niyogi 2009) or of a general metric space has been highly limited. That said, since the collection of halfspaces \( H \) is indexed by two points on the metric space, it may be reasonable to expect \( \text{VC}(H) \) to be finite if the geometry of \( M \) is regular enough. We establish the boundedness of VC dimensions for the collections of halfspaces on the sphere \( S^m \) and the space \( \mathbb{T}^3 \) of phylogenetic trees with 3 leaves.

**Proposition 3.** The following holds:

(a) On an \( m \)-dimensional sphere \( M = S^m \), \( \text{VC}(H) \leq m + 3 \).

(b) On the space of phylogenetic trees \( M = \mathbb{T}^3 \) with 3 leaves, \( \text{VC}(H) = 4 \).
By Theorem 2, Proposition 3 implies \( n^{1/2} \)-convergence for the empirical metric halfspace depth on these spaces.

A deepest point w.r.t. the sample is a consistent estimator of the population deepest point by \( M \)-estimation theory.

**Proposition 4.** Suppose that \( M \) is a complete and locally compact geodesic space, \( D(\cdot) \) has a unique maximum \( \theta = \arg\max_{x \in M} D(x) \), and the conditions of Theorem 2 hold. Let \( \theta_n \) be an arbitrary point in the deepest set \( S_n := \arg\max_{x \in M} D_n(x) \). Then

\[
d(\theta_n, \theta) \to 0 \quad \text{a.s.}
\]
as \( n \to \infty \).

By Theorem 1(a), the deepest set \( S_n \) is invariant to halfspace preserving transformations. In the asymptotic limit, the deepest set shrinks to the population deepest point if the latter is unique, so any sample deepest point is near-invariant.

### 3.3. Robustness

A depth median is defined as an estimator \( T(\cdot) \) that takes a point cloud \( Z = \{z_1, \ldots, z_n\} \) on \( M \) to a choice of point \( T(Z) \in \arg\max_{x \in M} D(x; P_Z) \) in the deepest set w.r.t. metric halfspace depth, where \( P_Z \) is the empirical measure placing equal point mass on each point in \( Z \). Given a sample, a depth median yields a (unique) point as output, but there exists potentially more than one depth medians (as estimators) in general if the deepest set is non-singleton. The depth median of a point cloud is interpreted as the most representative point of the data and can be used as a location descriptor/estimator. In the Euclidean case, Tukey's median is a generalization of the classical median on the real line. Robustness and asymptotic properties were investigated in Donoho and Gasko (1992) and Massé (2004), respectively.

The breakdown property of the metric halfspace median, which is the depth median based on our metric halfspace depth, is analyzed next. Intuitively, the breakdown point is the smallest fraction of contamination that brings an estimator to infinity. Formally, let \( X(n) = \{X_1, \ldots, X_n\} \) be a sample of \( n \) observations and \( Y(l) = \{Y_1, \ldots, Y_l\} \) be \( l \) contamination points. The breakdown point \( \epsilon^* \) of a metric halfspace median \( T(\cdot) \) in a sample \( X(n) \) is the smallest fraction of contamination to bring the estimate in the contaminated sample arbitrarily far away from that of the uncontaminated sample. The finite-sample (addition) breakdown point is defined as

\[
\epsilon^* = \epsilon^*(T; X(n)) := \min_l \left\{ \frac{l}{n+1} \left| \sup_{Y(l)} d(T(X(n)), T(X(n), Y(l))) = \infty \right. \right\},
\]

where we set \( \epsilon^* = 1 \) if the set being minimized is empty. The next proposition and its corollary analyze the finite-sample and asymptotic behavior of the breakdown point, respectively.

**Proposition 5.** Let \( M \) be an arbitrary metric space. For any metric halfspace median \( T(\cdot) \), it holds that

\[
\epsilon^* \geq \frac{D_n(\theta_n)}{1 + D_n(\theta_n)},
\]

where \( \theta_n = T(X(n)) \) is a deepest point w.r.t. sample \( X(n) \).

**Corollary 1.** If (N1) or (N2) holds, then as \( n \to \infty \),

\[
\epsilon^* \geq \frac{D(\theta)}{1 + D(\theta)} \quad \text{a.s.,}
\]

where \( \theta = \arg\max_{x \in M} D(x) \) is any deepest point w.r.t. the distribution \( P_X \) of \( X \).

Corollary 1 implies that the breakdown point for the metric halfspace median for any halfspace symmetric distribution is at least 1/3 regardless of the dimensions, extending the results for Tukey’s median in the Euclidean case (Donoho and Gasko 1992).

