Statistics and Its Interface Volume 16 (2023) 565–578

Metric distributional discrepancy in metric space

Wenliang Pan*, Yujue Li*, Jianwu Liu, Pei Dang, and Weixiong Mai† for the Alzheimer's Disease Neuroimaging Initiative

Independence analysis is an indispensable step before regression analysis to find out the essential factors that influence the objects. With many applications in machine Learning, medical Learning and a variety of disciplines, statistical methods of measuring the relationship between random variables have been well studied in vector spaces. However, there are few methods developed to verify the relation between random elements in metric spaces. In this paper, we present a novel index called metric distributional discrepancy (MDD) to measure the dependence between a random element X and a categorical variable Y, which is applicable to the medical image and related variables. The metric distributional discrepancy statistics can be considered as the distance between the conditional distribution of X given each class of Y and the unconditional distribution of X. MDD enjoys some significant merits compared to other dependence-measures. For instance, MDD is zero if and only if X and Y are independent. MDD test is a distribution-free test since there is no assumption on the distribution of random elements. Furthermore, MDD test is robust to the data with heavy-tailed distribution and potential outliers. We demonstrate the validity of our theory and the property of the MDD test by several numerical experiments and real data analysis.

Keywords and phrases: Metric distributional discrepancy, Random element, Metric space, Distribution-free test.

1. INTRODUCTION

In view of the ever-growing and complex data in today's scientific world, there is an increasing need for a generic method to deal with these datasets in diverse application scenarios. Non-Euclidean data, such as brain imaging, computational biology, computer graphics and computational social sciences and among many others [24, 21, 4, 31, 2, 30], arises in many domains. For instance, images and time-varying data can be presented as functional data that are in the form of functions [7, 6, 19]. The sphere, Matrix groups, Positive-Definite Tensors and shape spaces are also included as manifold examples [39, 43, 27]. It is of great interest to discover the associations in those data. Nevertheless, many traditional analysis methods that cope with data in Euclidean spaces become invalid since non-Euclidean spaces are inherently nonlinear space without inner product. The analysis of these Non-Euclidean data presents many mathematical and computational challenges.

One major goal of statistical analysis is to understand the relationship among random vectors, such as measuring a linear/ nonlinear association between data, which is also a fundamental step for further statistical analysis (e.g., regression modeling). Correlation is viewed as a technique for measuring associations between random variables. A variety of classical methods has been developed to detect the correlation between data. For instance, Pearson's correlation and canonical correlation analysis (CCA) are powerful tools for capturing the degree of linear association between two sets of multi-variate random variables [3, 15, 22]. In contrast to Pearson's correlation, Spearman's rank correlation coefficient, as a non-parametric measure of rank correlation, can be applied in non-linear conditions [42].

However, statistical methods for measuring the association between complex structures of Non-Euclidean data have not been fully accounted in the methods above. In metric space, [14] proposed a generalized mean in metric spaces and a corresponding variance that may be used to quantify the spread of the distribution of metric spaces valued random elements. However, in order to guarantee the existence and uniqueness of Fréchet mean, it requires the space should be with negative sectional curvature. While in a positive sectional curvature space, the extra conditions such as bound on the support and radial distribution are required [5]. Following this, more nonparametric methods for manifold-valued data was developed. For instance, A Riemannian CCA model was proposed by [22], measuring an intrinsically linear association between two manifold-valued objects. Recently, [45] proposed distance correlation to measure the association between random vectors. After that, [46] introduced Brownian covariance and showed it to be the same as distance covariance and [28] extended the distance covariance to metric space under the condition that the space should be of strong negative type. Pan et.al [36, 37] introduced the notions of ball divergence and ball covariance for Banach-valued random vectors. These two notions can

[*] These authors contributed equally to this work.

[†] Corresponding author.

also be extended to metric spaces but by a less direct approach. Note that a metric space is endowed with a distance, it is worth studying the behaviors of distance-based statistical procedures in metric spaces.

In this paper, we extend the method in [8] and propose a novel statistics in metric spaces based on [49], called metric distributional discrepancy (MDD), which considers a closed ball with the defined center and radius. We perform a powerful independence test which is applicable between a random vector X and a categorical variable Y based on MDD. The MDD statistics can be regarded as the weighted average of Cramér-von Mises distance between the conditional distribution of X given each class of Y and the unconditional distribution of X. Our proposed method has the following major advantages, (i) X is a random element in metric spaces. (ii) MDD is zero if and only if X and Y are statistically independent. (iii) It works well when the data is in heavy-tailed distribution or extreme value since it does not require any moment assumption. Compared to distance correlation, MDD as an alternative dependence measure can be applied in the metric space which is not of strong negative type. Unlike ball correlation, MDD gets rid of unnecessary ball set calculation of Y and has higher test power, which is shown in the simulation and real data analysis.

The organization of the rest of this paper is as follows. In section 2, we give the definition and theoretical properties of MDD, and present results of monte carlo simulations in section 3 and experiments on two real data analysis in section 4. Finally, we draw a conclusion in section 5.

2. METRIC DISTRIBUTIONAL DISCREPANCY

2.1 Metric distributional discrepancy statistics

For convenience, we first list some notations in the metric space. The order pair $(\mathcal{M}, d)$ denotes a *metric space* if $\mathcal{M}$ is a set and d is a *metric* or *distance* on $\mathcal{M}$. Given a metric space $(\mathcal{M}, d)$, let $\bar{B}(x, y) = \{v : d(x, v) \leq r\}$ be a closed ball with the center x and the radius $r = d(x, y)$.

Next, we define and illustrate the metric distributional discrepancy (MDD) statistics for a random element and a categorical one. Let X be a random element and Y be a categorical random variable with R classes where $Y = \{y_1, y_2, \ldots, y_R\}$. Then, we let X'' be a copy of random element X, $F(x, x') = P_{X''}\{X'' \in \bar{B}(x, x')\}$ be the unconditional distribution function of X, and $F_r(x, x') = P_{X''}\{X'' \in \bar{B}(x, x')|Y = y_r\}$ be the conditional distribution function of X given $Y = y_r$. The $MDD(X|Y)$ can be represented as the following quadratic form between $F(x, x')$ and $F_r(x, x')$,

(1)

$$MDD(X|Y) = \sum_{r=1}^{R} p_r \int [F_r(x, x') - F(x, x')]^2 d\nu(x) d\nu(x'),$$

where $p_r = P(Y = y_r) > 0$ for $r = 1, \ldots, R$.

We now provide a consistent estimator for $MDD(X|Y)$. Suppose that $\{(X_i, Y_i) : i = 1, \ldots, n\}$ with the sample size n are $i.i.d.$ samples randomly drawn from the population distribution of (X, Y). $n_r = \sum_{i=1}^{n} I(Y_i = y_r)$ denotes the sample size of the rth class and $\hat{p}_r = n_r/n$ denotes the sample proportion of the rth class, where $I(\cdot)$ represents the indicator function. Let $\hat{F}_r(x, x') = \frac{1}{n_r} \sum_{k=1}^{n} I(X_k \in \bar{B}(x, x'), Y_k = y_r)$, and $\hat{F}(x, x') = \frac{1}{n} \sum_{k=1}^{n} I(X_k \in \bar{B}(x, x'))$. The estimator of $MDD(X|Y)$ can be obtained by the following statistics

(2)

$$\widehat{MDD}(X|Y) = \frac{1}{n^2} \sum_{r=1}^{R} \sum_{i=1}^{n} \sum_{j=1}^{n} \hat{p}_r [\hat{F}_r(X_i, X_j) - \hat{F}(X_i, X_j)]^2.$$

2.2 Theoretical properties

In this subsection, we discuss some sufficient conditions for the metric distributional discrepancy and its theoretical properties in Polish space, which is a separable completely metric space. First, to obtain the property of X and Y are independent if and only if $MDD(X \mid Y) = 0$, we introduce an important concept named directionally (ϵ, η, L)-limited [13].

Definition 1. *A metric d is called directionally (ϵ, η, L)-limited at the subset $\mathcal{A}$ of $\mathcal{M}$, if $\mathcal{A} \subseteq \mathcal{M}$, $\epsilon > 0, 0 < \eta \leq 1/3$, 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 $b, c \in D (b \neq c), x \in M$ 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.

There are many metric spaces satisfying directionally (ϵ, η, L)-limited, such as a finite dimensional Banach space, a Riemannian manifold of class ≥ 2. However, not all metric spaces satisfy Definition 1, such as an infinite orthonormal set in a separable Hilbert space H, which is verified in [49].

Theorem 1. *Given a probability measure ν with its support $supp\{\nu\}$ on $(\mathcal{M}, d)$. Let X be a random element with probability measure ν on $\mathcal{M}$ and Y be a categorical random variable with R classes $\{y_1, y_2, \ldots, y_R\}$. Then X and Y are independent if and only if $MDD(X \mid Y) = 0$ if the metric d is directionally (ϵ, η, L)-limited at $supp\{\nu\}$.*

Theorem 1 introduces the necessary and sufficient conditions for $MDD(X|Y) = 0$ when the metric is directionally (ϵ, η, L)-limited at the support set of the probability measure. Next, we extend this theorem and introduce Corollary 1, which presents reasonable conditions on the measure or on the metric.

Corollary 1. *(a) **Measure Condition**: For $\forall\ \epsilon > 0$, there exists $\mathcal{U} \subset \mathcal{M}$ such that $\nu(\mathcal{U}) \geq 1 - \epsilon$ and the metric d is directionally (ϵ, η, L)-limited at $\mathcal{U}$. Then X and Y are* *independent if and only if* $MDD(X \mid Y) = 0$.

