Two sample test for the equality of distribution using projections

Xiangyu Shi[a], Jiang Du[a,b,*]

[a]*School of Mathematics, Statistics and Mechanics, Beijing University of Technology, Beijing, 100124, China*
[b]*Beijing Institute of Scientific and Engineering Computing, Beijing, 100124, China*

Abstract

In this paper,

Keywords:

1. Introduction

Testing whether samples are identically distributed is a fundamental problem in statistical inference theory and plays an important role in many scientific applications. For example, the difference between two phenotypes is examined in the human microbiome (Zhang and Dao [23]). Comparison of formal white matter tract profiles between healthy individuals and patients with multiple sclerosis (Pomann et al. [12]). Let x and y be two d-dimensional random vectors from two distributions P and Q supported on a common topological space $\mathcal{X}$, respectively. There are two independent samples $X := (x_1, \ldots, x_n)$ drawn from a distribution P and $Y := (y_1, \ldots, y_m)$ drawn from a distribution Q, where $x_i = (x_{i1}, \ldots, x_{id})^\top$ and $y_i = (y_{i1}, \ldots, y_{id})^\top$, n and m are the sample sizes. We consider testing

$$H_0 : P = Q \quad \text{v.s.} \quad H_1 : P \neq Q. \tag{1}$$

When the samples drawn from the normal distribution, the two sample test is simplified to testing mean vector differences or covariance matrix or both. Two classical tests are the student's t test (Student [18]) and Hotellings T^2 test (Hotelling [8]). Other methods include Li and Chen [10]; Cai et al. [3]; Yu et al. [22], etc. However, it is well known that

*Corresponding author
Email address: dujiang84@163.com (Jiang Du)

the first two moments are not sufficient to characterize the distribution. Such methods may no longer be applicable when the normality assumption is violated.

Many scholars have studied the problem of multivariate two-sample test with unknown distribution. Broadly speaking, these can be divided into three categories. The first is the nonparametric test constructed from the concept of a graph. Friedman and Rafsky [4] proposed the minimum spanning tree in Wald and Wolfowitz [21] as a multivariate generalization of univariate sorting tables. Following Schilling [15] and Henze [7], Mondal et al. [11] presented a multivariate two-sample test based on K nearest neighbors applicable to high-dimensional data. Rosenbaum [14] proposed a distribution-free test in finite samples based on cross-matches. More recently, Bhattacharya [1] has developed a unified theoretical framework based on two-sample tests without distribution graphs.

The second type is to construct a nonparametric two-sample test by measuring the distribution, density, or distance between the characteristic functions of x and y. For example, Gretton et al. [6] measured the distance between two distributions by embedding the maximum mean difference (MMD) into a regenerated nuclear Hilbert space. Székely et al. [19] used Euclidean distance to define a class of energy statistics for testing equal distributions. Further, Sejdinovic et al. [16] proved that the energy statistic is a special case of MMD.

The third class is to construct Cramér-von Mises(CvM) type test by random projection method. Zhu et al. [24] proposed the use of projected correlation to measure the dependence between two random vectors. Huang and Huo [9] introduced a two-sample test based on power statistics and random projection. To test for overall homogeneity in a high-dimensional setting, Qiu et al. [13] projected randomly selected subspaces onto a one-dimensional space and constructed nonparametric statistics using the projected CvM distances.

We propose to use random projection to characterize the distance between the distributions of multivariate random vectors x and y. Random projection first projects the multivariate random vector into a series of univariate random variables, and then calculates the MMD between the dichotomous univariate random variables to assess the difference in the distributions of the two samples. The projection measure between the distributions of x and y, denoted by $\Gamma(\mathrm{PMMD}^2)$, is nonnegative and zero if and only if x and y are identi-

cally distributed. The effective time complexity algorithm of single variable is extended to multivariate random vector by random projection method, which greatly reduces the time complexity. Under the null hypothesis, we prove that the test statistics are asymptotically converging to a linear combination of independent chi-square variables, which are related to the process of data generation. Therefore, we use the permutation test to obtain the critical value of the statistic. The sample estimators of the projection measure are n-consistent if x and y are identically distributed, and $\sqrt{n}$-consistent otherwise. Moreover, we simulate the finite sample performance of the statistics.

The rest of this paper is organized as follows. In Section 2, we reviewed the definition of MMD and its fast algorithm in the univariate case. Section 3 presents the random projection based MMD statistics, the detailed algorithm for the statistics and the corresponding critical values. In Section 4, we give some theoretical properties on the population metric and the asymptotic distribution of this estimator. The finite sample performance of the obtained test statistics is verified by the simulation experiments in Section 5. Section 6 gives the data study. For concluding observations, see section 7. All technical details are in the appendix.

2. Review of Maximum Mean Discrepancy: Definition and Fast Algorithm

In this section, we review some related existing works. We recall the concept of Maximum Mean Discrepancy(MMD) in Section 2.1. In Section 2.2, we discuss MMD estimators, and their computation.

2.1. Definition of Maximum Mean Discrepancy

Gretton et al. [6] used the difference between two mean function values as a test statistic to determine whether two random vectors were from different distributions. They called this statistic the MMD. We follow the definition of MMD.

