Submitted to the Annals of Statistics
arXiv: arXiv:0000.0000

BALL DIVERGENCE: NONPARAMETRIC TWO SAMPLE TEST*

By Wenliang Pan[‡], Yuan Tian[‡], Xueqin Wang[‡], Heping Zhang[‡,§]

Sun Yat-sen University[‡] and Yale University[§]

In this paper, we first introduce Ball Divergence, a novel measure of the difference between two probability measures in separable Banach spaces, and show that the Ball Divergence of two probability measures is zero if and only if these two probability measures are identical without any moment assumption. Using Ball Divergence, we present a metric rank test procedure to detect the equality of distribution measures underlying independent samples. It is therefore robust to outliers or heavy-tail data. We show that this multivariate two sample test statistic is consistent with the Ball Divergence, and it converges to a mixture of χ^2 distributions under the null hypothesis and a normal distribution under the alternative hypothesis. Importantly, we prove its consistency against a general alternative hypothesis. Moreover, this result does not depend on the ratio of the two imbalanced sample sizes, ensuring that can be applied to imbalanced data. Numerical studies confirm that our test is superior to several existing tests in terms of Type I error and power. We conclude our paper with two applications of our method: one is for virtual screening in drug development process and the other is for genome wide expression analysis in hormone replacement therapy.

1. Introduction. To distinguish two unknown samples among multivariate data is important but can be difficult. Student's t is the classic test for the equality of the means in two normally distributed samples. Hotelling's T^2 test emerged as the multivariate analog to the simple t test. The normality assumption for multivariate data, however, is usually difficult to validate. Efforts have been made to extend Hotelling's T^2 to relax the normality assumption. Van Der Laan and Bryan [24] and Kosorok and Ma [14] proved the uniformity of population mean and the convergence of p dimensional marginal statistics. Chen and Qin [5] proposed a two-sample test for high-dimensional data.

*Zhang's work is supported in part by grant R01 DA016750 from the National Institute of Drug Abuse and the Chinese 1000-talent scholarship.

†We thank the co-editor, the Associate Editor, and two referees for helpful comments.

AMS 2000 subject classifications: Primary Two sample test, Two sample test based on rank distance; secondary 60K35

Keywords and phrases: Ball divergence; Finite dimensional Banach space; Metric rank; Permutation procedure

We should note that the statistics extended from Hotelling's T^2 focus on the equality of means of two populations, but not the distributions that may depend on parameters of interest besides the means. To this end, rank-based methods such as generalized Wilcoxon test [9], multivariate Kolmogorov-Smirnov test [13], or multivariate Cramér-Von Mises test [6] have been developed.

Intuitively, like the one-dimensional case, it is ideal if we can characterize the difference between two sample distributions of any dimension and use it as the basis of a two-sample test. Divergence is such a concept that is often used in both statistical learning and information theory. It measures the discrepancy between two probabilities. The so-called F-divergence is the most commonly used family of divergence measures, and it includes Kullback-Leibler divergence, Jeffreys divergence, and exponential divergence. In particular, Kullback-Leibler divergence plays a fundamental role in two-sample test of homogeneity and association [18]. Unlike distance or metric, divergence does not need to be symmetric or satisfy the triangle inequality. The weaker conditions for divergence makes it more broadly applicable while more challenging to study its properties.

Generally, the existing two-sample multivariate tests require the moment assumption and overlook the extreme imbalance cases in which one sample size is disproportionately larger than the other. Since imbalanced data arise from a variety of applications, and have attracted a great deal of attention and interests in recent studies. It is important to develop a powerful two-sample multivariate test that takes into account the imbalanced designs. Chen, Dou and Qiao [4] proposed an ensemble sub-sampling nonparametric multivariate test using the nearest neighborhoods (ESS-NN) for imbalanced data. This method copes well for imbalanced data, and the power of the test increases as the size of the larger group increases by fixing the size of the smaller group. Gretton et al. [10] introduced the maximum mean discrepancy (MMD) for the two-sample problem. The MMD is particularly appealing because it can distinguish multivariate distributions for graph data and is robust for imbalanced data. However, the efficiency of ESS-NN and MMD tests are limited by various factors such as the number of neighbors k in ESS-NN and the kernel parameter in MMD despite the efforts to gauge the tuning parameters for optimal performance [4, 10].

In this paper, we introduce Ball Divergence (BD), which is a new concept to measure the difference between two probability distributions in separable Banach space. This concept relies on the fact that two Borel probability measures are identical if they agree on all balls in a separable Banach space [17]. Like Energy distance [23], the BD of two probability measures is shown

to be zero if and only if they are identical. However, BD is preferable to Energy distance because the latter requires the moment condition and hence not robust to heavy-tail data or outliers. Importantly, many Banach spaces are not of strong negative type, or even negative type. For example, $\mathbb{R}^n$ with ℓ^p metric is not of negative type whenever $3 \leq n \leq \infty$ and $2 < p \leq \infty$. This fact limits the application of Energy distance. Thus, we use BD to derive a two-sample metric rank test that can deal with the heavy-tail data. We shall demonstrate that this test performs well with respect to both the power and the type I error.

Specifically, our empirical BD statistic is defined from the difference between the averages of the metric ranks. Thus, we avoid estimating the multivariate probability function, which is a major difficulty in utilizing divergence based methods. We show that the empirical BD statistic converges to the BD. We also obtain the asymptotic distributions of our test statistic under both the null and alternative hypotheses. Importantly, we prove its consistency against alternative hypothesis, setting the stage for testing imbalanced data. Furthermore, through simulation studies, we verify the desirable properties of the BD statistic in a variety of representative settings.

In the following section, we introduce the BD and its sample version – the empirical BD statistic. In Section 3, we examine the properties of the BD statistic. Monte Carlo studies supporting the performance of the BD statistic are presented in Section 4, followed by the analysis of two real datasets in Section 5. A brief conclusion is provided in Section 6. Some of the technical details are deferred to the Appendix.

2. Nonparametric Two-Sample Test based on Ball Divergence.

Let $(\mathrm{V}, \|\cdot\|)$ be a Banach space, where the norm $\|\cdot\|$ induces a metric ρ via $\rho(u, v) = \|u - v\|$ for two points $u, v \in \mathrm{V}$. And let $\mathcal{B}$ be the Borel σ-algebra, which is the smallest σ-algebra in V that contains all closed (or open) subsets of V. Denoted by $\mathcal{B}_0$ the collection: $\{\bar{B}(u, r) : u \in \mathrm{V}, r \geq 0\}$, where $\bar{B}(u, r)$ is a closed ball with the center u and the radius r. In order for measures μ and ν to coincide on $\mathcal{B}$ in μ and ν on a separable Banach space V, it is sufficient that they are coincide on $\mathcal{B}_0$ [17]. This leads us to define some Cramér-Von Mises type criteria for testing the difference between μ and ν over these balls, although additional conditions are needed for those criteria such that they equal zero if and only if μ and ν are equal. We will give a sufficient condition, which is referred to as Ball Divergence.

2.1. *Ball Divergence.*

Definition 2.1. *The Ball Divergence of two Borel probability measures*

μ and ν is defined as an integral of the square of the measure difference between μ and ν over a given closed ball collection $\mathcal{B}$ as following,

$$D(\mu,\nu)=\iint_{\mathrm{V}\times\mathrm{V}}[\mu-\nu]^2(\bar{B}(u,\rho(u,v)))(\mu(du)\mu(dv)+\nu(du)\nu(dv)). \tag{2.1}$$

The following two theorems are the keystone to our testing procedure.

THEOREM 1. *Given two Borel probability measures μ and ν on a finite dimensional Banach space* V*, then $D(\mu,\nu)\geq 0$ where the equality holds if and only if $\mu=\nu$.*

The proof of this theorem depends on the covering theorem (See Theorem 3.1 in Jackson and Mauldin [12]). With the covering theorem, we can prove the following lemma which is analogous to Corollary 5.8.2 in Bogachev [3].

