Electronic Journal of Statistics
Vol. 18 (2024) 145–166
ISSN: 1935-7524
https://doi.org/10.1214/23-EJS2203

Manifold energy two-sample test*

Lynna Chu[1] **and Xiongtao Dai**[2]

[1]*Department of Statistics, Iowa State University, Ames, Iowa 50010, U.S.A.*
e-mail: lchu@iastate.edu

[2]*Division of Biostatistics, University of California, Berkeley, Berkeley, CA 94720, U.S.A.*
e-mail: xdai@berkeley.edu

Abstract: We consider the problem of two-sample testing for data generated under the manifold setting, namely where potentially high-dimensional observations are made for underlying objects concentrated near a low-dimensional manifold. Existing two-sample tests typically suffer from a loss of power under high-dimensionality; under the manifold setting, these tests largely ignore the underlying geometric structure of the data, resulting in misleading representations of similarity. Instead, we avoid these issues and propose a non-parametric two-sample test for general data objects which takes into account the intrinsic geometry of the data. A data-driven metric is utilized to characterize the distance between points while respecting the manifold structure. The test statistic behaves like a distance metric between distributions and is shown to be consistent against all alternatives where the two distributions have a positive energy distance. Empirical studies and data analysis of speech recordings demonstrate the test's superior performance for manifold data.

Keywords and phrases: Geodesic distance, high-dimensional data, manifold data, permutation test.

Received August 2022.

Contents

*Authors are listed alphabetically and contributed equally to this work.

1. Introduction

Two-sample tests for detecting a difference in distributions are a classic problem in statistics and appear across a wide-range of applications. Consider that we have two samples $\{X_1, \ldots, X_n\}$ and $\{Y_1, \ldots, Y_m\}$ that are independently and identically distributed from unknown distribution F_X and F_Y, respectively. The two-sample problem aims to distinguish the hypothesis $H_0 : F_X = F_Y$ from an omnibus alternative $H_1 : F_X \neq F_Y$. Advancements in data collection technology have produced data sets of increasingly complexity such that observations may be high-dimensional data objects, making it intractable to express or estimate F_X and F_Y directly due to the curse of dimensionality. Omnibus tests perform well for low-dimensional data but may incur loss of power [29] or become completely powerless [41] for high-dimensional data. These high-dimensional and non-standard data types can pose immense challenges from a practical and theoretical perspective for two-sample testing.

In practice, imaging, video, and audio data have exceedingly high dimensionality, yet humans and many learning techniques are able to distinguish classes among these types of objects with exceptional accuracy. This phenomenon is explained by the manifold hypothesis, which states that data may appear to reside in a high-dimensional space with a large number p of features, but in truth these observation live on or near a low-dimensional manifold with intrinsic dimensionality $d \ll p$ [17]. Data satisfying the manifold hypothesis commonly possess geometric structures that impose constraints on the data, resulting in inherent low-dimensionality. The intrinsic dimension is of interest in many manifold learning problems [21, 5, 12]. Other works have been devoted to learning a low-dimensional representation of the manifold, or more generally the underlying structure [37, 10, 9, 34]. In the context of hypothesis testing, non-parametric two-sample tests for non-Euclidean data designed to exploit the manifold feature are largely unexplored.

Existing non-parametric two-sample tests applicable to non-Euclidean and high-dimensional data are commonly equipped with a similarity measure, usually Euclidean distance. Notable non-parametric contributions include test statistics based on inter-point distances [35, 2, 30, 23], kernel maximum mean discrepancy (MMD) tests [19, 18], and graph-based two-sample tests [15, 31, 7, 6]. In particular, tests based on energy distance and MMD have been extensively studied. For example, [33] provided a framework establishing the equivalence between the (generalized) energy distance and the MMD. In the high-dimensional setting, the power of these tests have been found to decrease as dimensionality increases under various scenarios [29, 41]. Under the manifold hypothesis, [8] established a consistency result for the MMD test when the kernel bandwidth is adapted to the intrinsic dimensionality. However, existing tests that apply

a global Euclidean distance ignore any geometric structure in the data, and tests that apply local Euclidean distance [15, 31, 7, 8] respect the neighborhood and smooth manifold structure but not the data similarity measures in term of manifold distance between points.

We propose a non-parametric two-sample test for general data types lying in arbitrary dimensions that takes advantage of the intrinsic geometry of the data. Conscious of the manifold hypothesis, the proposed manifold energy test utilizes a data-driven metric [ISOMAP, 37] to tell points and distributions apart on the underlying low-dimensional manifold. As a result, the performance of the test is independent of the ambient space where the underlying geometrical data are embedded into, as long as data are intrinsically generated from a fixed manifold. This phenomenon is confirmed by our simulations and theory. Our theory suggests that the estimation error of the proposed test statistic under an unknown manifold structure scales with the intrinsic dimensionality d but not the ambient dimensionality p. This result supposes a noiseless situation and we also discuss the situation when the observations are manifold data convoluted with ambient noise. While the method and theory for estimating the inter-point manifold distance are drawn from ISOMAP [37, 1], our main contribution lies in devising the new hypothesis test that is aware of the manifold distance.

For the test of equality in distribution to be powerful, the energy distance between the two distributions needs to be positive. A commonly considered sufficient condition is that the underlying manifold is of strong negative type [32], which implies, *a fortiori*, the energy test is powerful against any alternative hypothesis. Hilbert spaces and other examples of strong negative type spaces have been found in the literature [25, 26], though it remains an open problem how to characterize spaces of negative type [14]. The permutation test is consistent whenever the energy distance between distributions is positive. The strong negative type condition is only sufficient but not necessary, and our numerical examples generate data on spaces not necessarily of strong negative type (e.g., sphere), yet the energy test demonstrates powerful results. An application of our test to distinguish voice command utterances yield much improved results than omnibus tests unaware of the manifold structure.

A simple motivating example illustrates the need for a test that leverages information from the data's geometric structure even when the ambient dimensionality p is small. Consider data generated uniformly from a swiss roll in $\mathbb{R}^p$ with $p = 3$. The inner spiral (see Figure 1 for an illustration) constitutes Sample 1 and the outer spiral constitutes Sample 2, and to ensure the samples are not well-separated, 30% of observations from Sample 1 are swapped with the same amount of observations from Sample 2 (uniformly randomly selected and reassigned to Sample 2, and vice versa). The number of trials, out of 100, that can successfully reject the null hypothesis at $\alpha = 0.05$ are reported. We examine the performance of two non-parametric tests, namely the two-sample energy statistic [35] and generalized edge-count two-sample test, known as GET [7], both of which are constructed from pairwise Euclidean distance and are applicable to data in arbitrary dimension. We compare this to our proposed test, $\hat{T}_{nm}$, which is constructed from geodesic distances that take into account the

TABLE 1
Number of trials out of 100 that reject H_0 at $\alpha = 0.05$. GET, generalized edge-count test; $\hat{T}_{nm}$, proposed manifold test.

energy test	GET	$\hat{T}_{nm}$
62	48	**80**

intrinsic geometry of the data. Significance is obtained via 1,000 permutations for each test statistic, respectively. It is clear that $\hat{T}_{nm}$ works well for data concentrated near or on a manifold and has improved power compared to the energy statistic and GET.

