Nonparametric Statistical Inference via Metric Distribution Function in Metric Spaces

Xueqin Wang, Jin Zhu, Wenliang Pan, Junhao Zhu & Heping Zhang

View supplementary material

Published online: 06 Nov 2023.

Submit your article to this journal

Article views: 805

View related articles

View Crossmark data

Taylor & Francis
Taylor & Francis Group

⌾ OPEN ACCESS | Check for updates

Nonparametric Statistical Inference via Metric Distribution Function in Metric Spaces

Xueqin Wang[a] ⬤, Jin Zhu[b,c], Wenliang Pan[d], Junhao Zhu[b,*], and Heping Zhang[e]

[a]University of Science and Technology of China, Hefei, China; [b]Sun Yat-Sen University, Guangzhou, China; [c]London School of Economics and Political Science, London, UK; [d]Chinese Academy of Sciences, Beijing, China; [e]Yale University, New Haven, CT

ABSTRACT

The distribution function is essential in statistical inference and connected with samples to form a directed closed loop by the correspondence theorem in measure theory and the Glivenko-Cantelli and Donsker properties. This connection creates a paradigm for statistical inference. However, existing distribution functions are defined in Euclidean spaces and are no longer convenient to use in rapidly evolving data objects of complex nature. It is imperative to develop the concept of the distribution function in a more general space to meet emerging needs. Note that the linearity allows us to use hypercubes to define the distribution function in a Euclidean space. Still, without the linearity in a metric space, we must work with the metric to investigate the probability measure. We introduce a class of metric distribution functions through the metric only. We overcome this challenging step by proving the correspondence theorem and the Glivenko-Cantelli theorem for metric distribution functions in metric spaces, laying the foundation for conducting rational statistical inference for metric space-valued data. Then, we develop a homogeneity test and a mutual independence test for non-Euclidean random objects and present comprehensive empirical evidence to support the performance of our proposed methods. Supplementary materials for this article are available online.

1. Introduction

Nowadays, many statistical applications study non-Euclidean data. Typical data examples include symmetric positive definite (SPD) matrices (Smith et al. 2013), the Grassmann manifold (Hong et al. 2016), the shape representation of corpus callosum (Cornea et al. 2017), samples of probability density functions in Wasserstein spaces (Petersen, Liu, and Divani 2021), paleomagnetic directional data in hyperspheres (Scealy and Wood 2019).

Analysis of non-Euclidean objects is challenging (Cornea et al. 2017; Scealy and Wood 2019; Petersen, Liu, and Divani 2021). A common strategy is to embed non-Euclidean data objects into a Hilbert or more general metric space before the analysis. When the non-Euclidean data objects can be embedded in a metric space but not a Euclidean space, metric (distance)—based methods can be applied. Many methods exist or are being developed in statistics and machine learning, for example, Székely and Rizzo (2004), Székely, Rizzo, and Bakirov (2007), Böttcher, Keller-Ressel, and Schilling (2019), Pan et al. (2018), Dubey and Muller (2019), and Dai and Lopez-Pintado (2022). Assessing the uncertainty following the use of the existing methods to analyze non-Euclidean data is important but difficult due to the absence of a fundamental concept in metric spaces analogous to the distribution function (DF) in Euclidean spaces.

DF relates theory to the real world in statistical inference, allowing us to conclude the data (Efron 1979). The DF is defined to uniquely determine the Borel probability measure of a random vector (or a scalar) according to the correspondence theorem (Halmos 1956). Given observed data, the DF can be well estimated by the empirical distribution function (EDF). As illustrated in Figure 1, the DF and observed sampled data are linked to form a directed closed loop by the correspondence theorem in measure theory and the Glivenko-Cantelli theorem in statistical inference. This connection creates a paradigm for statistical inference.

The properties and applications of EDF have been systematically investigated as a prominent field in mathematical statistics for a century, and many statistical methods are, in fact, functional of EDF. Examples include the Cramér-von Mises test (Darling 1957) for the equality of two unknown DFs and Hoeffding's independence test (Hoeffding 1948) for two random data samples. Thus, it is reasonable to anticipate the importance of generalizing the concepts of DF and EDF to metric spaces to have a basic foundation for the methods we may use to analyze non-Euclidean data objects.

In this article, we introduce a quasi-DF to serve as the cornerstone of nonparametric statistical inference for metric space-valued data objects. We consider several important problems in statistical inference to show the utility of the quasi-DF. Note that the DF in Euclidean space has the correspondence theorem because its definition is closely relative to the Euclidean metric topology. Indeed, the DF is the Borel probability measure of the Cartesian products of left closed rays, which is a base of the Euclidean metric topology. While a metric space is equipped

Figure 1. Conceptual diagram of statistical inference paradigm. The solid arrows indicate a conceptual deduction, and the dashed ones are a statistical approximation.

with a naturally metric topology that contains all open balls as a base, balls are generally not ordered, but concentric balls are. This ordering is essential for us to define the DFs in the same way as the ordered topology of one-dimensional Euclidean spaces, provided that we fix the center first. Using this center as the second variable, we can define the metric distribution function (MDF) in metric spaces as the counterpart of DF in Euclidean spaces. (Figure 1).

The rest of this article is organized as follows. We introduce the concepts of the MDF and the empirical MDF (EMDF) in Section 2, and present their theoretical properties in Section 3. In Section 4, based on the MDF and the EMDF, we develop several nonparametric statistical inference procedures. To demonstrate the MDF's effectiveness in practice, we employ the MDF-based methods on the synthetic and real-world datasets in Sections 5 and 6, respectively. Finally, we summarize our work for the MDF in Section 7. Technical proofs and some properties of EMDF are deferred to the "supplementary material".

2. MDF and EMDF

2.1. Notations

An order pair $(\mathcal{M}, d)$ is called *metric* space if $\mathcal{M}$ is a set and d is a *metric* or *distance* on $\mathcal{M}$. Many spaces we have encountered are metric spaces. Examples include Euclidean space, Banach space, and connected Riemannian manifold. A metric space $(\mathcal{M}, d)$ is called *separable* if it has a countable dense subset for the metric topology. A metric space $(\mathcal{M}, d)$ is said to be *complete* if every Cauchy sequence converges in $\mathcal{M}$. A completely separable metric space is sometimes called a *Polish* space. Given a metric space $(\mathcal{M}, d)$, let $\bar{B}(u, r) = \{v : d(u, v) \leq r\}$ be the closed ball with the center u and the radius $r \geq 0$, $B(u, r) = \{v : d(u, v) < r\}$ be the open ball and $\partial B(u, r) = \bar{B}(u, r) \setminus B(u, r)$ be the sphere.

If $(\mathcal{M}_k, d_k)$, $k = 1, \ldots, K$, are metric spaces, let $\mathcal{M}$ be the Cartesian product of $\mathcal{M}_k$, denoted by $\prod_{k=1}^{K} \mathcal{M}_k$. Here, Cartesian products of metric spaces are considered because they are useful for defining the independence measure in Section 4. For any $\mathbf{u} =$

$(u_1, \ldots, u_K)$ and $\mathbf{v} = (v_1, \ldots, v_K)$ in $\mathcal{M}$, we can define a metric vector $e(\mathbf{u}, \mathbf{v})$ on the product space $\mathcal{M}$:

$$e(\mathbf{u}, \mathbf{v}) = \Big(d_1(u_1, v_1), \ldots, d_K(u_K, v_K) \Big).$$

We also define $\bar{B}(\mathbf{u}, \mathbf{r}) := \prod_{k=1}^{K} \bar{B}(u_k, r_k)$ be the joint ball on the product space for a center vector $\mathbf{u} \in \mathcal{M}$ and a nonnegative radius vector $\mathbf{r} = (r_1, \ldots, r_K) \succeq \mathbf{0}$. For this product space, we can also assign a metric such that it is a metric space. For example, if we define

$$d(\mathbf{u}, \mathbf{v}) = \|e(\mathbf{u}, \mathbf{v})\|_p, \tag{1}$$

where $p \geq 1$ and $\| \cdot \|_p$ means the ℓ_p norm in $\mathbb{R}^K$. We can verify that $d(\cdot, \cdot)$ is a metric on $\mathcal{M}$. Given a point $\mathbf{u} \in \mathcal{M}, \pi_k(\cdot) : \mathcal{M} \to \mathcal{M}_k$ is called the projection on $\mathcal{M}_k$ if $\pi_k(\mathbf{u}) = u_k$. For a set $A \subset \mathcal{M}$, we also define $\pi_k(A) = \bigcup_{\mathbf{u} \in A} \{\pi_k(\mathbf{u})\}$.

Let μ be a (Borel) probability measure associated with an ordered K-tuple of random objects $\mathbf{X} = (X_1, \ldots, X_K)$ taking values in $\mathcal{M}(= \prod_{k=1}^{K} \mathcal{M}_k)$, and define $\mu \otimes \mu$ as the product measure on the measurable space.

2.2. MDF and EMDF

Denote the indicator function by $I(\cdot)$ and the radius vector $\mathbf{r} = e(\mathbf{u}, \mathbf{v})$ for $\mathbf{u}, \mathbf{v} \in \mathcal{M}$. We first define the metric distribution function (MDF) of μ on $\mathcal{M}$ that is the foundation of our proposed framework. For $\forall \mathbf{u}, \mathbf{v} \in \mathcal{M}$, let

$$\delta(\mathbf{u}, \mathbf{v}, \mathbf{x}) = \prod_{k=1}^{K} I\{x_k \in \bar{B}(u_k, r_k)\} = \prod_{k=1}^{K} I\{x_k \in \bar{B}(u_k, d_k(u_k, v_k))\}.$$

Definition 1. Given a probability measure μ, we define the metric distribution function $F_\mu^M(u, v)$ of μ on $\mathcal{M}$: $\forall \mathbf{u}, \mathbf{v} \in \mathcal{M}$,

$$F_\mu^M(\mathbf{u}, \mathbf{v}) = \mu \left[\prod_{k=1}^{K} \bar{B}(u_k, r_k) \right] = E \left[\delta(\mathbf{u}, \mathbf{v}, \mathbf{X}) \right].$$
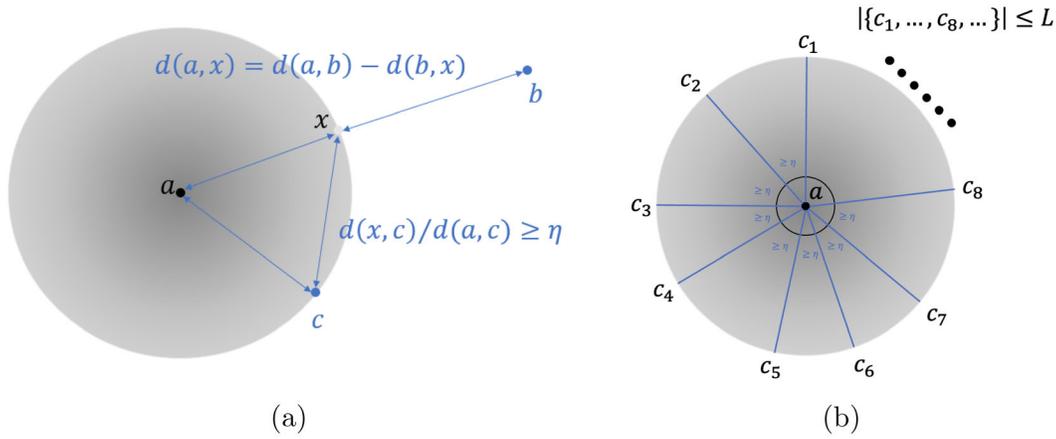
Figure 2. (a) Visualization of the direction in metric space by a 2-d Euclidean space example. (b) Visualization of the directionally (ϵ, η, L)-limited condition in the 2-d Euclidean space. For a given $\eta > 0$, consider a circle $\mathcal{N}$ with the radius r such that for any two points c_i and c_j in $\mathcal{N}$ we have $d(c_i, c_j)/r \geq \eta$. The directionally (ϵ, η, L)-limited condition means that there exists an L such that the cardinality of $\{c_1, \ldots, c_8, \ldots\}$ is always less than L.