(b) ***Metric Condition****: Given a point* $x \in \mathcal{M}$*, we define a projection on* $\mathcal{M}_k$*,* $\pi_k(\cdot) : \mathcal{M} \to \mathcal{M}_k$ *and* $\pi_k(x) = x_k$*. For a set* $A \subset \mathcal{M}$*, define* $\pi_k(A) = \bigcup_{x \in A} \{\pi_k(x)\}$*. There exist* $\{(\mathcal{M}_l, d)\}_{l=1}^{\infty}$ *which are the increasing subsets of* $\mathcal{M}$*, where each* $\mathcal{M}_l$ *is a Polish space satisfying the directionally-limited condition and their closure* $\overline{\bigcup_{l=1}^{\infty} \mathcal{M}_l} = \mathcal{M}$*. For every* $x \in \mathcal{M}$*,* $\pi_k(x)$ *is unique such that* $d(x, \pi_l(x)) = \inf_{z \in \mathcal{M}_l} d(x, z)$ *and* $\pi_l|_{\mathcal{M}_{l'}} \circ \pi_{l'} = \pi_l$ *if* $l' > l$*. Then* X *and* Y *are independent if and only if* $MDD(X \mid Y) = 0$*.*

Theorem 2. $\widehat{MDD}(X|Y)$ *almost surely converges to* $MDD(X|Y)$*.*

Theorem 2 demonstrates the consistency of proposed estimator for $MDD(X \mid Y)$. Hence, $\widehat{MDD}(X|Y)$ is consistent to the metric distributional discrepancy index. Due to the property, we propose an independence test between X and Y based on the MDD index. We consider the following hypothesis testing:

$$\begin{aligned} &H_0 : X \text{ and } Y \text{ are statistically independent.} \\ vs\ &H_1 : X \text{ and } Y \text{ are not statistically independent.} \end{aligned}$$

Theorem 3. *Under the null hypothesis* H_0*,*

$$n\widehat{MDD}(X|Y) \xrightarrow{d} \sum_{j=1}^{\infty} \lambda_j \mathcal{X}_j^2(1),$$

where $\mathcal{X}_j^2(1)$*'s,* $j = 1, 2, \ldots,$ *are independently and identically chi-square distribution with* 1 *degrees of freedom, and* $\xrightarrow{d}$ *denotes the convergence in distribution.*

Remark 1. *In practical application, we can estimate the null distribution of MDD by the permutation when the sample sizes are small, and by the Gram matrix spectrum in [16] when the sample sizes are large.*

Theorem 4. *Under the alternative hypothesis* H_1*,*

$$n\widehat{MDD}(X|Y) \xrightarrow{a.s.} \infty,$$

where $\xrightarrow{a.s.}$ *denotes the almost sure convergence.*

Theorem 5. *Under the alternative hypothesis* H_1*,*

$$\sqrt{n}[\widehat{MDD}(X|Y) - MDD(X|Y)] \xrightarrow{d} N(0, \sigma^2),$$

where σ^2 *is given in the appendix.*

3. NUMERICAL STUDIES

3.1 Monte Carlo simulation

In this section, we perform four simulations to evaluate the finite sample performance of MDD test by comparing with other existing tests: The distance covariance test (DC) [45] and the HHG test based on pairwise distances [18]. We consider the directional data on the unit sphere R^p, which is denoted by $S^{p-1} = \{x \in R^p : ||x||_2 = 1\}$ for x and n-dimensional data independently from Normal distribution. For all of methods, we use the permutation test to obtain p-value for a fair comparison and run simulations to compute the empirical Type-I error rate in the significance level of $\alpha = 0.05$. All numerical experiments are implemented by R language. The DC test and the HHG test are conducted respectively by R package *energy* [41] and R package HHG [20].

Simulation 1 In this simulation, we test independence between a high-dimensional variable and a categorical one. We randomly generate three different types of data X listed in three columns in Table 1. For the first type in column 1, we set $p = 3$ and consider coordinate of S^2, denoted as (r, θ, ϕ) where $r = 1$ as radial distance, θ and ϕ were simulated from the Uniform distribution $U(-\pi, \pi)$. For the second type in column 2, we generate three-dimensional variable from von Mises-Fisher distribution $M_3(\mu, k)$ where $\mu = (0, 0, 0)$ and $k = 1$. For the third type in column 3, each dimension of X is independently formed from $N(0, 1)$ where $X = (X_1, X_2, X_3), X_i \sim N(0, 1)$. We generate the categorical random variable Y from R classes $1, 2, \ldots, R$ with the unbalanced proportion $p_r = P(Y = r) = 2[1 + (r-1)/(R-1)]/3R,\ \ r = 1, 2, \ldots, R$. For instance, when Y is binary, $p_1 = 1/3$ and $p_2 = 2/3$ and when $R = 5, Y = 1, 2, 3, 4, 5$. Simulation times is set to 200. The sample size n are chosen to be 40, 60, 80, 120, 160. The results summarized in Table 1 show that all three tests perform well in independence testing since empirical Type-I error rates are close to the nominal significance level even in the condition of small sample size.

Simulation 2 In this simulation, we test the dependence between a high-dimensional variable and a categorical random variable when $R = 2$ or 5 with the proportion proposed in simulation 1. In column 1, we generate X and Y representing radial data as follows:

$$\begin{aligned} (1)\quad & Y = \{1, 2\}, (a)\ X = (1, \theta, \phi_1 + \epsilon), \\ & \theta \sim U(-\pi, \pi),\ \phi_1 \sim U(-\pi, \pi), \epsilon = 0, \\ & (b)\ X = (1, \theta, \phi_2 + \epsilon), \\ & \theta \sim U(-\pi, \pi), \phi_2 \sim U(1/5\pi, 4/5\pi), \epsilon \sim t(0, 1). \end{aligned}$$

$$\begin{aligned} (2)\quad & Y = \{1, 2, 3, 4, 5\}, X = (1, \theta, \phi_r + \epsilon), \\ & \phi_r \sim U([-1 + 2(r-1)/5]\pi, (-1 + 2r/5)\pi), \\ & r = 1, 2, 3, 4, 5.\ \ \theta \sim U(-\pi, \pi),\ \epsilon \sim t(0, 1). \end{aligned}$$

For column 2, we consider von Mises-Fisher distribution. Then we set $n = 3$ and $k = 1$, and the simulated data sets are generated as follows:

$$\begin{aligned} &(1) R = 2, \mu = (\mu_1, \mu_2) = (1, 2), \\ &(2) R = 5, \mu = \{\mu_1, \mu_2, \mu_3, \mu_4, \mu_5\} = \{4, 3, 1, 5, 2\}. \end{aligned}$$

Table 1. Empirical Type-I error rates at the significance level 0.05 in Simulation 1

		$X = (1, \theta, \phi)$			$X \sim M_3(\mu, k)$			$X = (X_1, X_2, X_3)$		
R	n	MDD	DC	HHG	MDD	DC	HHG	MDD	DC	HHG
2	40	0.080	0.085	0.070	0.040	0.050	0.030	0.060	0.060	0.030
	60	0.025	0.035	0.030	0.075	0.085	0.065	0.070	0.045	0.045
	80	0.030	0.045	0.065	0.035	0.035	0.030	0.035	0.035	0.045
	120	0.040	0.050	0.050	0.055	0.050	0.055	0.025	0.025	0.025
	160	0.035	0.035	0.055	0.045	0.040	0.050	0.050	0.075	0.065
5	40	0.015	0.025	0.045	0.050	0.045	0.050	0.050	0.060	0.055
	60	0.045	0.025	0.025	0.050	0.030	0.035	0.050	0.050	0.070
	80	0.035	0.030	0.050	0.060	0.070	0.060	0.060	0.066	0.065
	120	0.040	0.040	0.035	0.050	0.050	0.050	0.045	0.030	0.070
	160	0.030	0.065	0.045	0.050	0.060	0.040	0.050	0.035	0.035

Table 2. Empirical powers at the significance level 0.05 in Simulation 2

		$X = (1, \theta, \phi)$			$X \sim M_3(\mu, k)$			$X = (X_1, X_2, X_3)$		
R	n	MDD	DC	HHG	MDD	DC	HHG	MDD	DC	HHG
2	40	0.385	0.240	0.380	0.595	0.575	0.590	0.535	0.685	0.375
	60	0.530	0.385	0.475	0.765	0.745	0.650	0.730	0.810	0.590
	80	0.715	0.535	0.735	0.890	0.880	0.775	0.865	0.930	0.755
	120	0.875	0.740	0.915	0.965	0.955	0.905	0.970	0.990	0.925
	160	0.965	0.845	0.995	1.000	1.000	0.985	0.995	0.995	0.980
5	40	0.925	0.240	0.940	0.410	0.230	0.185	0.840	0.825	0.360
	60	0.995	0.350	0.995	0.720	0.460	0.330	0.965	0.995	0.625
	80	1.000	0.450	1.000	0.860	0.615	0.540	0.995	0.990	0.850
	120	1.000	0.595	1.000	0.990	0.880	0.820	1.000	0.995	0.990
	160	1.000	0.595	1.000	0.995	0.975	0.955	1.000	1.000	1.000

For column 3, X in each dimension was separately generated from normal distribution $N(\mu, 1)$ to represent data in the Euclidean space. There are two choices of R:

$$(1) R = 2, \mu = (\mu_1, \mu_2) = (0, 0.6),$$
$$(2) R = 5, \mu = \{\mu_1, \mu_2, \mu_3, \mu_4, \mu_5\} = \{4, 3, 1, 5, 2\}/3.$$

Table 2 based on 200 simulations shows that the MDD test performs well in most settings with Type-I error rate approximating 1 especially when the sample size exceeds 80. When data contains extreme value, the Type-I error rate of DC deteriorate quickly while the MDD test performs more stable. Moreover, the MDD test performs better than both DC test and HHG test in spherical space, especially when the number of class R increases.

Simulation 3 In this simulation, we set X with different dimensions, with the range of $\{3, 6, 8, 10, 12\}$, to test independence and dependence between a high-dimensional random variable and a categorical random variable. Respectively, I represents independence test, and II represents dependence one. We use empirical type-I error rates for I, empirical powers for II. We let sample size $n = 60$ and classes $R = 2$. Three types of data are shown in Table 3 for three columns.