### 4. Efficient Computation

#### 4.1. Approximation Algorithms

In the Euclidean space, exact computation of Tukey’s depth and deepest point are prohibitively slow if the dimension is higher than 3 even with efficient algorithms (Dyckerhoff and Mozharovskyi 2016). On a general metric space, the evaluation of the metric halfspace depth as an infimum faces additional difficulty and would require optimization algorithms that adapt to specific manifolds (Yang 2007). Moreover, the search for the deepest point requires difficult optimization of a discontinuous function \( D_n(\cdot) \). This motivates us to develop fast approximation algorithms for the metric halfspace depth and deepest point.

Let \( X = \{X_1, \ldots, X_n\} \subset M \) be the collection of observations, and also denote \( A \subset M \) as the anchor set containing \( |A| = n_A \) anchor points of halfspaces. We approximate \( D_n(x) \) w.r.t. \( M \) by taking the infimum over only halfspaces anchored at points in \( A \). The proposed metric halfspace depth approximation is

\[
\tilde{D}_n(x) = \tilde{D}_n(x; A) = \inf_{x_1 \neq x_2 \in A : d(x_1, x_2) \leq d(x_1, x_2)} \sum_{i=1}^{n-1} l(d(X_i, x_1) \leq d(X_i, x_2)).
\]

The infimum is taken over at most \( n_A(n_A - 1) \) ordered pairs of anchors. The number of anchors controls the tradeoff between computational cost and accuracy, in that using a larger number of anchors results in a better approximation but at a higher cost. In most applications, the anchor points \( A \) can be set to the sample points \( X \), and for improving approximation, one can enlarge the set of anchor points by including “jiggled” versions of these points; more information is included in Section S3.2, supplementary material. The deepest point is approximated by the in- and out-of-sample points with the largest approximate depth, defined, respectively, by

\[
\tilde{\theta} = \arg\max_{x \in A} \tilde{D}_n(x), \quad \hat{\theta} = \arg\max_{x \in M} \tilde{D}_n(x).
\]

The in-sample deepest point \( \tilde{\theta} \) can serve as a good initial value in numerical optimization procedures to search for the out-of-sample deepest \( \hat{\theta} \), where the latter is a more accurate approximation of \( \hat{\theta} \).

Like their population and sample versions, the approximate depth and deepest points incorporate the geometry of \( M \) through the metric \( d \) and thus, avoids the choice of a parameterization of the metric space or linearization onto the tangent
space, both of which could be ill-defined. The approximate depth is defined as long as the discrete graph of pairwise geodesic distances is given, and thus, the proposed depth is applicable to a wide range of scenarios where the available data are nodes and edges of a graph (Small 1997) or where the pairwise geodesics are estimated from a point cloud using a graph-based method (Tenenbaum, De Silva, and Langford 2000). Algorithms for computing depth and the deepest point are summarized in Algorithm 1 and Algorithm 2, respectively.

Algorithm 1: Evaluate depth at points in \( \mathcal{Y} \) w.r.t. \( \mathcal{X} \)

**Data:** Random sample \( \mathcal{X} \), depth evaluation points \( \mathcal{Y} \), and halfspace anchors \( \mathcal{A} \)

**Result:** Depths \( D(y) \) for \( y \in \mathcal{Y} \)

1. for \( x_1 \neq x_2 \in \mathcal{A} \) do
2. \[ p_{x_1,x_2} \leftarrow P_{\mathcal{X}}(H_{x_1,x_2}) \]
3. end
4. for \( y \in \mathcal{Y} \) do
5. \[ Q \leftarrow \emptyset \]
6. for \( x_1 \neq x_2 \in \mathcal{A} \) do
7. if \( d(y,x_1) \leq d(y,x_2) \) then
8. \[ \text{Add} \ p_{x_1,x_2} \text{ to } Q \]
9. end
10. end
11. \( D(y) \leftarrow \min Q \)
12. end

The complexity of Algorithm 1 is \( O(n_\mathcal{Y}n_\mathcal{A}^2 + n^3) \) for evaluating depth at points in \( \mathcal{Y} \) w.r.t. sample \( \mathcal{X} \) and anchor points \( \mathcal{A} = \mathcal{X} \), where \( n_\mathcal{Y} = |\mathcal{Y}| \). The rate of complexity does not have an exponent involving dimension \( m \), similar to those of the approximation algorithms (e.g., Bogićević and Mérk 2018; Zuo 2019, and the references therein) for computing Tukey’s depth in the Euclidean space. This contrasts with the exact algorithms (e.g., Dyckerhoff and Mozharovskyi 2016; Zuo 2019) for computing Tukey’s depth where the complexity is typically \( O(n_\mathcal{Y}n_\mathcal{X}^m) \) or \( O(n_\mathcal{Y}n_\mathcal{X}^{m-1}\log(n)) \). Algorithm 2 takes \( O(n^3) \) since \( n_\mathcal{Y} = n \).