Suppose that $\{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$ are iid samples generated from a probability measure μ on a product metric space $\mathcal{M} = \prod_{k=1}^{K} \mathcal{M}_k$. We define the empirical metric distribution function (EMDF) associated with μ by the following formula naturally:

$$F_{\mu,n}^M(\mathbf{u}, \mathbf{v}) = \frac{1}{n} \sum_{l=1}^{n} \delta(\mathbf{u}, \mathbf{v}, \mathbf{X}_l).$$

3. Theoretical Analysis of MDF and EMDF

In this section, we first discuss some sufficient conditions for reconstructing probability measures from the MDFs and exhibit the properties of the convergence of the EMDFs. Additional properties of the EMDF are presented in the third part of the "supplementary material."

3.1. Fundamental Reconstruction Theorems of MDF

Here we investigate whether a Borel probability measures μ on a separable metric space $(\mathcal{M}, d)$ can be uniquely determined by the MDF $F_\mu^M(u, v)$. We shall see that the answer depends on the probability measure and metric space. For separable metric spaces, Federer (2014) introduced the following geometrical condition on the metric, named *directionally (ϵ, η, L)-limited*, to characterize the correspondence property of the MDF.

Definition 2. (Federer 2014) A metric d is called directionally (ϵ, η, L)-limited at the subset A of $\mathcal{M}$, if $\epsilon > 0, 0 < \eta \leq 1/3, L$ is a positive integer, and the following condition holds: if for each $a \in A, D \subseteq A \cap \bar{B}(a, \epsilon)$ such that $d(x, c) \geq \eta d(a, c)$ whenever $b, c \in D$ $(b \neq c), x \in \mathcal{M}$ with

$$d(a, x) = d(a, c), d(x, b) = d(a, b) - d(a, x),$$

then the cardinality of D is no larger than L.

Definition 2 ensures that the covering theorem holds; namely, given a "thorough" covering of a set by closed balls, there is a subcollection of pairwise disjoint balls that almost covers the set. The counterpart of this result in Euclidean spaces is the so-called

Vitali covering theorem, which is important to the proof of the correspondence theorem in Euclidean spaces. Likewise, we need a similar condition for the correspondence theorem in metric spaces. Figure 2 intuitively illustrates this directionally limited assumption. Panel (a) visualizes the definition of direction in metric space. The ratio of chord length and radius can measure the direction between two lines in a metric space. Panel (b) illustrates the directionally limited assumption, which requires that the directions of every local point a are finite. This concept of "directionally (ϵ, η, L)-limited" is essential to our reconstruction theory. We examine a few examples to understand the implications of this condition.

First, if $(\mathcal{M}, \|\cdot\|)$ with the norm $\|\cdot\|$ is a Banach space, then the above definition implies

$$x = a + \frac{\|a - c\|}{\|a - b\|}(b - a),$$

thus $d(x, c) \geq \eta d(a, c)$ is equivalent to

$$\frac{d(x, c)}{d(a, c)} = \left\| \frac{b - a}{\|b - a\|} - \frac{c - a}{\|c - a\|} \right\| \geq \eta.$$

If $\mathcal{M}$ is a finite-dimensional Banach space, owing to the compactness of the unit sphere in $\mathcal{M}$, there exists a suitable L for each $\eta > 0$ such that the condition of directionally limited metric space holds.

Another case is when $\mathcal{M}$ is a finite-dimensional Riemannian manifold with bounded sectional curvature and A is any compact subset of $\mathcal{M}$. Let $B(a, \epsilon)$ be a normal ball of $a \in A$, and $\text{Exp}_a(\cdot)$ be the Riemannian exponential map and $\text{Log}_a(\cdot)$ be the Riemannian log map. By the bounded sectional curvature condition and Topogonov's theorem (Do Carmo and Flaherty Francis 1992), we can find a universal constant $\lambda > 0$ such that the for any $b = \text{Exp}_a(\beta)$ and $c = \text{Exp}_a(\gamma)$ satisfying

$$d(a, b) = \|\beta\|_a, d(a, c) = \|\gamma\|_a,$$
$$x = \text{Exp}_a\left[\frac{\|\gamma\|_a}{\|\beta\|_a} \beta \right], \text{ and } d(x, c)/d(a, c) \geq \eta,$$

the inequality $\|\frac{\beta}{\|\beta\|_a} - \frac{\gamma}{\|\gamma\|_a}\|_a \geq \lambda\eta > 0$ holds. Thus, if we associate each $b \in D$ with the direction $\text{Log}_a(b)$, then by the

compactness of unit sphere in the tangent space of a, there exists a suitable $L \in \mathbb{N}$ for each $\eta > 0$.

The last but important case is when $(\mathcal{M}, d)$ is the metric space of a binary phylogenetic tree with τ leaves, where $\tau \in \mathbb{N}^+$ is fixed. The space $\mathcal{M}$ is a Polish space and cubical complex (Lin and Müller 2021). Let $\mathcal{N}(\eta\epsilon; \bar{B}(x, \epsilon), d)$ be the $\eta\epsilon$-packing of $\bar{B}(x, \epsilon) \subseteq \mathcal{M}$ such that, for any $x, x' \in \mathcal{N}(\eta\epsilon; \bar{B}(x, \epsilon), d)$, the geodesic distance $d(x, x') \geq \eta\epsilon$. Denote $\omega(\tau) = (2\tau - 3) \times (2\tau - 5) \times \cdots \times 3$, for $\forall \eta \leq \frac{1}{3}$ and $\forall \epsilon \leq 1$, the space satisfies

$$\sup_{x \in \mathcal{M}} card(\mathcal{N}(\eta\epsilon; \bar{B}(x, \epsilon), d)) \leq C\omega(\tau) \left(\frac{2}{\eta}\right)^{\psi(\tau)},$$

for some constant $C < +\infty$, $\psi(\tau) < +\infty$. This implies that the whole space is directionally-limited with $(\frac{1}{3}, 1, C\omega(\tau)6^{\psi(\tau)})$.

Next, we provide an example of metric space that is not directionally limited. An infinite orthonormal base $B = \{e_1, e_2, \ldots\}$ in a separable Hilbert space H is not directionally (ϵ, η, L)-limited. Let $a \in H$, and $b = a + e_i, c = a + e_j \in a + B$, then by the above discussion for Banach space, we have

$$x = a + \frac{\|e_j\|}{\|e_i\|}e_i \quad \text{and} \quad \frac{d(x, c)}{d(a, c)} = \|e_j - e_i\| = \sqrt{2} \geq \eta$$

for all $0 < \eta \leq \frac{1}{3}$ and the cardinality of $a + B$ is infinite.

Remark 1. Metric entropy and directionally (ϵ, η, L)-limited are related concepts, and in certain classical metric spaces such as finite-dimensional Banach spaces and Riemannian manifolds with bounded sectional curvature, finite metric entropy implies directionally (ϵ, η, L)-limited. Specifically, for a given δ-covering number N, setting $\xi = \infty$, $\eta = 2\delta$, and $\zeta = N$ guarantees that the directional limitability condition holds for any point c and subset $B \subseteq A \setminus a$, as shown by the following inequality:

$$\frac{d(x, c)}{d(a, c)} = \left\| \frac{b - a}{\|b - a\|} - \frac{c - a}{\|c - a\|} \right\| \geq \eta = 2\delta,$$

where d denotes the metric, and a, b, c, x are points in the metric space. Following the definition of metric entropy, the cardinality of D is no larger than N. This result also holds for the Riemannian manifold with bounded sectional curvature. Let $\xi = \infty$, $\eta = 2\delta\lambda$, and $\zeta = N$, then the cardinality of D is no larger than N following inequality:

$$d(x, c)/d(a, c) \leq \lambda |\beta|/|\beta| - \gamma/|\gamma||.$$

Thus, the finite-dimensional Banach space and Riemannian manifold with its usual metric are directionally (ϵ, η, L)-limited if the metric entropy is finite for $\forall \delta$.

In a Euclidean space, two Borel probability measures $\mu = \nu$ if and only if their associated random objects $\mathbf{X}$ and $\mathbf{Y}$ share the common DF by the correspondence theorem (Halmos 1956). This correspondence lays the theoretical foundation for statistical inference. However, DF depends on the linear structure and the order of real numbers. We do not have this structure in general metric space, and DF can no longer be defined. The following theorems delineate how MDF overcomes this major challenge. Theorem 1 shows that $d(u, \mathbf{X})$ and $d(u, \mathbf{Y})$ share the same DF for each location $u \in \text{supp}\{\mu\}$ if and only if $\mu = \nu$.

Theorem 1 (The fundamental correspondence theorem of MDF). Denote $S = \{(u, v) \in \mathcal{M} \times \mathcal{M} : F_\mu^M(u, v) = F_\nu^M(u, v)\}$ for two given Borel probability measures, μ and ν, with their respective supports, $supp\{\mu\}$ and $supp\{\nu\}$, on $(\mathcal{M}, d)$. Suppose that $(\mathcal{M}, d)$ is a Polish space and the metric d is directionally (ϵ, η, L)-limited at $supp\{\mu\}$ and $supp\{\nu\}$, then $\mu \otimes \mu(S) = 1$ (or $\nu \otimes \nu(S) = 1$) if and only if $\mu = \nu$.