In column 1:

$$X_{dim} = (1, \theta, \phi_1 + \epsilon, \ldots, \phi_d + \epsilon),$$
$$d = 1, 4, 6, 8, 10, \theta \sim U(-\pi, \pi)$$

For ϕ_i of two classes,

$$(a) \phi_i \sim U(-\pi, \pi), \epsilon = 0,$$

$$(b) \phi_i \sim U(-1/5\pi, 4/5\pi), \epsilon \sim t(0, 1),$$

In column 2,

$$X \sim M_d(\mu, k), d = 3, 6, 8, 10, 12, \mu \in \{0, 2\},$$

where $k = 1$.
In column 3, we set

$$X_i = (x_{i1}, x_{i2}, \ldots, x_{id}), d = 3, 6, 8, 10, 12 \text{ and } x_{id} \sim N(\mu, 1),$$

where $\mu \in \{0, \frac{3}{5}\}$ in dependence test. Table 3 based on 300 simulations at $\alpha = 0.05$ shows that the DC test performs well in normal distribution but is conservative for testing the dependence between a radial spherical vector and a categorical variable. The HHG test works well for extreme value but is conservative when it comes to von Mises-Fisher distribution and normal distribution. It also can be observed that

Table 3. Empirical Type-I error rates / Empirical powers at the significance level 0.05 in Simulation 3

		$X_{dim} = (1, \theta, \phi_d)$			$X \sim M_{dim}(\mu, k)$			$X = (x_1, x_2, \ldots, x_{dim})$		
	dim	MDD	DC	HHG	MDD	DC	HHG	MDD	DC	HHG
I	3	0.047	0.067	0.050	0.047	0.047	0.053	0.047	0.060	0.033
	6	0.053	0.050	0.050	0.050	0.053	0.037	0.047	0.050	0.030
	8	0.037	0.040	0.050	0.037	0.047	0.050	0.040	0.037	0.030
	10	0.050	0.053	0.040	0.050	0.056	0.033	0.037	0.050	0.060
	12	0.020	0.070	0.050	0.053	0.047	0.047	0.040	0.067	0.043
II	3	0.550	0.417	0.530	0.967	0.950	0.923	0.740	0.830	0.577
	6	1.000	0.583	0.997	0.983	0.957	0.927	0.947	0.963	0.773
	8	1.000	0.573	1.000	0.967	0.950	0.920	0.977	0.990	0.907
	10	1.000	0.597	1.000	0.977	0.950	0.887	0.990	0.993	0.913
	12	1.000	0.590	1.000	0.957	0.940	0.843	0.997	1.000	0.963

Table 4. Empirical powers at the significance level 0.05 in Simulation 4

	landmark =20			landmark =50			landmark =70		
corr	MDD	DC	HHG	MDD	DC	HHG	MDD	DC	HHG
0	0.050	0.050	0.063	0.047	0.047	0.050	0.040	0.050	0.050
0.05	0.090	0.080	0.057	0.097	0.073	0.087	0.093	0.067	0.467
0.10	0.393	0.330	0.300	0.347	0.310	0.237	0.350	0.303	0.183
0.15	0.997	0.957	0.997	0.993	0.943	0.970	0.993	0.917	0.967
0.20	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

MDD test performs well in circumstance of high dimension.

Simulation 4 In this simulation, we use the $(\cos(\theta + d/2) + \epsilon/10, \cos(\theta - d/2) + \epsilon/10)$ parametrization of an ellipse where $\theta \in (0, 2\pi)$ to run our experiment, let X be an ellipse shape and Y be a categorical variable. The $\cos(d)$ is the parameter of correlation, which means when $\cos(d) = 0$, the shape is a unit circle and when $\cos(d) = 1$, the shape is a straight line. We set the noise $\epsilon \sim t(2)$ and $R = 2$ where $y_1 = 1$ represents that shape X is a circle with $\cos(d) = 0$ and $y_2 = 2$ represent that shape X is an ellipse with correlation $corr = \cos(d)$. It is intuitive that, when $corr = 0$, the MDD statistics should be zero. In our experiment, we set corr = 0,0.05,0.1,0.15,0.2 and let the number of landmark comes from $\{20, 50, 70\}$. Sample size is set to 60 and R package *shapes* [11] is used to calculate the distance between shapes. Table 4 summarizes the empirical Type-I error based on 300 simulations at $\alpha = 0.05$. It shows that the DC test and HHG test are conservative for testing the dependence between an ellipse shape and a categorical variable while the MDD test works well in different number of landmarks.

4. A REAL-DATA ANALYSIS

4.1 The hippocampus data analysis

Alzheimer's disease (AD) is a disabling neurological disorder that afflicts about 11% of the population over age 65 in United States. It is an aggressive brain disease that cause dementia – a continuous decline in thinking, behavioral and social skill that disrupts a person's ability to function independently. As the disease progresses, a person with Alzheimer's disease will develop severe memory impairment and lost ability to carry out the simplest tasks. There is no treatment that cure Alzheimer's diseases or alter the disease process in the brain.

The hippocampus, a complex brain structure, plays an important role in the consolidation of information from short-term memory to long-term memory. Humans have two hippocampi, each side of the brain. It is a vulnerable structure that gets affected in a variety of neurological and psychiatric disorders [1]. In Alzheimer's disease, the hippocampus is one of the first regions of the brain to suffer damage [12]. The relationship between hippocampus and AD has been well studied for several years, including volume loss of hippocampus [32, 51], pathology in its atrophy [17] and genetic covariates [48]. For instance, shape changes [47, 26, 29] in hippocampus are served as a critical event in the course of AD in recent years.

We consider the radical distances of hippocampal 30000 surface points on the left and right hippocampus surfaces. In geometry, radical distance, denoted r, is a coordinate in polar coordinate systems (r, θ). Basically, it is the scalar Euclidean distance between a point and the origin of the system of coordinates. In our data, the radical distance is the distance between medial core of the hippocampus and the corresponding vertex on the surface. The dataset obtained from the ADNI (The Alzheimer's Disease Neuroimaging Initiative) contains 373 observations (162 MCI individuals transformed to AD and the 212 MCI individuals who are not converted to AD) where Mild Cognitive Impairment

(MCI) is a transitional stage between normal aging and the development of AD [38] and 8 covariates in our dataset. Considering the large dimension of original functional data, we firstly use SVD (Singular value decomposition) to extract top 30 important features that can explains 85% of the total variance.

We first apply the MDD test to detect the significant variables associated with two sides of hippocampus separately at significance level $\alpha = 0.05$. Since 8 hypotheses are simultaneously tested, the Benjamini–Hochberg (BH) correction method is used to control false discover rate at 0.05, which ranks the p-value from the lowest to the highest. The statistics in (2) is used to do dependence test between hippocampus functional data and categorical variables. The categorical variables include Gender (1=Male; 2=Female), Conversion (1=converted, 0=not converted), Handedness (1=Right; 2=Left), Retirement (1=Yes; 0=No). Then, we apply DC test and HHG test for the dataset. Note that the p-value obtained in the three methods all used permutation test with 500 times. Table 5 summarizes the results that the MDD test, compared to other methods, are able to detect the significance on both side of hippocampus, which agrees with the current studies [33, 26, 29] that conversion and age are critical elements to AD disease. Then, we expanded our method to continuous variables, the Age and the ADAS-Cog score, which are both important to diagnosing AD disease [33, 24]. We discretize age and ADAS-Cog score into categorical ones by using the quartile. For instance, the factor level labels of Age are constructed as "$(54, 71], (71, 75], (75, 80], (80, 90]$", labelled as $1, 2, 3, 4$. The result of p-values in Table 5 agrees with the current research.

Next, we step forward to expand our method to genetic variables to further check the efficiency of MDD test. Some genes in hippocampus are found to be critical to cause AD, such as The apolipoprotein E gene (APOE). The three major human alleles (ApoE2, ApoE3, ApoE4) are the by-product of non-synonymous mutations which lead to changes in functionality and are implicated in AD. Among these alleles, ApoE4, named ϵ_4, is accepted as a factor that affect the event of AD [34, 23, 44]. In our second experiment, we test the correlation between ApoE2(ϵ_2), ApoE3(ϵ_3), ApoE4(ϵ_4) and hippocampus. The result in the Table 6 agrees with the idea that ϵ_4 is a significant variable to hippocampus shape. Besides, due to hippocampal asymmetry, the ϵ_4 is more influential to the left side one. The MDD test performs better than DC test and HHG test for both sides of hippocampus.

From the two experiments above, we can conclude that our method can be used in Eu variables, such as age, gender. It can also be useful and even better than other popular methods when it comes to genetic covariates. The correlation between genes and shape data(high dimensional data) is an interesting field that hasn't been well studied so far. There are much work to be done in the future.

Finally, we apply logistic regression to the hippocampus dataset by taking the conversion to AD as the response variable and the gender, age, hippocampus shape as predict variables. The result of regression shows that the age and hippocampus shape are significant and we present the coefficients of hippocampus shape in Figure 1 where a blue color indicates positive regions.

Table 5. The p-values for correlating hippocampi data and covariates after BH correction

	left	right
covariates	MDD DC HHG	MDD DC HHG
Gender	0.032 0.014 0.106	0.248 0.036 0.338
Conversion to AD	0.012 0.006 0.009	0.008 0.006 0.009
Handedness	0.600 0.722 0.457	0.144 0.554 0.045
Retirement	0.373 0.198 0.457	0.648 0.267 0.800
Age	0.012 0.006 0.009	0.009 0.006 0.009
ADAS-Cog Score	0.012 0.006 0.012	0.012 0.006 0.012

Table 6. The p-values for correlating hippocampi data and APOE covariates after BH correction

	left	right
APOE	MDD DC HHG	MDD DC HHG
ϵ_2	0.567 0.488 0.223	0.144 0.696 0.310
ϵ_3	0.357 0.198 0.449	0.157 0.460 0.338
ϵ_4	**0.022** 0.206 0.036	0.072 0.083 0.354

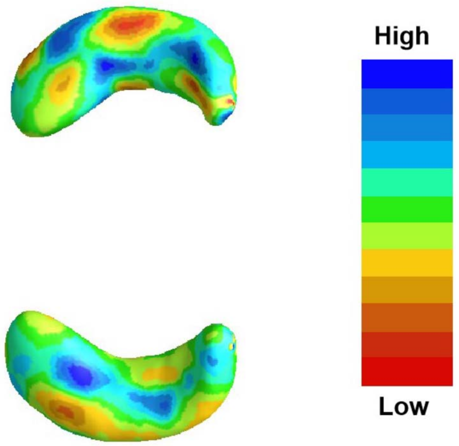

Figure 1. The Hippocampus 3D Shape Surface with coefficients.