Definition 2.1 (Gretton et al. [6], Definition 2)**.** *Let x and y be two d-dimensional random vectors from two distributions P and Q supported on a common topological space $\mathcal{X}$, respectively. Let $\mathcal{F}$ be a class of functions $f : \mathcal{X} \to \mathbb{R}$. Then, the maximum mean discrepancy*

(MMD) is defined as

$$\mathrm{MMD}(P, Q) := \sup_{f \in \mathcal{F}} (\mathbf{E}_{x \sim P}[f(x)] - \mathbf{E}_{y \sim Q}[f(y)]). \tag{2}$$

$\mathcal{F}$ is typically chosen to be a unit ball in a characteristic RKHS $\mathcal{H}$, defined on the metric space $\mathbb{R}^d$ with associated kernel $k(\cdot, \cdot)$ and feature mapping $\phi(\cdot)$. This is simplified by the fact that in an RKHS, function evaluations can be written $f(x) = \langle \phi(x), f \rangle$, where $\phi(x) = k(x, \cdot)$. Assume $k(\cdot, \cdot)$ is measurable and $\mathbf{E}_{x \sim P} \sqrt{k(x, x)} < \infty$, we denote $\mu_P := \mathbf{E}_{x \sim P(x)}[\phi(x)]$ as the expectation of $\phi(x)$.

Definition 2.2. *Assume the above condition for the existence of the mean embeddings μ_P and μ_Q are satisfied. Then, we have*

$$\mathrm{MMD}(P, Q) = \sup_{\|f\|_{\mathcal{H}} \leq 1} \langle \mu_P - \mu_Q, f \rangle = \|\mu_P - \mu_Q\|_{\mathcal{H}}. \tag{3}$$

The following alternative representation of the squared population MMD is also known (Gretton et al. [6], Lemma 6),

$$\mathrm{MMD}^2(P, Q) = \mathbf{E}_{x,x'}[k(x, x')] - 2\mathbf{E}_{x,y}[k(x, y)] + \mathbf{E}_{y,y'}[k(y, y')], \tag{4}$$

where x' is an independent copy of x with the same distribution, and y' is an independent copy of y.

2.2. Fast Algorithm in the Univariate Cases

Suppose we observe two independent samples $X := (x_1, \ldots, x_n)$ drawn from a distribution P and $Y := (y_1, \ldots, y_m)$ drawn from a distribution Q, where $x_i = (x_{i1}, \ldots, x_{id})^\top$ and $y_i = (y_{i1}, \ldots, y_{id})^\top$, n and m are the sample sizes. It has been proven (Gretton et al.

[5], Gretton et al. [6]) that

$$\begin{aligned}\mathrm{MMD}^2_{n,m}(X,Y) =& \frac{1}{n(n-1)} \sum_{1\le i\neq i'\le n} k(x_i, x_{i'}) + \frac{1}{m(m-1)} \sum_{1\le j\neq j'\le m} k(y_j, y_{j'}) \\ & - \frac{2}{nm} \sum_{i=1}^{n}\sum_{j=1}^{m} k(x_i, y_j)\end{aligned} \tag{5}$$

is an unbiased estimator of $\mathrm{MMD}^2(P,Q)$. Assume $n = m$ and $d = 1$, a fast algorithm has been propose (Bodenham and Kawahara [2]) for the MMD test statistic for the univariate case in $O(n \log n)$ time and $O(n)$ space using the Laplacian kernel.

Lemma 2.1 (Bodenham and Kawahara [2], Proposition 8)**.** *Given two samples* $(x_i, y_i) \in \mathbb{R}\times\mathbb{R}, i = 1,\ldots,n$*, the* euMMD *algorithm computes the exact* $\mathrm{MMD}^2_{n,m}(X,Y)$ *statistic defined in Eq.* (5) *using the Laplacian kernel in* $O(n\log n)$ *time and* $O(n)$ *space.*

Bodenham and Kawahara [2] extended the euMMD to an approximation method for calculating MMD^2 statistics of multivariate data by a random projection method. They introduced an $O(nK\log n)$ complexity algorithm in the multivariate case. However, as far as we know there are no theoretical results, this paper gives an asymptotic theory of test statistics in Section 4 to fill this gap.

3. Numerically Efficient Method for Random Vectors

3.1. Test statistics

Although euMMD is an exact method for computing the MMD^2 statistic of univariate data, this method does not easily extend to an exact compute in multivariate settings. Hence, we use the random projection in Huang and Huo [9] to make observations from $\mathbb{R}^d \to \mathbb{R}$ and then perform univariate tests. Let $\mathcal{S}^{d-1}$ denote the unit sphere in $\mathbb{R}^d$, i.e.

$$\mathcal{S}^{d-1} = \{\beta \in \mathbb{R}^d : \|\beta\|_2 = 1\}.$$

Along every direction $\beta \in \mathcal{S}^{d-1}$, let $u = \beta^\top x$ and $v = \beta^\top y$ be the random projections of x and y, respectively. Then, we construct new sample sets $U = \{u_1,\ldots,u_n\}$ and $V = \{v_1,\ldots,v_m\}$ by randomly projecting all the samples $X = \{x_1,\ldots,x_n\}$ and $Y =$

$\{y_1, \ldots, y_m\}$. The following defines the random projection Maximum Mean Discrepancy (PMMD).