Theorem 1 ensures that the MDF has a one-to-one correspondence with a probability measure when the metric is directionally (ϵ, η, L)-limited at the support of the probability measure. The conditions of Theorem 1 may not be satisfied if $\mathcal{M}$ is a separable Hilbert space of infinite dimension. Corollary 1 presents a reasonable condition so that the probability measure μ can still be determined by MDF in infinite dimension space.

Corollary 1. For $\forall \varepsilon > 0$, suppose that there exists $\mathcal{M}_l \subseteq \mathcal{M}$ such that $\mu(\mathcal{M}_l) \geq 1 - \varepsilon$ (or $\nu(\mathcal{M}_l) \geq 1 - \varepsilon$) and the metric d is directionally $(\epsilon(\mathcal{M}_l), \eta(\mathcal{M}_l), L(\mathcal{M}_l))$-limited at $\mathcal{M}_l$, then $\mu \otimes \mu(S) = 1$ (or $\nu \otimes \nu(S) = 1$) if and only if $\mu = \nu$.

Corollary 1 includes separable Hilbert spaces as a special case. For example, a random function, or a random curve $f(t)$ in a separable Banach space with unconditional Schauder base functions $\{\phi_j\}_{j=1}^\infty$, can be expanded as $f(t) = \sum_{j=1}^\infty \beta_j \phi_j(t)$, where the probability measure of the coefficients $\{\beta\}_{j=1}^\infty$, denoted as μ, satisfies a sparse condition $\lim_{S \to \infty} \mu(\exists j > S, \beta_j \neq 0) = 0$. This setting is similar to the sparse priors used in Castillo, Schmidt-Hieber, and Van der Vaart (2015) and O'Hara and Sillanpää (2009), but it is important to note that we allow the underlying function space to be infinite-dimensional. In this example, the "measure condition" of Corollary 1 is satisfied. The metric condition of Corollary 1 implies that if μ and ν are two Borel probability measures on $\ell^2 = \{(a_1, a_2, \ldots, a_i, \ldots) : \sum_i a_i^2 < \infty\}$, and they share the common metric distribution function on any finite subspace of ℓ^2, then we have $\mu = \nu$. According to the previous statement, many metric spaces, including the space of smooth functions, Riemannian manifold space, shape space, τ-leaves binary phylogenetic spaces, satisfy the conditions of Corollary 1.

In general, if the metric space is not linear, the geometric condition of "directionally (ϵ, η, L)-limited" cannot be induced from the compactness. Davies (1971) gives a counter-example that there exists a compact metric space $(\mathcal{M}, d)$ and two distinct Borel probability measures μ and ν on $\mathcal{M}$, such that μ and ν agree on all closed balls.

Next, we extend the 1–1 correspondence theorem to product metric spaces. This extension is challenging because the topological structure of a product metric space may not be as simple as that of Euclidean space. For example, the product of two circles $S^1 \times S^1$ is topologically not a sphere S^2 anymore. Let μ and ν be two Borel probability measures on $\mathcal{M} = \prod_{k=1}^K \mathcal{M}_k$, and

$\mathbf{M}_1 = \{\mu : \mu$ is a discrete Borel probability measure on $\mathcal{M}\}$,

$\mathbf{M}_2 = \{\mu : \mu$ is a Borel probability measure on $\mathcal{M}$ such that $\forall \mathbf{u} \in \mathcal{M}$ and $\mathbf{v} \sim \mu$, $e(\mathbf{u}, \mathbf{v})$ has continuous density function$\}$.

We have the following fundamental correspondence theorem of the joint metric distribution function in a product metric space.

Theorem 2 (The fundamental correspondence theorem of joint MDF). Given two Borel probability measures, μ and ν, on a product Polish space $\mathcal{M} = \prod_{k=1}^{K} \mathcal{M}_k$, let $\mathbf{S} = \{(\mathbf{u}, \mathbf{v}) \in \mathcal{M} \times \mathcal{M} : F_\mu^M(\mathbf{u}, \mathbf{v}) = F_\nu^M(\mathbf{u}, \mathbf{v})\}$. Suppose that d_k is directionally-(ξ_k, η_k, L_k) limited at $\mathcal{M}_k$ and $\mu = \alpha\eta + (1 - \alpha)\gamma$ for some $\alpha \in [0, 1]$, $\eta \in \mathbf{M}_1$ and $\gamma \in \mathbf{M}_2$, then $\mu \otimes \mu(\mathbf{S}) = 1$ (or $\nu \otimes \nu(\mathbf{S}) = 1$) if and only if $\mu = \nu$.

Similar to Corollary 1, we have the following corollary for the product space.

Corollary 2. For $\forall \, \varepsilon > 0$, suppose that there exists $\mathcal{M}_l \subseteq \mathcal{M}$ such that $\mu(\mathcal{M}_l) \geq 1 - \varepsilon$ (or $\nu(\mathcal{M}_l) \geq 1 - \varepsilon$) and the metric d_k is directionally $\big(\epsilon(\pi_k(\mathcal{M}_l)), \eta(\pi_k(\mathcal{M}_l)), L(\pi_k(\mathcal{M}_l))\big)$-limited at $\pi_k(\mathcal{M}_l)$. If the combination $\mu = \alpha\eta + (1 - \alpha)\gamma$ for $\eta \in M_1$ and $\eta \in M_2$, then $\mu \otimes \mu(\mathbf{S}) = 1$ (or $\nu \otimes \nu(\mathbf{S}) = 1$) if and only if $\mu = \nu$.

3.2. Main Properties of EMDF

Here, we provide EMDF's Glivenko-Cantelli property and Donsker property. First, we define the collection of the indicator functions of closed balls on $\mathcal{M}$:

$$\mathcal{F} = \{\delta(\mathbf{u}, \mathbf{v}, \cdot) : \mathbf{u} \in \mathcal{M} \text{ and } \mathbf{v} \in \mathcal{M}\}.$$

The uniform convergence property of EMDF is given as follows.

Theorem 3 (The Glivenko-Cantelli type property of EMDF). Let $\mathcal{M} = \prod_{k=1}^{K} \mathcal{M}_k$ be a product space and μ be a probability measure on it. Suppose that $\{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$ is a sample of iid observations from μ. Define $\mathcal{F}(\mathbf{X}_1^n) := \{(f(\mathbf{X}_1), \ldots, f(\mathbf{X}_n)) | f \in \mathcal{F}\}$. If μ satisfies that

$$\frac{1}{n} E_{\mathbf{X}} \left[\log(card(\mathcal{F}(\mathbf{X}_1^n))) \right] \to 0,$$

where $card(\cdot)$ is the cardinality of a set, we have the Glivenko-Cantelli property of our empirical metric distribution function:

$$\lim_{n \to \infty} \sup_{\mathbf{u} \in \mathcal{M}, \mathbf{v} \in \mathcal{M}} |F_{\mu,n}^M(\mathbf{u}, \mathbf{v}) - F_\mu^M(\mathbf{u}, \mathbf{v})| = 0, \, a.s.$$

The conditions of Theorem 3 are often satisfied in practice. The first example is $\mathcal{M} = \mathbb{R}^q$ with the ℓ_p-norm (where p is a positive integer or ∞), and μ is an arbitrary probability measure because the set of ℓ_p ball has a finite VC-dimension. We also allow the dimension of $\mathcal{M}$ to increase as the sample size increases if $\mathcal{M}$ is a Euclidean space. Since the VC-dimension of closed balls in Euclidean space R^q is $q + 2$ (see Example 4.14 in Wainwright 2019), if $q = o(\frac{n}{\log n})$ the Glivenko-Cantelli property still holds (Lemma 4.14 in Wainwright (2019)). The second example is that $\mathcal{M}$ is a smooth regular curve in Euclidean space or a sphere in $\mathbb{R}^q$ with the geodesic distance, and μ is an arbitrary probability measure. In this case, we can reparameterize $\mathcal{M}$ to be a unit speed curve such that every geodesic ball in $\mathcal{M}$ is mapped to an interval in $\mathbb{R}$ and the set of intervals in $\mathbb{R}$ has a finite VC-dimension. The third example is that $\mathcal{M}$ is a set of polygonal

curves in $\mathbb{R}^d$ with the Hausdorff distance for the Fréchet distance (Driemel et al. 2021) and μ is an arbitrary probability measure. Another example is that $\mathcal{M}$ is a separable Hilbert space with a probability measure μ with support on a finite-dimensional subspace because the set of balls on the support of μ has a finite VC-dimension.

Based on the two reconstruction theorems, whether the two probability measures are identical depends on whether their MDFs are the same over their support sets but not the whole space. This leads us to consider the Glivenko-Cantelli type property for the MDF over the sample set because the sample set contains the information that supports the underlying unknown probability measure.

Corollary 3 (A concentration inequality of EMDF). Let $\mathcal{M} = \prod_{k=1}^{K} \mathcal{M}_k$ be a product space and μ be a probability measure on it. For each $t > 0$, there exists a universal constant $N(t) \in \mathbb{N}$ such that for all $n \geq N(t)$, we have

$$P(\max_{1 \leq i,j \leq n} |F_{\mu,n}^M(\mathbf{X}_i, \mathbf{X}_j) - F_\mu^M(\mathbf{X}_i, \mathbf{X}_j)| > t) \leq 2n \exp(-\frac{nt^2}{32}),$$

which leads to

$$\max_{1 \leq i,j \leq n} |F_{\mu,n}^M(\mathbf{X}_i, \mathbf{X}_j) - F_\mu^M(\mathbf{X}_i, \mathbf{X}_j)| \xrightarrow{a.s.} 0, \text{ as } n \to \infty.$$

Without restriction on metric spaces and probability measures, Theorem 3 shows that the EMDF has the concentration phenomenon at an exponential convergence rate for a sufficiently large sample. This uniform convergence result over the sample set is essential when we apply the EMDF to analyze data objects in metric spaces as it is the data analysis in a Euclidean space. The other important convergence property of the EMDF is the convergence in distribution, called the Donsker property, which is similar to the central limits theorem.