4.2 The corpus callosum data analysis

We consider another real data, the corpus callosum (CC), which is the largest white matter structure in the brain. CC has been a structure of high interest in many neuroimaging studies of neuro-developmental pathology. It helps the hemispheres share information, but it also contributes to the spread of seizure impulses from one side of the brain to the other. Recent research [40, 50] has investigated the individual differences in CC and their possible implications regarding interhemispheric connectivity for several years.

Table 7. The p-values for correlating CC contour data and five categorical covariates after BH correction

covariates	MDD	DC	HHG
Gender	0.015	0.018	0.222
Handedness	0.482	0.499	0.461
Marital Status	0.482	0.744	0.773
Retirement	0.482	0.482	0.461
Diagnosis	0.015	0.018	0.045

We consider the CC contour data obtained from the ADNI study to test the dependence between a high-dimensional variable and a random variable. In the ADNI dataset, the segmentation of the T1-weighted MRI and the calculation of the intracranial volume were done by using $FreeSurfer$ package created by [9], whereas the midsagittal CC area was calculated by using CCseg package. The CC data set includes 409 subjects with 223 healthy controls and 186 AD patients at baseline of the ADNI database. Each subject has a CC planar contour Y_j with 50 landmarks and five covariates. We treat the CC planar contour Y_j as a manifold-valued response in the Kendall planar shape space and all covariates in the Euclidean space. The Riemannian shape distance was calculated by R package *shapes* [11].

It is of interest to detect the significant variable associated with CC contour data. We applied the MDD test for dependence between five categorical covariates, gender, handedness, marital status, retirement and diagnosis at the significance level $\alpha = 0.05$. We also applied DC test and HHG test for the CC data. The result is summarized in Table 7. It reveals that the shape of CC planar contour are highly dependent on gender, AD diagnosis. It may indicate that gender and AD diagnosis are significant variables, which agree with [50, 35]. This result also demonstrated that the MMD test performs better to test the significance of variable gender than HHG test.

We plot the mean trajectories of healthy controls (HC) and Alzheimer's disease (AD). The similar process is conducted on the Male and Female. Both of the results are shown in Figure 2. It can be observed that there is an obvious difference of the shape between the AD disease and healthy controls. Compared to healthy controls, the spleen of AD patients seems to be less thinner and the isthmus is more rounded. Moreover, the splenium can be observed that it is thinner in male groups than in female groups. This could be an intuitive evidence to agree with the correlation between gender and the AD disease.

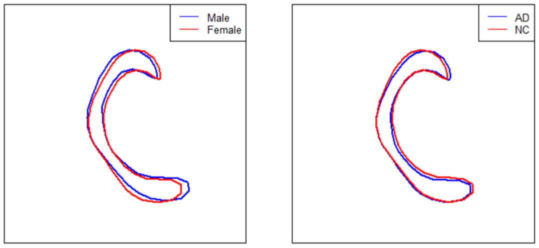

Figure 2. The corpus callosum Data Surface.

5. CONCLUSION

In this paper, we propose the MDD statistics of correlation analysis for Non-Euclidean data in metric spaces and give some conditions for constructing the statistics. Then, we proved the mathematical preliminaries needed in our analysis. The proposed method is robust to outliers or heavy tails of the high dimensional variables. Depending on the results of simulations and real data analysis in hippocampus dataset from the ADNI, we demonstrate its usefulness for detecting correlations between the high dimensional variable and different types of variables (including genetic variables). We also demonstrate its usefulness in another manifold-valued data, CC contour data. We plan to explore our method to variable selection methods and other regression models.

APPENDIX: TECHNICAL DETAILS

Proof of Theorem 1. It is obvious that if X and Y are independent, $F_r(x,x') = F(x,x')$ for $\forall x, x' \in \mathcal{M}$, then $MDD(X \mid Y) = 0$. Next, we need to prove that if $MDD(X \mid Y) = 0$, then X and Y are independent.

According to the definition of MDD, $MDD(X \mid Y) = \sum_{r=1}^{R} p_r \int [F_r(u,v) - F(u,v)]^2 d\nu(u) d\nu(v)$. It is obvious that $MDD(X \mid Y) \geq 0$, so if $MDD(X \mid Y) = 0$, we have $F_r(u,v) = F(u,v)$, $a.s.$ $\nu \otimes \nu$.

Given $Y = y_r$, define ϕ_r is a Borel probability measure of $X \mid Y = y_r$, and we have

$$\phi_r[\bar{B}(u, d(u,v))] := F_r(u,v) = P(X \in \bar{B}(u,v) \mid Y = y_r).$$

Because $(\mathcal{M}, d)$ is a Polish space that d is directionally (ϵ, η, L)-limited and we have $F_r(u,v) = F(u,v)$, $a.s.$ $\nu \otimes \nu$. Next, we can apply Theorem 1 in [49] to get the conclusion that $\nu = \phi_r$ for $r = 1, \ldots, R$.

Therefore, we have $F_r(x,x') = F(x,x')$ for $\forall x, x' \in \mathbb{R}$. That is, for every x, x' and every r, we have

$$P(X \in \bar{B}(x,x') \mid Y = y_r) = P(X \in \bar{B}(x,x')),$$

$i.e.$ X and Y are independent. □

Proof of Corollary 1. (a) We have $\nu(\mathcal{U}) \geq 1 - \epsilon$ and the metric d is directionally (ϵ, η, L)-limited at $\mathcal{U}$, we can obtain the result of independence according to Theorem 1 and Corollary 1(a) in [49].

(b) Similarly, we know that each $\mathcal{M}_l$ is a Polish space satisfying the directionally-limited condition, we also can obtain the result of independence according to Corollary 1(b) in [49]. □

Proof of Theorem 2. Consider

$$\begin{aligned}
&\widehat{MDD}(X|Y)\\
=&\frac{1}{n^2}\sum_{r=1}^{R}\sum_{i=1}^{n}\sum_{j=1}^{n}\hat{p}_r[\hat{F}_r(X_i,X_j)-\hat{F}(X_i,X_j)]^2\\
=&\frac{1}{n^2}\sum_{r=1}^{R}\sum_{i=1}^{n}\sum_{j=1}^{n}\hat{p}_r\left[\frac{1}{n}\sum_{k=1}^{n}\frac{I(X_k\in\bar{B}(X_i,X_j),Y_k=y_r)}{\hat{p}_r}\right.\\
&\left.-\frac{1}{n}\sum_{k=1}^{n}I(X_k\in\bar{B}(X_i,X_j)\right]^2\\
=&\sum_{r=1}^{R}\left[\frac{1}{n^4\hat{p}_r}\sum_{i,j,k,k'=1}^{n}I(X_k\in\bar{B}(X_i,X_j),Y_k=y_r)\right.\\
&I(X_{k'}\in\bar{B}(X_i,X_j),Y_{k'}=y_r)\\
&+\frac{\hat{p}_r}{n^4}\sum_{i,j,k,k'=1}^{n}I(X_k\in\bar{B}(X_i,X_j))I(X_{k'}\in\bar{B}(X_i,X_j))\\
&-\frac{1}{n^4}\sum_{i,j,k,k'=1}[I(X_k\in\bar{B}(X_i,X_j),Y_k=y_r)\\
&I(X_{k'}\in\bar{B}(X_i,X_j))\\
&-\frac{1}{n^4}\sum_{i,j,k,k'=1}I(X_{k'}\in\bar{B}(X_i,X_j),Y_{k'}=y_r)\\
&\left.I(X_k\in\bar{B}(X_i,X_j))\right]\\
=:&\,Q_1+Q_2+Q_3.
\end{aligned}$$

For Q_1, $\frac{1}{n^4}\sum_{i,j,k,k'=1}^{n}I(X_k \in \bar{B}(X_i,X_j),Y_k = y_r)I(X_{k'}\in\bar{B}(X_i,X_j),Y_{k'}=y_r)$ is a V-statistic of order 4. We can verify that

$$\begin{aligned}
&E[I(X_k\in\bar{B}(X_i,X_j),Y_k=y_r)\\
&\times I(X_{k'}\in\bar{B}(X_i,X_j),Y_{k'}=y_r)]\\
=&E[E[I(X_k\in\bar{B}(X_i,X_j),Y_k=y_r)|X_i,X_j]\\
&\times E[I(X_{k'}\in\bar{B}(X_i,X_j),Y_{k'}=y_r)|X_i,X_j]]\\
=&E[E^2[I(X_k\in\bar{B}(X_i,X_j),Y_k=y_r)|X_i,X_j]]\\
=&p_r^2E[\frac{1}{p_r^2}E^2[I(X_k\in\bar{B}(X_i,X_j),Y_k=y_r)|X_i,X_j]]\\
=&p_r^2E[F_r^2(X_i,X_j)].
\end{aligned}$$

Since $E|I(X_k \in \bar{B}(X_i,X_j),Y_k = y_r)I(X_{k'} \in \bar{B}(X_i,X_j),Y_{k'} = y_r)| \leq 1 < \infty$, according to Theorem 3 of [25], $\frac{1}{n^4}\sum_{i,j,k,k'=1}^{n}I(X_k \in \bar{B}(X_i,X_j),Y_k = y_r)I(X_{k'}\in\bar{B}(X_i,X_j),Y_{k'}=y_r)$ almost surely converges to $p_r^2E[F_r^2(X_i,X_j)]$. And we have $p_r \to \hat{p}_r$, and we can draw conclusion that $Q_1 \xrightarrow{a.s.} \sum_{r=1}^{R}p_rE[F_r^2(X_i,X_j)]$, as $n\to\infty$.

Similarly, we also have $Q_2 \xrightarrow{a.s.} \sum_{r=1}^{R}p_rE[F^2(X_i,X_j)]$ and $Q_3 \xrightarrow{a.s.} -2\sum_{r=1}^{R}p_rE[F_r(X_i,X_j)F(X_i,X_j)]$, as $n\to\infty$.

Because $MDD(X|Y) = E[\sum_{r=1}^{R}p_r(F_r(X_i,X_j) - F(X_i,X_j))^2]$, we have

$$\widehat{MDD}(X|Y)\xrightarrow{a.s.}MDD(X|Y),\ n\to\infty. \qquad\square$$

Before we prove Theorem 3, we give an lemma and some notations as follows [25].