LEMMA 2.1. *Let μ be a probability measure on a finite dimensional Banach space* V*, $\mathcal{B}_C$ a collection of non-degenerate closed balls, and V_C the set of their centers such that for every $v\in V_C$ and every $\varepsilon>0$, $\mathcal{B}_C$ contains a ball $\bar{B}(v,r)$ with $r<\varepsilon$. Then, for every nonempty open set $U\subset\mathrm{V}$, there is at most a countable collection of disjoint balls $\bar{B}_j\in\mathcal{B}_C$ such that*

$$\bigcup_{j=1}^{\infty}\bar{B}_j\subset U \quad \textit{and} \quad \mu((V_c\cap U)\backslash\bigcup_{j=1}^{\infty}\bar{B}_j)=0.$$

Next is a proof of Theorem 1.

PROOF. It is obvious that $D(\mu,\nu)\geq 0$ and if $\mu=\nu$, then $D(\mu,\nu)=0$. Next we shall verify that if $D(\mu,\nu)=0$, then $\mu=\nu$; that is, $\mu(B)=\nu(B)$ for $B\in\mathcal{B}_0$.

Let S_μ be the support of μ consisting of the points such that every of their open neighborhoods has positive measure, and S_ν the support of ν. Then $S_\mu^c=\mathrm{V}/S_\mu$ is the union of all μ-null open sets. Also, since V is separable, $\mu(S_\mu^c)=0$. $D(\mu,\nu)=0$ implies that

$$\iint_{\mathrm{S}_\mu\times\mathrm{S}_\mu}[\mu-\nu]^2(\bar{B}(u,\rho(u,v)))\mu(du)\mu(dv)=0.$$

Since $[\mu-\nu]^2(\bar{B}(u,\rho(u,v)))$ is nonnegative, we have

$$[\mu-\nu]^2(\bar{B}(u,\rho(u,v)))=0. \quad a.s.$$

According to the definition of the support set, we know that no μ-null set is contained in S_μ. Therefore, $\mu = \nu$ on $\bar{B}(u, \rho(u, v))$ if $u, v \in S_\mu$. The equality also holds for $u, v \in S_\nu$.

Next for $u \in S_\mu$ and $r \geq 0$, let $r_\mu = \sup\{\rho(u, v) : v \in S_\mu \cap \bar{B}(u, r)\}$. Since $S_\mu \cap \bar{B}(u, r)$ is closed, there exists $v_0 \in S_\mu \cap \bar{B}(u, r)$ such that $r_\mu = \rho(u, v_0)$, thus $\mu(\bar{B}(u, r)) = \mu(\bar{B}(u, r_\mu)) = \nu(\bar{B}(u, r_\mu)) \leq \nu(\bar{B}(u, r))$. Analogously, $\nu(\bar{B}(u, r)) \leq \mu(\bar{B}(u, r))$ if $u \in S_\nu$.

Finally, we shall show that $\mu = \nu$ on $\bar{B}(u, r)$ if $u \in \mathrm{V}$ and $r \geq 0$. Given $\epsilon > 0$, there exists an open cover U_ϵ such that $\mu(U_\epsilon \backslash (S_\mu \cap \bar{B}(u, r))) < \epsilon$, $\nu(U_\epsilon \backslash (S_\mu \cap \bar{B}(u, r))) < \epsilon$. by Lemma 2.1, we can find an at most countable collection of disjoint balls $\bar{B}_{\epsilon,j}$ with the ball center in $S_\mu \cap \bar{B}(u, r)$ such that

$$\bigcup_{j=1}^{\infty} \bar{B}_{\epsilon,j} \subset U_\epsilon \quad \text{and} \quad \mu(S_\mu \cap \bar{B}(u, r) \backslash \bigcup_{j=1}^{\infty} \bar{B}_{\epsilon,j}) = 0.$$

Because $\mu(\bar{B}_{\epsilon,j}) \leq \nu(\bar{B}_{\epsilon,j})$, let $\varepsilon \to 0$, we obtain that $\mu(S_\mu \cap \bar{B}(u, r)) \leq \nu(S_\mu \cap \bar{B}(u, r))$. Thus $\mu(\bar{B}(u, r)) = \mu(S_\mu \cap \bar{B}(u, r)) \leq \nu(S_\mu \cap \bar{B}(u, r)) \leq \nu(\bar{B}(u, r))$. Similarly, we can show that $\nu(\bar{B}(u, r)) \leq \mu(\bar{B}(u, r))$. Therefore, $\nu(\bar{B}(u, r)) = \mu(\bar{B}(u, r))$.

This completes the proof.

□

Note that Lemma 2.1 is unnecessary if $S_\mu = \mathrm{V}$ or $S_\nu = \mathrm{V}$ in the proof of Theorem 1, due to [17]. Hence, we could extend Theorem 1 from finite dimensional Banach spaces to separable Banach spaces.

THEOREM 2. *Suppose μ and ν are two Borel probability measures in a separable Banach space V. Denote their support sets by S_μ and S_ν, if $S_\mu = \mathrm{V}$ or $S_\nu = \mathrm{V}$, then we have $D(\mu, \nu) \geq 0$ where the equality holds if and only if $\mu = \nu$.*

REMARK 2.1. *Ball Divergence can be extended to more general divergence. More specifically, define a generalized version of $D(\mu, \nu)$ as follows:*

$$D^{(\alpha)}(\mu, \nu)$$
$$= [\iint_{\mathrm{V} \times \mathrm{V}} |\mu - \nu|^\alpha (\bar{B}(u, \rho(u, v)))(\mu(du)\mu(dv) + \nu(du)\nu(dv))]^{1/\alpha}, 0 < \alpha < \infty$$

and

$$D^{(\infty)}(\mu, \nu) = \sup_{u,v \in S_\mu, u,v \in S_\nu} |\mu - \nu|(\bar{B}(u, \rho(u, v)))$$

Theorems 1 and 2 still hold for $D^{(\alpha)}(\mu,\nu)$ and $D^{(\infty)}(\mu,\nu)$. By the definitions of $D^{(\alpha)}(\mu,\nu)$ and $D^{(\infty)}(\mu,\nu)$, we observe

$$|\mu-\nu|^{\alpha}(\bar{B}(u,\rho(u,v)))=0$$

if $D^{(\alpha)}(\mu,\nu)=0$ or $D^{(\infty)}(\mu,\nu)=0$, for $u,v\in S_\mu$ or $u,v\in S_\nu$. Therefore, $\mu=\nu$ on $\bar{B}(u,\rho(u,v))$ if $u,v\in S_\mu$ or $u,v\in S_\nu$. The rest follows from the proof of Theorem 1.

REMARK 2.2. *It's worth noting that the square root of Ball Divergence is a symmetric divergence but not a metric, because it does not satisfy the triangle inequality. For example, let $\mu(X=0)=1$, $\nu(Y=1)=1$ and $\theta(Z=0)=\theta(Z=1)=0.5$, then we have $\sqrt{D(\mu,\nu)}=\sqrt{2}$ and $\sqrt{D(\mu,\theta)}+\sqrt{D(\nu,\theta)}=\sqrt{1.5}$.*

2.2. *Nonparametric Two-Sample Test based on Ball Divergence.* Next, we find the sample version of the above mentioned $D(\mu,\nu)$. Given two independent samples $\mathcal{X}_n=\{X_1,\dots,X_n\}$ with the associated probability measure μ and $\mathcal{Y}_m=\{Y_1,\dots,Y_m\}$ with ν, where the observations in each sample are *i.i.d.* For convenience, we decompose the Ball Divergence into two parts:

$$A=\iint_{\mathrm{V}\times\mathrm{V}}[\mu-\nu]^2(\bar{B}(u,\rho(u,v)))\mu(du)\mu(dv),$$

and

$$C=\iint_{\mathrm{V}\times\mathrm{V}}[\mu-\nu]^2(\bar{B}(u,\rho(u,v)))\nu(du)\nu(dv).$$