Theorem 4 (The Convergence of Metric Distribution Process). Let $\mathcal{M} = \prod_{k=1}^{K} \mathcal{M}_k$ be a product space and μ be a probability measure on it. Define

$$\mathbb{G}_n(\mathbf{u}, \mathbf{v}) = \sqrt{n}(F_{\mu,n}^M(\mathbf{u}, \mathbf{v}) - F_\mu^M(\mathbf{u}, \mathbf{v})), \; \mathbf{u}, \mathbf{v} \in \mathcal{M}.$$

If $\mathcal{F}$ is a VC class with VC-dimension $V(\mathcal{F}) < \infty$, then we have the Donsker property of the metric distribution process: $\{\mathbb{G}_n(\mathbf{u}, \mathbf{v}) : \mathbf{u}, \mathbf{v} \in \mathcal{M}\}$ converges in distribution to a Gaussian process $\{\mathbb{G}_\mu(\mathbf{u}, \mathbf{v}), \mathbf{u} \in \mathcal{M} \text{ and } \mathbf{v} \in \mathcal{M}\}$, with zero mean and the covariance function:

$$E\mathbb{G}_\mu(\mathbf{u}_1, \mathbf{v}_1)\mathbb{G}_\mu(\mathbf{u}_2, \mathbf{v}_2)$$
$$= \mu\big(\bar{B}(\mathbf{u}_1, e(\mathbf{u}_1, \mathbf{v}_1)) \cap \bar{B}(\mathbf{u}_1, e(\mathbf{u}_2, \mathbf{v}_2))\big)$$
$$- F_\mu^M(\mathbf{u}_1, \mathbf{v}_1)F_\mu^M(\mathbf{u}_2, \mathbf{v}_2).$$

The conditions in Theorem 4 also imply the Glivenko-Cantelli property of the EMDF. It is noteworthy that EMDF has the Glivenko-Cantelli and Donsker properties in the infinite-dimensional cases if we impose some entropy and continuity conditions on the probability measures:

Corollary 4. Assume the following conditions hold:

1. $\int_0^\infty \sqrt{\frac{\log N(\epsilon, \mathcal{M}, t)}{t}} dt < \infty$, where $N(\epsilon, \mathcal{M}, t)$ is the covering number.
2. The CDF of $d(\mathbf{u}, \mathbf{X})$, has probability density function $f(\mathbf{u}, \mathbf{r})$ for all $\mathbf{u} \in \mathcal{M}$, and satisfies $\sup_{\mathbf{u}, \mathbf{r}} f(\mathbf{u}, \mathbf{r}) \le C_f < \infty$ for some constant C_f.

Then, both Glivenko-Cantelli and Donsker properties hold.

Remark 2. The preceding result can be extended without significant additional effort to replace $\mathcal{M}_k$ in the first condition with the projection of the support of μ onto $\mathcal{M}_k$. The first condition on $\mathcal{M}_k$ is mild and can be satisfied by any space satisfying $\log N(t, \mathcal{M}, d) = O(t^{-c})$ for some constant $c < 1$, including

1. Bounded subsets of the α-times continuously differentiable function space $C_1^\alpha(\mathcal{X})$ defined on $\mathcal{X}$ equipped with the $\|\cdot\|_\infty$-norm or $L_r(Q)$- norm for a certain probability measure Q on $\mathcal{X}$. Here, $\mathcal{X}$ is a bounded convex subset of $\mathbb{R}^p$ and $\alpha > d$, as proved by Theorem 2.7.1 in Wellner and Vaart (2013).
2. Bounded subsets of Riemannian manifold with bounded sectional curvature by the Bishop-Gromov packing lemma (Petersen 2006), such as a bounded subset of SPD matrices manifold equipped with the affine-invariant metric and p-dimensional sphere.
3. Bounded subsets or balls of $\mathbb{R}^p$ equipped with $\|\cdot\|_q$-norm for $1 \le q \le \infty$ (Wainwright 2019).
4. Bounded subsets of the binary phylogenetic tree with τ leaves, where $\tau \in \mathbb{N}^+$ is fixed.

4. MDF Based Statistical Methods

In this section, we discuss using the MDF to conduct statistical inference in a few important and common problems.

4.1. Homogeneity Test

A common and basic hypothesis testing problem in statistical inference is whether samples are generated from the same distribution. Suppose we have data objects from unknown Borel probability measures, $\mu_1, \mu_2, \ldots, \mu_K$, on a metric space $(\mathcal{M}, d)$ and need to check whether they are homogeneous, that is, testing $H_0 : \mu_1 = \mu_2 = \cdots = \mu_K$.

We introduce a homogeneity measure based on MDF, called metric Cramér-von Mises (MCVM). Let $F_{\mu_k}^M(\mathbf{u}, \mathbf{v})$ be the MDFs for μ_k, μ is the mixture distribution of $\mu_1, \ldots, \mu_K$ with proportions $p_1, \ldots, p_K$, and $F_\mu^M(\mathbf{u}, \mathbf{v})$ be the MDF of μ, we use some Cramér-von Mises-type criteria to evaluate the distinction of $F_{\mu_k}^M(\mathbf{u}, \mathbf{v})$ and $F_\mu^M(\mathbf{u}, \mathbf{v})$ for at $\mathbf{u}, \mathbf{v}$ from $\mu_1, \ldots, \mu_K$:

$$\text{MCVM}(\mu_k \| \mu) = \int_{(\mathbf{u}, \mathbf{v}) \in \mathcal{M} \times \mathcal{M}} p_k^2 w(\mathbf{u}, \mathbf{v})$$
$$\left(F_{\mu_k}^M(\mathbf{u}, \mathbf{v}) - F_\mu^M(\mathbf{u}, \mathbf{v}) \right)^2 d\mu_k(\mathbf{u}) d\mu_k(\mathbf{v}),$$

where $w(\mathbf{u}, \mathbf{v}) = \exp\{-\frac{(d(\mathbf{u}, \mathbf{v}))^2}{2\sigma^2}\}$. We aggregate $\text{MCVM}(\mu_k \| \mu)$ by defining

$$\text{MCVM}(\mu_1, \ldots, \mu_K) = \sum_{k=1}^K p_k^2 \text{MCVM}(\mu_k \| \mu).$$

For each $k = 1, \ldots, K$, let $\mathcal{X}_k$ be the kth sample set of $\mathbf{X}_1^{(k)}, \ldots, \mathbf{X}_{n_k}^{(k)} \stackrel{iid}{\sim} \mu_k$. Then $\text{MCVM}(\mu_k \| \mu)$ can be estimated on the basis of EMDF:

$$\widehat{\text{MCVM}}(\mu_k \| \mu) = \frac{1}{n_k^2} \sum_{\mathbf{X}_i^{(k)}, \mathbf{X}_j^{(k)} \in \mathcal{X}_k} (\hat{p}_k)^2 w(\mathbf{X}_i^{(k)}, \mathbf{X}_j^{(k)})$$
$$\left(F_{\mu_k, n_k}^M(\mathbf{X}_i^{(k)}, \mathbf{X}_j^{(k)}) - F_{\mu, n}^M(\mathbf{X}_i^{(k)}, \mathbf{X}_j^{(k)}) \right)^2,$$

where $\hat{p}_k = n_k / \sum_{k=1}^K n_k$. Thus, $\widehat{\text{MCVM}}(\mu_1, \ldots, \mu_K) = \sum_{k=1}^K (\hat{p}_k)^2 \widehat{\text{MCVM}}(\mu_k \| \mu)$. We use the median heuristic for choosing the σ^2 as the median of $\{d(\mathbf{X}, \mathbf{X}') : \mathbf{X}, \mathbf{X}' \in \cup_{k=1}^K \mathcal{X}_k\}$.

Following the above Theorems, we can obtain the theoretical properties of MCVM.

Proposition 1. (a) Suppose that the conditions of Theorem 1 or Corollary 1 hold and $p_k > 0, k = 1, \ldots, K$, then $\text{MCVM}(\mu_1, \ldots, \mu_K) = 0$ if and only if $\mu_1 = \mu_2 = \cdots = \mu_K$.
(b) Suppose that the conditions of Theorem 3 hold and $p_k > 0, k = 1, \ldots, K$, then

$$\widehat{\text{MCVM}}(\mu_1, \ldots, \mu_K) \xrightarrow[n_1, \ldots, n_K \to \infty]{a.s.} \text{MCVM}(\mu_1, \ldots, \mu_K).$$

(c) Suppose $p_k > 0$ $(k = 1, \ldots, K)$, under the conditions of Theorem 3, when the null hypothesis holds,

$$n\widehat{\text{MCVM}}(\mu_1, \ldots, \mu_K) \xrightarrow[n_1, \ldots, n_K \to \infty]{d} \sum_{l=1}^\infty \lambda_l Z_l^2,$$

where Z_l are iid standard normal random variables and λ_l are the constants depending on $\mu_1, \ldots, \mu_K$, $l = 1, \ldots, \infty$. $n\widehat{\text{MCVM}}(\mu_1, \ldots, \mu_K)$ can serve as a test statistic for homogeneity, which is consistent against the alternatives.

To test H_0, we can use permutation to approximate the p-value directly. On the other hand, the asymptotic distribution of MCVM in Proposition 1(c) motivates us to test homogeneity by estimating $\{\lambda_i\}_{i \in \mathbb{N}}$ when n is sufficiently large. We name this test procedure as a spectrum-based test and study its numerical performance in Section 5.1. We defer its implementation details and theoretical property in Section 4 of supplementary materials.

4.2. Mutual Independence Test

Another fundamental problem in statistical inference is testing the mutual independence among several elements of a random object. Suppose $\mathbf{X} = (X_1, \ldots, X_K)$ is a random object of K-tuple random objects ($K \ge 2$) on a metric space $(\mathcal{M}, d)$, in which $\mathbf{X}$ is associated with probability measure μ, and X_k is associated with probability measure μ_k on $\mathcal{M}_k$ for $k = 1, \ldots, K$. The study of mutual independence is formulated as testing $H_0 : \mu = \mu_1 \otimes \cdots \otimes \mu_K$.

It is very convenient to use the MDF to measure mutual dependence because of the definition of the MDF in product metric spaces. Following Hoeffding's dependence paradigm (Hoeffding 1948), we integrate the difference between the joint MDF associated with $\mathbf{X}$ and the product of marginal MDFs

associated with X_k's. We then obtain our metric association (MA) measure:

$$MA(\mu, \otimes_{k=1}^K \mu_k)$$
$$= \int \left(F_\mu^M(\mathbf{u}, \mathbf{v}) - \prod_{k=1}^K F_{\mu_k}^M(u_k, v_k) \right)^2 d\mu(\mathbf{u}) d\mu(\mathbf{v}).$$

When $K = 2$, $MA(\mu, \otimes_{k=1}^2 \mu_k)$ is the square of ball covariance in Pan et al. (2020). When the entries of $\mathbf{X}$ are dependent, then Theorem 2 implies that $MA(\mu, \otimes_{k=1}^K \mu_k) > 0$. Suppose that $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are iid observations of $\mathbf{X}$ associated with the Borel probability measures μ. The consistent estimator for $MA(\mu, \otimes_{k=1}^K \mu_k)$ is given by

$$\widehat{MA}(\mu, \otimes_{k=1}^K \mu_k) = \frac{1}{n^2} \sum_{i,j=1}^n \left(F_{\mu,n}^M(\mathbf{X}_i, \mathbf{X}_j) - \prod_{k=1}^K F_{\mu_k,n}^M(X_{ik}, X_{jk}) \right)^2.$$

The following proposition demonstrates the practical use of Theorems 1–3.