Definition 3.1. *Let $x_1, \ldots, x_n$ and $y_1, \ldots, y_m$ be two independent samples from P and Q, respectively. For every direction $\beta \in \mathcal{S}^{d-1}$, let $u = \beta^\top x$ and $v = \beta^\top y$ are the one-dimensional projections of x and y, respectively. The population PMMD^2 is*

$$\mathrm{PMMD}^2(P, Q; \beta) = \mathbf{E}_{x,x'}[k(u, u')] - 2\mathbf{E}_{x,y}[k(u, v)] + \mathbf{E}_{y,y'}[k(v, v')], \tag{6}$$

where $u' = \beta^\top x'$ and $v' = \beta^\top y'$. x' and y' are i.i.d. copy of x and y, respectively. Using sample mean $\mu_n(U) := \frac{1}{n}\sum_{i=1}^n \phi(u_i)$ and $k(u, u') = \langle \phi(u), \phi(u') \rangle$, an unbiased estimate of PMMD^2 is

$$\begin{aligned}\mathrm{PMMD}^2_{n,m}(X, Y; \beta) =& \frac{1}{n(n-1)} \sum_{1 \le i \ne i' \le n} k(u_i, u_{i'}) + \frac{1}{m(m-1)} \sum_{1 \le j \ne j' \le m} k(v_j, v_{j'}) \\ & - \frac{2}{nm} \sum_{i=1}^n \sum_{j=1}^m k(u_i, v_j).\end{aligned} \tag{7}$$

Assume $n = m$. Let $Z := (z_1, \ldots, z_n)$ be n i.i.d. random variables, where $z_i := (x_i, y_i)$. An unbiased empirical estimate of PMMD^2 is

$$\mathrm{PMMD}^2_n(X, Y; \beta) = \frac{1}{n(n-1)} \sum_{i \ne j}^n h(\beta^\top z_i, \beta^\top z_j), \tag{8}$$

which is a one-sample U-statistic with $h(\beta^\top z_i, \beta^\top z_j) := k(u_i, u_j) + k(v_i, v_j) - k(u_i, v_j) - k(u_j, v_i)$.

To enhance the power of the PMMD test, it is natural to base the tests on a distance from $\mathrm{PMMD}_{n,m}$ to zero, i.e., on a norm $\Gamma(\mathrm{PMMD}_{n,m})$. The most used norms are the Cramér-von Mises (CvM) and Kolmogorov-Smirnov (KS) functionals.

First, we consider the best separating direction, that is, $\beta_{\max} := \arg\max_{\beta \in \mathcal{S}^{d-1}} |\mathrm{PMMD}(\beta)|$, along which the two distributions differ most. To test the null hypothesis, we only substitute $\beta_{\max}$ into (6) to get the oracle test statistic $PKS := \mathrm{PMMD}(\beta_{\max})$. Since the distribution function P is unknown, this is not actually feasible. Hence, we propose the following

feasible projection KS test statistic (PKS) for testing H_0,

$$PKS_{n,m} := \mathrm{PMMD}_{n,m}(X, Y; \beta_{\max}) = \sup_{\beta \in \mathcal{S}^{d-1}} |\mathrm{PMMD}_{n,m}(X, Y; \beta)|, \tag{9}$$

where $\mathrm{PMMD}_{n,m}(X, Y; \beta)$ is defined in Equation (7) (or Equation (8)). The null hypothesis is rejected when the $PKS_{n,m}$ is large enough.

Next, we consider Cramér-von Mises(CvM) function. The choice of the optimal direction depends on the process of data generation. However, the integral function avoids the influence of the data on the projection direction. Then, we define the new PCvM test as

$$PCvM_{n,m} = \int_{\mathbb{S}^{d-1}} \mathrm{PMMD}^2_{n,m}(X, Y; \beta) d\beta, \tag{10}$$

where $d\beta$ is the uniform density on the unit sphere. We reject the null hypothesis for large values of $PCvM_{n,m}$. There is no explicit form of integration in Eq. (10). Therefore, to solve the problem of computational complexity, we average MMD from different projection directions.

More specifically, our estimator can be computed as follows. Consider the multivariate samples $x_1, \ldots, x_n \in \mathbb{R}^d$ and $y_1, \ldots, y_m \in \mathbb{R}^d$, let L be a predetermined number of random projections, we do:

(1) For each $\ell (1 \leq \ell \leq L)$, randomly generate projecting direction β_ℓ from $U(\mathcal{S}^{d-1})$, Where $U(\cdot)$ represents the uniform distribution.

(2) Let $\beta_\ell^\top x$ and $\beta_\ell^\top y$ denote the projections of x and y. That is,

$$u_\ell = (\beta_\ell^\top x_1, \ldots, \beta_\ell^\top x_n) \quad v_\ell = (\beta_\ell^\top y_1, \ldots, \beta_\ell^\top y_m).$$

(3) Using the euMMD algorithm in Bodenham and Kawahara [2], we compute the statistic $\mathrm{PMMD}^{2(\ell)}_{n,m}$ in Eq. (7).

(4) Repeat above steps for L times. The final estimator is

$$\overline{PCvM}_{n,m} = \frac{1}{L}\sum_{\ell=1}^{L} \mathrm{PMMD}^{2(\ell)}_{n,m},$$

which averages across the L projections.

The multivariate efficient approximate MMD using random projections is outlined in Algorithm 1.

Algorithm 1 MEA-MMD-Proj

Input: The observed n-sample X and m-sample Y, number of random projections L.