2. Methods

2.1. Background: the energy test

The energy test [35, 2] provides a general framework to conduct two-sample test for non-Euclidean random objects and multivariate data. Let $\mathcal{M}$ be the space where the observations assume values, and μ and ν be two probability measures supported on $\mathcal{M}$. Let $X_1, \ldots, X_n \sim \mu$ and $Y_1, \ldots, Y_m \sim \nu$ be two independent samples of non-Euclidean objects. The hypothesis being tested is

$$H_0 : \mu = \nu \quad \text{vs} \quad H_A : \mu \neq \nu. \tag{1}$$

Given a choice of metric ρ on $\mathcal{M}$, the energy test targets the population energy distance between the two distributions, defined as

$$D(\mu, \nu) = 2E\rho(X, Y) - E\rho(X, X') - E\rho(Y, Y'), \tag{2}$$

where X, X' and Y, Y' are independent random variables following the law of μ and ν, respectively. Intuitively, the energy distance measures the difference of the between-sample and within-sample difference, and should be small when the two distributions are equal and large when they are different.

The energy statistic [2, 35] is the sample estimate of $D(\mu, \nu)$ and is defined as

$$T_{nm} = \frac{2}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} \rho(X_i, Y_j) - \frac{1}{n^2} \sum_{i,j=1}^{n} \rho(X_i, X_j) - \frac{1}{m^2} \sum_{i,j=1}^{m} \rho(Y_i, Y_j).$$

The energy test has been proven to be a generally powerful test [2, 35, 3] in low dimensions and has been applied in many scientific contexts [11, 28].

In practice, the energy distance between two distinct distributions are usually positive, endowing power to hypothesis tests based on this distance. A sufficient condition to guarantee the positiveness is that $(\mathcal{M}, \rho)$ is a metric space of strong negative type [33, see also Definition 2.1]; in this case, the energy distance is a proper distance between probability measures and equals 0 if μ and ν agree and is positive otherwise. In fact, this condition implies the universal consistency of

the energy test, namely the test is powerful against all alternative hypotheses. Universal consistency hinges on the intuition to hold true that the between-sample distances tend to be larger than the within-sample distances, which is formalized by the notion of negative type spaces [32]. A metric space $(\mathcal{M}, \rho)$ has *negative type* if for all $n \geq 1$, $x_1, \ldots, x_n \in \mathcal{M}$, and real numbers $a_1, \ldots, a_n$ with $\sum_{i=1}^n a_i = 0$, we have $\sum_{i,j} a_i a_j \rho(x_i, x_j) \leq 0$. On a metric space of negative type, any set of n red points x_i and n blue points x_i' satisfies that the sum of distances between different colored points is larger than that within points of the same color, namely

$$2 \sum_{i,j=1}^{n} \rho(x_i, x_j') - \sum_{i,j=1}^{n} \rho(x_i, x_j) - \sum_{i,j=1}^{n} \rho(x_i', x_j') \geq 0.$$

As shown by [32, 24], negative type is equivalent to embeddability into a Hilbert space with a change of the metric, namely $\mathcal{M}$ has negative type if and only if there exists a Hilbert space with norm $\|\cdot\|$ and a map ϕ such that for all $x, x' \in \mathcal{M}$, $\rho(x, x') = \|\phi(x) - \phi(x')\|^2$. Examples for metric spaces of negative type can be found in [27].

The notion of *strong negative type* was first defined in [42] as follows.

Definition 2.1. The metric space $(\mathcal{M}, \rho)$ has *strong negative type* if for any two probability measure μ and ν with finite first moments

$$\iint \rho(x_1, x_2) d\mu(x_1) d\mu(x_2) + \iint \rho(x_1, x_2) d\nu(x_1) d\nu(x_2) - 2 \iint \rho(x_1, x_2) d\mu(x_1) d\nu(x_2) \leq 0$$

and the LHS equals 0 if and only if $\mu = \nu$.

Immediately, on a space of strong negative type the energy distance (2) distinguishes distributions.

Proposition 2.1 (Proposition 3 in [36])**.** *If the metric space $(\mathcal{M}, \rho)$ has strong negative type, then the energy distance (2) is a distance metric between distributions. In particular, $D(\mu, \nu) = 0$ if and only if $\mu = \nu$.*

Spaces of strong negative type have been extensively discussed and applied by [24, 33] in the context of energy statistics. Examples for spaces of strong negative type include separable Hilbert spaces [24], hyperbolic spaces [25], and subsets of a sphere containing at most one pair of antipodal points such as an open hemisphere [26]. If $(\mathcal{M}, \rho)$ has negative type, then $(\mathcal{M}, \rho^r)$ has strong negative type when $0 < r < 1$ [22]. Characterizing Riemannian manifolds that have strong negative type has been an ongoing interest [36] but is out of scope of this work.

2.2. Our setting: data lying on an unknown manifold

Practical evidence supports the manifold hypothesis that high-dimensional data such as audios and images often lie on or close to a low-dimensional manifold.

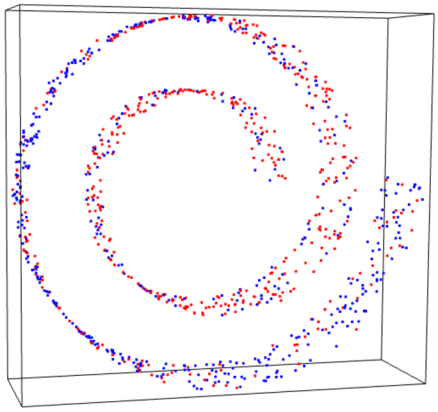

FIG 1. *Observations generated uniformly from a swiss roll in* $\mathbb{R}^3$. *Observations from Sample 1 are represented by red dots and observations from Sample 2 are represented by blue dots.*

This could be because data objects are subject to intrinsic geometrical constraints or are connected to each other via transformations [37]. The extrinsic Euclidean distance fails to represent the inter-relationship between objects. In contrast, the geometry is respected and well-represented by the geodesic distance for measuring inter-point distances. This can be seen in the Swiss roll example illustrated in Figure 1. Here, the distributions μ and ν are induced by sampling from observations on different layers of the swiss roll. Observations from μ consist of data sampled from the inner spiral (Sample 1) while observations from ν are sampled from the outer spiral (Sample 2). To make the setting more challenging, 30% of the observations in each sample are reassigned to the other sample. Explicitly, 30% of the observations from Sample 1 are uniformly randomly selected and re-labelled as Sample 2 and 30% of the observations from Sample 2 are uniformly randomly selected and re-labelled as Sample 1. As illustrated, the overall between- and within-sample distances do not differ much in terms of Euclidean distance. However, if one follows the flow of the data and does not travel across any space where there is no data, it is clear that the between-sample difference is much larger than within-sample difference. This example illustrates that the two samples drastically differ in terms of the intrinsic geodesic distance.

This phenomenon prompts us to study the energy test when the metric is appropriately chosen and estimated in a data-driven fashion depending on the geometry of the dataset. Our setting is that we have available two samples of multivariate observations $\{X_i\}_{i=1}^n$ and $\{Y_j\}_{j=1}^m$ lying on a compact smooth d-dimensional submanifold $\mathcal{M}$ of the Euclidean space $\mathbb{R}^p$, for $p \geq d$. This reflects the situation where p-dimensional features are available for geometrical objects that can be represented using d parameters. The geometry of the data is determined *locally* by the Euclidean distance between the p-dimensional features, while *globally* the inter-point distances are not Euclidean, as illustrated in Figure 1.

In what follows, we consider, in particular, the geodesic distance on $\mathcal{M}$ as the metric ρ, defined for $x, y \in \mathcal{M}$ by

$$\rho(x, y) = \inf_{\gamma} \int \|\gamma'(t)\| \, dt,$$

where the infimum is taken over all piecewise smooth paths $\gamma : [0, 1] \to \mathcal{M}$ that starts at x and ends at y, and $\|\cdot\|$ is the ambient Euclidean norm. The geodesic distance ρ must be estimated since the manifold $\mathcal{M}$ is unknown. We propose to first obtain an estimate $\hat{\rho}(z_1, z_2)$ of the geodesic distance between any two observations z_1, z_2 by the graph distance on a neighborhood graph, adopting the method proposed by [37], as described in Algorithm 1. Denote the pooled random sample as $\mathcal{Z} = \{X_1, \ldots, X_n, Y_1, \ldots, Y_m\}$.