Proposition 2. (a) Under the conditions in Theorem 2 or Corollary 2, $MA(\mu, \otimes_{k=1}^K \mu_k) = 0$ if and only if $\mu = \otimes_{k=1}^K \mu_k$.
(b) Under the conditions in Theorem 3, we have $\widehat{MA}(\mu, \otimes_{k=1}^K \mu_k) \xrightarrow[n\to\infty]{a.s.} MA(\mu, \otimes_{k=1}^K \mu_k)$.
(c) Suppose the conditions in Theorem 3 hold, then under the null hypothesis

$$n\widehat{MA}(\mu, \otimes_{k=1}^K \mu_k) \xrightarrow[n\to\infty]{d} \sum_{l=1}^\infty \lambda_l Z_l^2,$$

where $\{Z_l\}_{l \in \mathbb{N}}$ is a countable sequence of iid standard normal random variables and $\{\lambda_l\}_{l \in \mathbb{N}}$ is a descending ordered. Thus, $\widehat{MA}(\mu, \otimes_{k=1}^K \mu_k)$ is a consistent test for any fixed alternative hypothesis.

Motivated by Proposition 2(c), we can derive an estimator for the $\{\lambda_l\}_{l \in \mathbb{N}}$ and give a spectrum-based test by following the similar procedure for MCVM (See Section 4 in supplementary materials). Also, we can approximate p-values by permutation when the sample size is relatively small. The numerical comparison between the permutation-based and spectrum-based mutual independence test is conducted in Section 5.1.

5. Monte Carlo Studies

5.1. Consistency of Tests: Large-Sample Regime

We investigate the consistency of the permutation-based and spectrum-based tests proposed above. We simulate datasets drawn from the multivariate Gaussian distribution $N(\mu, I_{2\times2})$, the von-Miser Fisher distribution $\mathcal{V}(w)$ with concentration parameter $\|w\|$, and the Wishart distribution $\mathcal{W}(\Sigma)$ with degree of freedom 8. They are common distributions in Euclidean, spherical, and SPD matrices space. Here, we vary sample size n from 200 to 600. The significance level is fixed at 0.05.

For the assessment of the homogeneity test, we design the following models:

- Euclidean: $X \sim N(\mathbf{0}, I_{2\times2})$, $Y \sim N(\mu^y, I_{2\times2})$, $Z \sim N(\mu^z, I_{2\times2})$.

- Sphere: Let $X \sim \mathcal{V}((1,0)^\top)$, $Y \sim \mathcal{V}(w^y)$ and $Z \sim \mathcal{V}(w^z)$.
- SPD: $X \sim \mathcal{W}(I_{2\times2})$, $Y \sim \mathcal{W}(\Sigma^y)$, $Z \sim \mathcal{W}(\Sigma^z)$, where Σ^y, Σ^z are 2×2 matrices.

We set $\mu^y = \mu^z = \mathbf{0}$, $w^y = w^z = (1,0)^\top$, and $\Sigma^y = \Sigma^z = I_{2\times2}$ to examine Type-I errors for the above settings, respectively. To check the consistency of the proposed tests, we set $\mu^y = (0.2, 0.2)^\top$, $\mu^z = (-0.2, -0.2)^\top$, $w^y = \frac{1}{\sqrt{5}}(2,1)^\top$, $w^z = \frac{1}{\sqrt{2}}(1,1)^\top$, $\Sigma_{ij}^y = 0.1^{|i-j|}$, $\Sigma_{ij}^z = (-0.1)^{|i-j|}$.

To detect the mutual dependence among (X, Y, Z), we consider the three models below.

- Euclidean: $(X, Y, Z) \sim N(\mathbf{0}_3, \Sigma)$.
- Spherical: X', Y', Z' are random variables in $\mathbb{R}^2$ and $(X', Y', Z') \sim N(\mathbf{0}_6, \Sigma)$. And $X \sim \mathcal{V}(X')$, $Y \sim \mathcal{V}(Y')$, $Z \sim \mathcal{V}(Z')$.
- SPD: Draw (X', Y', Z') from $N(\mathbf{0}, \Sigma)$. Let $\mathcal{W}(a)$ be a 2×2 matrix with diagonal value 10 and non-diagonal value a, we set $X \sim \mathcal{W}(X')$, $Y \sim \mathcal{W}(Y')$, $Z \sim \mathcal{W}(Z')$.
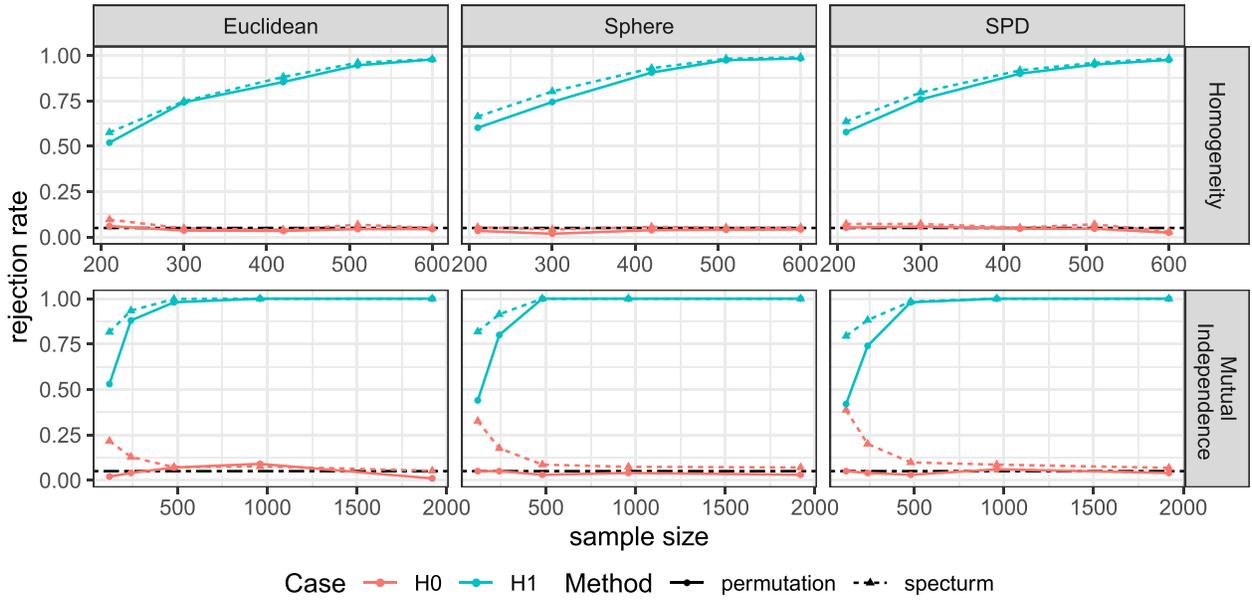
We set Σ as the identity matrix to assess Type-I errors. As for power analysis, we set $\Sigma_{ij} = 0.2^{|i-j|}$ for Euclidean data and $\Sigma_{ij} = 0.6^{|i-j|}$ for spherical and SPD datasets. We now introduce the distance measure. We employ Euclidean/geodesic distance for data in Euclidean/spherical space. We use the Cholesky distance (Dryden, Koloydenko, and Zhou 2009) to measure the difference between two SPD matrices P_1, P_2, which is defined as $\|chol(P_1) - chol(P_2)\|_F$, where $\| \cdot \|_F$ is Frobenius norm and $chol(\cdot)$ is the Cholesky decomposition.

Figure 3 displays the rejection rate of the homogeneity and mutual-independence tests under 500 Monte Carlo runs. When the null hypotheses hold, both permutation and spectrum tests control the rejection rate well around the significance level. When n is not sufficiently large, the spectrum-based test may have an excessive rejection rate than the permutation test. Besides, controlling the Type-I error of the mutual independence test requires more samples than the homogeneity test. This is because the form of h_2 involves multiple terms depending on the unknown probability measure $\{\mu_k\}_{k=1}^K$, implying more samples are required to control the approximation error of h_2. When the alternative hypotheses hold, the empirical powers of the permutation and spectrum tests are close, and they both increase to 1 as n goes to infinity, reflecting the consistency of the two tests.

5.2. Power Analysis: Finite-Sample Regime

We first depict the setting for testing the homogeneity of X, Y, Z. Let $\Sigma(a)$ be a p-by-p matrix whose non-diagonal entries are a, and diagonal entries are 1.

- Euclidean: (i) $X \sim N(\mathbf{0}, I_{2\times2})$, $Y \sim N(\kappa \times \mathbf{1}, I_{2\times2})$, $Z \sim N(-\kappa \times \mathbf{1}, I_{2\times2})$; (ii) X, Y, Z are drawn from zero-mean multivariate t-distributions with degree of freedom 3, $3 - \kappa$, and $3 + \kappa$.
- Spherical: (i) X, Y, Z comes from the von Miser-Fisher distributions with concentration parameter 2.5, with directions $(\cos\frac{\pi}{4}, \sin\frac{\pi}{4})^\top$, $(\cos(\frac{\pi}{4}(1 + \tanh\kappa)), \sin(\frac{\pi}{4}(1 + \tanh\kappa)))^\top$, and $(\cos(\frac{\pi}{4}(1 - \tanh\kappa)), \sin(\frac{\pi}{4}(1 - \tanh\kappa)))^\top$, respec-

Figure 3. Rejection rate of the proposed homogeneity and mutual independence tests. The line type distinguishes permutation-based and spectrum-based tests; the color distinguishes the H_0 and H_1. The black dashed line is the significance level.

tively. (ii) X is the equiv-probability mixture of von-Miser Fisher distributions with directions $(\cos\frac{\pi}{4}, -\sin\frac{\pi}{4})^\top$, $(\cos\frac{\pi}{4}, \sin\frac{\pi}{4})^\top$ and concentration parameter 2.5. Y, Z are the same as X except that directions are replaced with $(\cos\alpha, \sin\alpha)^\top, (\cos\alpha, -\sin\alpha)^\top$ and $(\cos\beta, \sin\beta)^\top$, $(\cos\beta, -\sin\beta)^\top$, respectively. Here, $\alpha = \frac{\pi}{4}(1 + \tanh\kappa)$ and $\beta = \frac{\pi}{4}(1 - \tanh\kappa)$.