1: **for** $\ell = 1, \ldots, L$ **do**
2: Randomly generate β_ℓ from $U(\mathcal{S}^{d-1})$.
3: Compute $u_\ell = (\beta_\ell^\top x_1, \ldots, \beta_\ell^\top x_n)$ and $v_\ell = (\beta_\ell^\top y_1, \ldots, \beta_\ell^\top y_n)$.
4: Compute the statistics with euMMD: $\mathrm{PMMD}^{2(\ell)}_{n,m}$.
5: **end for**

Output: $\overline{PCvM}_{n,m} = \frac{1}{L}\sum_{\ell=1}^{L} \mathrm{PMMD}^{2(\ell)}_{n,m}$.

3.2. Critical Values

The randomly projected MMD statistics could be used to test whether two distributions are equal. Theorem 3 yields the asymptotic null distribution of the statistic, converging to a linear combination of independent chi-square variables. The weight in this linear combination is function of the eigenvalue of the operator defined in (12) and the zero expected value of the associated eigenfunction, which cannot be computed since P is unknown. Hence, the critical values of the test statistic can be obtained by the permutation method. Recall that we have multivariate samples $x_1, \ldots, x_n, y_1, \ldots, y_m \in \mathbb{R}^d$ and projecting direction $\beta \in \mathcal{S}^{d-1}$. Set $u = \beta^\top x$ and $v = \beta^\top y$. Let $\boldsymbol{u} = \{u_1, \ldots, u_n, u_{n+1}, \ldots, u_{n+m}\}$ be the pooled sample such that $u_i = u_i, i \in \{1, \ldots n\}$ and $u_{n+i} = v_i, i \in \{1, \ldots m\}$. Let $\pi(\boldsymbol{u})$ be the vector $\boldsymbol{u}$ rearranged according to a permutation π, that is, $\pi(\boldsymbol{u}) = (u_{\pi(1)}, \ldots, u_{\pi(n+m)})$. For a sample $\boldsymbol{u}$, let $u_i^\star = u_{\pi(i)}, i \in \{1, \ldots, n\}$, and $v_j^\star = u_{\pi(n+j)}, j \in \{1, \ldots, m\}$. Let $\mathrm{PMMD}^{2\star}_{n,m}$ denote the permutation version of $\mathrm{PMMD}^2_{n,m}$, that is,

$$\mathrm{PMMD}^{2\star}_{n,m} = \frac{1}{n^2}\sum_{i,j=1}^{n} k(u_i^\star, u_j^\star) - \frac{2}{nm}\sum_{i,j=1}^{n,m} k(u_i^\star, v_j^\star) + \frac{1}{m^2}\sum_{i,j=1}^{m} k(v_i^\star, v_j^\star). \tag{11}$$

The following algorithm describes a two-sample test using the PKS test statistic to generate critical values by permutation.

(1) Calculate $PKS_{n,m,obs}$ for the raw data with Eq. (9).

(2) For each $b(1 \leq b \leq B)$, randomly generate a permutation of observation: let

$$(x^{\star,b}, y^{\star,b}) = (x_1^{\star,b}, \ldots, x_n^{\star,b}, y_1^{\star,b}, \ldots, y_m^{\star,b})$$

be a random permutation of $(x_1, \ldots, x_n, y_1, \ldots, y_m)$.

(3) Let $\beta^\top x^{\star,b}$ and $\beta^\top y^{\star,b}$ denote the projections of x and y. That is,

$$u^{\star,b} = (\beta^\top x_1^{\star,b}, \ldots, \beta^\top x_n^{\star,b}), \quad v^{\star,b} = (\beta^\top y_1^{\star,b}, \ldots, \beta^\top y_m^{\star,b}).$$

(4) Plug Eq. (11) into Eq. (9) and compute the statistics $D^{(b)} = PKS_{n,m}(x^{\star,b}, y^{\star,b}, \beta)$.

(5) Reject the null hypotheses if and only if

$$\frac{1 + \sum_{b=1}^{B} I(PKS_{n,m,obs} > D^{(b)})}{1 + B} > \alpha.$$

Two-sample test using PKS statistic based on permutations is summarized in Algorithm 2.

Algorithm 2 Two-Sample Test using PKS Statistic Based on Permutations.

Input: The observed n-sample X and m-sample Y, significance level α, number of permutations B.

1: Calculate the test statistic $PKS_{n,m,obs}$ for the raw data.
2: Generate B permutations $\pi_1, \ldots, \pi_B$.
3: **for** $b = 1, \ldots, B$ **do**
4: Compute the test statistic for the permuted dataset, $D^{(b)} = PKS_{n,m}(x^{\star,b}, y^{\star,b}, \beta)$.
5: **end for**

Output: $p_{value} = \frac{1+\sum_{b=1}^{B} I(PKS_{n,m,obs} > D^{(b)})}{1+B}$.

The following algorithm describes a two-sample test using the PCvM test statistic to generate critical values by permutation.

(1) Calculate $\overline{PCvM}_{n,m,obs}$ for the raw data with Algorithm 1.

(2) For each $b(1 \leq b \leq B)$, randomly generate a permutation of observation: let

$$(x^{\star,b}, y^{\star,b}) = (x_1^{\star,b}, \ldots, x_n^{\star,b}, y_1^{\star,b}, \ldots, y_m^{\star,b})$$

be a random permutation of $(x_1, \ldots, x_n, y_1, \ldots, y_m)$.