### 4.2. Theoretical Properties for the Approximation

We establish that the approximate depth converges to the truth if the anchor points are dense enough in \( \mathcal{M} \). Halfspace \( H_{z_1,z_2} \) is said to be a minimizing halfspace at \( x \) if \( x \in H_{z_1,z_2} \) and \( P_{\mathcal{X}}(H_{z_1,z_2}) = D(x) \). The following theorem derives the rate of convergence for the approximation if a minimizing halfspace exists, and the consistency result otherwise. To obtain the rate of convergence, for \( z_2 \in \mathcal{M} \) let \( D_j = d(X, z_j), j = 1, 2 \) and assume the following conditions.

1. For some \( \epsilon > 0 \) and \( c_1 > 0 \), \( D_j \) has a small ball probability near 0 satisfying \( P(D_j \leq t) \geq c_1t^{m-1} \) for \( j = 1, 2 \) and \( t \leq \epsilon \).
2. For some \( \epsilon > 0 \) and \( c_2 > 0 \), \( P(|D_1 - D_2| \leq t) \leq c_2\epsilon \) holds for \( t \leq \epsilon \).

**Theorem 3.** Suppose that either (N1) or (N2) holds, and the approximation algorithm uses the sample points \( \mathcal{X} \) as the anchor points \( \mathcal{A} \). Let \( x \) be a point on \( \mathcal{M} \).

(a) If the infimum in \( D(x) = \inf_{H \in \mathcal{H}} P_{\mathcal{X}}(H) \) is achieved by a halfspace \( H_{z_1,z_2} \), that is, \( D(x) = P_{\mathcal{X}}(H_{z_1,z_2}) \), and (P1) and (P2) hold for \( (z_1, z_2) \), then as \( n \to \infty \),

\[
\left| \tilde{D}_n(x) - D(x) \right| = O_p n^{-1/m_0}.
\]

(b) Suppose that the infimum of \( D(x) = \inf_{H \in \mathcal{H}} P_{\mathcal{X}}(H) \) is not achieved by any halfspace. If \( P(d(z,X) < r) > 0 \) for all \( z \in \mathcal{M} \) and \( r > 0 \) and \( P_{\mathcal{X}}(E_{z_1,z_2}) = 0 \) for all \( z_1,z_2 \in \mathcal{M} \), then as \( n \to \infty \),

\[
\left| \tilde{D}_n(x) - D(x) \right| = o_p(1).
\]

The idea of proof for Theorem 3 is to approximate the minimizing halfspace probabilities by random halfspaces. The halfspace where the infimum is attained does not need to be unique. Conditions (P1) and (P2) are requirements on both the distribution of \( X \) and on the geometry of \( \mathcal{M} \). They ensure that if the random anchor points lie close enough to the anchor points of a minimizing halfspace, then the halfspace probabilities are close. If \( \mathcal{M} \) is a Riemannian manifold and \( X \) has a density bounded away from 0 w.r.t. the Riemannian volume measure, then \( m_0 \in \{P1\} \) is the intrinsic dimension \( m \) of \( \mathcal{M} \). Thus, the rate of convergence of the approximation algorithm given by Theorem 3(a) is as fast as \( O_p(n^{-1/m}) \) on an \( m \)-dimensional Riemannian manifold. Conditions (P1) and (P2) hold in a Euclidean space if \( X \) has a finite first moment and density bounded away from zero and infinity, and (P2) is violated if the distribution of \( d(X, z_1) \) is overly concentrated around \( d(X, z_2) \). To give details, we describe two examples when (P1) and (P2) are satisfied and a counter-example in Section S8, supplementary material, and additional properties of the approximate depth in Section S9, supplementary material.

### 5. Numerical Experiments

We investigate the performance of the metric halfspace median as a robust estimate for the center of a distribution. Three Riemannian manifolds were considered for the data space \( \mathcal{M} \), namely the \( k \times k \) symmetric positive definite matrices \( SPD(k) \) with the affine invariant metric; the \( k \)-dimensional unit sphere \( S^k \); and the rotational group \( SO_k(k) \) of \( k \times k \) orthogonal matrices with determinant 1. The intrinsic dimensions \( m \) for these manifolds equal, respectively, \( k(k+1)/2, k, \) and \( k(k-1)/2 \). The definitions of the tangent spaces and their bases, and exponential maps are described in Sections S1 and S2, supplementary material.