- SPD matrix: (i) X, Y, Z are drawn from the Wishart distributions with the degree of freedom 8. The scale matrices of X, Y, Z are $I_{3\times3}$, $\Sigma_3(\kappa)$, $\Sigma_3(-\kappa)$; (ii) X, Y, Z are the Wishart distributions with the degree of freedom $8 - \kappa$, 8, $8 + \kappa$, respectively. The scale matrices of X, Y, Z are $\frac{8}{8-\kappa} \times \Sigma_3(0.1)$, $\Sigma_3(0.1)$, and $\frac{8}{8+\kappa} \times \Sigma_3(0.1)$.

For each metric space, case (i) makes distributions only have a difference on Fréchet mean, while case (ii) only has a difference on Fréchet variance.

Next, we describe the setting for testing the mutual independence among (X, Y, Z).

- Euclidean: (i) $(X, Y, Z) \sim N(\mathbf{0}, \Sigma_3(\kappa))$; (ii) (X, Y, Z) follows the equiv-probability mixture of $N(\mathbf{0}, \Sigma_3(\kappa))$ and $N(\mathbf{0}, \Sigma_3(-\kappa))$.
- Spherical: first sample (X', Y', Z') from $N(\mathbf{0}, \Sigma_3(\kappa))$. (i) X, Y, Z come from three von-Miser distributions with concentration parameters 2.8 and directions: $(\cos(X'), \sin(X'), 0, 0)^\top$, $(\cos(Y'), \sin(Y'), 0, 0)$ and $(\cos(Z'), \sin(Z'), 0, 0)$, respectively. (ii) X, Y, Z come from three von-Miser distributions whose directions are $(1, 0, 0, 0)^\top$ and concentration parameters $|X'|, |Y'|, |Z'|$, respectively.
- SPD matrix: $(X', Y', Z') \sim N(\mathbf{0}, \Sigma_3(\kappa))$, then we generate X, Y, Z from three Wishart distributions with parameters: (i) scale matrix $\Sigma_3(0.1)$ and degree-of-freedom are $3 + 8|X'|$, $3 + 8|Y'|$ and $3 + 8|Z'|$; or (ii) scale matrices $\frac{1}{3+12|X'|} \times \Sigma_3(0.1)$, $\frac{1}{3+12|Y'|} \times \Sigma_3(0.1)$, $\frac{1}{3+12|Z'|} \times \Sigma_3(0.1)$ and degree-of-freedom are $3 + 12|X'|$, $3 + 12|Y'|$, $3 + 12|Z'|$.

For each metric space, X, Y, Z have mean dependence in (i) and variance dependence in (ii).

We study the empirical power of the proposed tests when the distribution discrepancy/dependence strength κ varies, but n is fixed. We compare our proposed method with energy distance (ED, Székely and Rizzo 2004), Fréchet variance analysis (FVA, Dubey and Muller 2019) for homogeneity test with $n_1 = n_2 = n_3 = n$. For testing mutual independence, we compare our proposed test to the total multivariance (TM) method (Böttcher, Keller-Ressel, and Schilling 2019). The significance level is fixed at 0.05. We use 399 permutation replications to compute p-values. 500 Monte Carlo runs are performed to estimate the power. The results are presented in Figure 4. We also conduct experiments where the sample size n increases but κ is fixed, whose results are deferred to supplementary materials.

From Figure 4(a), the power of the MCVM and ED increases to 1 as the gap between distributions enlarges; their power also approaches the significance level as the gap closes. From the upper panel of Figure 4(a), when the distributions have a Fréchet mean difference, ED outperforms the others, followed by MCVM and FVA that have competitive performance. On the other hand, MCVM is superior in detecting Fréchet variance difference. These observations coincide with the finding that the MDF-based method is good at detecting scale differences (Kim, Balakrishnan, and Wasserman 2020). Notice that the FVA also has a remarkable power at detecting Fréchet variance difference for the SPD data, which coincides with the report of Dubey and Muller (2019). Unfortunately, this fact relies on the well-estimation for the Fréchet mean, which may be difficult or even impossible for heavy-tailed data, as we can see in the left-bottom panel in Figure 4(a). Moreover, the (approximate) violation of the uniqueness assumption of the Fréchet mean hinders the power of the FVA increases. This fact can be witnessed in the middle-bottom panel of Figure 4(a)—the power of FVA cannot improve to 1.0 when κ exceeds 1.0, where Y approaches a
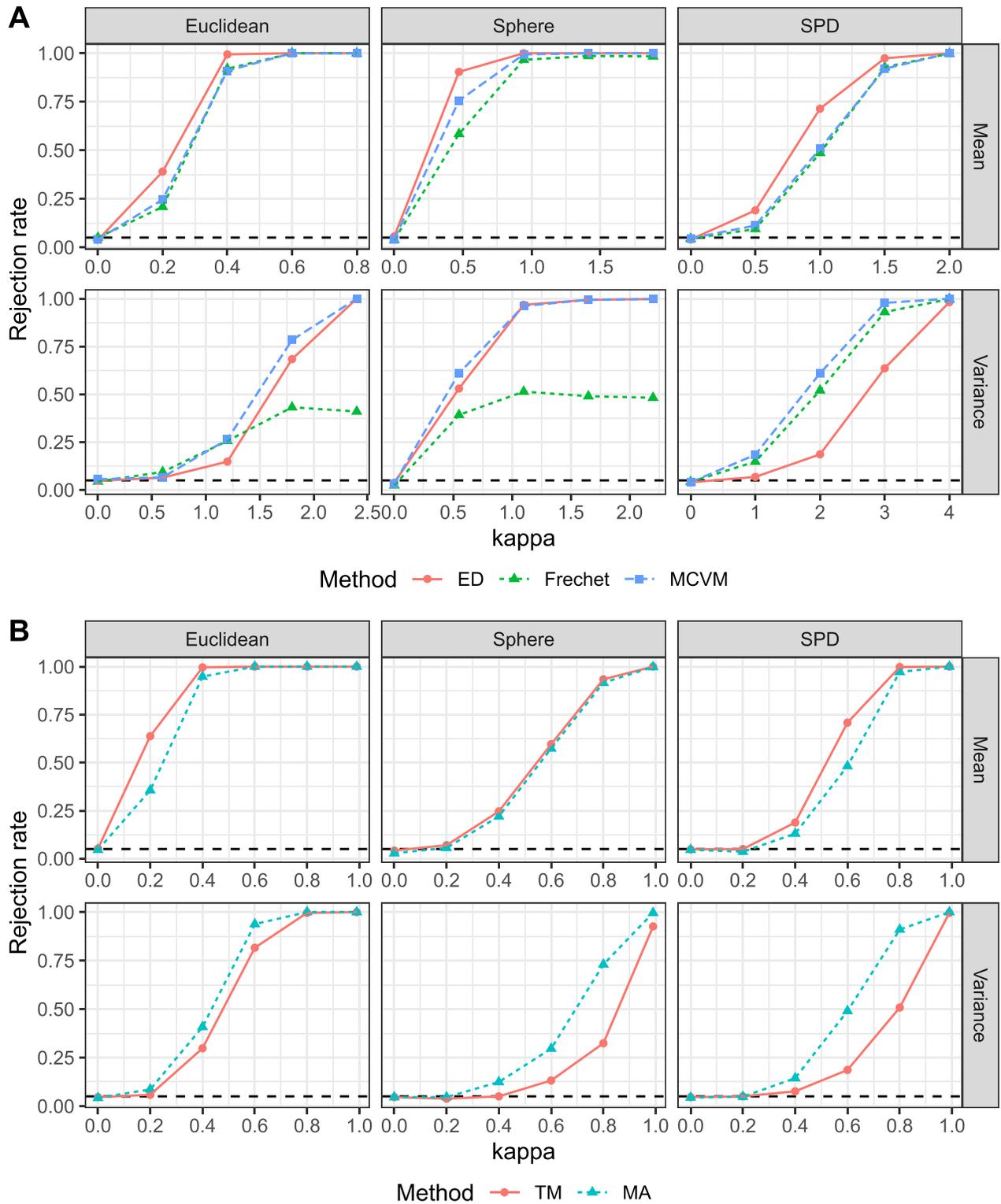
Figure 4. Rejection rate of hypothesis tests when κ increases. (a) the homogeneity test. (b) the mutual independence test. Methods are distinguished by point and line. The black dashed line is the nominal significance level.

mixture of von-Miser distributions with directions $(0,1)$ and $(0,-1)$ whose Fréchet mean does not exist.

Figure 4(b) displays the results of the mutual independence test. From Figure 4(b), the power functions of the two tests monotonously increase to 1 as the dependence strength κ increases. Notably, when random objects are mutually inde-

pendent ($\kappa = 0$), the empirical power of the two tests is around the nominal significance level. Moreover, TM is better than MA when random objects depend on Fréchet mean, but the MA test still enjoys a competitive power. On the other hand, MA is more powerful than TM when random objects have a dependence on variance, which is even more visibly in spherical and SPD

data. Lastly, it is noteworthy that the advantages of the MDF-based tests persist in the complex scenarios presented in the final section of the Supplementary Materials. This underscores the promise of our proposed tests for real-world data that may possess complexity.

6. Real Data Analysis

6.1. Alzheimer's Disease Neuroimaging Initiative Data

The Alzheimer's Disease Neuroimaging Initiative (ADNI) is a multisite study that aims to improve the prevention and treatment of Alzheimer's disease (AD). AD is a neurodegenerative disease, resulting in the decline of some cognitive impairments that cause behavioral deficits. Data including magnetic resonance images, demographic variables, genetic markers, and AD assessment scale cognitive score (ADASCS) were collected to study AD and the human brain. In this study, we focus on a critical brain region: the hippocampus, which is typically firstly damaged by AD, leading to the first clinical manifestations in the form of episodic memory deficits (Weiner et al. 2013). By applying our method for the data preprocessed by Kong et al. (2018), we are interested in factors that affect the hippocampus.

The preprocessed data contain the left and right hippocampus of 373 individuals, each of which is characterized by 15,000 radial distances on the left and right hippocampus surfaces, where the radial distance is defined as the Euclidean distance between the corresponding vertex on the surface and the medial core of the hippocampus (see Figure 5(a)). From a functional curve example exhibited in Figure 5(b), we see that the functional curve has an obvious fluctuation and periodicity. And thus, to capture the main variation of functional curves, like

Kong et al. (2018), we apply smoothing and functional principal component analysis (Ramsay and Silverman 1997) on the functional curves of left and right surfaces, respectively. We find that the top nine functional principal components for each hippocampus can explain 99% of the total variance.