Thus,

$$D(\mu,\nu)=A+C.$$

Also, let $\delta(x,y,z)=I(z\in\bar{B}(x,\rho(x,y)))$ and $\xi(x,y,z_1,z_2)=\delta(x,y,z_1)\cdot\delta(x,y,z_2)$, where $\delta(x,y,z)$ indicates whether z is located in the closed ball $\bar{B}(x,\rho(x,y))$. We denote

$$A_{ij}^X=\frac{1}{n}\sum_{u=1}^{n}\delta(X_i,X_j,X_u),\ A_{ij}^Y=\frac{1}{m}\sum_{v=1}^{m}\delta(X_i,X_j,Y_v),$$
$$C_{kl}^X=\frac{1}{n}\sum_{u=1}^{n}\delta(Y_k,Y_l,X_u),\ C_{kl}^Y=\frac{1}{m}\sum_{v=1}^{m}\delta(Y_k,Y_l,Y_v).$$

A_{ij}^X represents the proportion of samples from the probability measure μ located in the ball $\bar{B}(X_i,\rho(X_i,X_j))$ and A_{ij}^Y represents the proportion of

samples from the probability measure ν located in the ball $\bar{B}(X_i, \rho(X_i, X_j))$. Meanwhile, C_{kl}^X and C_{kl}^Y represent the corresponding proportions located in the ball $\bar{B}(Y_k, \rho(Y_k, Y_l))$. Therefore, we can use $A_{ij}^X, A_{ij}^Y, C_{kl}^X, C_{kl}^Y$ to construct the sample version of A and C, respectively. The sample versions of A and C are as follows:

$$A_{n,m} = \frac{1}{n^2}\sum_{i,j=1}^{n}(A_{ij}^X - A_{ij}^Y)^2,\ C_{n,m} = \frac{1}{m^2}\sum_{k,l=1}^{m}(C_{kl}^X - C_{kl}^Y)^2.$$

Consequently, we can define our test statistic as:

$$D_{n,m} = A_{n,m} + C_{n,m}.$$

Note that nA_{ij}^X is the rank of $\rho(X_i, X_j)$ among $\{\rho(X_i, X_u), u = 1, \cdots, n\}$ and mA_{ij}^Y is the rank of $\rho(X_i, X_j)$ among $\{\rho(X_i, Y_v), v = 1, \cdots, m\} \cup \{\rho(X_i, X_j)\}$. Thus, our test belongs to a class of metric rank tests and possesses the properties of a general rank test such as robustness.

REMARK 2.3. *BD can be generalized to the K-sample problem. Suppose that $\mu_1, \ldots, \mu_K$ are $K(\geq 2)$ measures on Banach space $(\mathrm{V}, \|\cdot\|)$. We can define the K-sample BD as*

$$D(\mu_1, \ldots, \mu_K) = \sum_{1\leq k<l\leq K}\iint_{\mathrm{V}\times\mathrm{V}}[\mu_k - \mu_l]^2(\bar{B}(u, \rho(u, v)))(\mu_k(du)\mu_k(dv) + \mu_l(du)\mu_l(dv)).$$

It follows from Theorems 1 and 2 that $D(\mu_1, \ldots, \mu_K) = 0$ if and only if $\mu_1 = \cdots = \mu_K$.

REMARK 2.4. *BD uses the ball, and hence differs from the classic Kolmogorov-Smirnov test and Cramér-von Mises test. It is, however, related to the energy distance and MMD. Since the energy distance and MMD are actually equivalent (Sejdinovic et al. [22]), it suffices to explain how BD and MMD are related. BD, Energy distance and MMD can be unified in the framework of the variogram:*

$$D(\mu, \nu) = E(U(X) - U(Y))^2,$$

where U is a Gaussian process with mean zero and $X \sim \mu, Y \sim \nu$. Let $k(z_1, z_2)$ denote the covariance function of U.

(1) If $k(z_1, z_2) = \|z_1\| + \|z_2\| - \|z_1 - z_2\|$, then $D(\mu, \nu)$ is Energy distance.

(2) If $k(z_1, z_2)$ is a positive definite kernel function, then $D(\mu, \nu)$ is MMD.
(3) If $k_\mu(z_1, z_2) = E\xi(X_1, X_2, z_1, z_2)$ and $k_\nu(z_1, z_2) = E\xi(Y_1, Y_2, z_1, z_2)$, then $D_\mu(\mu, \nu) + D_\nu(\mu, \nu)$ is Ball divergence.

Next we can verify that $k_\mu(z_1, z_2)$ is positive definite (Sejdinovic et al. [22]) as follows. Let $z_0 \notin \bar{B}(X, \rho(X_1, X_2))$ (e.g., $z_0 = \infty$).

$$\begin{aligned} 2k_\mu(z_1, z_2) =& 2E\xi(X_1, X_2, z_1, z_2) \\ =& E|\delta(X_1, X_2, z_1) - \delta(X_1, X_2, z_0)|^2 + E|\delta(X_1, X_2, z_2) - \delta(X_1, X_2, z_0)|^2 \\ & - E|\delta(X_1, X_2, z_1) - \delta(X_1, X_2, z_2)|^2. \end{aligned}$$

Since $E|\delta(X_1, X_2, z_1) - \delta(X_1, X_2, z_2)|^2$ satisfies Schoenberg's condition (Schoenberg [20, 21]), it is negative type. Thus, $k_\mu(z_1, z_2)$ is positive definite. A similar argument holds for $k_\nu(z_1, z_2)$.

3. Asymptotic Properties. The first theorem guarantees that the sample version $D_{n,m}$ converges to $D(\mu, \nu)$ as the sample sizes increase. Note that $D_{n,m}$ itself is not a V-Statistic but both of its two components $A_{n,m}$ and $C_{n,m}$ are. Indeed, $A_{n,m}$ is a two-sample V-statistic of order $(4, 2)$ while $C_{n,m}$ is a two-sample V-statistic of order $(2, 4)$. Thus, the consistency of $D_{n,m}$ follows from the theory on V-statistic as stated below.

THEOREM 3. (*Consistency*) *We have*

$$D_{n,m} \xrightarrow[n,m\to\infty]{a.s.} D(\mu, \nu),$$

where $\frac{n}{n+m} \to \tau$ for some $\tau \in [0, 1]$.

As the sample version of $D(\mu, \nu)$, $D_{n,m}$ converges to $D(\mu, \nu)$ when n, m increase to infinity. According to Theorem 1 and 2, μ and ν are identical if and only if $D(\mu, \nu) = 0$. Thus, $D_{n,m}$ can be used to detect the difference between μ and ν.

We further investigate the asymptotic properties of $D_{n,m}$ under the null and alternative hypotheses. In particular, we consider the limiting distributions when n and m tend to infinity at different rates. Theorem 4 states that under the null hypothesis, the asymptotic distribution is a mixture of χ^2 distributions for any τ. In contrast, Theorem 5 shows that under the alternative hypothesis, the statistic converges in distribution to a normal distribution with mean 0 and the variance as a function of τ. This is because that $A_{n,m}$ and $C_{n,m}$ are degenerate V-statistics under the null hypothesis, but not under the alternative hypothesis. The H-decomposition (Lee [15], Section 1.6) will be used to derive the asymptotic distributions.

We define the symmetric function

$$Q(x,y;x',y') = \big(\phi_A^{(2,0)}(x,x') + \phi_A^{(1,1)}(x,y) + \phi_A^{(1,1)}(x',y') + \phi_A^{(0,2)}(y,y')\big),$$

where

$$\begin{aligned}
\phi_A^{(2,0)}(x,x') =& E[\xi(X_1,X_2,x,x')] + E[\xi(X_1,X_2,Y,Y_3)] \\
&- E[\xi(X_1,X_2,x,Y)] - E[\xi(X_1,X_2,x',Y_3)], \\
\phi_A^{(1,1)}(x,y) =& E[\xi(X_1,X_2,x,X_3)] + E[\xi(X_1,X_2,y,Y_3)] \\
&- E[\xi(X_1,X_2,x,y)] - E[\xi(X_1,X_2,X_3,Y_3)], \\
\phi_A^{(0,2)}(y,y') =& E[\xi(X_1,X_2,X,X_3)] + E[\xi(X_1,X_2,y,y')] \\
&- E[\xi(X_1,X_2,X,y)] - E[\xi(X_1,X_2,X,y')].
\end{aligned}$$