(3) Let $\beta^\top x^{\star,b}$ and $\beta^\top y^{\star,b}$ denote the projections of x and y. That is,

$$u^{\star,b} = (\beta^\top x_1^{\star,b}, \ldots, \beta^\top x_n^{\star,b}), \quad v^{\star,b} = (\beta^\top y_1^{\star,b}, \ldots, \beta^\top y_m^{\star,b}).$$

(4) Using the euMMD algorithm, we compute the estimator $T^{(b)} = \overline{PCvM}_{n,m}(x^{\star,b}, y^{\star,b}, \beta)$.

(5) Reject the null hypotheses if and only if

$$\frac{1 + \sum_{b=1}^{B} I(\overline{PCvM}_{n,m,obs} > T^{(b)})}{1 + B} > \alpha.$$

A stand-alone description of the above procedure can be found in Algorithm 3.

Algorithm 3 Two-Sample Test using PCvM Statistic Based on Permutations.

Input: The observed n-sample X and m-sample Y, number of random projections L, significance level α, number of permutations B.

1: Calculate the test statistic $\overline{PCvM}_{n,m,obs}$ for the raw data with Algorithm 1.
2: Generate B permutations $\pi_1, \ldots, \pi_B$.
3: **for** $b = 1, \ldots, B$ **do**
4: Compute the test statistic for the permuted dataset, $T^{(b)} = \overline{PCvM}_{n,m}(x^{\star,b}, y^{\star,b}, \beta)$.
5: **end for**

Output: $p_{value} = \frac{1+\sum_{b=1}^{B} I(\overline{PCvM}_{n,m,obs} > T^{(b)})}{1+B}$.

4. Assumptions and asymptotic distributions

In this section, we investigate some properties of random projection MMD and its estimators. The following theorem gives a necessary and sufficient condition for equality of distributions.

Theorem 4.1. *Let $\mathcal{F}$ be a unit ball in a universal RKHS $\mathcal{H}$, defined on the compact metric space $\mathcal{X}$, with associated kernel $k(\cdot,\cdot)$. Assume β is some random point on unite sphere $\mathcal{S}^{d-1}$. Then, we have $\mathrm{PMMD} = 0$ for any $\beta \in \mathcal{S}^{d-1}$ if and only if $P = Q$.*

Remark 1. $\mathrm{MMD}^2(P,Q) = \mathcal{E}(x,y)$ when we use kernel $k(x,y) := -\|x-y\|_2$ for $x, y \in \mathbb{R}^d$ (Sejdinovic et al. [16]), where $\mathcal{E}(x,y)$ represents the energy distance. Assume random vectors x, y with $\mathbf{E}[\|x\|_2] < \infty$, $\mathbf{E}[\|y\|_2] < \infty$. Let μ denote the uniform probability measure on $\mathcal{S}^{d-1}$. According to Lemma 4.2 of Huang and Huo [9], we have

$$\mathrm{MMD}^2(P,Q) = c_d \int_{\mathcal{S}^{d-1}} \mathrm{PMMD}^2\left(P,Q;\beta\right) d\mu(\beta),$$

where $c_d = \frac{\sqrt{\pi}\Gamma((d+1)/2)}{\Gamma(d/2)}$. Similarly, for PMMD statistics, we have

$$\mathrm{MMD}^2_{n,m}(X,Y) = c_d \int_{\mathcal{S}^{d-1}} \mathrm{PMMD}^2_{n,m}\left(X,Y;\beta\right) d\mu(\beta).$$

The metric of multivariate random variables is equivalent to the weighted integration of metric of univariate random variables.

For the rest of this section, assume $n = m$. Recall $Z := (z_1, \ldots, z_n)$ be n i.i.d. random variables, where $z_i := (x_i, y_i)$. Under the null hypothesis, it is obvious that the test statistic PMMD^2_n is a degenerate U statistic for any direction β. According to the properties of the degenerate U statistic (Section 5.5.2 of Serfling [17]), the following asymptotic null distribution is obtained.

Theorem 4.2. *Assume $\mathbf{E}(h^2) < \infty$. Let $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$ be two independent random samples from P and Q, respectively. For any direction $\beta \in \mathcal{S}^{d-1}$, under H_0, the one-sample U-statistic is degenerate, meaning $\mathbf{E}_{z'} h(\beta^\top z, \beta^\top z') = 0$. For each $\iota = 1, 2, \ldots$, let λ_ι be the eigenvalue with the corresponding eigenfunction ψ_ι satisfying the integral equation*

$$\mathbf{E}_x[\tilde{k}(u,u')\psi_\iota(u)] = \lambda_\iota \psi_\iota(u'), \tag{12}$$

and $\tilde{k}(u_i,u_j) := k(u_i,u_j) - \mathbf{E}_x k(u_i,u) - \mathbf{E}_x(u,u_j) + \mathbf{E}_{x,x'} k(u,u')$. Then, PMMD^2_n has

the limiting null distribution given by

$$n\mathrm{PMMD}_n^2 \xrightarrow{d} Q, \tag{13}$$

with $Q \overset{d}{=} \sum_{\iota=1}^{\infty} \lambda_\iota(\xi_\iota^2 - 2)$*, where* $\xrightarrow{d}$ *stands for "convergence in distribution",* $\xi_1, \xi_2, \ldots$ *are independent normal random variables with mean* 0 *and variance* 2*.*

According to Theorem 4.2, when x and y are equally distributed, the statistic PMMD_n^2 converges to a weighted sum of independent chi-squared random variables instead of a normal distribution. λ_ι in Eq. (12) is related to the process of data generation. Therefore, the critical value is obtained by the permutation test in Section 3.2. Theorem 4.2 and the continuous mapping theorem yield the asymptotic null distribution of the test statistics PKS_n and PCvM_n.