Lemma A.1. *Let V_n be a V-statistic of order m, where $V_n = n^{-m}\sum_{i_1=1}^{n}\cdots\sum_{i_m=1}^{n}h(X_{i_1},\ldots,X_{i_m})$ and $h(X_{i_1},\ldots,X_{i_m})$ is the kernel of V_n. For all $1\leq i_1\leq\cdots\leq i_m\leq m$, $E[h(X_{i_1},\ldots,X_{i_m})]^2<\infty$. We have the following conclusions.*

(i) If $\zeta_1=\mathrm{Var}(h_1(X_1))>0$, then

$$\sqrt{n}(V_n-E(h(X_1,\ldots,X_m)))\xrightarrow{d}N(0,m^2\zeta_1),$$

where $\zeta_k=\mathrm{Var}(h_k(X_1,\ldots,X_K))$ and

$$\begin{aligned}
h_k(X_1,\ldots,X_K)=&E[h(X_1,\ldots,X_m)|X_1=x_1,\ldots,X_k=x_k]\\
=&E[h(x_1,\ldots,x_k,X_{k+1},\ldots,X_m)].
\end{aligned}$$

(ii) If $\zeta_1=0$ but $\zeta_2=\mathrm{Var}(h_2(X_1,X_2))>0$, then $n(V_n-E(h(X_1,\ldots,X_m)))\xrightarrow{d}\frac{m(m-1)}{2}\sum_{j=1}^{\infty}\lambda_j\mathcal{X}_j^2(1)$, where $\mathcal{X}_j^2(1)$'s, $j=1,2,\ldots$, are independently and identically distributed $\mathcal{X}^2$ random variables with 1 degree of freedom and λ_j's meet the condition $\sum_{j=1}^{\infty}\lambda_j^2=\zeta_2$.

Proof of Theorem 3. Denote $Z_i = (X_i,Y_i)$, $Z_j = (X_j,Y_j)$, $Z_k=(X_k,Y_k)$, $Z_{k'}=(X_{k'},\ Y_{k'})$. We consider the statistic with the known parameter p_r

$$\begin{aligned}
I_n=&\frac{1}{n^2}\sum_{r=1}^{R}\sum_{i=1}^{n}\sum_{j=1}^{n}p_r\left[\frac{1}{n}\sum_{k=1}^{n}\frac{I(X_k\in\bar{B}(X_i,X_j),Y_k=y_r)}{p_r}\right.\\
&\left.-\frac{1}{n}\sum_{k=1}^{n}I(X_k\in\bar{B}(X_i,X_j)\right]^2\\
=&\frac{1}{n^4}\sum_{r=1}^{R}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k,k'=1}^{n}\left[\frac{1}{p_r}I(X_k\in\bar{B}(X_i,X_j),Y_k=y_r)\right.\\
&\times I(X_{k'}\in\bar{B}(X_i,X_j),Y_{k'}=y_r)\\
&+p_rI(X_k\in\bar{B}(X_i,X_j))I(X_{k'}\in\bar{B}(X_i,X_j))\\
&-I(X_k\in\bar{B}(X_i,X_j),Y_k=y_r)I(X_{k'}\in\bar{B}(X_i,X_j))\\
&\left.-I(X_{k'}\in\bar{B}(X_i,X_j),Y_{k'}=y_r)I(X_k\in\bar{B}(X_i,X_j))\right].
\end{aligned}$$

Let $V_n^{p_r} = \frac{1}{n^4}\sum_{i,j,k,k'=1}^n \Psi^{(r)}(Z_i, Z_j, Z_k, Z_{k'})$ and

$$\begin{aligned}
&\Psi^{(r)}(Z_i, Z_j, Z_k, Z_{k'})\\
=&\frac{1}{p_r} I(X_k \in \bar{B}(X_i, X_j), Y_k = y_r)\\
&\times I(X_{k'} \in \bar{B}(X_i, X_j), Y_{k'} = y_r)\\
&+ p_r I(X_k \in \bar{B}(X_i, X_j)) I(X_{k'} \in \bar{B}(X_i, X_j))\\
&- I(X_k \in \bar{B}(X_i, X_j), Y_k = y_r) I(X_{k'} \in \bar{B}(X_i, X_j))\\
&- I(X_{k'} \in \bar{B}(X_i, X_j), Y_{k'} = y_r) I(X_k \in \bar{B}(X_i, X_j)),
\end{aligned}$$

then we have

$$\begin{aligned}
I_n =& \sum_{r=1}^R V_n^{(p_r)}\\
=& \sum_{r=1}^R \frac{1}{n^4} \sum_{i,j,k,k'=1}^n \Psi^{(r)}(Z_i, Z_j, Z_k, Z_{k'}).
\end{aligned}$$

We would like to use V statistic to obatain the asymptotic properties of $\widehat{MDD}(X|Y)$. We symmetrize the kernel $\Psi^{(r)}(Z_i, Z_j, Z_k, Z_{k'})$ and denote

$$\begin{aligned}
&\Psi_S^{(r)}(Z_i, Z_j, Z_k, Z_{k'})\\
=&\frac{1}{4!} \sum_{\tau \in \pi(i,j,k,k')} \Psi^{(r)}(Z_{\tau(1)}, Z_{\tau(2)}, Z_{\tau(3)}, Z_{\tau(4)}),
\end{aligned}$$

where $\pi(i,j,k,k')$ are the permutations of $\{i,j,k,k'\}$. Now, the kernel $\Psi_S^{(r)}(Z_i, Z_j, Z_k, Z_{k'})$ is symmetric, and $\frac{1}{n^4}\sum_{i,j,k,k'=1}^n \Psi_S^{(r)}(Z_i, Z_j, Z_k, Z_{k'})$ should be a V-statistic. By using the denotation of Lemma A.1, we should consider $E[\Psi_S^{(r)}(z_i, Z_j, Z_k, Z_{k'})]$, that is, the case where only one random variable fixed its value. And, we have to consider $E[\Psi^{(r)}(z_i, Z_j, Z_k, Z_{k'})]$, $E[\Psi^{(r)}(Z_i, z_j, Z_k, Z_{k'})]$, $E[\Psi^{(r)}(Z_i, Z_j, z_k, Z_{k'})]$ and $E[\Psi^{(r)}(Z_i, Z_j, Z_k, z_{k'})]$.

We consider

$$\begin{aligned}
&E[\Psi^{(r)}(z_i, Z_j, Z_k, Z_{k'})]\\
=&\frac{1}{p_r} E[I(X_k \in \bar{B}(x_i, X_j), Y_k = y_r)\\
&\quad I(X_{k'} \in \bar{B}(x_i, X_j), Y_{k'} = y_r)]\\
&+ p_r E[I(X_k \in \bar{B}(x_i, X_j)) I(X_{k'} \in \bar{B}(x_i, X_j))]\\
&- E[I(X_k \in \bar{B}(x_i, X_j), Y_k = y_r) I(X_{k'} \in \bar{B}(x_i, X_j))]\\
&- E[I(X_{k'} \in \bar{B}(x_i, X_j), Y_{k'} = y_r) I(X_k \in \bar{B}(x_i, X_j))]\\
=&\frac{1}{p_r} P_{j,k,k'}[X_k, X_{k'} \in \bar{B}(x_i, X_j), Y_k = Y_{k'} = y_r)]\\
&+ p_r P_{j,k,k'}[X_k, X_{k'} \in \bar{B}(x_i, X_j)]\\
&- P_{j,k,k'}[X_k, X_{k'} \in \bar{B}(x_i, X_j), Y_k = y_r]\\
&- P_{j,k,k'}[X_k, X_{k'} \in \bar{B}(x_i, X_j), Y_{k'} = y_r],
\end{aligned}$$

where $P_{j,k,k'}$ means the probability of Z_j, Z_k and $Z_{k'}$.

Under the null hypothesis H_0, X and Y are independent. Then we have

$$\begin{aligned}
&\frac{1}{p_r} P_{j,k,k'}[X_k, X_{k'} \in \bar{B}(x_i, X_j), Y_k = Y_{k'} = y_r]\\
=&\frac{1}{p_r} P_{j,k,k'}[X_k, X_{k'} \in \bar{B}(x_i, X_j), Y_k = y_r] P_{k'}[Y_{k'} = y_r]\\
=&P_{j,k,k'}[X_k, X_{k'} \in \bar{B}(x_i, X_j), Y_k = y_r],
\end{aligned}$$

and

$$\begin{aligned}
&p_r P_{j,k,k'}[X_k, X_{k'} \in \bar{B}(x_i, X_j)]\\
=&P_{j,k,k'}[X_k, X_{k'} \in \bar{B}(x_i, X_j), Y_{k'} = y_r],
\end{aligned}$$

Thus, $E[\Psi^{(r)}(z_i, Z_j, Z_k, Z_{k'})] = 0$.

Similarly, we have $E[\Psi^{(r)}(Z_i, z_j, Z_k, Z_{k'})] = 0$, $E[\Psi^{(r)}(Z_i, Z_j, z_k, Z_{k'})] = 0$ and $E[\Psi^{(r)}(Z_i, Z_j, Z_k, z_{k'})] = 0$ because of the independence of X and Y under the null hypothesis H_0.