For each metric space \( \mathcal{M} \), we considered four cases where iid data were generated according to either an uncontaminated distribution \( P = P_1 \) for Case 1 or contaminated distribution \( P = 0.9P_1 + 0.1P_2 \) for Cases 2 to 4. Under Case 1, independent
samples $X_i$, $i = 1, \ldots, n$ were generated according to $P = P_1$, where $P_1$ is the law of random variable $X = \exp_{\theta_1} V_1$; $\exp_{\theta_1}: T_{\theta_1} M \to M$ is the exponential map at the center $\theta_1 \in M$ of the uncontaminated distribution; $T_{\theta_1} M$ is the tangent space at $\theta_1$; and $V_1$ is a nonisotropic normal random variable lying on $T_{\theta_1} M$. Let $B_{ij}, j = 1, \ldots, m$ be an orthonormal basis on $T_{\theta_1} M$, and set $V_i = \sum_{j=1}^m Z_j B_{ij}$ where $Z_j$ follows independent $N(0, \sigma_j^2)$ with $\sigma_j/\sigma_{j+1} = 3$ for $j = 1, \ldots, m - 1$, having a total variance $\sum_{j=1}^m \sigma_j^2 = 1$. For Cases 2 to 4, iid data $X_i$ were generated under mixture distributions $P = 0.9P_1 + 0.1P_2$ with 10% of data coming from the contaminating distribution $P_2$ that varied between different cases. In Case 2, $P_2$ was set to a location contamination with the same distribution as $\exp_{\theta_2} V_2$, where $\theta_2 = \exp_{\theta_1} U$ is a random location at a unit distance away from $\theta_1$, $U$ is sampled (once per Monte Carlo repeat) from the uniform distribution on the unit sphere on $T_{\theta_1} M$, $V_2 = V_2' := \sum_{j=1}^m Z_j' B_{2j}$, $Z_j'$ follows independent $N(0, \sigma_j^2)$, and $B_{2j}$ is an orthonormal basis of $T_{\theta_2} M$, $j = 1, \ldots, m$; in Case 3, $P_2$ was a scale contamination sharing the same distribution as $\exp_{\theta_2} V_2$ where $\theta_2 = \theta_1$ and $V_2 = S := \sum_{j=1}^m W_j B_{1j}$ is a zero-mean multivariate normal distribution on $T_{\theta_1} M$, and $W_j$ follows independent $N(0, \sigma_j^2)$ with variance $\sigma_j = \sigma_{m-j+1}$, $j = 1, \ldots, m$, differing from that for $P_1$; in Case 4, $P_2$ was a location-and-scale contamination sharing the same distribution as $\exp_{\theta_2} V_2$ where $\theta_2$ is the same as in Case 2 and $V_2$ is the same as in Case 3.

Our target is to estimate robustly the center $\theta_1$ of the uncontaminated distribution $P_1$, with the center and random tangent vector varying between simulation cases. The contamination distribution $P_2$ was set to a distribution that differed from $P_1$, in the location, scale, and location-and-scale for Case 2, 3, and 4, respectively. To summarize, the simulation scenarios considered were

- Case 1, uncontaminated distribution centered at $\theta_1$,
- Case 2, contaminated distribution with location outliers,
- Case 3, contaminated distribution with scale outliers, and
- Case 4, contaminated distribution with location-and-scale outliers.

For example, on $S^2$, a location outlier is centered around $\theta_2$ that lies far away from the center $\theta_1$ of the uncontaminated distribution $P_1$. A scale outlier is generated from $V_2$ which has a different covariance matrix than $V_1$; therefore, a scale outlier may lie away from its center in a direction uncommon to the inliers. We varied the sample size $n \in \{50, 100, 200\}$ and the manifold parameter $k \in \{2, 3, 4\}$ in each case.