Corollary 4.3. *Under the assumptions of Theorem 4.2, for any continuous functional* $\Gamma(\cdot)$*,*

$$\Gamma(n\mathrm{PMMD}_n^2) \xrightarrow{d} \Gamma(Q). \tag{14}$$

Furthermore,

$$\sqrt{n}PKS_n \xrightarrow{d} \sup_{\mathbb{S}^{d-1}} |Q|^{1/2}, \tag{15}$$

and

$$nPCvM_n \xrightarrow{d} \int_{\mathbb{S}^{d-1}} Q d\beta. \tag{16}$$

Next, we discuss the asymptotic properties of the test statistics when the distributions of x and y are not equal. According to the properties of the one-sample U statistic, the statistic PMMD_n^2 converges to a normal distribution (Section 5.5.1 of Serfling [17]).

Theorem 4.4. *Assume* $\mathbf{E}(h^2) < \infty$*. Let us denote*

$$\sigma_n^2 = 4\left(\mathbf{E}_z[(\mathbf{E}_{z'}h(\beta^\top z, \beta^\top z'))^2] - [\mathbf{E}_{z,z'}(h(\beta^\top z, \beta^\top z'))]^2\right).$$

Then under H_1*, we have*

$$\sqrt{n}(\mathrm{PMMD}_n^2 - \mathrm{PMMD}^2) \xrightarrow{d} N(0, \sigma_n^2). \tag{17}$$

Corollary 4.5. *Under H_1, for any continuous function $\Gamma(\cdot)$*

$$\sqrt{n}(\Gamma(\mathrm{PMMD}_n^2) - \Gamma(\mathrm{PMMD}^2)) \xrightarrow{d} N(0, \tilde{\sigma}_n^2), \tag{18}$$

where $\tilde{\sigma}_n^2 = [\Gamma(\mathrm{PMMD}^2)']^2\sigma_n^2$ and $\Gamma(\mathrm{PMMD}^2)'$ is the derivative of $\Gamma(\mathrm{PMMD}^2)$. Furthermore,

$$\sqrt{n}(PKS_n - PKS) \xrightarrow{d} N(0, \tilde{\sigma}_n^2), \tag{19}$$

and

$$\sqrt{n}(PCvM_n - PCvM) \xrightarrow{d} N(0, \tilde{\sigma}_n^2). \tag{20}$$

Theorem 4.4 and Corollary 4.5 ensure that our proposed tests have nontrivial power performance under H_1. To be precise, the power under H_1 is

$$\delta := \Phi(-z_\alpha + \sqrt{n}\Gamma(\mathrm{PMMD}^2)/\tilde{\sigma}_n),$$

where z_α is the $100(1-\alpha)\%$ quantile of standard normal.

5. Numerical analysis

$B = 200$, 1000 times,

Example 1. *x_i are independently generated from $N_d(\mu_1, \Sigma_1)$, while y_i are independently generated from $N_d(\mu_2, \Sigma_2)$, where $\mu_1 = \mathbf{0}_d$, $\Sigma_1 = I_{d\times d}$, $\mu_2 = \delta\mathbf{1}_d$ and $\Sigma_2 = (1-\rho)I_{d\times d} + \rho e_d e_d^\top$. The parameter values of (δ, ρ) are set to $(0,0)$, $(0.1, 0)$, and $(0, 0.1)$, which correspond to the null hypothesis, location shift, and scale difference, respectively. The number of random projections, K, is varied from 10 to 200 to evaluate the performance of the proposed test.*

$n = m = 40$

Example 2. *In this example, we consider three scenarios. The data x_{ij} $(i \in \{1, \ldots, n\}, j \in \{1, \ldots, d\})$ and $y_{k\ell}$ $(k \in \{1, \ldots, m\}, \ell \in \{1, \ldots, d\})$ are generated independently in the*

Table 1: The empirical sizes and powers of Example 1. It is designed to evaluate how the number of random projections, K, affects the size and power performance of our proposed test. The sample sizes are fixed at $n = m = 40$.

(δ, ρ)	n	d	The number K of random projections				
			10	50	100	150	200
$(0, 0)$	40	20	0.035	0.033	0.044	0.054	0.047
		40	0.050	0.036	0.055	0.052	0.046
		80	0.046	0.052	0.043	0.037	0.053
		100	0.063	0.063	0.044	0.063	0.060
$(0.1, 0)$	40	20	0.291	0.289	0.281	0.294	0.290
		40	0.533	0.516	0.545	0.532	0.501
		80	0.810	0.793	0.794	0.809	0.784
		100	0.885	0.891	0.880	0.880	0.881
$(0, 0.1)$	40	20	0.386	0.370	0.365	0.377	0.381
		40	0.779	0.767	0.780	0.762	0.776
		80	0.990	0.986	0.988	0.990	0.987
		100	0.998	0.991	0.992	0.994	0.993

manner shown below.