Algorithm 1: Estimate geodesic distance

Data: Observations $\mathcal{Z}$, a pair of points $z_1, z_2 \in \mathcal{Z}$, and neighborhood size r
Result: Geodesic distance estimate $\hat{\rho}(z_1, z_2)$
Build a graph connecting all the pairs $a, b \in \mathcal{Z}$ of observations for which the edges are the segments (a, b) with $\|a - b\| \leq r$.
Set $\hat{\rho}(z_1, z_2)$ to be the length of the shortest path in the graph.

Next, to test the hypothesis (1), we propose the manifold energy statistic

$$\hat{T}_{nm} = \hat{T}(X_1, \ldots, X_n, Y_1, \ldots, Y_m), \tag{3}$$

where for $x_1, \ldots, x_n, y_1, \ldots, y_m \in \mathcal{M}$ define

$$\begin{aligned}&\hat{T}(x_1, \ldots, x_n, y_1, \ldots, y_m) \\ &\quad = \frac{2}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} \hat{\rho}(x_i, y_j) - \frac{1}{n^2} \sum_{i,j=1}^{n} \hat{\rho}(x_i, x_j) - \frac{1}{m^2} \sum_{i,j=1}^{m} \hat{\rho}(y_i, y_j).\end{aligned}$$

Here $\hat{\rho}$ is constructed using Algorithm 1 applied on $\{x_1, \ldots, x_n, y_1, \ldots, y_m\}$.

The null hypothesis is rejected if $\hat{T}_{nm}$ is large.

We propose to approximate the null distribution using the permutation distribution. We applied the permutation null distribution to derive the p-values. Let $\mathcal{Z} = \{Z_1, \ldots, Z_n, Z_{n+1}, \ldots, Z_{n+m}\} = \{X_1, \ldots, X_n, Y_1, \ldots, Y_m\}$ denote the pooled sample. The permutation null distribution is the distribution of $\hat{T}_{\boldsymbol{\pi}} = \hat{T}(Z_{\pi_1}, \ldots, Z_{\pi_{n+m}})$ given $\mathcal{Z}$, where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_{n+m})$ is a random permutation uniformly distributed over all permutations of $\{1, \ldots, n+m\}$. The nominal α-level permutation test rejects H_0 if the permutation p-value

$$p_{\text{perm}} = P(\hat{T}_{\boldsymbol{\pi}} \geq \hat{T}_{nm} \mid \mathcal{Z})$$

is at most α. In practice, the permutation null distribution is approximated by the empirical distribution of $\hat{T}_{\boldsymbol{\pi}_1}, \ldots, \hat{T}_{\boldsymbol{\pi}_B}$ for a large number B of permutations, which is constructed using random permutation $\boldsymbol{\pi}_1, \ldots \boldsymbol{\pi}_B$ uniformly sampled with replacement from the collection of all permutations.

3. Theoretical results

3.1. Notations

Let a_n, b_n be non-zero real sequences, and A_n, B_n be sequences of real-valued random variables, $n = 1, 2, \ldots$. We write $a_n = o(b_n)$ if $\lim_{n\to\infty} a_n/b_n = 0$; $a_n = O(b_n)$ if $\limsup_{n\to\infty} |a_n/b_n| < \infty$; $a_n \asymp b_n$ if $a_n = O(b_n)$ and $b_n = O(a_n)$.

We also denote $A_n = o_P(1)$ if A_n converges in probability to zero; $A_n = O_P(1)$ if A_n is bounded in probability, namely for every $\epsilon > 0$, there exists $M > 0$ such that $\limsup_n P(|A_n| > M) < \epsilon$; $A_n = o_P(b_n)$ if $A_n/b_n = o_P(1)$; $A_n = O_P(b_n)$ if $A_n/b_n = O_P(1)$. We say $A_n = o_P(b_n)$ conditional on B_n in probability if for any $\epsilon > 0$, the conditional probability $P(|A_n/b_n| > \epsilon \mid B_n)$ converges in probability to 0; also, $A_n = O_P(b_n)$ conditionally on B_n in probability if for any $\epsilon_0, \epsilon_1 > 0$, there exists $M = M(\epsilon_0, \epsilon_1) > 0$ such that $\limsup_{n\to\infty} P(P(|A_n/b_n| > M \mid B_n) \geq \epsilon_0) < \epsilon_1$.

3.2. Asymptotic results

The next proposition provides the rate of convergence for the approximation of the energy distance (2) by the version (3), where the geodesic distance are estimated from data. For the theory, we assume data are sampled from the manifold and thus lies exactly on it, and we later discuss the case with ambient noise added. The proof utilizes a result by [1] for the uniform approximation of the geodesic distance by the graph distance; see also [4]. The rate for the approximation depends, as expected, only on the intrinsic dimensionality d but not the ambient dimensionality p. The following conditions for estimating the geodesic distance [1] are needed.

(A1) $\mathcal{M}$ is a d-dimensional compact manifold, $d < p$, isometrically and $\mathcal{C}^2$-smoothly embedded in $\mathbb{R}^p$ without boundary.
(A2) Distributions μ and ν respectively has a continuous density bounded below on $\mathcal{M}$ by a positive constant c_0.
(A3) The neighborhood graph in Algorithm 1 is constructed with $r = r_n = c(\max_{x\in\mathcal{Z}}(\min_{y\neq x,\, y\in\mathcal{Z}} \|x - y\|))^{2/3}$ for some constant $c > 0$.
(A4) $n, m \to \infty$ such that $n/(n+m) \to \lambda \in (0, 1)$.

Proposition 3.1. *Suppose that Conditions (A1)–(A4) hold. Then*

$$\left|\hat{T}_{nm} - T_{nm}\right| = O\left(\left(\frac{\log n}{n}\right)^{2/(3d)}\right) \quad a.s.$$

Proof. We verify the conditions of Corollary 2.2 in [1] which is restated as Theorem A.1 in the Appendix. The manifold condition is verified by (A1), the density condition by (A2) since an observation sampled uniformly from

$\{X_1, \ldots, X_n, Y_1, \ldots, Y_m\}$ has a continuous density uniformly bounded below by c_0, and the neighborhood condition by (A3). By (A4),

$$\max_{x,y\in\mathcal{Z}} |\hat{\rho}(x,y) - \rho(x,y)| = O\left(\left(\frac{\log n}{n}\right)^{2/(3d)}\right) \quad a.s. \tag{4}$$

For $x, y \in \mathcal{Z}$, write $\hat{\delta}(x,y) = \hat{\rho}(x,y) - \rho(x,y)$. Then

$$\begin{aligned}\left|\hat{T}_{nm} - T_{nm}\right| &= \left|\frac{2}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\hat{\delta}(X_i,Y_j) - \frac{1}{n^2}\sum_{i,j=1}^{n}\hat{\delta}(X_i,X_j) - \frac{1}{m^2}\sum_{i,j=1}^{m}\hat{\delta}(Y_i,Y_j)\right| \\ &\le 4\max_{x,y\in\mathcal{Z}}\left|\hat{\delta}(x,y)\right| \\ &= O\left(\left(\frac{\log n}{n}\right)^{2/3d}\right) \quad a.s.,\end{aligned}$$

where the inequality is due to the triangle inequality, and the last equality is due to (4). □

The permutation test satisfies the following properties on size and power.