Next, we consider the case when two random elements are fixed.

$$\begin{aligned}
&E[\Psi^{(r)}(Z_i, Z_j, z_k, z_{k'})]\\
=&\frac{1}{p_r} E[I(x_k \in \bar{B}(X_i, X_j), y_k = y_r)\\
&\quad \times I(x_{k'} \in \bar{B}(X_i, X_j), y_{k'} = y_r)]\\
&+ p_r E[I(x_k \in \bar{B}(X_i, X_j)) I(x_{k'} \in \bar{B}(X_i, X_j))]\\
&- E[I(x_k \in \bar{B}(X_i, X_j), y_k = y_r) I(x_{k'} \in \bar{B}(X_i, X_j))]\\
&- E[I(x_{k'} \in \bar{B}(X_i, X_j), y_{k'} = y_r) I(x_k \in \bar{B}(X_i, X_j))]\\
=&\frac{1}{p_r} P_{i,j}[x_k, x_{k'} \in \bar{B}(X_i, X_j), y_k = y_{k'} = y_r)]\\
&+ p_r P_{i,j}[x_k, x_{k'} \in \bar{B}(X_i, X_j)]\\
&- P_{i,j}[x_k, x_{k'} \in \bar{B}(X_i, X_j), y_k = y_r]\\
&- P_{i,j}[x_k, x_{k'} \in \bar{B}(X_i, X_j), y_{k'} = y_r],
\end{aligned}$$

$E[\Psi^{(r)}(Z_i, Z_j, z_k, z_{k'})]$ is a non-constant function related to $z_k, z_{k'}$. In addition, we know

$$\begin{aligned}
&E[\Psi^{(r)}(Z_i, Z_j, Z_k, Z_{k'})]\\
=&\frac{1}{p_r} E[I(X_k \in \bar{B}(X_i, X_j), Y_k = y_r)\\
&\quad \times I(X_{k'} \in \bar{B}(X_i, X_j), Y_{k'} = y_r)]\\
&+ p_r E[I(X_k \in \bar{B}(X_i, X_j)) I(X_{k'} \in \bar{B}(X_i, X_j))]\\
&- E[I(X_k \in \bar{B}(X_i, X_j), Y_k = y_r) I(X_{k'} \in \bar{B}(X_i, X_j))]\\
&- E[I(X_{k'} \in \bar{B}(X_i, X_j), Y_{k'} = y_r) I(X_k \in \bar{B}(X_i, X_j))]\\
=&\frac{1}{p_r} P_{i,j,k,k'}[X_k, X_{k'} \in \bar{B}(X_i, X_j), Y_k = Y_{k'} = y_r)]\\
&+ p_r P_{i,j,k,k'}[X_k, X_{k'} \in \bar{B}(X_i, X_j)]\\
&- P_{i,j,k,k'}[X_k, X_{k'} \in \bar{B}(X_i, X_j), Y_k = y_r]\\
&- P_{i,j,k,k'}[X_k, X_{k'} \in \bar{B}(X_i, X_j), Y_{k'} = y_r] = 0,
\end{aligned}$$

The last equation is derived from the independence of X

and Y. By Lemma A.1 (*ii*), $V_n^{(p_r)}$ is a limiting χ^2-type V statistic.

Now, we consider $V_n^{(\hat{p}_r)}$. Let $t=(t_1,t_2)$. In showing the conditions of Theorem 2.16 in [10] hold, we use

$$h(z_1,z_2;p_r)=p_r\int g(z_1,t;p_r)g(z_2,t;p_r)dM(t),$$

where $g(z,t;\gamma)=\sqrt{p_r}(I(z\in\bar{B}(t_1,t_2),y_1=y_r)/\gamma-I(z\in\bar{B}(t_1,t_2)))$ and $M(t)=\nu\otimes\nu$ is the product measure of ν with respect to X.

Thus,

$$\begin{aligned}\mu(t;\gamma)=&Eg(Z,t;\gamma)\\=&\sqrt{\gamma}(P(X\in\bar{B}(t_1,t_2))p_r/\gamma-P(X\in\bar{B}(t_1,t_2)))\end{aligned}$$

and

$$\mathbf{d}_1\mu(t;p_r)=p_r^{-\frac{1}{2}}P(X\in\bar{B}(t_1,t_2)).$$

The condition of Theorem 2.16 in [10] can be shown to hold in this case using

$$\begin{aligned}&h_*(Z_1,Z_2)\\=&\int[g(Z_1,t;p_r)+\mathbf{d}_1\mu(t;p_r)(I(Y_1=y_r)-p_r)]\\&[g(Z_2,t;p_r)+\mathbf{d}_1\mu(t;p_r)(I(Y_2=y_r)-p_r)]dM(t).\end{aligned}$$

Let $\{\lambda_i\}$ denote the eigenvalues of the operator A defined by

$$Aq(x)=\int h_*(z_1,z_2)q(y)d\nu(x_2)dP(y_2),$$

then

$$nV_n^{(\hat{p}_r)}\xrightarrow{d}\sum_{j=1}^{\infty}\lambda_j^{(r)}\mathcal{X}_j^2(1),$$

where $\mathcal{X}_j^2(1)'s$, $j=1,2,\ldots$, are independently and identically distributed chi-square distribution with 1 degree of freedom.

Notice that $\widehat{MDD}(X|Y)=\sum_{r=1}^{R}V_n^{(r)}$, according to the independence of the sample and the additivity of chi-square distribution,

$$n\widehat{MDD}(X|Y)\xrightarrow{d}\sum_{j=1}^{\infty}\lambda_j\mathcal{X}_j^2(1),$$

where $\lambda_j=\sum_{r=1}^{R}\lambda_j^{(r)}$. □

***Proof of Theorem 4*.** Under the alternative hypothesis H_1, $\widehat{MDD}(X|Y)\xrightarrow{a.s.}MDD(X|Y)>0$, as $n\to\infty$. Thus, we have $n\widehat{MDD}(X|Y)\xrightarrow{a.s.}\infty$, as $n\to\infty$. □

***Proof of Theorem 5*.** We consider

$$\widehat{MDD}(X|Y)=I_n+J_n,$$

where

$$\begin{aligned}I_n=&\sum_{r=1}^{R}\frac{1}{n^4}\sum_{i,j,k,k'=1}^{n}\frac{1}{p_r}I(X_k\in\bar{B}(X_i,X_j),Y_k=y_r)\\&\times I(X_{k'}\in\bar{B}(X_i,X_j),Y_{k'}=y_r)\\&+\sum_{r=1}^{R}\frac{1}{n^4}\sum_{i,j,k,k'=1}^{n}p_rI(X_k\in\bar{B}(X_i,X_j))I(X_{k'}\in\bar{B}(X_i,X_j))\\&-\sum_{r=1}^{R}\frac{1}{n^4}\sum_{i,j,k,k'=1}^{n}I(X_k\in\bar{B}(X_i,X_j),Y_k=y_r)\\&\times I(X_{k'}\in\bar{B}(X_i,X_j))\\&-\sum_{r=1}^{R}\frac{1}{n^4}\sum_{i,j,k,k'=1}^{n}I(X_{k'}\in\bar{B}(X_i,X_j),Y_{k'}=y_r)\\&\times I(X_k\in\bar{B}(X_i,X_j))\end{aligned}$$

and

$$\begin{aligned}J_n=&\sum_{r=1}^{R}(\frac{1}{\hat{p}_r}-\frac{1}{p_r})\frac{1}{n^4}\sum_{i,j,k,k'=1}^{n}I(X_k\in\bar{B}(X_i,X_j),Y_k=y_r)\\&\times I(X_{k'}\in\bar{B}(X_i,X_j),Y_{k'}=y_r)\\&+\sum_{r=1}^{R}(\hat{p}_r-p_r)\frac{1}{n^4}\sum_{i,j,k,k'=1}^{n}I(X_k\in\bar{B}(X_i,X_j))\\&\times I(X_{k'}\in\bar{B}(X_i,X_j)).\end{aligned}$$

We consider $E[\Psi^{(r)}(z_i,Z_j,Z_k,Z_{k'})]$ in the proof of Theorem 3,

$$\begin{aligned}&E[\Psi^{(r)}(z_i,Z_j,Z_k,Z_{k'})]\\=&\frac{1}{p_r}P_{j,k,k'}[X_k,X_{k'}\in\bar{B}(x_i,X_j),Y_k=Y_{k'}=y_r)]\\&+p_rP_{j,k,k'}[X_k,X_{k'}\in\bar{B}(x_i,X_j)]\\&-P_{j,k,k'}[X_k,X_{k'}\in\bar{B}(x_i,X_j),Y_k=y_r]\\&-P_{j,k,k'}[X_k,X_{k'}\in\bar{B}(x_i,X_j),Y_{k'}=y_r].\end{aligned}$$

Under the alternative hypothesis H_1, X and Y is not independent of each other, i.e.

$$\begin{aligned}&\frac{1}{p_r}P_{j,k,k'}[X_k,X_{k'}\in\bar{B}(x_i,X_j),Y_k=Y_{k'}=y_r]\\\neq&P_{j,k,k'}[X_k,X_{k'}\in\bar{B}(x_i,X_j),Y_k=y_r],\end{aligned}$$

and

$$\begin{aligned}&p_rP_{j,k,k'}[X_k,X_{k'}\in\bar{B}(x_i,X_j)]\\\neq&P_{j,k,k'}[X_k,X_{k'}\in\bar{B}(x_i,X_j),Y_{k'}=y_r],\end{aligned}$$

so $E[\Psi^{(r)}(z_i, Z_j, Z_k, Z_{k'})]$ is a non-constant function related to z_i, and we know $h_1^{(r)}(Z_1) = \frac{1}{4}[E[\Psi^{(r)}(z_i, Z_j, Z_k, Z_{k'})] + E[\Psi^{(r)}(Z_i, z_j, Z_k, Z_{k'})] + E[\Psi^{(r)}(Z_i, Z_j, z_k, Z_{k'})] + E[\Psi^{(r)}(Z_i, Z_j, Z_k, z_{k'})]]$, then we have $\mathrm{Var}[h_1^{(r)}(Z_1)] > 0$. We apply Lemma A.1 (i), we have

$$\sqrt{n}[V_n^{(p_r)} - E[\Psi^{(r)}(Z_i, Z_j, Z_k, Z_{k'})]] \xrightarrow{d} N(0, 16\mathrm{Var}[h_1^{(r)}(Z_1)]),$$