As estimators of the center, we compared the proposed metric halfspace median $\hat{\mu}_{\text{MHD}} = \theta$ as defined in (7) and the Fréchet mean $\hat{\mu}_{\text{FM}}$. The Fréchet mean (Fréchet 1948) of the sample $X_i$, $i = 1, \ldots, n$ under distance $d$ is $\hat{\mu}_{\text{FM}} = \arg \min_{\mu \in M} n^{-1} \sum_{i=1}^n d(x, X_i)$, which is a generalization of the classical mean. For calculating the metric halfspace median, 10 jiggled points were added to the anchor set around each sample point. We also compared with the Fréchet median $\hat{\mu}_{\text{FMd}} = \arg \min_{x \in M} n^{-1} \sum_{i=1}^n d(x, X_i)$, which coincides with the deepest point w.r.t. the geodesic distance depth proposed by Chau et al. (2019) on $M = \text{SPD}(k)$. These location estimators $\hat{\mu}$ were evaluated according to the median geodesic distance to the true mean $d(\hat{\mu}, \mu)$ out of 1024 Monte Carlo repeats.

Results for $M = \text{SPD}(k)$ displayed in Table 1 show that the proposed metric halfspace median performs well in general. In Case 1 without contamination, the Fréchet mean was the most efficient overall, especially for smaller sample sizes $n = 50$ and 100, while the metric halfspace median and the Fréchet median are competitive. In the presence of contamination, both deepest points $\hat{\mu}_{\text{MHD}}$ and $\hat{\mu}_{\text{FMd}}$ dominated $\hat{\mu}_{\text{FM}}$ and demonstrated robustness by producing estimates that were close in performance to those in Case 1 without contamination. The proposed metric halfspace median outperformed the Fréchet median in the contaminated scenarios. A reason for this is that the Fréchet median only considers the sum $\sum_{i=1}^n d(x, X_i)$ of geodesic distances from the data points to $x$, disregarding the relative locations of the data points within the point clouds and thus, having weaker invariant properties than the metric halfspace depth. The advantage of $\hat{\mu}_{\text{MHD}}$ over $\hat{\mu}_{\text{FMd}}$ becomes more significant when the sample size is larger, in which case the approximation of the metric halfspace depth through $D_n$ is improved.

Results for two bounded manifolds are shown in Table 2, where the exponential maps are not injective on these manifolds and thus, depth concepts cannot be defined in general through mapping data onto the tangent space. The metric halfspace median is overall superior to the Fréchet mean in the presence of contamination, especially when the intrinsic dimension is large. Even in Case 3, where the scale-only outliers do not affect the true center, the metric halfspace median was, in many cases, more efficient than the Fréchet mean on spheres. This could be due to the low rate of convergence of the sample Fréchet mean for data that extends the entire manifold (Eltzner and Huckemann 2019). On the bounded manifolds, the metric halfspace median was overall comparable with the Fréchet median, slightly outperforming the latter on $S^k$ in Case 4 when $n = 200$, and slightly under-performing it on $SO(k)$. The advantage of the metric halfspace median over the Fréchet median is clearly more significant on $\text{SPD}(k)$ than on $S^k$ and $SO(k)$. Some possible

---

**Table 1.** Median distances over the replicates between the estimated center and the actual center of $P_1$ for data being symmetric positive definite matrices on $M = \text{SPD}(k)$.

| $n$   | $k = 2$ |   | $k = 3$ |   | $k = 4$ |   |
|-------|---------|---|---------|---|---------|---|
|       | MHD     | FM | FMd     |   | MHD     | FM | FMd     |   |
| Case 1|         |   |         |   |         |   |         |   |
| 50    | 0.117   | 0.103| 0.122   |   | 0.116   | 0.103| 0.121   |   |
| 100   | 0.075   | 0.071| 0.081   |   | 0.076   | 0.071| 0.080   |   |
| 200   | 0.053   | 0.049| 0.054   |   | 0.053   | 0.048| 0.054   |   |
| Case 2|         |   |         |   |         |   |         |   |
| 50    | 0.124   | 0.140| 0.136   |   | 0.114   | 0.140| 0.124   |   |
| 100   | 0.091   | 0.120| 0.101   |   | 0.084   | 0.121| 0.097   |   |
| 200   | 0.070   | 0.108| 0.084   |   | 0.064   | 0.110| 0.079   |   |
| Case 3|         |   |         |   |         |   |         |   |
| 50    | 0.104   | 0.107| 0.108   |   | 0.102   | 0.108| 0.104   |   |
| 100   | 0.072   | 0.075| 0.074   |   | 0.064   | 0.075| 0.071   |   |
| 200   | 0.051   | 0.054| 0.051   |   | 0.042   | 0.053| 0.050   |   |
| Case 4|         |   |         |   |         |   |         |   |
| 50    | 0.123   | 0.142| 0.133   |   | 0.122   | 0.145| 0.124   |   |
| 100   | 0.087   | 0.124| 0.102   |   | 0.084   | 0.124| 0.094   |   |
| 200   | 0.065   | 0.110| 0.086   |   | 0.061   | 0.112| 0.078   |   |

**NOTE:** The dimensions of the manifold for parameters $k = 2, 3, 4$ are 3, 6, and 10, respectively. The standard errors of the reported median distances for $n = 50, 100$, and 200 were less than 0.003, 0.002, and 0.002, respectively. MHD, the proposed metric halfspace depth median; FM, Fréchet mean; FMd, Fréchet median.