Theorem 3.1. *Let $C_{nm} = C_{nm}(Z_1, \ldots, Z_{n+m})$ be the α-upper quantile of the permutation distribution $\hat{T}_{\boldsymbol{\pi}}$ given $\mathcal{Z}$.*

(a) Suppose that n and m are large enough so that the exact α-upper quantile C_{nm} of $\hat{T}_{\boldsymbol{\pi}}$ given $\mathcal{Z}$ is well-defined. If H_0 is true, then the size of the test is exactly α. Namely,

$$P(\hat{T}_{nm} \ge C_{nm}) = \alpha.$$

(b) Suppose that Conditions (A1)–(A4) hold. If H_A is true for some μ, ν with $D(\mu,\nu) > 0$, then the test is consistent against this alternative. Namely, under H_A,

$$\lim_{n,m\to\infty} P(\hat{T}_{nm} \ge C_{nm}) = 1.$$

Our main theorem shows that the proposed permutation procedure has an exact size if the null hypothesis is correct, and is otherwise consistent against an alternative hypothesis where $D(\mu,\nu) > 0$, as the sample sizes diverge. The power of the test critically depends on whether the energy distance $D(\mu,\nu) > 0$ under the alternative. A widely applied condition to verify this is that $(\mathcal{M}, \rho)$ is a metric space of has strong negative type. This is, however, a sufficient but not necessary condition for the proposed test to be powerful. The energy distance between populations are often positive even if data lie on a space not of strong negative type, which is the case in the spherical simulation in Section 4. The proof relies on U-statistics theory and an estimation of the size of the permutation test statistic.

Our permutation test can handle noise-contaminated manifold data. Suppose that the observations now take the form of $X_i = X_i' + \epsilon_i$ where the latent $X_i' \sim \mu$ lies on the manifold and $\epsilon_i \sim N(0, \sigma^2 I)$ is ambient Gaussian noise where I is

the $p \times p$ identity matrix (and similar definition for the noise contaminated Y_i in the other sample). Theorem 3.1(a) holds without modification and the size of the test is always exact because of a property of the permutation test. We conjecture that Theorem 3.1(b) still holds if $\hat{\rho}$ converges to a metric ρ' in the ambient space $\mathbb{R}^d$ that makes this space of strong negative type. Relatedly, [8] showed that the kernel MMD test can handle small additive ambient Gaussian noise that shrinks as the bandwidth shrinks. Recently, approaches [16, 13, 40] on denoising manifold data have become available.

Proof. (a) Under H_0 and the given conditions,

$$P(\hat{T}_{\boldsymbol{\pi}} \geq C_{nm} \mid \mathcal{Z}) = \alpha,$$

so unconditionally,

$$P(\hat{T}_{\boldsymbol{\pi}} \geq C_{nm}) = \alpha.$$

Note that $\hat{T}_{\boldsymbol{\pi}}$ has the same marginal distribution as $\hat{T}_{nm}$ under H_0, and $C_{nm}(Z_1, \ldots, Z_{n+m})$ stays the same if we permute its inputs. So $(\hat{T}_{\boldsymbol{\pi}}, C_{nm})$ shares the same distribution as $(\hat{T}_{nm}, C_{nm})$, and

$$P(\hat{T}_{nm} \geq C_{nm}) = \alpha.$$

(b) Under H_A, by Proposition 3.1 and by adapting U-statistics theory to T_{nm} (Theorem 12.3 and 12.6 in [38]), we have that

$$\hat{T}_{nm} = D(\mu, \nu) + o_P(1) \tag{5}$$

as $n, m \to \infty$, where $D(\mu, \nu) > 0$. Write for $x_1, \ldots, x_n, y_1, \ldots, y_m \in \mathcal{M}$,

$$\begin{aligned} &T(x_1, \ldots, x_n, y_1, \ldots, y_m) \\ &\quad = \frac{2}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} \rho(x_i, y_j) - \frac{1}{n^2} \sum_{i,j=1}^{n} \rho(x_i, x_j) - \frac{1}{m^2} \sum_{i,j=1}^{m} \rho(y_i, y_j). \end{aligned}$$

Inspecting the proof of Proposition 3.1, we have

$$\left|\hat{T}_{\boldsymbol{\pi}} - T(Z_{\pi_1}, \ldots, Z_{\pi_{n+m}})\right| \leq \max_{x,y \in \mathcal{Z}} \left|\hat{\rho}(x, y) - \rho(x, y)\right|$$

where the RHS is $O((\log n/n)^{2/(3d)})$ a.s. and is a function of $\mathcal{Z}$. Combine the last display with Lemma A.2 in the Appendix, conditional on $\mathcal{Z}$, we have

$$\hat{T}_{\boldsymbol{\pi}} = O_P\left(\frac{1}{n+m} + \left(\frac{\log n}{n}\right)^{2/(3d)}\right) = o_P(1)$$

in probability and thus $C_{nm} = o_P(1)$ unconditionally. This and (5) imply the desired result. □

TABLE 2
The number of trials (out of 100) to reject H_0 of equal distribution, $\alpha = 0.05$. The underlying manifold varies between scenarios but all has intrinsic dimension $d = 2$ and ambient dimension p.

	S-surface ($\delta = 0.15$)		fish bowl ($\delta = 0.08$)		sphere ($\delta = 0.1$)	
	$p = 3$	$p = 10$	$p = 3$	$p = 10$	$p = 3$	$p = 10$
energy statistic	**81**	38	**63**	44	**56**	16
GET (L_2)	18	7	24	21	5	5
GET (ρ)	18	7	24	21	5	5
$\hat{T}_{nm}$	80	**75**	41	**54**	55	**54**

Theorem 3.1(b) can be strengthened to obtain a rate of convergence for the test being arbitrarily close to $(\log n/n)^{2/3d}$. More precisely,

$$\lim_{n,m\to\infty} P_{nm}(\hat{T}_{nm} \geq C_{nm}) = 1,$$

where P_{nm} denotes the joint distribution of the X_i and the $Y_{i'}$, which are respectively iid observations from μ_n and ν_m such that $\lim_{n,m\to\infty} D(\mu_n, \nu_m)/(\log n/n)^{2/3d} = \infty$. The line of proof is highly similar to that of Theorem 3.1(b). To bound the exception probability along P_{nm}, we apply a Lindeberg CLT for the U-statistics and also extend Proposition 2.1 to an in-probability version (see the proof of Corollary 2.1 in [1]).

4. Simulations

We compare the new manifold energy tests with other commonly used nonparametric two-sample tests that are widely applicable to data in arbitrary dimension/object data and straightforward to implement. We compare our approach to the energy test statistic based on Euclidean distance and the generalized edge-count test statistic (GET). Both are implemented using their respective R packages `energy` and `gTests`. For GET, the similarity graph can be constructed based on a user-selected similarity measure. We carry out the test using Euclidean distance (L_2) and geodesic distance (ρ) and compare the results. As per the recommendation of [7], the similarity graph used is the 5-minimum spanning tree (MST). In each simulation setting, we let $n = m = 1000$ and sampled uniformly from a manifold with ambient noise ($\sigma = 0.3$ for the S-surface, $\sigma = 0.2$ for fish bowl, and $\sigma = 0.5$ for sphere). To make the simulation scenarios more challenging for the tests, only one dimension in Sample 2 is shifted by amount δ. We see in Table 2 the tests based on geodesic distance clearly have stable power for these simple manifolds even as the ambient dimensionality p increases.