$Q(x,y;x',y')$ is the second component of random vectors X and Y in the H-decomposition of $A_{n,m}$ and $C_{n,m}$. It has the following spectral decomposition

$$Q(x,y;x',y') = \sum_{k=1}^{\infty} \lambda_k f_k(x,y) f_k(x',y'),$$

where λ_k and f_k are the eigenvalues and eigenfunctions of Q. For $k = 1,2,\cdots$, Z_{1k}, Z_{2k} are $i.i.d. N(0,1)$, and

$$\begin{gathered}
a_k^2(\tau) = (1-\tau)E_X[E_Y f_k(X,Y)]^2, \quad b_k^2(\tau) = \tau E_Y[E_X f_k(X,Y)]^2, \\
\theta = 2E[E(\delta(X_1,X_2,X)(1-\delta(X_1,X_2,Y))|X_1,X_2)].
\end{gathered}$$

THEOREM 4. (*Asymptotic distribution under the null hypothesis*) *Suppose that both n and $m \to \infty$ in such a way that $\frac{n}{n+m} \to \tau$, $0 \le \tau \le 1$. Under the null hypothesis, we have*

$$\frac{nm}{n+m} D_{n,m} \xrightarrow[n\to\infty]{d} \sum_{k=1}^{\infty} 2\lambda_k[(a_k(\tau)Z_{1k} + b_k(\tau)Z_{2k})^2 - (a_k^2(\tau) + b_k^2(\tau))] + \theta.$$

Under the alternative hypothesis, $A_{n,m}$ and $C_{n,m}$ are non-degenerate V-statistics. They are asymptotically normal because their projections are asymptotically normal. Let $g^{(1,0)}(X)$ and $g^{(0,1)}(Y)$ be the first component of random vectors X and Y in the H-decomposition of $A_{n,m}$ and $C_{n,m}$, respectively, and let

$$\delta_{1,0}^2 = Var(g^{(1,0)}(X)) \quad and \quad \delta_{0,1}^2 = Var(g^{(0,1)}(Y)). \tag{3.1}$$

We obtain the following limit distribution of $\sqrt{nm}(D_{n,m} - D(\mu,\nu))/\sqrt{n+m}$.

THEOREM 5. (*Distribution under the alternative hypothesis*) *Suppose that both* n *and* $m \to \infty$ *in such a way that* $\frac{n}{n+m} \to \tau$, $0 \leq \tau \leq 1$. *Under the alternative hypothesis, we have*

$$\sqrt{\frac{nm}{n+m}}(D_{n,m} - D(\mu,\nu)) \xrightarrow[n\to\infty]{d} N(0,(1-\tau)\delta_{1,0}^2 + \tau\delta_{0,1}^2).$$

From Theorem 5, we know that the distribution under the alternative hypothesis is determined by the random vector in the group with the smaller sample size.

Let $\eta = \frac{n}{m}(\leq 1)$ be the ratio of the smaller to the larger sample size. Without loss of generality, we assume that $n \leq m$. As in Chen, Dou and Qiao [4], the consistency theorem does not depend on the ratio η as presented below.

THEOREM 6. *The test based on* $D_{n,m}$ *is consistent against any general alternative* H_1. *More specifically,*

$$\lim_{n\to\infty} Var_{H_1}(D_{n,m}) = 0,$$

and

$$\Delta(\eta) := \liminf_{n\to\infty}(E_{H_1}D_{n,m} - E_{H_0}D_{n,m}) > 0.$$

More importantly, $\Delta(\eta)$ *can also be expressed as*

$$\Delta(\eta) \equiv D(\mu,\nu),$$

which is independent of η.

Theorem 6 shows that the consistency result of our new test statistic is independent of the ratio η, making it possible for $D_{n,m}$ to cope with imbalance data.

4. Simulation. In this section, we present the numerical performance of the proposed BD and compare it with other tests, including Energy distance (ED), ESS-NN, Hotelling's T^2 and MMD, in testing distribution equality for random samples. We use permutation test to obtain an empirical distribution and derive the p-value of the BD statistic. For MMD, we used the Gaussian kernel whose bandwidth was set at the median distance between points in the aggregate samples.

The comparisons are based on the Type-I error and power. We fix the smaller sample size at 30 and let the sample size ratio be 1, 4, or 16. Each simulation is replicated 400 times.

For Type I error evaluation, we consider the three models below.

Model 1: Multivariate normal distribution. The mean vectors are $(\mu_x, \ldots, \mu_x)_d$, $(\mu_y, \ldots, \mu_y)_d$ and the covariance matrices are $\sigma_x I_d$, $\sigma_y I_d$, where I_d is the identity matrices.

(1.1) $\{d = 1, \mu_x = 0, \mu_y = 0, \sigma_x = 1, \sigma_y = 1\}$

(1.2) $\{d = 5, \mu_x = 0, \mu_y = 0, \sigma_x = 1, \sigma_y = 1\}$

Model 2: The marginal distribution of the multivariate random variable follows a log-normal distribution. That is, $\log(X) \sim N(\mu_x, 1)$ and $\log(Y) \sim N(\mu_y, 1)$.

(2.1) $\{d = 1, \mu_x = 1, \mu_y = 1\}$

(2.2) $\{d = 5, \mu_x = 1, \mu_y = 1\}$

Model 3: Multivariate t distribution with freedom degree 1. The location parameter vectors are $(\mu_x, \ldots, \mu_x)_d, (\mu_y, \ldots, \mu_y)_d$ and the scale parameter matrices are identity matrices $\sigma_x I_d, \sigma_y I_d$, where I_d is the identity matrices.

(3.1) $\{d = 1, \mu_x = 0, \mu_y = 0, \sigma_x = 1, \sigma_y = 1\}$

(3.2) $\{d = 5, \mu_x = 0, \mu_y = 0, \sigma_x = 1, \sigma_y = 1\}$

For the power evaluation, we examine three simulation models by varying the location parameter shift and/or scale parameter shift in each model. The following three models are used to evaluate the power.

Model 4: Multivariate normal distribution. In the location shift case, the covariance matrices for both distributions are the scaled identity matrix σI_d and only different in the means $(\mu_x, \ldots, \mu_x)_d$ and $(\mu_y, \ldots, \mu_y)_d$.

(4.1) $\{d = 1, \mu_x = 0, \mu_y = 1, \sigma_x = 1, \sigma_y = 1\}$

(4.2) $\{d = 5, \mu_x = 0, \mu_y = 0.5, \sigma_x = 1, \sigma_y = 1\}$

In the scale shift case, both distributions have zero mean and their covariance matrices are scaled identity matrices.

(4.3) $\{d = 1, \mu_x = 0, \mu_y = 0, \sigma_x = 1, \sigma_y = 2\}$

(4.4) $\{d = 5, \mu_x = 0, \mu_y = 0, \sigma_x = 1, \sigma_y = 1.4\}$

Model 5: The marginal distribution of the multivariate random variable follows a log-normal distribution. $\log(X) \sim N(\mu_x, 1)$ and $\log(Y) \sim N(\mu_y, 1)$.

(5.1) $\{d = 1, \mu_x = 0, \mu_y = 0.5\}$

(5.2) $\{d = 5, \mu_x = 0, \mu_y = 0.4\}$

Model 6: Multivariate t distribution with freedom degree 1. In the location shift case, both distributions have the identity scale and differ only in the locations.

(6.1) $\{d = 1, \mu_x = 0, \mu_y = 1, \sigma_x = 1, \sigma_y = 1\}$

(6.2) $\{d = 5, \mu_x = 0, \mu_y = 1, \sigma_x = 1, \sigma_y = 1\}$

In the scale shift case, both distributions have zero mean and scaled identity covariance matrix σI_d.

(6.3) $\{d = 1, \mu_x = 0, \mu_y = 0, \sigma_x = 0.45, \sigma_y = 1\}$

(6.4) $\{d = 5, \mu_x = 0, \mu_y = 0, \sigma_x = 0.45, \sigma_y = 1\}$

Table 1 evaluates the Type-I error. For the aforementioned models, BD, Energy distance, ESS-NN and MMD control the Type-I error well around 0.05. The Type-I error of Hoteling's T^2 is slightly unstable in Model (3.1). Tables 2, 3, and 4 compare the power of all tests. A common pattern is that the power increases as the ratio of the sample sizes increases.