Because $I_n = \sum_{r=1}^R V_n^{(p_r)}$, $MDD(X|Y) = E[\sum_{r=1}^R p_r(F_r(X_i, X_j) - F(X_i, X_j))^2]$ and

$$\begin{aligned}
&E[\Psi^{(r)}(Z_i, Z_j, Z_k, Z_{k'})]\\
=&E[E[\Psi^{(r)}(Z_i, Z_j, Z_k, Z_{k'})|Z_i, Z_j]]\\
=&\frac{1}{p_r}E[E[I(X_k \in \bar{B}(X_i, X_j), Y_k = y_r)|Z_i, Z_j]\\
&\quad E[I(X_{k'} \in \bar{B}(X_i, X_j), Y_{k'} = y_r)|Z_i, Z_j]]\\
&+ p_r E[E[I(X_k \in \bar{B}(X_i, X_j))|Z_i, Z_j]\\
&\quad E[I(X_{k'} \in \bar{B}(X_i, X_j))|Z_i, Z_j]]\\
&- E[E[I(X_k \in \bar{B}(X_i, X_j), Y_k = y_r)|Z_i, Z_j]\\
&\quad E[I(X_{k'} \in \bar{B}(X_i, X_j))|Z_i, Z_j]]\\
&- E[E[I(X_{k'} \in \bar{B}(X_i, X_j), Y_{k'} = y_r)|Z_i, Z_j]\\
&\quad E[I(X_k \in \bar{B}(X_i, X_j))|Z_i, Z_j]]\\
=&\frac{1}{p_r}E[E^2[I(X_k \in \bar{B}(X_i, X_j), Y_k = y_r)|Z_i, Z_j]]\\
&+ p_r E[E^2[I(X_k \in \bar{B}(X_i, X_j))|Z_i, Z_j]]\\
&- 2E[E[I(X_k \in \bar{B}(X_i, X_j), Y_k = y_r)|Z_i, Z_j]\\
&\quad E[I(X_k \in \bar{B}(X_i, X_j))|Z_i, Z_j]]\\
=&p_r E[(\frac{1}{p_r}E[I(X_k \in \bar{B}(X_i, X_j), Y_k = y_r)|Z_i, Z_j]\\
&- E[I(X_k \in \bar{B}(X_i, X_j))|Z_i, Z_j])^2]\\
=&E[p_r(F_r(X_i, X_j) - F(X_i, X_j))^2],
\end{aligned}$$

we have

$$E[\Psi^{(r)}(Z_i, Z_j, Z_k, Z_{k'})]] = \sum_{r=1}^R MDD(X|Y)$$

and

$$\sqrt{n}[I_n - MDD(X|Y)] \xrightarrow{d} N(0, \sigma_I^2),$$

where $\sigma_I^2 = \sum_{r=1}^R 16\mathrm{Var}[h_1^{(r)}(Z_i)] + 2n\sum_{i<j} Cov(V_n^{(i)}, V_n^{(j)})$. We explain here that $Cov(V_n^{(i)}, V_n^{(j)})$ is the covariance of V-statistic $V_n^{(i)}$ and $V_n^{(j)}$ of order 4, which can be written as $Cov(V_n^{(i)}, V_n^{(j)}) = \frac{1}{n^4}\sum_{c=1}^4 \binom{4}{c}(n-4)^{4-c}\sigma_{c,c}^2$, where $\sigma_{c,c}^2 = Cov(h_c^{(p)}(Z_1, \ldots, Z_c), h_c^{(q)}(Z_1, \ldots, Z_c))$ and $h_c^{(p)}(Z_1, \ldots, Z_c)$ represents $h_c(Z_1, \ldots, Z_c)$ of Lemma A.1 when $r = p$.

Next, we would like to know the asymptotic distribution of $\sqrt{n}[\widehat{MDD}(X|Y) - MDD(X|Y)]$, we need to consider the asymptotic distribution of $\sqrt{n}J_n$ as follows

$$\begin{aligned}
&\sqrt{n}[\widehat{MDD}(X|Y) - MDD(X|Y)]\\
=&\sqrt{n}[I_n + J_n - MDD(X|Y)].
\end{aligned}$$

Then, we consider J_n. Denote $V_1^{(r)} = \frac{1}{n^4}\sum_{i,j,k,k'=1}^n I(X_k \in \bar{B}(X_i, X_j), Y_k = y_r)I(X_{k'} \in \bar{B}(X_i, X_j), Y_{k'} = y_r)$ and $V_2 = \frac{1}{n^4}\sum_{i,j,k,k'=1}^n I(X_k \in \bar{B}(X_i, X_j))I(X_{k'} \in \bar{B}(X_i, X_j))$, we have

$$J_n = \sum_{r=1}^R (\hat{p}_r - p_r)(V_2 - \frac{1}{\hat{p}_r p_r}V_1^{(r)}).$$

The asymptotic distribution of $V_2 - \frac{1}{\hat{p}_r p_r}V_1^{(r)}$ is a constant, because we know $E[V_1^{(r)}] = E[I(X_k \in \bar{B}(X_i, X_j), Y_k = y_r)I(X_{k'} \in \bar{B}(X_i, X_j), Y_{k'} = y_r)] = p_r^2 E[F_r^2(X_i, X_j)]$, $E[V_2] = E[I(X_k \in \bar{B}(X_i, X_j)I(X_{k'} \in \bar{B}(X_i, X_j)] = E[F^2(X_i, X_j)]$ and $\hat{p}_r \to p_r$. Then

$$V_2 - \frac{1}{\hat{p}_r p_r}V_1^{(r)} \to E[F^2(X_i, X_j)] - E[F_r^2(X_i, X_j)].$$

According to the Central Limit Theorem (CLT), for arbitrary $r = 1, \ldots, R$, we have

$$\sqrt{n}(\hat{p}_r - p_r)(V_2 - \frac{1}{\hat{p}_r p_r}V_1^{(r)}) \to N(0, \sigma_r^2),$$

where $\sigma_r^2 = p_r(1 - p_r)(E[F^2(X_i, X_j)] - E[F_r^2(X_i, X_j)])$.

Let $\hat{\mathbf{p}}^{(i)} = (I(Y_i = y_1), I(Y_i = y_2), \ldots, I(Y_i = y_R))^T$, where $\hat{\mathbf{p}}^{(i)}$ is a R-dimensional random variable and $\hat{\mathbf{p}}^{(1)}, \hat{\mathbf{p}}^{(2)}, \ldots, \hat{\mathbf{p}}^{(n)}$ is dependent of each other. Therefore, according to multidimensional CLT, $\sqrt{n}(\frac{1}{n}\sum_{i=1}^n \hat{\mathbf{p}}^{(i)} - E[\hat{\mathbf{p}}^{(i)}]) = (\sqrt{n}(\hat{p}_1 - p_1), \ldots, \sqrt{n}(\hat{p}_R - p_R))^T$ asymptotically obey the R-dimensional normal distribution. In this way we get the condition of the additivity of the normal distribution, then we have

$$\sqrt{n}J_n \to N(0, \sigma_J^2),$$

where $\sigma_J^2 = \sum_{r=1}^R \sigma_r^2 + 2\sum_{i<j} Cov(\hat{p}_i, \hat{p}_j) = \sum_{r=1}^R \sigma_r^2 - \frac{2}{n}\sum_{i<j} p_i p_j = \sum_{r=1}^R p_r(1 - p_r)(E[F^2(X_i, X_j)] - E[F_r^2(X_i, X_j)]) - \frac{2}{n}\sum_{i<j} p_i p_j$.

Similarly, we can use multidimensional CLT to prove $\sqrt{n}J_n$ and $\sqrt{n}[I_n - MV(X|Y)]$ are bivariate normal distribution. Then, we make a conclusion,

$$\sqrt{n}[\widehat{MDD}(X|Y) - MDD(X|Y)] \xrightarrow{d} N(0, \sigma^2),$$

where $\sigma^2 = \sigma_I^2 + \sigma_J^2 + 2nCov(I_n, J_n)\sigma_I\sigma_J$. We explain here that $Cov(I_n, J_n)$ is the covariance of V-statistic I_n of order 4 and J_n of order 1, which can be written as $Cov(I_n, J_n) =$

$Cov(\sum_{r=1}^R V_n^{(r)}, \sum_{r=1}^R (\hat{p}_r - p_r)(V_2 - \frac{1}{\hat{p}_r p_r} V_1^{(r)})) = \sum_{r=1}^R \sum_{r'=1}^R Cov(V_n^{(r)}, (\hat{p}_{r'} - p_{r'})(V_2 - \frac{1}{\hat{p}_{r'} p_{r'}} V_1^{(r)})) = \sum_{r=1}^R \sum_{r'=1}^R \frac{1}{n}\binom{4}{1}(n - 4)^3 Cov(h_1^{(r)}(Z_1), I(Y_k = y_{r'}))(E[F^2(X_i, X_j)] - E[F_r^2(X_i, X_j)]).$ □

ACKNOWLEDGEMENTS

W.-L. Pan was supported by the Science and Technology Development Fund, Macau SAR (Project code: 0002/2019/APD), National Natural Science Foundation of China (12071494), Hundred Talents Program of the Chinese Academy of Sciences and the Science and Technology Program of Guangzhou, China (202002030129,202102080481). P. Dang was supported by the Science and Technology Development Fund, Macau SAR (Project code: 0002/2019/APD), the Science and Technology Development Fund, Macau SAR(File no. 0006/2019/A1). W.-X. Mai was supported by NSFC Grant No. 11901594.

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen Idec Inc.; Bristol-Myers Squibb Company; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd. and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Medpace, Inc.; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Synarc Inc.; and Takeda Pharmaceutical Company. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California. We are grateful to Prof. Hongtu Zhu for generously sharing preprocessed ADNI dataset to us.

Received 13 December 2021

REFERENCES

[1] Anand, K. and Dhikav, V. (2012). Hippocampus in health and disease: An overview. *Annals of Indian Academy of Neurology* **15** 239–46.

[2] Balasundaram, B., Butenko, S. and Hicks, I. V. (2011). Clique Relaxations in Social Network Analysis: The Maximum k-Plex Problem. *Operations Research* **59** 133–142. MR2814224

[3] Benesty, J., Chen, J., Huang, Y. and Cohen, I. (2009). Pearson correlation coefficient. In *Noise reduction in speech processing* 1–4. Springer.

[4] Bookstein, F. L. (1996). Biometrics, biomathematics and the morphometric synthesis. *Bulletin of Mathematical Biology* **58** 313–365. MR1348660

[5] Charlier, B. (2013). Necessary and sufficient condition for the existence of a Fréchet mean on the circle. *ESAIM: Probability and Statistics* **17** 635–649. MR3126155

[6] Şentürk, D. and Müller, H.-G. (2010). Functional varying coefficient models for longitudinal data. *Journal of the American Statistical Association* **105** 1256–1264. MR2752619

[7] Şentürk, D. and Nguyen, D. V. (2011). Varying coefficient models for sparse noise-contaminated longitudinal data. *Statistica Sinica* **21** 1831. MR2896001

[8] Cui, H. and Zhong, W. (2018). A Distribution-Free Test of Independence and Its Application to Variable Selection. *arXiv: Methodology*.