To further illustrate the performance of our test statistic, we check its behavior when the ambient dimension p increases even more dramatically, but the intrinsic dimension remains low. Here, we simulate data from a 9-dimensional hypersphere with $n = m = 500$ and ambient gaussian noise ($\sigma = 0.05$). The first dimension of observations from Sample 2 are shifted by amount $\delta = 0.25$. The observations are then embedded into p-dimensional space. The results can be

TABLE 3
The number of trials (out of 100) to reject H_0 of equal distribution, with $\alpha = 0.05$. Both samples are uniformly generated from 9-dimensional hypersphere. Observations from Sample 2 are shifted by amount $\delta = 0.25$. Observations are embedded into a p-dimensional space.

p	10	20	75	100	150
energy statistic	**100**	**100**	29	17	16
GET (L_2)	70	51	57	46	37
GET (ρ)	70	51	58	46	38
$\hat{T}_{nm}$	**100**	**100**	**100**	**99**	**98**

TABLE 4
The proportion of trials (out of 1000) to reject H_0 of equal distribution, with $\alpha = 0.05$. Both samples are uniformly generated from 9-d hypersphere. Observations are embedded into a p-dimensional space.

p	10	20	75	100	150
energy statistic	0.04	0.05	0.04	0.06	0.06
GET (L_2)	0.06	0.06	0.05	0.04	0.04
GET (ρ)	0.06	0.06	0.05	0.04	0.04
$\hat{T}_{nm}$	0.06	0.04	0.05	0.03	0.05

seen in Table 3, which shows that as the ambient dimension grows, the proposed manifold energy test still retains power but other tests have diminished power.

To demonstrate that the manifold energy test is conservative and not simply rejecting random noise generated by the embedding, we simulate data under the same manifold settings as in Table 3 but with no difference between the two populations. Results reported in Table 4 show that the manifold test is near exact and rejects almost as often as the nominal level.

5. Real data application

To illustrate the practical utility of the geodesic two-sample tests, we apply it to the speech commands dataset [39]. The dataset consists of 35 keywords spoken for a one second duration by different contributors, as well as recordings of background noises such as running water and exercise machinery. Since different utterances of the same keyword are highly alike in nature and differ only in terms of pitch, tone, volume, and speed, etc, the audio recording of a keyword can be viewed as data lying on a low-dimension manifold embedded into a high-dimensional space.

A raw utterance is a sound snippet of 1 second sampled at 16 kHz and is denoted as a time series $u(j)$ for $j = 1, \ldots, T$ where $T = 16000$. It is standard to convert these raw audio files to spectrograms and this pre-processing step is widely used in speech recognition [20]. Standard time series techniques would not be applicable here since the audio sample is non-stationary and not easily modeled by a few parameters. For example, the same word can be spoken in different pitches or intonations, which results in drastically different time series. The spectrogram $s_u(\cdot, \cdot)$ of $u(\cdot)$ is a time-frequency representation of the sound obtained as the magnitude of the short-time Fourier transform, defined for time

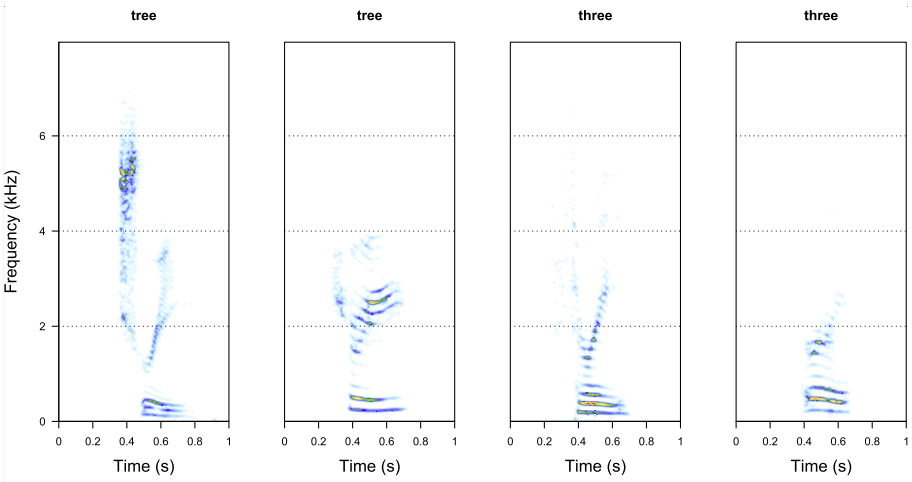

FIG 2. *Spectrograms of two "tree" and two "three" utterances. Blue to red spectral color stands for low to high magnitude.*

$t \in [0,1]$ and frequency $\omega \in [0,8]$ kHz, as

$$s_u(t,\omega) = \left| \sum_{k=-T}^{T} u(Tt-k) W\left(\frac{2k}{M}\right) \exp(-ik2\pi\omega/16) \right|$$

where $W(\cdot)$ was chosen to be the Hann window supported on $[-1,1]$ and $M = 512$ is the window length. It is understood that we set $u(t) = 0$ if $t \leq 0$ or $t > T$. The spectrogram of the input signal was evaluated at 122 time points and 256 frequencies equally spaced in $[0,1] \times [0,8]$, where examples are shown in Figure 2 as 2D surfaces over time and frequency. Each spectrogram is then vectorized such that each observation is treated as a 31232-dimensional vector.

We focus on comparing the distribution of two different spoken words. All the observations from Sample 1 are audio recordings of one word (for example, tree) and the observations from Sample 2 are the audio recordings of a different word (for example, three). Intuitively, the tests should reject H_0 that the two samples are equal in distribution. To improve computational speed, we randomly sample 5% of the available recordings. The following word pairs are considered: forward ($n = 79$) vs. follow ($m = 78$), tree ($n = 88$) vs. three ($m = 186$), backward (83) vs. forward ($n = 78$), and stop ($n = 193$) vs. go ($m = 194$). To make the setting more challenging, we include randomly sampled recordings of background noise. The amount of background noise is chosen so that the tests have moderate power to be comparable. Specifically, each observation is a weighted sum of signal and noise where 40% of the weight is given to the signal (word utterance) and 60% of the weight is given the random Gaussian background noise. In effect, each recording sounds like the word stated with background noise. We repeat this process 100 times and power is estimated to be the number of trials (out of 100)

TABLE 5
Speech commands data with randomly sampled background noise.

	forward vs. follow	tree vs. three	backward vs. forward	stop vs. go
spectral	11	14	10	11
energy test	7	6	8	10
GET (L_2)	43	31	32	45
GET (ρ)	41	34	32	45
$\hat{T}_{nm}$	**64**	**45**	**55**	**58**

to reject the H_0 at the 0.05 significance level. Results can be see in Table 5. We compare our proposed test ($\hat{T}_{nm}$) to the energy statistic and GET. We also compare it to a test statistic that treats each word utterance as a stationary time series and estimates the average spectral density of each sample of words. An L_2 statistic is then calculated of the difference between spectral densities, with a permutation test done to assess significance. It is clear that in this setting, the manifold energy test enjoys substantial power gain compared to the energy statistic based on Euclidean distance. For the graph-based tests, following the recommendations of [7], the similarity graph used is the 5-MST.

6. Discussion

6.1. Generalizing manifold energy test

The concept of manifold energy test can be extended from two closely related perspectives. For example, the metric ρ on the manifold can be chosen to be the diffusion distance [9] which can be estimated from the data. More generally, the energy test can be defined for any general metric after estimating the manifold. Equivalently, by the equivalence of the energy test and kernel MMD test [33], the energy test could be generalized to any semimetric resulting from the equivalent kernel defined on the manifold. Choosing a metric to make the energy distance large, or a kernel to make the MMD large under the alternative is crucial for the power. Again, negative type space/positive kernel is a sufficient but not necessary condition for the test to be powerful against any alternative hypothesis. When this condition cannot be satisfied or verified in practice, test methods need to be developed and validated based on extensive numerical simulations.