Table 2 presents the simulation results of Model 4. In this table, we can know that Hotelling's T^2 test enjoys the highest power in detecting the location shift in multivariate normal distribution (Model 4.1 & 4.2), which makes sense since Hotelling's T^2 test is a kind of parametric method that based on the multivariate normal distribution assumption and maximizes the usage of known information about the data. BD shares a comparable power with Hotelling's T^2 and increases to almost 1 as the sample sizes ratio climbs to 64. However, the advantage of Hotelling's T^2 vanishes when detect the scale difference (Model 4.3 & 4.4). BD turns to be the most powerful one within five methods. Both ED and ESS-NN meet an obvious decrement when the underlying difference changes from location shift to scale difference. The power of MMD is more robust ignoring the changes of difference types.

In log normal distribution, ED test performs best in the location shift in the univariate balanced design. With the increasing of sample sizes ratio, BD outperforms ED and becomes the most powerful in univariate case (Model 5.1). MMD also enjoys a more sharp increment than ESS-NN and Hotelling's T^2 but less than BD. Under multivariate case, Hotelling's T^2 becomes the most powerful, whereas BD shares a comparative performance with Hotelling's T^2.

Table 4 presents the results of Model 6. In this table, MMD has a remarkable performance in detecting the location shift in both univariate and multivariate cases as well as scale difference in the multivariate case (Model 6.1, 6.2, 6.4). In univariate scale difference test, BD exceeds MMD (Model 6.3). It's worth noting that BD almost catch up with MMD in other cases, which also indicate a powerful performance when the underlying distribution is heavy tailed. The performance of ESS-NN is also noticeable, whereas the performance of ED and Hotelling T^2 share a decreasing trend with the increasing of sample sizes ratio.

TABLE 1
Performance of Type I error in Model 1,2,3

Model	ratio	BD	ED	ESS-NN	hotelling	MMD
(1.1)	1	0.0400	0.0325	0.0475	0.0500	0.0575
	4	0.0425	0.0450	0.0350	0.0600	0.0375
	16	0.0525	0.0400	0.0300	0.0525	0.0475
(1.2)	1	0.0325	0.0325	0.0475	0.0575	0.0375
	4	0.0550	0.0400	0.0450	0.0550	0.0525
	16	0.0325	0.0425	0.0500	0.0375	0.0375
(2.1)	1	0.0375	0.0525	0.0250	0.0625	0.0425
	4	0.0475	0.0275	0.0350	0.0300	0.0400
	16	0.0350	0.0275	0.0250	0.0325	0.0450
(2.2)	1	0.0350	0.0450	0.0350	0.0575	0.0325
	4	0.0350	0.0550	0.0600	0.0650	0.0525
	16	0.0325	0.0450	0.0350	0.0375	0.0600
(3.1)	1	0.0550	0.0350	0.0575	0.0250	0.0400
	4	0.0300	0.0400	0.0400	0.1025	0.0325
	16	0.0300	0.0425	0.0300	0.0700	0.0525
(3.2)	1	0.0425	0.0475	0.0475	0.0425	0.0450
	4	0.0325	0.0600	0.0400	0.0625	0.0400
	16	0.0375	0.0250	0.0425	0.0450	0.0400

Next, we consider the performance in the mixture of k normal distributions with probability equals $p_1, \ldots, p_k$, separately.

Model 7: Mixture of distributions.

(7.1) $\{d = 1, \mu_x = 0, \mu_{y_1} = -1, \mu_{y_2} = 1, \sigma_x = \sigma_{y_1} = \sigma_{y_2} = 1,$
$p_{y_1} = p_{y_2} = 0.5\}$

(7.2) $\{d = 1, \mu_{x_1} = 0.3, \mu_{x_2} = -0.3, \mu_{y_1} = 1.3, \mu_{y_2} = -1.3,$
$\sigma_{x_1} = \sigma_{x_2} = \sigma_{y_1} = \sigma_{y_2} = 1, p_{x_1} = p_{x_2} = p_{y_1} = p_{y_2} = 0.5\}$

(7.3) $\{d = 5, \mu_x = 0, \mu_{y_1} = -0.5, \mu_{y_2} = 0.5, \mu_{y_3} = 0,$
$\sigma_x = \sigma_{y_1} = \sigma_{y_2} = 1, \sigma_{y_3} = 2, p_{y_1} = p_{y_2} = 0.25, p_{y_3} = 0.5\}$

Table 5 displays the performance under the mixture of distributions. BD maintains a desirable performance and keeps being the most powerful one disregard the specific mixture components. MMD shares a robust performance with BD. The Hotelling's T^2 seems to lose its power in such cases.

In summary, BD is a powerful two-sample test in many settings, and particularly powerful for the cases where the underlying distributions differ in scale or are mixture distributions. It remains competitive in other cases.

5. Real data analysis.

5.1. *Virtual drug screening.* Virtual screening plays an important role in drug development process. Non-toxic compounds can be separated from the

TABLE 2
Performance of Power under Model 4. The highest power is hightlighted in bold. The last four columns refers to the power ratio between the four methods to BD.

Model	Ratio	BD	ED/BD	ESS-NN/BD	T^2/BD	MMD/BD
	1	0.8625	1.0957	0.8928	**1.1333**	0.9420
(4.1)	4	0.9700	1.0206	0.9587	**1.0258**	0.9845
	16	0.9950	**1.0050**	0.9548	**1.0050**	0.9874
	1	0.7675	1.1629	1.0098	**1.2932**	0.8404
(4.2)	4	0.9450	1.0344	0.9630	**1.0582**	0.9259
	16	0.9825	1.0153	0.9542	**1.0178**	0.9186
	1	**0.6500**	0.6000	0.7038	0.0808	0.8731
(4.3)	4	**0.9275**	0.5202	0.7547	0.0081	0.9461
	16	**0.9775**	0.5141	0.7826	0.0000	0.9591
	1	**0.9475**	0.2718	0.3272	0.0501	0.6992
(4.4)	4	**1.0000**	0.2100	0.4500	0.0175	0.9225
	16	**1.0000**	0.1925	0.4850	0.0050	0.9500

TABLE 3
Performance of Power in Model 5. The highest power is hightlighted in bold. The last four columns refers to the power ratio between the four methods to BD.

Model	ratio	BD	ED/BD	ESS-NN/BD	hotelling/BD	MMD/BD
	1	0.3125	**1.1600**	0.7600	1.0400	0.9440
(5.1)	4	**0.4625**	0.9568	0.6811	0.7946	0.9946
	16	**0.5125**	0.8878	0.7512	0.6878	0.8927
	1	0.5375	1.1023	0.8186	**1.4419**	0.9814
(5.2)	4	0.7525	0.9635	0.9037	**1.2060**	0.9801
	16	0.8175	0.8930	0.9113	**1.1368**	0.9664

toxic ones and be developed in further drug development procedure by combining screening techniques such as High-Throughput Screening and other computational techniques. The computational techniques involved in this screening process consist of discriminate analysis, such as traditional discrimination methods and machine learning approaches. Models are built so that active compounds can be discriminated from inactive ones. In such experiments, data tend to be imbalanced where the ratio of active compounds to inactive compounds is 0.001 on average, making traditional discriminate analysis methods ineffective.

We re-analyze two data sets reported in Schierz [19], which are available on http://www.biomedcentral.com. It is clear in Table 6 that both datasets are highly imbalanced in the comparison groups. We use BD, ESS-KNN, Energy distance, Hotelling's T^2 and MMD for these two datasets and report the results in Table 7.

According to Table 7, BD detected significant differences between the inactive and active groups in both datasets, Energy distance and ESS-NN

TABLE 4
Performance of Power in Model 6. The highest power is hightlighted in bold. The last four columns refers to the power ratio between the four methods to BD.