We consider gender, age, handedness, marital status, education length, retirement, Apolipoprotein E (APOE) 3-allele haplotype (i.e., the $\epsilon 2$, $\epsilon 3$, and $\epsilon 4$ variants), and the ADASCS. The ADASCS is quantitatively evaluated behavioral deficits caused by AD, and the higher the ADASCS is, the more severe the deficits are.

The metric space for left/right hippocampus is (L^2, d_h), where L^2 is the collection of square integral functions and $d_h(u, v) = \|u - v\|_2$ for $u, v \in L^2$. To jointly consider left and right hippocampi, we set their product metric space as $(L^2 \times L^2, d)$, where $d(\mathbf{u}, \mathbf{v}) = \|(d_h(u_1, v_1), d_h(u_2, v_2))\|_2$ for $\mathbf{u} = (u_1, u_2), \mathbf{v} = (v_1, v_2) \in L^2 \times L^2$.

To answer the above question, we use the MA and TM tests to evaluate whether demographic and genetic factors affect the hippocampus. The results are displayed in Table 1. As seen from Table 1, both MA and TM detect associations of the hippocampus with age and ADASCS. The increase in age accumulates the abnormal deposition of β amyloid fibrils, which starts the neural damage with hippocampus atrophy. Furthermore, the atrophy of the hippocampus causes behavior deficits (Jack et al. 2010), and the ADASCS are expected to be related to the hippocampus.

Notably, the MA test also detects APOE$\epsilon 4$, but TM does not. This difference is important because APOE$\epsilon 4$ is a well-known major genetic risk factor for AD and has repeatedly been reported to affect the hippocampus (Jack et al. 2010). Figure 6 displays the difference in the mean functional curves between the APOE$\epsilon 4$ carriers and non-carriers and indicates

(a)   (b)
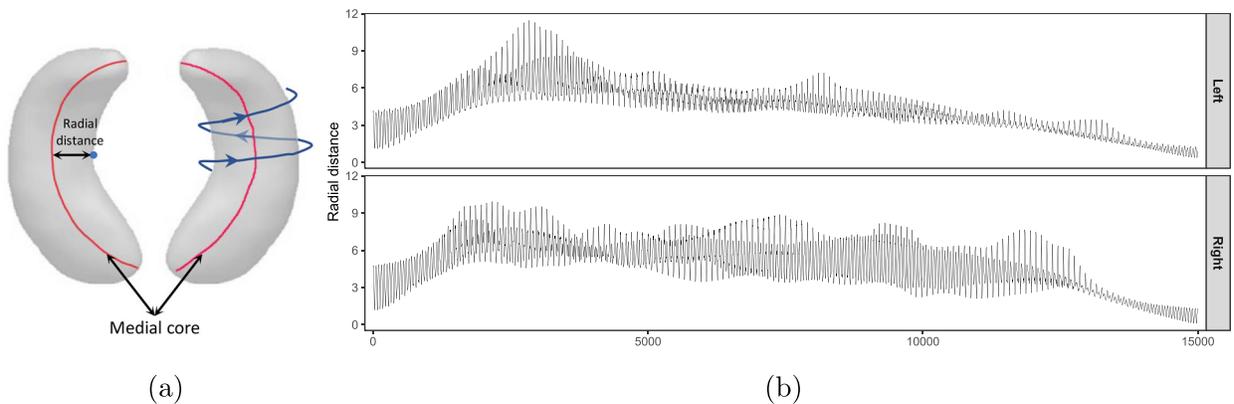
Figure 5. (a) Visualization for medial core, radial distance, and the order of 15,000 landmarks on each hippocampus surface. (b) One subject's functional curve of radial distance on the left and right hippocampus surfaces. The x-axis corresponds to 15,000 landmarks that whirlingly surround the hippocampus.

Table 1. The p-values (adjusted p-values under Holm's correction) of independence tests for the ADNI dataset.

Test	Gender	Handedness	MS	EL	Retirement
TM	0.012 (0.096)	0.760 (1.000)	0.628 (1.000)	0.020 (0.140)	0.318 (1.000)
MA (Permute)	0.038 (0.259)	0.589 (1.000)	0.830 (1.000)	0.079 (0.395)	0.507 (1.000)
MA (Spectrum)	0.046 (0.295)	0.546 (1.000)	0.811 (1.000)	0.077 (0.386)	0.499 (1.000)
Test	Age	APOE$\epsilon 2$	APOE$\epsilon 3$	APOE$\epsilon 4$	ADASCS
TM	**0.001 (0.010)**	0.871 (1.000)	0.228 (1.000)	0.023 (0.140)	**0.001 (0.010)**
MA (Permute)	**0.001 (0.010)**	0.226 (0.904)	0.037 (0.259)	**0.004 (0.032)**	**0.001 (0.010)**
MA (Spectrum)	**0.0003 (0.001)**	0.229 (0.918)	0.042 (0.295)	**0.002 (0.018)**	**0.0002 (0.002)**

NOTE: MS and EL are abbreviations of marital status and educational length. The p-values under 0.05 after Holm's correction are bolded.
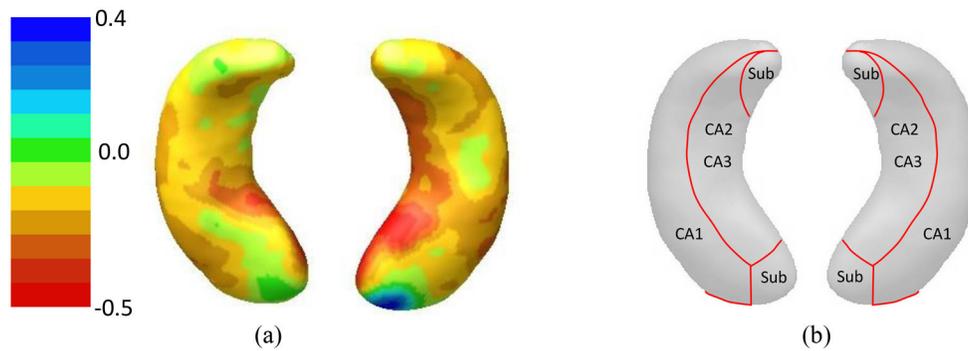
Figure 6. (a) Visualization for mean radial distance difference between APOEϵ4 and non-APOEϵ4 carriers on hippocampi surfaces. The color bar depicts the mean radial distance difference. (b) Hippocampal subfields. CA and Sub are abbreviations for cornu amonis and subiculum, respectively.

Table 2. The p-values (adjusted p-values under Holm's correction) of (mutual) independence tests for the ADHD-200 dataset.

Test	Joint	(G, H)-FC	(G, FC)-H	(H, FC)-G	FC-H	FC-G	H-G
TM	0.206	0.018 (0.054)	0.321 (0.321)	0.037 (0.074)	0.020 (0.060)	0.057 (0.114)	0.755 (0.755)
MA$_1$	**0.017**	**0.006 (0.018)**	0.642 (0.642)	**0.017 (0.035)**	**0.005 (0.016)**	0.112 (0.224)	0.989 (0.989)
MA$_2$	**0.022**	**0.005 (0.015)**	0.701 (0.701)	**0.017 (0.034)**	**0.008 (0.024)**	0.139 (0.278)	0.876 (0.876)

NOTE: The p-values under 0.05 after Holm's correction are bolded. G, H, and FC are abbreviations of gender, handedness, and functional connectome, respectively. MA$_1$: spectrum-based test; MA$_2$: permutation-based test. Note that the results are rounded to three digits.

that APOEϵ4 shrinks the hippocampus. Moreover, Figure 6 suggests that the atrophy caused by APOEϵ4 is more severe in the right hippocampus, and cornu amonis 2 (CA2) and CA3 are heavily shrunk by APOEϵ4 following by CA1 and subiculum. O'Dwyer et al. (2012) also found this phenomenon by studying left and right hippocampi volumes.

6.2. ADHD-200 Dataset

The correlated spontaneous fluctuations manifest different patterns in different brain regions during sleep or under anesthesia. This phenomenon is called functional connectome (FC), where a node represents a part of the brain, and an edge represents the direct correlation between two nodes. It is interesting to evaluate the relationship between FC phenotypes and other variables. For example, we investigate the mutual dependence among FC, gender, and handedness with ADHD-200 dataset (Bellec et al. 2017).

The ADHD-200 dataset includes 162 individuals, with 63 males and 99 females. In this dataset, handedness is measured as a continuous score—a person with large handedness implies that he/she is dextromanuality. For each subject, the ADHD-200 dataset provides a preprocessed resting-state functional magnetic resonance imaging (rfMRI) that repeatedly records the blood-oxygen levels on 111 disjoint regions in the brain. As suggested by Smith et al. (2013), we can reconstruct the functional connectome by computing a 111×111 partial correlation matrix of the disjoint regions. Then, the Cholesky distance is chosen to measure the difference among SPDs due to its favorable performance in the simulation studies.

We applied the MA-based and TM tests to this dataset to answer our question. The test results are displayed in Table 2. As can be seen from the first column of Table 2, only the MA-based tests reject the null hypothesis of the mutual independence among FC, gender, and handedness, while the TM test does not. In addition, we divide the three factors into two groups

and study the dependence of the two groups. The results are presented in the third to fifth columns in Table 2. The third column shows that gender and handedness together may affect the FC by our MA-based methods. While the null hypothesis of the independence between (gender, FC) and handedness is not rejected in the fourth column, the null hypothesis of the independence between (handedness, FC) and gender is hard to judge in the fifth column. We further investigate the pairwise dependence among three factors. From the sixth column in Table 2, we see that only the MA-based tests reject the hypothesis of the independence between handedness and FC. A recent study found that the FCs of left- and right-handed individuals extend across every brain region (Tejavibulya et al. 2022). They discovered that connections between and within the cerebellum have distinct connectivity patterns. Besides, from the seventh to eighth columns in Table 2, both the MA-based and TM tests suggest insufficient evidence to assert gender influences handedness or FC. In conclusion, our MA-based method for mutual independence is very useful for exploring the high-order dependence among variables, including non-Euclidean ones.

7. Conclusion and Discussion

To address the need for non-Euclidean data analysis, we characterize Borel probability measures in metric space. This is analogous to how the DF represents Borel probability measures in Euclidean space. We propose a metric distribution function. To a large extent, the MDF retains the DF's desirable properties, including the 1–1 correspondence theorem, under certain mild conditions. Moreover, the EMDF provides a simple and reliable method for estimating the MDF, which coincides with the weak mode of calculating the probability measure in Vapnik (2010). The outer-directed loop in Figure 1 is closed thanks to the EMDF's Glivenko-Cantelli property and the Donsker property; this paves the way for a new paradigm in statistical inference in metric spaces. As a result, we can perform nonparametric statis-

tical inference for non-Euclidean data with the MDF, similar to that for Euclidean data with the DF.