6.2. Relationship between our manifold energy test and kernel MMD test

A well-known result by [33] has shown that the energy test and the maximum mean discrepancies (MMD) test are equivalent. Given a nondegenerate kernel k ($z \mapsto k(\cdot, z)$ is injective), a semimetric ρ can be defined which makes $(\mathcal{Z}, \rho)$ of negative type; vise versa, given a metric ρ, a distance-induced kernel k can be defined. The MMD test using the distance-induced kernel k is equivalent to the energy test using the corresponding metric ρ. It is thus natural to wonder

TABLE 6
The number of trials (out of 100) to reject H_0 of equal distribution, with $\alpha = 0.05$. Observations from Sample 2 are shifted by amount δ. Observations are embedded into a p-dimensional space.

	MMD ($\gamma = 0.7$)	MMD ($\gamma = 0.5$)	MMD ($\gamma = 0.1$)	MMD ($\gamma = 0.05$)	$\hat{T}_{nm}$
hypersphere ($\delta = 0.25, p = 50$)	71	80	78	65	100
S-surface ($\delta = 0.15, p = 10$)	39	26	26	25	75
fish bowl ($\delta = 0.08, p = 10$)	5	5	4	4	54
sphere ($\delta = 0.10, p = 10$)	8	8	6	5	54

whether the proposed manifold energy test is equivalent to an MMD test. It was shown in [8] that if data lie on a low-dimensional manifold, applying the kernel MMD test designed for Euclidean data with a local bandwidth is able to adapt to the manifold data and achieve consistency, regardless of the dimensionality of the ambient space.

While this method targets the same scenario as we consider, the two methods are quite different in both the approach and the quantity they target, and the consistency result of neither method implies the other. Our manifold energy test specifically targets the manifold geometry by estimating the geodesic distance, while the kernel MMD test [8] applies kernel designed for Euclidean data and adapt to the manifold case via appropriately tuned bandwidth. The equivalent energy test of the kernel MMD test [8] incorporates a distance metric defined in the ambient space (which could utilize the Euclidean distance between points far away, though assigning very small weight only). It is also interesting to find that the asymptotic limit of our energy statistics is the energy distance $D(\mu, \nu) = \int_{\mathcal{M}} \int_{\mathcal{M}} \rho(x, y)(2p(x)q(y) - p(x)p(y) - q(x)q(y))dV(x)dV(y)$, while the main term of the kernel MMD test approaches the (squared) L^2 distance $\int_{\mathcal{M}}(p(x)-q(x))^2 dV(x)$ (not an energy distance!), where p and q are the densities of μ and ν, respectively, w.r.t. the volume measure V of the manifold. So the two tests are drastically different.

For numerical comparisons, we present in Table 6 the power performance of the kernel MMD test and our proposed test ($\hat{T}_{nm}$). Here, we simulate data under the same settings as Table 2 and Table 3. The first dimension of observations from Sample 2 are shifted by amount δ and p represents the ambient dimension. For the kernel MMD test, we use the Gaussian Radial Basis function (RBF) kernel and report the power of the MMD test over a range of kernel bandwidth parameters γ. The threshold of the MMD test is obtained via 1000 bootstraps.

6.3. Computational speed

To compute geodesic distance, standard algorithms such as Dijkstra's and Floyd-Warshall all-pairs shortest path run in $O(n^3)$ time. For practical sample sizes, we report in Table 7 the average time (in seconds) it takes to implement our geodesic test ($\hat{T}_{nm}$) under the same simulation setup as Table 4. The total sample size is denoted by n and the extrinsic dimensionality of each observation is denoted

TABLE 7
Average computation time (secs) for 100 trials.

		$p = 10$	$p = 20$	$p = 75$	$p = 100$	$p = 150$
$\hat{T}_{nm}$	$n = 1000$	0.796	0.805	0.848	0.879	0.925
	$n = 2000$	3.673	3.716	3.912	3.998	4.161
	$n = 3000$	9.613	9.685	10.103	10.306	10.679
	$n = 10,000$	112.456	113.248	116.826	119.256	126.004
energy Test	$n = 1000$	0.492	0.495	0.514	0.524	0.550
	$n = 2000$	2.391	2.500	2.572	2.614	2.693
	$n = 3000$	6.992	7.013	7.176	7.300	7.261
	$n = 10,000$	76.528	78.276	79.797	80.69	83.25
GET (L_2)	$n = 1000$	3.728	3.764	3.788	3.91	3.806
	$n = 2000$	12.661	13.520	13.695	13.779	13.847
	$n = 3000$	31.933	31.530	31.160	31.092	31.424
	$n = 10,000$	437.786	439.82	440.962	446.508	452.396

by p. The results were run on a Macbook M1 Pro with 32 GB of memory and 8 cores. For comparison, we also report the average time (in seconds) to implement the energy test (based on Euclidean distance) and the graph-based test (GET). For the graph-based test, we construct an MST (minimum spanning tree) using Euclidean distance. For fair comparison, all tests use 1,000 permutations. The `R` packages `energy` and `gTests` were used for the energy test and graph-based test, respectively. In comparison, the results for the geodesic test are not too terrible: For example, for sample size of 10,000 it takes approximately 2 minutes when $p = 150$ while the energy test requires a little under a minute and a half.

Appendix

Technical details

The following is a restatement of Corollary 2.2 in [1] listed here for clarity, which is invoked by our Proposition 3.1.

Theorem A.1 (Corollary 2.2 in [1]). *Let $\mathcal{M} \subset \mathbb{R}^p$, $p \geq 2$, be a d-dimensional compact manifold with $d < p$, of class $\mathcal{C}^2$ without boundary. Suppose that the observations $X_1, \ldots, X_n$ follow probability distribution $\mathbb{P}$ with continuous density w.r.t. the volume measure of $\mathcal{M}$ bounded from below on $\mathcal{M}$ by a positive constant c_0. Then, for any $c > 0$, setting $r = r_n = c(\max_i(\min_{j \neq i} \|X_i - X_j\|))^{2/3}$ in Algorithm 1, we have*

$$\max_{i,j} |\hat{\rho}(X_i, X_j) - \rho(X_i, X_j)| = O\left(\left(\frac{\log n}{n}\right)^{2/3d'}\right) \quad a.s.$$

as $n \to \infty$.

Lemma A.2. *Suppose that the conditions of Theorem 3.1(b) hold. Conditional on $\mathcal{Z}$, as $n, m \to \infty$,*

$$T_{nm}(Z_{\pi_1}, \ldots, Z_{\pi_{n+m}}) = O_P\left(\frac{1}{n+m}\right)$$

in probability.