Model	ratio	BD	ED/BD	ESS-NN/BD	hotelling/BD	MMD/BD
(6.1)	1	0.5100	0.6422	0.9706	0.1814	**1.1422**
	4	0.7550	0.3742	0.9338	0.1424	**1.1325**
	16	0.8250	0.1848	0.9242	0.0636	**1.0667**
(6.2)	1	0.7225	0.9239	0.8824	0.3080	**1.3737**
	4	0.9400	0.4149	1.0585	0.1596	**1.0638**
	16	0.9750	0.1641	1.0205	0.0897	**1.0256**
(6.3)	1	**0.4250**	0.5118	0.6059	0.0765	0.9882
	4	**0.6650**	0.0564	0.6692	0.0451	0.8722
	16	**0.7525**	0.0233	0.6777	0.0532	0.8605
(6.4)	1	0.7200	0.4757	0.8681	0.0660	**1.1354**
	4	0.9275	0.0647	0.8814	0.0323	**1.0377**
	16	0.9700	0.0180	0.8608	0.0258	**1.0077**

TABLE 5
Performance of Power in Model 7. The highest power is hightlighted in bold. The last four columns refers to the power ratio between the four methods to BD.

Model	ratio	BD	ED/BD	ESS-NN/BD	hotelling/BD	MMD/BD
(7.1)	1	**0.2550**	0.4412	0.7255	0.1569	0.9412
	4	**0.5025**	0.1990	0.4726	0.0199	0.8756
	16	**0.5225**	0.2105	0.6220	0.0096	0.8756
(7.2)	1	**0.5500**	0.4500	0.8045	0.3045	0.9273
	4	**0.8075**	0.3065	0.7523	0.0836	0.9164
	16	**0.8575**	0.3936	0.8367	0.0350	0.9475
(7.3)	1	**0.5950**	0.1681	0.2479	0.1092	0.4328
	4	**0.8800**	0.0824	0.2500	0.0511	0.4318
	16	**0.9475**	0.0580	0.2559	0.0211	0.4855

detected it in data AID1284 but not in data AID439. This suggests that BD performs well in highly imbalanced datasets. Since in Schierz [19], authors suggest that with the best classification model for AID439 and AID1284, the classification accurate can reach at least 77%, confirming that these exist difference between Active observations and Inactive ones both in AID1284 and AID439.

5.2. *Hormone replacement therapy.* Hormone replacement therapy (HRT) is a topic of many studies Andersen et al. [1], Denti [7], Hou et al. [11]. As the second application, we re-analyzed the dataset reported in Dumeaux [8]. The GEO accession number for this dataset is GSE3492 and it was downloaded from `http://www.ncbi.nlm.nih.gov`. This dataset was collected to explore the potential difference in the genome wide expression profiles of 22,153 probes between 47 non HRT users and 42 HRT users. The analysis

TABLE 6
Sample size comparison

Data	No. of Active observations	No. of Inactive observations	No. of Compounds
AID439	11	45	81
AID1284	46	244	103

TABLE 7
P-values from BD, Energy distance, ESS-KNN, Hotelling's T^2 and MMD

Method	BD	ED	ESS-KNN	Hotelling's T^2	MMD
AID439	0.038	0.108	0.074	0.781	0.686
AID1284	0.004	0.008	0.018	0.467	0.131

in Dumeaux [8] ignored the potential correlation in the expression levels among the genes investigated. In our analysis, we consider all gene expressions simultaneously. For convenience, we deleted the missing data, and our working dataset ended up with expression levels from 2,759 genes. Since the dataset contains 89 samples, the statistical inference becomes a $p \gg n$ problem for which the classical Hotelling's T^2 has little power (Bai and Saranadasa [2]).

Table 8 presents the p-values by using the five tests. We can see that the BD reveals a significant difference in the gene expression profiles between the HRT group and non HRT group, whereas the other tests failed to detect a significant difference. It is worth mentioning that the original analysis [8] did not detect any significant difference in gene expression profiles between the HRT and non-HRT groups, either.

TABLE 8
P-values of the comparison between HRT and non HRT from BD, Energy distance, ESS-NN, Hotelling T and MMD

Method	BD	ED	ESS-NN	Hotelling's T^2	MMD
p-value	0.025	0.477	0.568	0.164	0.313

To further evaluate the distribution distinction between HRT and no HRT group, we select the following two probes that show significant difference with the p-value of BD < 0.005, and plot the histogram of the two probes as below.

From Figure 1, we find that the distribution of the 5771-th probe in HRT group is symmetric with a sharp peak whereas that in non-HRT group is also

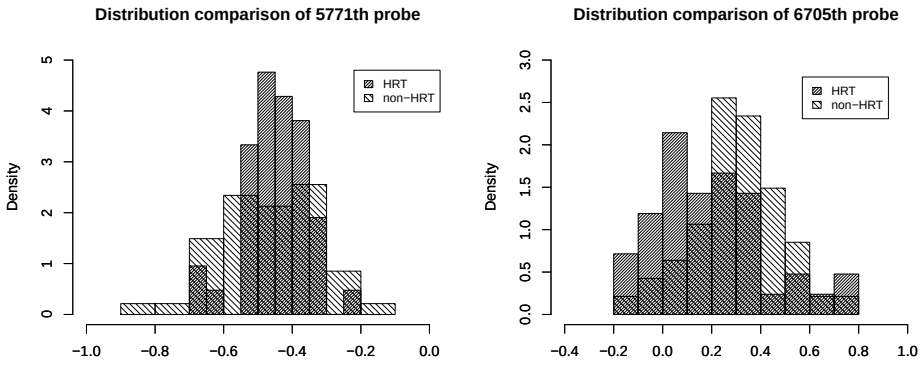

FIG 1. *The distribution distinctions of the selected probes between HRT and non-HRT group.*

symmetric but more smooth. This indicates that the marginal distributions of the probes in HRT and non-HRT group may have similar mean values but the variance is different. The distribution of the 6705-th probe in HRT group is positively skewed whereas in non-HRT group is negatively skewed. Their means are different. Both of the graphics of these two probes result in distribution distinction and prove the reasonability of BD test.

6. Conclusion. The two-sample Cramér-Von Mises criterion measures the difference between two univariate probability distributions. Developing such measures to compare multivariate probability distributions is challenging. We filled this gap by proposing and investigating Ball Divergence. This breakthrough was possible because the Ball Divergence relies on the difference of the probabilities over balls, whereas Cramér-Von Mises criterion depends directly on the difference in the distribution functions. The Ball Divergence provides a nonparametric measure as it doesn't require any assumption on the moments, and hence is easy and flexible to apply and/or extend.

We not only introduced the BD as a novel concept, but also defined its sample version—the BD based nonparametric two-sample test statistic—for practical use. The BD statistic can be expressed as a function of metric ranks or V-statistics, and as we proved theoretically and demonstrated numerically, the BD statistic can cope with highly imbalanced design. The ability to deal with highly imbalanced designs is an obvious advantage of the BD statistic, while it still performs well for a balanced design, including

high dimension data. Specifically, we showed that the power of the BD test increases as the sample size of the larger group increases while the sample size of the smaller group is fixed.

As a proof of concept, we re-analyzed several datasets that are publicly available and have been used as benchmarks for methodological developments. In a separate effort, we applied the BD for important biomedical applications such as change point detection for time series (Wang et al., preprint [25]). One major implication of the BD test is its potential to address the multiple comparison issue. In the existing literature, a common practice is to examine two-sample differences in individual response variables whether they are gene expressions or drug responses. Correcting the Type-I error rates due to the multiple comparison issue is necessary and often not effective. We have difficulties to identify true signals while unable to eliminate false discoveries. The BD test can be used similarly to the omnibus in MANOVA by providing a global p-value for all responses.

Appendix.

Proofs of Asymptotic Properties.