[9] Dale, A. M., Fischl, B. and Sereno, M. I. (1999). Cortical Surface-Based Analysis: I. Segmentation and Surface Reconstruction. *NeuroImage* **9** 179–194.

[10] de Wet, T. and Randles, R. H. (1987). On the Effect of Substituting Parameter Estimators in Limiting $\chi^2 U$ and V Statistics. *The Annals of Statistics* **15** 398–412. MR0885745

[11] Dryden, I. L. (2019). shapes: Statistical Shape Analysis R package version 1.2.5.

[12] Dubois, B., Hampel, H., Feldman, H., Scheltens, P., Aisen, P., Andrieu, S., Bakardjian, H., Benali, H., Bertram, L., Blennow, K., Broich, K., Cavedo, E., Crutch, S., Dartigues, J.-F., Duyckaerts, C., Epelbaum, S., Frisoni, G., Gauthier, S., Genthon, R. and Jack, C. (2016). Preclinical Alzheimer's disease: Definition, natural history, and diagnostic criteria. *Alzheimer's and Dementia* **12** 292–323.

[13] Federer, H. (2014). *Geometric measure theory*. Springer. MR0257325

[14] Fréchet, M. (1948). Les éléments Aléatoires de Nature Quelconque dans une Espace Distancié. *Ann. Inst. H. Poincaré* **10** 215–310. MR0027464

[15] Fukumizu, K., Bach, F. and Gretton, A. (2005). Consistency of kernel canonical correlation analysis. MR2320675

[16] Gretton, A., Fukumizu, K., Harchaoui, Z. and Sriperumbudur, B. K. (2009). A fast, consistent kernel two-sample test. *Advances in neural information processing systems* **22**.

[17] Halliday, G. (2017). Pathology and hippocampal atrophy in Alzheimer's disease. *The Lancet Neurology* **16** 862–864.

[18] Heller, R., Heller, Y. and Gorfine, M. (2012). A consistent multivariate test of association based on ranks of distances. *Biometrika* **100**. MR3068450

[19] Horváth, L. and Kokoszka, P. (2012). *Inference for functional data with applications* **200**. Springer Science & Business Media. MR2920735

[20] Kaufman, B. B. S., based in part on an earlier implementation by Ruth Heller and Heller, Y. (2019). HHG: Heller-Heller-Gorfine Tests of Independence and Equality of Distributions R package version 2.3.2.

[21] Kendall, D., Barden, D., Carne, T. and Le, H. (1999). Shape and Shape Theory. MR1891212

[22] Kim, H. J., Adluru, N., Bendlin, B. B., Johnson, S. C., Vemuri, B. C. and Singh, V. (2014a). Canonical correlation analysis on riemannian manifolds and its applications. In *European Conference on Computer Vision* 251–267. Springer. MR3444347

[23] Kim, Y., Cho, H., Kim, Y., Ki, C.-S., Chung, S., Ye, B. S., Kim, J.-H., Kim, S. T., Lee, K., Jeon, S., Lee, J., Chin, J., Kim, J., Na, D., Seong, J.-K. and Seo, S. (2014b). Apolipoprotein E4 Affects Topographical Changes in Hippocampal and Cortical Atrophy in Alzheimer's Disease Dementia: A Five-Year Lon-

gitudinal Study. *Journal of Alzheimer's disease: JAD* **44**.

[24] Kong, D., Ibrahim, J., Lee, E. and Zhu, H. (2017). FLCRM: Functional linear cox regression model: Functional Linear Cox Regression Model. *Biometrics* **74**. MR3777931

[25] Lee, A. J. (2019). *U-statistics: Theory and Practice*. Routledge.

[26] Li, S., Shi, F., Pu, F., Li, X., Jiang, T., Xie, S. and Wang, Y. (2007). Hippocampal Shape Analysis of Alzheimer Disease Based on Machine Learning Methods. *AJNR. American journal of neuroradiology* **28** 1339–45.

[27] Lin, L., St. Thomas, B., Zhu, H. and Dunson, D. B. (2017). Extrinsic local regression on manifold-valued data. *Journal of the American Statistical Association* **112** 1261–1273. MR3735375

[28] Lyons, R. (2013). Distance covariance in metric spaces. *The Annals of Probability* **41** 3284–3305. MR3127883

[29] Manning, E., Macdonald, K., Leung, K., Young, J., Pepple, T., Lehmann, M., Zuluaga, M., Cardoso, M. J., Schott, J., Ourselin, S., Crutch, S., Fox, N. and Barnes, J. (2015). Differential hippocampal shapes in posterior cortical atrophy patients: A comparison with control and typical AD subjects. *Human Brain Mapping* n/a-n/a.

[30] Marron, J. and Alonso, A. (2014). Overview of object oriented data analysis. *Biometrical journal. Biometrische Zeitschrift* **56**. MR3258083

[31] Müller, H.-G. (2016). Peter Hall, functional data analysis and random objects. *The Annals of Statistics* **44** 1867–1887. MR3546436

[32] Nobis, L., Manohar, S., Smith, S., Alfaro-Almagro, F., Jenkinson, M., Mackay, C. and Husain, M. (2019a). Hippocampal volume across age: Nomograms derived from over 19,700 people in UK Biobank. *NeuroImage: Clinical* **23** 101904.

[33] Nobis, L., Manohar, S., Smith, S., Alfaro-Almagro, F., Jenkinson, M., Mackay, C. and Husain, M. (2019b). Hippocampal volume across age: Nomograms derived from over 19,700 people in UK Biobank. *NeuroImage: Clinical* **23** 101904.

[34] 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** e48895.

[35] Pan, W., Wang, X., Wen, C., Styner, M. and Zhu, H. (2017). Conditional Local Distance Correlation for Manifold-Valued Data. 41–52.

[36] Pan, W., Tian, Y., Wang, X. and Zhang, H. (2018). Ball divergence: nonparametric two sample test. *Annals of statistics* **46** 1109. MR3797998

[37] Pan, W., Wang, X., Zhang, H., Zhu, H. and Zhu, J. (2019). Ball covariance: A generic measure of dependence in banach space. *Journal of the American Statistical Association*. MR4078465

[38] Petersen, R. (2004). Mild cognitive impairment as a diagnostic entity. *Journal of internal medicine* **256** 183–94.

[39] Rahman, I. U., Drori, I., Stodden, V. C., Donoho, D. L. and Schröder, P. (2005). Multiscale representations for manifold-valued data. *Multiscale Modeling & Simulation* **4** 1201–1232. MR2203850

[40] Raz, N., Ghisletta, P., Rodrigue, K. M., Kennedy, K. M. and Lindenberger, U. (2010). Trajectories of brain aging in middle-aged and older adults: regional and individual differences. *NeuroImage* **51** 501–511.

[41] Rizzo, M. and Szekely, G. (2019). energy: E-Statistics: Multivariate Inference via the Energy of Data R package version 1.7-7.

[42] Sedgwick, P. (2014). Spearman's rank correlation coefficient. *Bmj* **349**.

[43] Shi, X., Styner, M., Lieberman, J., Ibrahim, J. G., Lin, W. and Zhu, H. (2009). Intrinsic regression models for manifold-valued data. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* 192–199. Springer.

[44] Shi, J., Lepore, N., Gutman, B., Thompson, P., Baxter, L. and Caselli, R. (2014). Genetic Influence of Apolipoprotein E4 Genotype on Hippocampal Morphometry: An N=725 Surface-Based Alzheimer's Disease Neuroimaging Initiative Study. *Human Brain Mapping* **35**.

[45] Szekely, G., Rizzo, M. and Bakirov, N. (2008). Measuring and Testing Dependence by Correlation of Distances. *The Annals of Statistics* **35**. MR2382665

[46] Szekely, G. and Rizzo, M. (2010). Brownian Distance Covariance. *The Annals of Applied Statistics* **3**. MR2752127

[47] Tang, X., Holland, D., Dale, A., Younes, L. and Miller, M. (2015). The diffeomorphometry of regional shape change rates and its relevance to cognitive deterioration in mild cognitive impairment and Alzheimer's disease: Diffeomorphometry of Regional Shape Change Rates. *Human Brain Mapping* **36**.

[48] Telenti, A., Lippert, C., Chang, P.-C. and DePristo, M. (2018). Deep Learning of Genomic Variation and Regulatory Network Data. *Human molecular genetics* **27**.

[49] Wang, X., Zhu, J., Pan, W., Zhu, J. and Zhang, H. (2021). Nonparametric Statistical Inference via Metric Distribution Function in Metric Spaces. *arXiv preprint arXiv:2107.07317*.

[50] Witelson, S. F. (1989). Hand and sex differences in the isthmus and genu of the human corpus callosum: a postmortem morphological study. *Brain* **112** 799–835.

[51] Zarow, C., Wang, L., Chui, H., Weiner, M. and Csernansky, J. (2011). MRI Shows More Severe Hippocampal Atrophy and Shape Deformation in Hippocampal Sclerosis Than in Alzheimer's Disease. *International journal of Alzheimer's disease* **2011** 483972.

Wenliang Pan
Faculty of Innovation Engineering
Macau University of Science and Technology
Macao
China
Key Laboratory of Systems and Control
Academy of Mathematics and Systems Science
Chinese Academy of Sciences
Beijing 100190
China
E-mail address: panwliang@amss.ac.cn

Yujue Li
School of Mathematics
Sun Yat-sen University
Guangzhou
China
E-mail address: liyj255@mail2.sysu.edu.cn

Jianwu Liu
School of Mathematics
Sun Yat-sen University
Guangzhou
China
E-mail address: liujw86@mail2.sysu.edu.cn

Pei Dang
Faculty of Innovation Engineering
Macau University of Science and Technology
Macao
China
E-mail address: peidang@must.edu.mo

Weixiong Mai
Macao Center for Mathematical Sciences
Macau University of Science and Technology
Macao
China
E-mail address: wxmai@must.edu.mo