Proof. Let $N = n + m$ and

$$\phi_{nm}(u, u') = \begin{cases} -\frac{1}{n^2}, & u, u' \le \frac{n}{N+1} \\ -\frac{1}{m^2}, & u, u' > \frac{n}{N+1} \\ \frac{1}{nm}, & \text{otherwise.} \end{cases}$$

Denote $\vec{R} = (R_1, \ldots, R_N)$ as the rank corresponding to $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_N)$ (i.e., $R_i = j$ if $\pi_j = i$ for $i, j = 1, \ldots, N$). Let $D_{kk'} = \rho(Z_k, Z_{k'})$ and $a_{kk'} = \phi_{nm}(k/(N+1), k'/(N+1))$, $k, k' = 1, \ldots, N$. Then

$$\begin{aligned} T_{\boldsymbol{\pi}} &\coloneqq T(Z_{\pi_1}, \ldots, Z_{\pi_N}) = \sum_{k,k'=1}^{N} D_{kk'} a_{R_k R_{k'}} = \sum_{1 \le k \ne k' \le N} D_{kk'} a_{R_k R_{k'}} \\ &= \sum_{1 \le k \ne k' \le N} (D_{kk'} - \bar{D})(a_{R_k R_{k'}} - \bar{a}) + N(N-1)\bar{D}\bar{a}, \end{aligned}$$

where $\bar{D} = (N(N\text{–}1))^{-1} \sum_{1 \le k \ne k' \le N} D_{kk'}$, and $\bar{a} = (N(N\text{–}1))^{-1} \sum_{1 \le k \ne k' \le N} a_{kk'}$, and the second equality is due to $\rho(Z_k, Z_k) = 0$ for $k = 1, \ldots, N$. Taking expected values and variances conditional on $\mathcal{Z}$,

$$E[T_{\boldsymbol{\pi}} \mid \mathcal{Z}] = N(N-1)\bar{D}\bar{a}, \tag{6}$$

and

$$\begin{aligned} \operatorname{var}(T_{\boldsymbol{\pi}} \mid \mathcal{Z}) &= \operatorname{var}\left(\sum_{1 \le k \ne k' \le N} (D_{kk'} - \bar{D})(a_{R_k R_{k'}} - \bar{a}) \right) \\ &= \Bigg(\sum_{\substack{(k,k') \text{ and } (j,j') \text{ share 2 indices} \\ k \ne k' \\ j \ne j'}} + \sum_{\substack{(k,k') \text{ and } (j,j') \text{ share 1 index} \\ k \ne k' \\ j \ne j'}} \\ &\quad + \sum_{\substack{(k,k') \text{ and } (j,j') \text{ share 0 indices} \\ k \ne k' \\ j \ne j'}} \Bigg) \\ &\quad \times \Big((D_{kk'} - \bar{D})(D_{jj'} - \bar{D}) E[(a_{R_k R_{k'}} - \bar{a})(a_{R_j R_{j'}} - \bar{a})] \Big) \\ &= 2E[(a_{R_1 R_2} - \bar{a})^2] \sum_{k=1}^{N} \sum_{\substack{k'=1 \\ k' \ne k}}^{N} (D_{kk'} - \bar{D})^2 \\ &\quad + 4E[(a_{R_1 R_2} - \bar{a})(a_{R_1 R_3} - \bar{a})] \sum_{k=1}^{N} \sum_{\substack{k'=1 \\ k' \ne k}}^{N} \sum_{\substack{j'=1 \\ j' \ne k \\ j' \ne k'}}^{N} (D_{kk'} - \bar{D})(D_{kj'} - \bar{D}) \end{aligned}$$

$$+ E[(a_{R_1R_2}-\bar{a})(a_{R_3R_4}-\bar{a})]\sum_{k=1}^{N}\sum_{\substack{k'=1\\k'\neq k}}^{N}\sum_{\substack{j=1\\j\neq k\\j\neq k'}}^{N}\sum_{\substack{j'=1\\j'\neq k\\j'\neq k'\\j'\neq j}}^{N}(D_{kk'}-\bar{D})(D_{jj'}-\bar{D}), \tag{7}$$

where we utilized the independence of the ranks R_k and $\mathcal{Z}$ and the symmetry of $a_{kk'}$ and $D_{kk'}$ in their two indices. After algebraic computation, we have

$$\begin{aligned}
&E[a_{R_1R_2}] = \bar{a}\\
&\quad = \frac{1}{(n+m)(n+m-1)}\sum_{1\le k\neq k'\le n+m} a_{kk'}\\
&\quad = \frac{1}{mn(m+n-1)} \asymp N^{-3}, \qquad (8)\\
&E[(a_{R_1R_2}-\bar{a})^2]\\
&\quad = \frac{m^2(n-1)-n+mn(1+n)}{m^3n^3(m+n-1)} - \left(\frac{1}{mn(m+n-1)}\right)^2\\
&\quad = \frac{(m+n-1)\left(m^2(n-1)+mn(n+1)-n^2\right)-mn}{m^3n^3(m+n-1)^2} \asymp N^{-4}, \qquad (9)\\
&E[(a_{R_1R_2}-\bar{a})(a_{R_1R_3}-\bar{a})]\\
&\quad = E[(a_{R_1R_2}a_{R_1R_3}-\bar{a}a_{R_1R_2}-\bar{a}a_{R_1R_3}+\bar{a}^2)]\\
&\quad = E[(a_{R_1R_2}a_{R_1R_3}-\bar{a}^2)]\\
&\quad = \frac{m^2(n-2)+2n^2-mn(n+2)}{m^3n^3(m+n-2)(m+n-1)} - \left(\frac{1}{mn(m+n-1)}\right)^2\\
&\quad = -\frac{2m^2\left(n^2+1\right)+m^3(n-2)+mn\left(n^2-4\right)-2(n-1)n^2}{m^3n^3(m+n-2)(m+n-1)^2} \asymp N^{-5}, \qquad (10)\\
&E[(a_{R_1R_2}-\bar{a})(a_{R_3R_4}-\bar{a})]\\
&\quad = E[(a_{R_1R_2}a_{R_3R_4}-\bar{a}^2)]\\
&\quad = \frac{3(m^2(n-2)-2n^2+mn(2+n)}{m^3n^3(m+n-3)(m+n-2)(m+n-1)} - \left(\frac{1}{mn(m+n-1)}\right)^2\\
&\quad = \frac{2\left(m^2\left(2n^2+n+3\right)+m^3(n-3)+mn\left(n^2+n-6\right)-3(n-1)n^2\right)}{m^3n^3(m+n-3)(m+n-2)(m+n-1)^2}\\
&\quad \asymp N^{-6}. \qquad (11)
\end{aligned}$$

Unconditional on $\mathcal{Z}$, by the U-statistics CLT (Theorem 12.3 in [38]) and compactness of $\mathcal{M}$, we have

$$\bar{D} = \frac{1}{N(N-1)}\sum_{k=1}^{N}\sum_{\substack{k'=1\\k'\neq k}}^{N} D_{kk'} = E[D_{12}] + O_P(N^{-1/2}), \tag{12}$$

$$\frac{1}{N(N-1)}\sum_{k=1}^{N}\sum_{\substack{k'=1\\k'\neq k}}^{N}(D_{kk'}-\bar{D})^2 = E[D_{12}^2]-E[D_{12}]^2+O_P(N^{-1/2}), \tag{13}$$

$$\begin{aligned}&\frac{1}{N(N-1)(N-2)}\sum_{k=1}^{N}\sum_{\substack{k'=1\\k'\neq k}}^{N}\sum_{\substack{j'=1\\j'\neq k\\j'\neq k'}}^{N}(D_{kk'}-\bar{D})(D_{kj'}-\bar{D})\\ &\quad = E[D_{12}D_{13}]-E[D_{12}]^2+O_P(N^{-1/2}),\end{aligned} \tag{14}$$

$$\begin{aligned}&\frac{1}{N(N-1)(N-2)(N-3)}\sum_{k=1}^{N}\sum_{\substack{k'=1\\k'\neq k}}^{N}\sum_{\substack{j=1\\j\neq k\\j\neq k'}}^{N}\sum_{\substack{j'=1\\j'\neq k\\j'\neq k'\\j'\neq j}}^{N}(D_{kk'}-\bar{D})(D_{jj'}-\bar{D})\\ &\quad = O_P(N^{-1/2}).\end{aligned} \tag{15}$$

Combining (6), (8), and (12), we have

$$E[T_{\boldsymbol{\pi}} \mid \mathcal{Z}] = O_P(N^{-1}),$$

and combining (7), (9)–(11), and (13)–(15), we have

$$\operatorname{var}(T_{\boldsymbol{\pi}} \mid \mathcal{Z}) = O_P(N^{-2}).$$

By Chebychev's inequality, conditional on $\mathcal{Z}$,

$$T_{\boldsymbol{\pi}} = O_P(N^{-1})$$

in probability. □

References

[1] Aaron, C. and Bodart, O. (2018). Convergence rates for estimators of geodesic distances and Fréchet expectations. *Journal of Applied Probability* **55** 1001–1013. MR3899923

[2] Baringhaus, L. and Franz, C. (2004). On a new multivariate two-sample test. *Journal of Multivariate Analysis* **88** 190–206. https://doi.org/10.1016/S0047-259X(03)00079-4. MR2021870

[3] Baringhaus, L. and Franz, C. (2010). Rigid motion invariant two-sample tests. *Statistica Sinica* 1333–1361. MR2777328

[4] Bernstein, M., De Silva, V., Langford, J. C. and Tenenbaum, J. B. (2000). Graph approximations to geodesics on embedded manifolds Technical Report, Citeseer.