---
reasons for this are that the directionality of the outliers and the transformation invariance property of the proposed metric halfspace depth are more relevant on the unbounded SPD than on the bounded manifolds. Overall, results for the different manifolds demonstrate that the proposed metric halfspace median is in general a valid robust measure of centrality. Moreover, our proposed depth can be generally applied to rank general data objects in a center-outward fashion, as demonstrated in the real data applications.

6. Real Data Applications

6.1. Functional Connectivity in Alzheimer’s Disease Patients

The first data application considers symmetric positive definite (SPD) matrices that represent brain connectivity, which are widely used as a biomarker of brain function. The connectivity between defined regions of interest is calculated as the temporal association between their blood-oxygen-level-dependent (BOLD) signals in functional magnetic resonance imaging (fMRI) scans when the subjects are in a resting state. We analyzed fMRI scans recorded in the Alzheimer’s Disease Neuroimaging Initiative (ADNI) with the goal of making inference regarding brain connectivity in different dementia study groups. Our analysis included n = 181 subjects who, according to the severity of cognitive decline, were classified at enrollment as: cognitively normal (CN), early mild cognitive impaired (EMCI), late mild cognitive impaired (LMCI), or Alzheimer’s disease (AD) patients. The fMRI data were preprocessed by following a standard protocol to remove motion and timing artifacts, scaling effects, and trends, and we considered only the fMRI scans at the participants’ first visits. Problematic scans are not uncommon in fMRI studies as a result of imaging artifacts that come from head motion and cognitive state (Laumann et al. 2017). Statistical depth approaches are appealing for analyzing imaging data since they are fully nonparametric and robust to outliers. Here we compare the proposed metric halfspace depth with the geodesic distance depth (Chau et al. 2019).

For each subject, the average bold signals in each of the 10 defined brain regions (Buckner’s hubs) in a subject’s brain were first calculated, obtaining a 10-dimensional times series (Buckner et al. 2009). Next, brain connectivity is represented by the covariance (at lag 0) of the average bold signals, obtaining 10 × 10 covariance matrices as the data observations $X_i$. The left panel of Figure 3 illustrates the connectivity covariance matrices of four random subjects in the cognitively normal group. We analyzed the covariance matrices in $M = \text{SPD}(10)$ with the affine invariant metric. The deepest covariance matrix in the cognitively normal group with respect to metric halfspace depth (upper right panel of Figure 3) exhibits nonzero cross-covariances between different brain regions, resembling the original sample matrices; in contrast, the deepest image w.r.t. the geodesic distance depth (Chau et al. 2019) (lower right panel) has near 0 cross-covariances, which is not commonly observed in the sample.

We next investigated whether group differences exist among the four groups of patients studied. We applied the depth-based Kruskal–Wallis test proposed by Chenouri and Small (2012) based on both the proposed metric halfspace depth and the geodesic distance depth (Chau et al. 2019). The Kruskal–Wallis test is designed to be sensitive to both location and scale changes by calculating the depth of the observations with respect to each group and aggregating the depth ranks. Using the permutation null distribution, a $p$-value of 0.0194 was produced using the proposed metric halfspace depth, and a $p$-value of 0.0652 for the geodesic distance depth. Further, pairwise comparisons of the dementia groups using the depth-based Wilcoxon test (Chenouri and Small 2012) revealed that the most significant difference exists between the Alzheimer’s disease and the cognitively normal groups as shown in Table 3. This demonstrates the potential utility of fMRI-based connectivity measures and depth-based methods for studying Alzheimer’s disease.

6.2. Phylogenetic Tree Application

In evolutionary biology, the ancestral relationship among a fixed collection of species is represented by a tree structure. Each leaf corresponds to a species, each interior node a speciation
Figure 3. Left: Connectivity covariance matrices for four cognitively normal individuals. Right: Deepest matrices among cognitively normal individuals in terms of metric halfspace depth (MHD, upper panel) and geodesic distance depth (GDD, lower panel). Brain regions used for creating the connectivity matrices are indicated.