THE PROOF OF THEOREM 3. It can be verified that

$$\begin{aligned}
A_{n,m} =& \frac{1}{n^2}\sum_{i,j=1}^{n}(A_{ij}^X - A_{ij}^Y)^2 \\
=& \frac{1}{n^4m^2}\sum_{i,j,u,u'=1}^{n}\sum_{v,v'=1}^{m}\Big\{ \\
& \delta(X_i,X_j,X_u)\cdot\delta(X_i,X_j,X_{u'}) + \delta(X_i,X_j,Y_v)\cdot\delta(X_i,Y_j,Y_{v'}) \\
& - \delta(X_i,X_j,X_u)\cdot\delta(X_i,X_j,Y_v) - \delta(X_i,X_j,X_{u'})\cdot\delta(X_i,X_j,Y_{v'})\Big\} \\
=& \frac{1}{n^4m^2}\sum_{i,j,u,u'=1}^{n}\sum_{v,v'=1}^{m}\big\{\xi(X_i,X_j,X_u,X_{u'}) + \xi(X_i,X_j,Y_v,Y_{v'}) \\
& - \xi(X_i,X_j,X_u,Y_v) - \xi(X_i,X_j,X_{u'},Y_{v'})\big\}.
\end{aligned}$$

According to the definition of $A_{n,m}$, we denote the kernel of $A_{n,m}$ by $\psi_A(X_i,X_j,X_u,X_{u'};Y_v,Y_{v'})$:

$$\begin{aligned}
& \psi_A(X_i,X_j,X_u,X_{u'};Y_v,Y_{v'}) \\
=& \xi(X_i,X_j,X_u,X_{u'}) + \xi(X_i,X_j,Y_v,Y_{v'}) \\
& - \xi(X_i,X_j,X_u,Y_v) - \xi(X_i,X_j,X_{u'},Y_{v'}).
\end{aligned}$$

Similarly, denote the kernel of $C_{n,m}$ by $\psi_C(X_u, X_{u'}; Y_k, Y_l, Y_v, Y_{v'})$:

$$\begin{aligned}&\psi_C(X_u, X_{u'}; Y_k, Y_l, Y_v, Y_{v'})\\ =&\xi(Y_k, Y_l, Y_v, Y_{v'}) + \xi(Y_k, Y_l, X_u, X_{u'})\\ &- \xi(Y_k, Y_l, Y_v, X_u) - \xi(Y_k, Y_l, Y_{v'}, X_{u'}).\end{aligned}$$

We easily have $E(|\psi_A|) \le 4 < \infty$, $E(|\psi_C|) \le 4 < \infty$ and

$$\begin{aligned}&E\psi_A(X_1, X_2, X_3, X; Y_3, Y)\\ =&E[E(\psi_A(X_1, X_2, X_3, X; Y_3, Y)|X_1, X_2)]\\ =&E[E(\delta(X_1, X_2, X)|X_1, X_2)E(\delta(X_1, X_2, X_3)|X_1, X_2)\\ &+ E(\delta(X_1, X_2, Y)|X_1, X_2)E(\delta(X_1, X_2, Y_3)|X_1, X_2)\\ &- 2E(\delta(X_1, X_2, X)|X_1, X_2)E(\delta(X_1, X_2, Y)|X_1, X_2)]\\ =&E[E^2(\delta(X_1, X_2, X)|X_1, X_2) + E^2(\delta(X_1, X_2, Y)|X_1, X_2)\\ &- 2E(\delta(X_1, X_2, X)|X_1, X_2)E(\delta(X_1, X_2, Y)|X_1, X_2)]\\ =&E[E(\delta(X_1, X_2, X)|X_1, X_2) - E(\delta(X_1, X_2, Y)|X_1, X_2)]^2\\ =&\iint_{\mathrm{V}\times\mathrm{V}} [\mu - \nu]^2(\bar{B}(u, \rho(u, v)))\mu(du)\mu(dv)\\ =&A.\end{aligned}$$

Similarly, we have

$$\begin{aligned}&E\psi_C(X_3, X; Y_1, Y_2, Y_3, Y)\\ =&E[E(\delta(Y_1, Y_2, X)|Y_1, Y_2) - E(\delta(Y_1, Y_2, Y)|Y_1, Y_2)]^2\\ =&\iint_{\mathrm{V}\times\mathrm{V}} [\mu - \nu]^2(\bar{B}(u, \rho(u, v)))\nu(du)\nu(dv)\\ =&C.\end{aligned}$$

According to Theorem 3 (Lee [15], P.122), we have both $A_{n,m}$ and $C_{n,m}$ converge a.s. to A and C respectively. Due to the definition of $D_{n,m}$ and $D(\mu, \nu)$, we have

$$D_{n,m} \xrightarrow[n,m\to\infty]{a.s.} D(\mu, \nu).$$

□

The proof of Theorem 4. From the definition of $D_{n,m}$, we know that it is the sum of $A_{n,m}$ and $C_{n,m}$. Thus, $D_{n,m}$ is not a V-statistic and we can't apply the V-statistic theory to $D_{n,m}$ directly. Our proof is divided

into three parts. First, we decompose $A_{n,m}$ into the sum of U statistic with different degrees. The distribution of $A_{n,m}$ is determined by the U statistic with the same degree as $A_{n,m}$. Second, we can apply H-decomposition(Lee [15], Section 1.6) to the U statistic and obtain the primary component which can determine the distribution of the U statistic. Further, we deal with $C_{n,m}$ similarly. Third, we combine the primary component of $A_{n,m}$ and $C_{n,m}$ and obtain the asymptotic distribution in null hypothesis in different cases.

Step 1: Following the same step of the aforementioned proof, we begin with $A_{n,m}$ and then $C_{n,m}$.

Firstly, we symmetrize the kernel of $\psi_A(X_i, X_j, X_u, X_{u'}; Y_v, Y_{v'})$ and get

$$\begin{aligned}&\psi_A^s(X_i, X_j, X_u, X_{u'}; Y_v, Y_{v'})\\ &\quad= \frac{1}{4!2!}\sum_{\tau\in\pi(i,j,u,u')}\sum_{\gamma\in\pi(v,v')}\psi_A(X_{\tau_{(1)}}, X_{\tau_{(2)}}, X_{\tau_{(3)}}, X_{\tau_{(4)}}; Y_{\gamma_{(1)}}, Y_{\gamma_{(2)}}),\end{aligned}$$

where $\pi(i,j,u,u')$ and $\pi(v,v')$ are the permutations of $\{i,j,u,u'\}$ and $\{v,v'\}$, respectively. Since the kernel $\psi_A^s(X_i, X_j, X_u, X_{u'}; Y_v, Y_{v'})$ is symmetric, the corresponding V-statistic should be

$$A_{n,m} = \frac{1}{n^4m^2}\sum_{i,j,u,u'=1}^{n}\sum_{v,v'=1}^{m}\psi_A^s(X_i, X_j, X_u, X_{u'}; Y_v, Y_{v'}).$$

$A_{n,m}$ is a two-sample V-statistic of order $(4,2)$. Thus, it can be decomposed into the sum of U-statistic with different degrees, and the decomposition is as follow

$$\begin{aligned}A_{n,m} =&\frac{1}{n^4m^2}\bigg(48\binom{n}{4}\binom{m}{2}\hat{U}_A^{(4,2)} + 24\binom{n}{4}\binom{m}{1}\hat{U}_A^{(4,1)} + 72\binom{n}{3}\binom{m}{2}\hat{U}_A^{(3,2)}\\ &+ 36\binom{n}{3}\binom{m}{1}\hat{U}_A^{(3,1)} + 28\binom{n}{2}\binom{m}{2}\hat{U}_A^{(2,2)} + 14\binom{n}{2}\binom{m}{1}\hat{U}_A^{(2,1)}\\ &+ 2\binom{n}{1}\binom{m}{2}\hat{U}_A^{(1,2)} + \binom{n}{1}\binom{m}{1}\hat{U}_A^{(1,1)}\bigg).\end{aligned}$$

Then

$$\begin{aligned}A_{n,m} =&\frac{1}{n^4m^2}\bigg(48\binom{n}{4}\binom{m}{2}\hat{U}_A^{(4,2)} + 24\binom{n}{4}\binom{m}{1}\hat{U}_A^{(4,1)}\\ &+ 72\binom{n}{3}\binom{m}{2}\hat{U}_A^{(3,2)}\bigg) + O_p\Big(\frac{1}{n^2} + \frac{1}{nm}\Big),\end{aligned} \tag{6.1}$$

where

$$\hat{U}_A^{(4,2)} = \frac{1}{\binom{n}{4}\binom{m}{2}}\sum_{i<j<u<u'}^{n}\sum_{v<v'}^{m}\psi_A^s(X_i, X_j, X_u, X_{u'}; Y_v, Y_{v'})$$

$$=\frac{1}{\binom{n}{4}\binom{m}{2}}\sum_{i<j<u<u'}^{n}\sum_{v<v'}^{m}\frac{1}{4!2!}\sum_{\tau\in\pi(i,j,u,u')}\sum_{\gamma\in\pi(v,v')}\psi_A(X_{\tau_{(1)}},X_{\tau_{(2)}},X_{\tau_{(3)}},X_{\tau_{(4)}};Y_{\gamma_{(1)}},Y_{\gamma_{(2)}}).$$