(i) $x_{ij} \sim N(1, 1)$*, and* $y_{k\ell} \sim \exp(1)$*, where* $\exp(1)$ *is Exponential(1) distribution.*

(ii) $x_{ij} \sim N(1, 2)$*, and* $y_{k\ell} \sim \chi^2(1)$.

(iii) $x_{ij} \sim \text{Poisson}(1)$*, and* $y_{k\ell} \sim \exp(1)$.

In each of these three scenarios, x and y have the same first and second moments, but their distributions are different. The sample sizes are fixed at $n = m = 40$, and d is chosen from a range of $20, 40, 80, 100,$ and 200.

6. Real Data

7. Conclusion

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Nos. 11971045 and 12271014), Young Talent program of Beijing Municipal Education Commission (No. CIT&TCD201904021), and the Science and Technology Project of Beijing Municipal Education Commission (No. KM202210005012).

References

[1] Bhattacharya, B. B., 2019. A general asymptotic framework for distribution-free graph-based two-sample tests. Journal of the Royal Statistical Society Series B: Statistical Methodology 81 (3), 575–602.

[2] Bodenham, D. A., Kawahara, Y., 2023. euMMD: efficiently computing the MMD two-sample test statistic for univariate data. Statistics and Computing 33 (5), 110.

[3] Cai, T., Liu, W., Xia, Y., 2013. Two-sample covariance matrix testing and support recovery in high-dimensional and sparse settings. Journal of the American Statistical Association 108 (501), 265–277.

[4] Friedman, J. H., Rafsky, L. C., 1979. Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests. The Annals of Statistics 7 (4), 697–717.

[5] Gretton, A., Borgwardt, K., Rasch, M., Schölkopf, B., Smola, A., 2006. A kernel method for the two-sample-problem. Advances in neural information processing systems 19.

[6] Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., Smola, A., 2012. A kernel two-sample test. The Journal of Machine Learning Research 13 (1), 723–773.

[7] Henze, N., 1988. A multivariate two-sample test based on the number of nearest neighbor type coincidences. The Annals of Statistics 16 (2), 772–783.

[8] Hotelling, H., 1992. The generalization of Students ratio. In: Breakthroughs in statistics: Foundations and basic theory. Springer, pp. 54–65.

[9] Huang, C., Huo, X., 2017. An efficient and distribution-free two-sample test based on energy statistics and random projections. arXiv preprint arXiv:1707.04602.

[10] Li, J., Chen, S. X., 2012. Two sample tests for high-dimensional covariance matrices. The Annals of Statistics 40 (2), 908–940.

[11] Mondal, P. K., Biswas, M., Ghosh, A. K., 2015. On high dimensional two-sample tests based on nearest neighbors. Journal of Multivariate Analysis 141, 168–178.

[12] Pomann, G.-M., Staicu, A.-M., Ghosh, S., 2016. A two-sample distribution-free test for functional data with application to a diffusion tensor imaging study of multiple sclerosis. Journal of the Royal Statistical Society Series C: Applied Statistics 65 (3), 395–414.

[13] Qiu, T., Zhang, Q., Fang, Y., Xu, W., 2024. Testing homogeneity in high dimensional data through random projections. Journal of Multivariate Analysis 200, 105252.

[14] Rosenbaum, P. R., 2005. An exact distribution-free test comparing two multivariate distributions based on adjacency. Journal of the Royal Statistical Society Series B: Statistical Methodology 67 (4), 515–530.

[15] Schilling, M. F., 1986. Multivariate two-sample tests based on nearest neighbors. Journal of the American Statistical Association 81 (395), 799–806.

[16] Sejdinovic, D., Sriperumbudur, B., Gretton, A., Fukumizu, K., 2013. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. The annals of statistics 41 (5), 2263–2291.

[17] Serfling, R., 1980. Approximation Theorems of Mathematical Statistics. Wiley Series in Probability and Statistics.

[18] Student, 1908. The probable error of a mean. Biometrika 6 (1), 1–25.

[19] Székely, G. J., Rizzo, M. L., et al., 2004. Testing for equal distributions in high dimension. InterStat 5 (16.10), 1249–1272.

[20] Van der Vaart, A. W., 2000. Asymptotic Statistics. Vol. 3. Cambridge university press.

[21] Wald, A., Wolfowitz, J., 1940. On a test whether two samples are from the same population. The Annals of Mathematical Statistics 11 (2), 147–162.

[22] Yu, X., Li, D., Xue, L., Li, R., 2022. Power-enhanced simultaneous test of high-dimensional mean vectors and covariance matrices with application to gene-set testing. Journal of the American Statistical Association, 1–14.

[23] Zhang, Q., Dao, T., 2020. A distance based multisample test for high-dimensional compositional data with applications to the human microbiome. BMC bioinformatics 21 (9), 1–17.

[24] Zhu, L., Xu, K., Li, R., Zhong, W., 2017. Projection correlation between two random vectors. Biometrika 104 (4), 829–843.