Table 3. The \( p \)-values based on the metric halfspace depth-based Wilcoxon rank test for the pairwise comparisons between the four dementia groups.

|       | EMCI | LMCI | AD  |
|-------|------|------|-----|
| CN    | 0.644| 0.339| 0.021|
| EMCI  | -    | 0.350| 0.126|
| LMCI  | -    | -    | 0.074|

NOTE: CN, cognitively normal; EMCI, early mild cognitive Impairment; LMCI, late mild cognitive impairment; AD, Alzheimer’s disease.

It has been of great interest to construct a consensus tree that summarizes the individual trees to infer the evolutionary history. In addition to the complex structure of the trees, this task is complicated by the stark heterogeneity in the individual trees due to analytic artifacts such as sequence misalignment, remarkable biological variation, or low signal-to-noise ratio in the random subsample. Recently, tree space geometry-aware methods such as the Fréchet mean tree (e.g., Nye et al. 2017) have been proposed. These methods have been shown to produce reliable inference of tree topology and edge lengths. However, a preliminary outlier removal step (e.g., Weyenberg et al. 2014) is usually performed since the Fréchet mean is a nonrobust measure of location. Here, we apply the metric halfspace depth to obtain a “summary tree” that best represents the data and to identify potential outliers. We infer the phylogeny of 7 pathogenic Apicomplexan species relative to an outgroup species using \( n = 268 \) individual gene trees constructed by Kuo, Wares, and Kissinger (2008). The Apicomplexa phylum contains many important pathogenic parasites that are detrimental to humans and livestock. The Apicomplexan species included the infamous malaria pathogens \textit{Plasmodium falciparum} (Pt) and \textit{Plasmodium vivax} (Pv); tick-borne haemopathogens \textit{Babesia bovis} (Bb) and \textit{Theileria annulata} (Ta); and coccidian parasites \textit{Eimeria tenella} (Et), \textit{Toxoplasma gondii} (Tg), and \textit{Cryptosporidium parvum} (Cp) which infect intestines. The outgroup \textit{Tetrahymena thermophila} (Tt) is a remotely related model species included to root the phylogeny. We model the gene trees as rooted trees with the root placed as the point where the outgroup joins with the apicomplexanspecies.

To model the evolutionary divergence between all species and their ancestors, we consider \( T^8 \times R^8 \) with the product metric, where the BHV space \( T^8 \) models the tree topology and the interior edge lengths, and \( R^8 \) models the pendant edge lengths. The proposed metric halfspace depths were calculated at each of the individual trees, with 10 additional jiggled trees added as anchors per original tree for improving approximation. In the deepest tree as displayed in Figure 4, tick parasites \textit{B. bovis} and \textit{T. annulata} and malaria parasites \textit{P. falciparum} and \textit{P. vivax} are, respectively, monophyletic, that is, sharing the
same immediate ancestor; these haemoparasites descend from a common ancestor; coccidian species *E. tenella* and *T. gondii* form a sister group to the former; *C. parvum* is the deepest rooting species. The deepest tree we produced is congruent to the consensus tree identified by Kuo, Wares, and Kissinger (2008) constructed through maximum likelihood, maximum parsimony, and neighbor-joining methods, and also agree with the Fréchet mean tree found by Nye et al. (2017), who performed the analysis after removing 16 outliers. Our depth-based approach has the advantage of being robust to extreme values and does not require separate outlier identification and removal.

We also identified 27 gene trees with the least metric halfspace depth, indicating that they correspond to the most extreme trees. Among these trees four potential outliers are displayed in Figure 5 and the rest are included in Figure S11. Trees 488 and 546 have exceptionally long branches, and, in addition, the *Plasmodium* species in tree 488 (Pr and Pv, hard to distinguish in the figure due to the long branch) and tick parasites *B. bovis* and *T. annulata* in trees 625 and 703 are not monophyletic. These structures, which differ from what has been reported in the literature (Kuo, Wares, and Kissinger 2008), demonstrate the utility of the metric halfspace depth for highlighting outliers.

**Supplementary Materials**

The Supplementary material contains the following: a) technical proofs, b) background on Riemannian manifold, c) additional details and results for the numerical studies, d) comparisons between the metric halfspace depth and alternative depth concepts, and e) additional properties for the approximated depth.

**Funding**

Xiongtao Dai was partially supported by NSF grant DMS-2113713. Sara Lopez-Pintado was partially supported by NSF grant DMS-2113696 and by NIH grant 1R21 MH120534-01.

**Acknowledgement**

The authors would like to thank the Associate Editor and two reviewers for their constructive comments. They are also grateful to Dr Chih-Horng Kuo for providing the phylogenetic data.