We provide extensive empirical results for using the MDF in the homogeneity test and the (mutual) independence test for non-Euclidean data, including SPD matrices, shapes, and smooth functions, to highlight the great potential of MDF in practice. Using the Glivenko-Cantelli property and the Donsker property of the EMDF, we obtain estimators and their consistency for estimations and test procedures that are described in Section 4. Our simulation experiments demonstrate that the tests based on the MDF have good finite sample performance, are robust to various data settings, and are free of tuning parameters. Therefore, the applicability of MDF is expected to be broad and simple. We also reanalyzed the ADNI dataset and found evidence that the APOEϵ4 affected the hippocampus.

The MDF and the DF are connected. Denote $F(\mathbf{u}) = P(X \leq \mathbf{u})$ as the DF associated with probability measure μ defined in $\mathbb{R}$. Given $\mathbf{u}, \mathbf{v} \in \mathbb{R}$ such that $\mathbf{u} < \mathbf{v}$, we have

$$F_\mu^M(\mathbf{u}, \mathbf{v}) = P(|X - \mathbf{u}| \leq |\mathbf{v} - \mathbf{u}|)$$
$$= P(X \leq \mathbf{v}) - P(X < 2\mathbf{u} - \mathbf{v}).$$

Let $\mathbf{u} \to -\infty$, then we have $\lim_{\mathbf{u} \to -\infty} F_\mu^M(\mathbf{u}, \mathbf{v}) = F(\mathbf{u})$. Despite their connections, the MDF and the DF are not the same. The MDF describes the random object by the distribution of distances at different locations. Thus, the viewpoint of MDF is flexible and relative since the location variable $\mathbf{u}$ can be different. Such characteristics allow the MDF to grasp the information of probability measures in metric spaces.

Notably, the MDF is also related to another useful concept in nonparametric statistics—statistical depth. The statistical depth methods use a unique center, the so-called deepest point, to characterize the distribution. These methods can properly characterize distributions in Euclidean spaces, and recent progress enables the characteristics of the depth in metric space (Dai and Lopez-Pintado 2022; Liu, Wang, and Zhu 2022). We note that the method proposed by Liu, Wang, and Zhu (2022) is derived from the MDF, providing another example that the MDF establishes a unified framework for analyzing complex data objects.

Supplementary Materials

Supplementary document: contains technical proof, some properties of EMDF, a discussion about the spectrum-based tests, and additional simulation results. (.pdf file)

Acknowledgments

Data Availability Statement

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative database (*adni.loni.usc.edu*).

As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in the analysis or writing of this report. A complete listing of ADNI investigators can be found at: *adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf* .

Disclosure Statement

No potential conflict of interest was reported by the author(s).

ORCID

Xueqin Wang ⬡ https://orcid.org/0000-0001-5205-9950

References

Bellec, P., Chu, C., Chouinard-Decorte, F., Benhajali, Y., Margulies, D. S., and Craddock, R. C. (2017), "The Neuro Bureau adhd-200 Preprocessed Repository," *NeuroImage*, 144, 275–286. Data Sharing Part II. [11]

Böttcher, B., Keller-Ressel, M., and Schilling, R. L. (2019), "Distance Multivariance: New Dependence Measures for Random Vectors," *Annals of Statistics*, 47, 2757–2789. [1,8]

Castillo, I., Schmidt-Hieber, J., and Van der Vaart, A. (2015), "Bayesian Linear Regression with Sparse Priors," *Annals of Statistics*, 43, 1986–2018. [4]

Cornea, E., Zhu, H., Kim, P., Ibrahim, J. G., and Initiative, A. D. N. (2017), "Regression Models on Riemannian Symmetric Spaces," *Journal of the Royal Statistical Society*, Series B, 79, 463–482. [1]

Dai, X., and Lopez-Pintado, S. (2022), "Tukey's Depth for Object Data," *Journal of the American Statistical Association*, 118, 1760–1772. [1,12]

Darling, D. A. (1957), "The Kolmogorov-Smirnov, Cramér-von mises Tests," *The Annals of Mathematical Statistics*, 28, 823–838. [1]

Davies, R. O. (1971), "Measures not Approximable or not Specifiable by Means of Balls," *Mathematik*, 18, 157–160. [4]

Do Carmo, M. P., and Flaherty Francis, J. (1992), *Riemannian Geometry* (Vol. 6), Cham: Springer. [3]

Driemel, A., Nusser, A., Phillips, J. M., and Psarros, I. (2021), "The vc Dimension of Metric Balls under Fréchet and Hausdorff Distances," *Discrete & Computational Geometry*, 66, 1351–1381. [5]

Dryden, I. L., Koloydenko, A., and Zhou, D. (2009), "Non-Euclidean Statistics for Covariance Matrices, with Applications to Diffusion Tensor Imaging," *Annals of Applied Statistics*, 3, 1102–1123. [7]

Dubey, P., and Muller, H.-G. (2019), "Fréchet Analysis of Variance for Random Objects," *Biometrika*, 106, 803–821. [1,8]

Efron, B. (1979), "Bootstrap Methods: Another Look at the Jackknife," *Annals of Statistics*, 7, 1–26. [1]

Federer, H. (2014), *Geometric Measure Theory*, Berlin, Heidelberg: Springer. [3]

Halmos, P. R. (1956), *Measure Theory*, New York: D. Van Nostrand Company, Inc. [1,4]

Hoeffding, W. (1948), "A Non-parametric Test of Independence," *The Annals of Mathematical Statistics*, 19, 546–557. [1,6]

Hong, Y., Kwitt, R., Singh, N., Vasconcelos, N., and Niethammer, M. (2016), "Parametric Regression on the Grassmannian," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38, 2284–2297. [1]

Jack, C. R., Knopman, D. S., Jagust, W. J., Shaw, L. M., Aisen, P. S., Weiner, M. W., Petersen, R. C., and Trojanowski, J. Q. (2010), "Hypothetical Model of Dynamic Biomarkers of the Alzheimer's Pathological Cascade," *The Lancet Neurology*, 9, 119–128. [10]

Kim, I., Balakrishnan, S., and Wasserman, L. (2020), "Robust Multivariate Nonparametric Tests via Projection Averaging," *Annals of Statistics*, 48, 3417–3441. [8]

Kong, D., Ibrahim, J. G., Lee, E., and Zhu, H. (2018), "Flcrm: Functional Linear Cox Regression Model," *Biometrics*, 74, 109–117. [10]

Lin, Z., and Müller, H.-G. (2021), "Total Variation Regularized Fréchet Regression for Metric-Space Valued Data," *Annals of Statistics*, 49, 3510–3533. [4]

Liu, H., Wang, X., and Zhu, J. (2022), "Quantiles, Ranks and Signs in Metric Spaces," arXiv preprint arXiv:2209.04090. [12]

O'Dwyer, L., Lamberton, F., Matura, S., Tanner, C., Scheibe, M., Miller, J., Rujescu, D., Prvulovic, D., and Hampel, H. (2012), "Reduced Hippocampal Volume in Healthy Young apoe4 Carriers: An MRI Study," *PloS One*, 7, 1–10. [11]

O'Hara, R. B., and Sillanpää, M. J. (2009), "A Review of Bayesian Variable Selection Methods: What, How and Which," *Bayesian Analysis*, 4, 85–117. [4]

Pan, W., Tian, Y., Wang, X., and Zhang, H. (2018), "Ball Divergence: Nonparametric Two Sample Test," *Annals of Statistics*, 46, 1109–1137. [1]

Pan, W., Wang, X., Zhang, H., Zhu, H., and Zhu, J. (2020), "Ball Covariance: A Generic Measure of Dependence in Banach Space," *Journal of the American Statistical Association*, 115, 307–317. [7]

Petersen, A., Liu, X., and Divani, A. A. (2021), "Wasserstein f-tests and Confidence Bands for the Fréchet Regression of Density Response Curves," *Annals of Statistics*, 49, 590–611. [1]

Petersen, P. (2006), *Riemannian Geometry* (Vol. 171), New York: Springer. [6]

Ramsay, J., and Silverman, B. W. (1997), *Functional Data Analysis*, New York: Springer. [10]

Scealy, J. L., and Wood, A. T. A. (2019), "Scaled von Mises–Fisher Distributions and Regression Models for Paleomagnetic Directional Data," *Journal of the American Statistical Association*, 114, 1547–1560. [1]

Smith, S. M., Vidaurre, D., Beckmann, C. F., Glasser, M. F., Jenkinson, M., Miller, K. L., Nichols, T. E., Robinson, E. C., Salimi-Khorshidi, G., Woolrich, M. W., Barch, D. M., Uğurbil, K., and Essen, D. C. V. (2013), "Functional Connectomics from Resting-State fMRI," *Trends in Cognitive Sciences*, 17, 666–682. Special Issue: The Connectome. [1,11]

Székely, G. J., and Rizzo, M. L. (2004), "Testing for Equal Distributions in High Dimension," *InterStat*, 5, 1249–1272. [1,8]

Székely, G. J., Rizzo, M. L., and Bakirov, N. K. (2007), "Measuring and Testing Dependence by Correlation of Distances," *Annals of Statistics*, 35, 2769–2794. [1]

Tejavibulya, L., Peterson, H., Greene, A., Gao, S., Rolison, M., Noble, S., and Scheinost, D. (2022), "Large-Scale Differences in Functional Organization of Left-and Right-Handed Individuals Using Whole-Brain, Data-Driven Analysis of Connectivity," *NeuroImage*, 252, 119040. [11]

Vapnik, V. (2010), *The Nature of Statistical Learning Theory*, New York: Springer. [11]

Wainwright, M. J. (2019), *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, Cambridge: Cambridge University Press. [5,6]

Weiner, M. W., Veitch, D. P., Aisen, P. S., Beckett, L. A., Cairns, N. J., Green, R. C., Harvey, D., Jack, C. R., Jagust, W., Liu, E., Morris, J. C., Petersen, R. C., Saykin, A. J., Schmidt, M. E., Shaw, L., Shen, L., Siuciak, J. A., Soares, H., Toga, A. W., and Trojanowski, J. Q. (2013), "The Alzheimer's Disease Neuroimaging Initiative: A Review of Papers Published Since its Inception," *Alzheimer's & Dementia*, 9, e111–e194. [10]

Wellner, J., and Vaart, A. W. (2013), *Weak Convergence and Empirical Processes: With Applications to Statistics*, New York: Springer. [6]