PROOF (PROOF OF EQ. (6)).

$$\begin{aligned}\mathrm{PMMD}^2(P,Q;\beta) =& \|\mathbf{E}_x[\phi(u)] - \mathbf{E}_y[\phi(v)]\|^2_{\mathcal{H}} \\ =& \langle \mathbf{E}_x[\phi(u)], \mathbf{E}_{x'}[\phi(u)]\rangle_{\mathcal{H}} + \langle \mathbf{E}_y[\phi(v)], \mathbf{E}_{y'}[\phi(v)]\rangle_{\mathcal{H}} \\ & - 2\langle \mathbf{E}_x[\phi(u)], \mathbf{E}_y[\phi(v)]\rangle_{\mathcal{H}} \\ =& \mathbf{E}_{x,x'}\langle \phi(u), \phi(u')\rangle + \mathbf{E}_{y,y'}\langle \phi(v), \phi(v')\rangle - 2\mathbf{E}_{x,y}\langle \phi(u), \phi(v)\rangle \\ =& \mathbf{E}_{x,x'}[k(u,u')] - 2\mathbf{E}_{x,y}[k(u,v)] + \mathbf{E}_{y,y'}[k(v,v')].\end{aligned} \tag{21}$$

Directly replacing the population expectation with the corresponding U-statistic and sample mean yields an empirical estimate. This statistic, according to Serfling, is unbiased. □

PROOF (PROOF OF THEOREM 4.1). Let Φ_x and Φ_y denote the characteristic functions of x and y, respectively. It is known that random vector $x \in \mathbb{R}^d$ and $y \in \mathbb{R}^d$ have same distribution if and only if $\Phi_x = \Phi_y$, which by definition of the characteristic functions is equivalent to

$$\mathbf{E}[e^{ix^\top t}] = \mathbf{E}[e^{iy^\top t}], \quad \text{any} \quad t \in \mathbb{R}^p.$$

By variable change $t = \beta t'$, one has

$$\mathbf{E}[e^{i\beta^\top x t'}] = \mathbf{E}[e^{i\beta^\top y t'}], \quad \text{any} \quad \beta \in \mathcal{S}^{d-1} \quad \text{and} \quad t' \in [0,\infty),$$

or equivalently, the following

$$\Phi_{\beta^\top x} = \Phi_{\beta^\top y}, \quad \text{any} \quad \beta \in \mathcal{S}^{d-1}.$$

According to the definition and properties of MMD in Gretton et al. [5], we know that the previous is equivalent to

$$\mathrm{PMMD}(P,Q;\beta) = 0.$$

The proof of Theorem 4.1 is completed. □

PROOF (PROOF OF THEOREM 4.2). Under H_0, we know x and y have some distribution, which implies $\mathbf{E}_{z'}h(\beta^\top z, \beta^\top z') = 0$. Then, one has

$$\zeta_1 = \mathbf{E}_z[(\mathbf{E}_{z'}h(\beta^\top z, \beta^\top z'))^2] - [\mathbf{E}_{z,z'}(h(\beta^\top z, \beta^\top z'))]^2 = 0.$$

If $\mathbf{E}(h^2) < \infty$. By using Theorem in Section 5.5.2 of Serfling [17] to show the asymptotic distribution theory of degenerate U-statistic for the general case, we know

$$n\mathrm{PMMD}_n \xrightarrow{d} \frac{\tau(\tau-1)}{2} Q, \tag{22}$$

where τ is order of kernel h and Q is a random variable of the form

$$Q = \sum_{\iota=1}^{\infty} (\chi_{1\iota}^2 - 1),$$

where $\chi_{11}^2, \chi_{12}^2, \ldots$ are independent χ_1^2 variate.

According to Eq. (8), we know $\tau = 2$, which completes the proof of Theorem 4.2. □

PROOF (PROOF OF COROLLARY 4.3). By Eq. (13), $n\mathrm{PMMD}_n^2 \xrightarrow{d} Q$. According to the continuous-mapping theorem (Theorem 2.3 of Van der Vaart [20]), it is straightforward to prove Corollary 4.3, so details are omitted here. □

PROOF (PROOF OF THEOREM 4.4). Assume F is distribution function and $\mathbf{E}_F(h^2) < \infty$. Denote U_n is the corresponding U-statistic for estimation of $\theta = \mathbf{E}_F(h)$. By using Theorem A. in Section 5.5.1 of Serfling [17] to show the asymptotic distribution theory of U-statistic for the general case, we know

$$n^{1/2}(U_n - \theta) \xrightarrow{d} N(0, \tau^2 \zeta_1), \tag{23}$$

where τ is order of kernel h and $\zeta_1 = \mathbf{Var}_F h(X_1)$.

Recall $h(\beta^\top z_i, \beta^\top z_j) := k(u_i, u_j) + k(v_i, v_j) - k(u_i, v_j) - k(u_j, v_i)$ in Definition 3.1, we know $\tau = 2$. It is not difficult to find that

$$\zeta_1 = \mathbf{E}_z[(\mathbf{E}_{z'} h(\beta^\top z, \beta^\top z'))^2] - [\mathbf{E}_{z,z'}(h(\beta^\top z, \beta^\top z'))]^2.$$

Thus, $\zeta_1 > 0$, which completes the proof of Theorem 4.4. □

PROOF (PROOF OF COROLLARY 4.5). By Eq. (17), $\sqrt{n}(\mathrm{PMMD}_n^2 - \mathrm{PMMD}^2) \xrightarrow{d} N(0, \sigma_n^2)$. According to the continuous-mapping theorem (Theorem 2.3 of Van der Vaart [20]) and Delta-Method, it is straightforward to prove Corollary 4.5, so details are omitted here. □