**References**

Bai, Z.-D., and He, X. (1999), “Asymptotic Distributions of the Maximal Depth Estimators for Regression and Multivariate Location,” *The Annals of Statistics*, 27, 1616–1637. [1763]

Bhattacharya, R., and Patrangenaru, V. (2005), “Large Sample Theory of Intrinsic and Extrinsic Sample Means on Manifolds - II,” *Annals of Statistics*, 33, 1225–1259. [1760]

Billera, L. J., Holmes, S. P., and Vogtmann, K. (2001), “Geometry of the Space of Phylogenetic Trees,” *Advances in Applied Mathematics*, 27, 733–767. [1760,1761,1763,1764]

Bingham, M. A., Nordman, D. J., and Vandermar, S. B. (2009), “Modeling and Inference for Measured Crystal Orientations and a Tractable Class of Symmetric Distributions for Rotations in Three Dimensions,” *Journal of the American Statistical Association*, 104, 1385–1397. [1761]

Bogićević, M., and Merkle, M. (2018), “Approximate Calculation of Tukey’s Depth and Median with High-Dimensional Data,” *Yugosl Journal of Operations Research*, 28, 475–499. [1767]

Buckner, R. L., Sepulcre, J., Talukdar, T., Krienen, F. M., Liu, H., Hedden, T., Andrews-Hanna, J. K., Sperling, R. A., and Johnson, K. A. (2009), “Cortical Hubs Revealed by Intrinsic Functional Connectivity: Mapping, Assessment of Stability, and Relation to Alzheimer’s Disease,” *The Journal of Neuroscience*, 29, 1860–1873. [1769]

Caplin, A., and Nalebuff, B. (1988), “On 64%-Majority Rule,” *Econometronica*, 56, 787–814. [1762]

Carriozza, E. (1996), “A Characterization of Halfspace Depth,” *Journal of Multivariate Analysis*, 58, 21–26. [1761,1762]

Chau, J., Ombao, H., and von Sachs, R. (2019), “Intrinsic Data Depth for Hermitian Positive Definite Matrices,” *Journal of Computational and Graphical Statistics*, 28, 427–439. [1761,1768,1769]

Chen, M., Gao, C., and Ren, Z. (2018), “Robust Covariance and Scatter Matrix Estimation Under Huber’s Contamination Model,” *The Annals of Statistics*, 46, 1932–1960. [1761,1763]

Chenoury, S. and Small, C. G. (2012), “A Nonparametric Multivariate Multisample Test Based on Data Depth,” *Electronic Journal of Statistics*, 6, 760–782. [1760,1769]

Dai, X., and Müller, H.-G. (2018), “Principal Component Analysis for Functional Data on Riemannian Manifolds and Spheres,” *Annals of Statistics*, 46, 3334–3361. [1760]

Dai, X., Lin, Z., and Müller, H.-G. (2020), “Modeling Longitudinal Data on Riemannian Manifolds,” *Biometrics*, Accepted. [1761,1763]

Donoho, D. L., and Gasko, M. (1992), “Breakdown Properties of Location Estimates Based on Halfspace Depth and Projected Outlyingness,” *The Annals of Statistics*, 20, 1803–1827. [1760,1763,1764,1766]

Dryden, I. L., and Mardia, K. V. (2016), *Statistical Shape Analysis: With Applications in R*, Hoboken: Wiley. [1761]

Dykkeroff, R., and Mozharovskyi, P. (2016), “Exact Computation of the Halfspace Depth,” *Computational Statistics & Data Analysis*, 98, 19–30. [1766,1767]

Einmahl, J. H. J., Li, J., and Liu, R. Y. (2015), “Bridging Centrality and Extremity: Refining Empirical Data Depth Using Extreme Value Statistics,” *Annals of Statistics*, 43, 2738–2765. [1760]

Eltzner, B., and Huckemann, S. F. (2019), “A Smeary Central Limit Theorem for Manifolds with Application to High-Dimensional Spheres,” *The Annals of Statistics*, 47, 3360–3381. [1768]

Eltzner, B., Huckemann, S., and Mardia, K. V. (2018), “Torus Principal Component Analysis with Applications to RNA Structure,” *Annals of Applied Statistics*, 12, 1332–1359. [1761]

Feragen, A., and Nye, T. (2020), “Statistics on Stratified Spaces,” in *Riemannian Geometric Statistics in Medical Image Analysis*, eds. X. Pennec, S. Sommer and T. Fletcher, pp. 299–342, London: Academic Press. [1761]