[5] Block, A., Jia, Z., Polyanskiy, Y. and Rakhlin, A. (2021). Intrinsic dimension estimation. *arXiv preprint arXiv:2106.04018*.

[6] Chen, H., Chen, X. and Su, Y. (2018). A weighted edge-count two-sample test for multivariate and object data. *Journal of the American Statistical Association* 1–10. MR3862346

[7] Chen, H. and Friedman, J. H. (2017). A new graph-based two-sample test for multivariate and object data. *Journal of the American Statistical Association* **112** 397–409. MR3646580
[8] Cheng, X. and Xie, Y. (2021). Kernel MMD two-sample tests for manifold data. arXiv:2105.03425 [cs, math, stat].
[9] Coifman, R. R. and Lafon, S. (2006). Diffusion maps. *Applied and Computational Harmonic Analysis* **21** 5–30. MR2238665
[10] Donoho, D. L. and Grimes, C. (2003). Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. *Proceedings of the National Academy of Sciences* **100** 5591–5596. MR1981019
[11] Earnshaw, H. P., Roberts, T. P., Middleton, M. J., Walton, D. J. and Mateos, S. (2019). A new, clean catalogue of extragalactic non-nuclear X-ray sources in nearby galaxies. *Monthly Notices of the Royal Astronomical Society* **483** 5554–5573. https://doi.org/10.1093/mnras/sty3403
[12] Facco, E., d'Errico, M., Rodriguez, A. and Laio, A. (2017). Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific Reports* **7** 1–8.
[13] Fefferman, C., Ivanov, S., Kurylev, Y., Lassas, M. and Narayanan, H. (2018). Fitting a putative manifold to noisy data. In *Conference on Learning Theory* 688–720. PMLR.
[14] Feragen, A. and Hauberg, S. (2016). Open problem: Kernel methods on manifolds and metric spaces. What is the probability of a positive definite geodesic exponential kernel? In *Conference on Learning Theory* 1647–1650. PMLR.
[15] Friedman, J. H. and Rafsky, L. C. (1979). Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests. *The Annals of Statistics* 697–717. MR0532236
[16] Genovese, C. R., Perone-Pacifico, M., Verdinelli, I., Wasserman, L. and others (2012). Manifold estimation and singular deconvolution under Hausdorff loss. *The Annals of Statistics* **40** 941–963. MR2985939
[17] Goodfellow, I., Bengio, Y. and Courville, A. (2016). *Deep Learning*. MIT Press Google-Books-ID: omivDQAAQBAJ. MR3617773
[18] Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B. and Smola, A. (2012). A kernel two-sample test. *The Journal of Machine Learning Research* **13** 723–773. MR2913716
[19] 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**.
[20] Katti, A. and Sumana, M. (2022). Pipeline for pre-processing of audio data. In *IOT with Smart Systems: Proceedings of ICTIS 2022, Volume 2* 191–198. Springer.
[21] Levina, E. and Bickel, P. (2004). Maximum likelihood estimation of intrinsic dimension. In *Advances in Neural Information Processing Systems* **17**. MIT Press.
[22] Li, H. and Weston, A. (2010). Strict p-negative type of a metric space.

Positivity **14** 529–545. MR2680513

[23] Li, J. (2018). Asymptotic normality of interpoint distances for high-dimensional data with applications to the two-sample problem. *Biometrika* **105** 529–546. MR3842883

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

[25] Lyons, R. (2014). Hyperbolic space has strong negative type. *Illinois Journal of Mathematics* **58** 1009–1013. https://doi.org/10.1215/ijm/1446819297. MR3421595

[26] Lyons, R. (2020). Strong negative type in spheres. *Pacific Journal of Mathematics* **307** 383–390. https://doi.org/10.2140/pjm.2020.307.383. MR4149080

[27] Meckes, M. W. (2013). Positive definite metric spaces. *Positivity* **17** 733–757. MR3090690

[28] Naman, S. M., Rosenfeld, J. S., Neuswanger, J. R., Enders, E. C. and Eaton, B. C. (2019). Comparing correlative and bioenergetics-based habitat suitability models for drift-feeding fishes. *Freshwater Biology* **64** 1613–1626. https://doi.org/10.1111/fwb.13358

[29] Ramdas, A., Reddi, S. J., Póczos, B., Singh, A. and Wasserman, L. (2015). On the decreasing power of kernel and distance based nonparametric hypothesis tests in high dimensions. In *Proceedings of the AAAI Conference on Artificial Intelligence* **29**.

[30] Sarkar, S. and Ghosh, A. K. (2018). On some high-dimensional two-sample tests based on averages of inter-point distances. *Stat* **7** e187. MR3816902

[31] Schilling, M. F. (1986). Multivariate two-sample tests based on nearest neighbors. *Journal of the American Statistical Association* **81** 799–806. MR0860514

[32] Schoenberg, I. J. (1938). Metric spaces and positive definite functions. *Transactions of the American Mathematical Society* **44** 522–536. MR1501980

[33] Sejdinovic, D., Sriperumbudur, B., Gretton, A. and Fukumizu, K. (2013). Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *Annals of Statistics* **41** 2263–2291. https://doi.org/10.1214/13-AOS1140. MR3127866

[34] Singer, A. and Wu, H.-T. (2012). Vector diffusion maps and the connection Laplacian. *Communications on Pure and Applied Mathematics* **65** 1067–1144. MR2928092

[35] Székely, G. J. and Rizzo, M. L. (2004). Testing for equal distributions in high dimension. *InterStat* **5** 1249–1272.

[36] Székely, G. J. and Rizzo, M. L. (2017). The energy of data. *Annual Review of Statistics and Its Application* **4** 447–479. https://doi.org/10.1146/annurev-statistics-060116-054026

[37] Tenenbaum, J. B., De Silva, V. and Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science* **290** 2319-2323.

[38] VAN DER VAART, A. W. (2000). *Asymptotic Statistics*. *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge. MR1652247

[39] WARDEN, P. (2018). Speech commands: A dataset for limited-vocabulary speech recognition. *arXiv preprint arXiv:1804.03209*.

[40] YAO, Z. and XIA, Y. (2023). Manifold fitting under unbounded noise. arXiv:1909.10228 [cs, stat]. https://doi.org/10.48550/arXiv.1909.10228

[41] ZHU, C. and SHAO, X. (2021). Interpoint distance based two sample tests in high dimension. *Bernoulli* **27** 1189–1211. MR4255231

[42] ZINGER, A., KAKOSYAN, A. V. and KLEBANOV, L. B. (1992). A characterization of distributions by mean values of statistics and certain probabilistic metrics. *Journal of Soviet Mathematics* **59** 914–920. MR1163396