Step 2: Now we deal with the distribution of $\hat{U}_A^{(4,2)}$. Because $\hat{U}_A^{(4,2)}$ is a two-sample U-statistic with degrees $(4,2)$ and kernel $\psi_A^s(X_i,X_j,X_u,X_{u'};Y_v,Y_{v'})$, we can apply H-decomposition (Lee [15], Section 1.6) to $\hat{U}_A^{(4,2)}$:

(6.2)
$$\begin{aligned}\hat{U}_A^{(4,2)}=&\sum_{c=0}^{4}\sum_{d=0}^{2}\binom{4}{c}\binom{2}{d}H_A^{(c,d)}\\=&\sum_{c=0}^{4}\sum_{d=0}^{2}\binom{4}{c}\binom{2}{d}\frac{1}{\binom{n}{c}\binom{m}{d}}\sum_{(n,c)}\sum_{(m,d)}h_A^{(c,d)}(X_{i_1},\ldots,X_{i_c};Y_{j_1},\ldots,Y_{j_d}).\end{aligned}$$

(6.2) can be simplified as

$$\begin{aligned}\hat{U}_A^{(4,2)}=&\frac{1}{\binom{n}{2}\binom{m}{2}}\sum_{u<u'}^{n}\sum_{v<v'}^{m}Q_A(X_u,X_{u'};Y_v,Y_{v'})+R_n\\=&\frac{1}{\binom{n}{2}\binom{m}{2}}\sum_{u<u'}^{n}\sum_{v<v'}^{m}\\&\big(\phi_A^{(2,0)}(X_u,X_{u'})+\phi_A^{(1,1)}(X_u,Y_v)+\phi_A^{(1,1)}(X_{u'},Y_{v'})\\&+\phi_A^{(0,2)}(Y_v,Y_{v'})\big)+R_n\\=&\frac{4}{n(n-1)m(m-1)}\sum_{u<u'}^{n}\sum_{v<v'}^{m}\\&\big(\phi_A^{(2,0)}(X_u,X_{u'})+\phi_A^{(1,1)}(X_u,Y_v)+\phi_A^{(1,1)}(X_{u'},Y_{v'})\\&+\phi_A^{(0,2)}(Y_v,Y_{v'})\big)+R_n,\end{aligned}$$

where

$$\begin{aligned}&{\phi_A}^{(2,0)}(x,x')\\&=E[\xi(X_1,X_2,x,x')]+E[\xi(X_1,X_2,Y,Y_3)]\\&\quad-E[\xi(X_1,X_2,x,Y)]-E[\xi(X_1,X_2,x',Y_3)],\\&{\phi_A}^{(1,1)}(x,y)\\&=E[\xi(X_1,X_2,x,X_3)]+E[\xi(X_1,X_2,y,Y_3)]\end{aligned}$$

$$
\begin{aligned}
&- E[\xi(X_1, X_2, x, y)] - E[\xi(X_1, X_2, X_3, Y_3)],\\
&\phi_A{}^{(0,2)}(y, y')\\
&=E[\xi(X_1, X_2, X, X_3)] + E[\xi(X_1, X_2, y, y')]\\
&\quad - E[\xi(X_1, X_2, X, y)] - E[\xi(X_1, X_2, X, y')],
\end{aligned}
$$

$$
R_n = \sum_{c+d\geq 3} h_A^{(c,d)} \quad and \quad VarR_n = O(\frac{1}{n^c m^d}), c+d=3.
$$

On the other hand, we consider $C_{n,m}$ and get the corresponding U-statistic as

$$
\begin{aligned}
\hat{U}_C^{(2,4)} =&\frac{4}{n(n-1)m(m-1)} \sum_{u<u'}^{n} \sum_{v<v'}^{m} Q_C(X_u, X_{u'}; Y_v, Y_{v'}) + R_n,\\
=&\frac{4}{n(n-1)m(m-1)} \sum_{u<u'}^{n} \sum_{v<v'}^{m}\\
&\big(\phi_C^{(2,0)}(X_u, X_{u'}) + \phi_C^{(1,1)}(X_u, Y_v) + \phi_C^{(1,1)}(X_{u'}, Y_{v'})\\
&+ \phi_C^{(0,2)}(Y_v, Y_{v'})\big) + R_n,
\end{aligned}
$$

where

$$
\begin{aligned}
&\phi_C{}^{(2,0)}(x, x')\\
&=E[\xi(Y_1, Y_2, x, x')] + E[\xi(Y_1, Y_2, Y, Y_3)]\\
&\quad - E[\xi(Y_1, Y_2, x, Y)] - E[\xi(Y_1, Y_2, x', Y_3)],\\
&\phi_C{}^{(1,1)}(x, y)\\
&=E[\xi(Y_1, Y_2, x, X_3)] + E[\xi(Y_1, Y_2, y, Y_3)]\\
&\quad - E[\xi(Y_1, Y_2, x, y)] - E[\xi(Y_1, Y_2, X_3, Y_3)],\\
&\phi_C{}^{(0,2)}(y, y')\\
&=E[\xi(Y_1, Y_2, X, X_3)] + E[\xi(Y_1, Y_2, y, y')]\\
&\quad - E[\xi(Y_1, Y_2, X, y)] - E[\xi(Y_1, Y_2, X, y')],
\end{aligned}
$$

$$
R_n = \sum_{c+d\geq 3} h_C^{(c,d)} \quad and \quad VarR_n = O(\frac{1}{n^c m^d}), c+d=3.
$$

Step 3: Under the null hypothesis, Q_C actually has the following relation with Q_A:

$$
Q_C(x, y; x', y') = Q_A(x, y; x', y').
$$

Thus, representing $Q(x,y;x',y') = Q_C(x,y;x',y') = Q_A(x,y;x',y')$ under the null hypothesis, we can express $D_{n,m}$ as

$$\begin{aligned}
D_{n,m} =& A_{n,m} + C_{n,m} \\
=& \frac{1}{n^4m^2}\left(48\binom{n}{4}\binom{m}{2}\hat{U}_A^{(4,2)} + 24\binom{n}{4}\binom{m}{1}\hat{U}_A^{(4,1)} + 72\binom{n}{3}\binom{m}{2}\hat{U}_A^{(3,2)}\right) \\
&+ \frac{1}{n^2m^4}\left(48\binom{n}{2}\binom{m}{4}\hat{U}_C^{(2,4)} + 24\binom{n}{1}\binom{m}{4}\hat{U}_C^{(1,4)} + 72\binom{n}{2}\binom{m}{3}\hat{U}_C^{(2,3)}\right) \\
&+ O_p\Big(\frac{1}{n^2} + \frac{1}{nm} + \frac{1}{m^2}\Big) \\
=& \frac{2}{\binom{n}{2}\binom{m}{2}}\sum_{u<u'}^{n}\sum_{v<v'}^{m} Q(X_u, X_{u'}; Y_v, Y_{v'}) + \frac{1}{n^4m^2}\left(24\binom{n}{4}\binom{m}{1}\hat{U}_A^{(4,1)}\right. \\
&\left. + 72\binom{n}{3}\binom{m}{2}\hat{U}_A^{(3,2)}\right) + \frac{1}{n^2m^4}\left(24\binom{n}{1}\binom{m}{4}\hat{U}_C^{(1,4)} + 72\binom{n}{2}\binom{m}{3}\hat{U}_C^{(2,3)}\right) \\
&+ O_p\Big(\frac{1}{n^2} + \frac{1}{nm} + \frac{1}{m^2}\Big).
\end{aligned}$$

Suppose that n and $m \to \infty$ in such a way that $n/(n+m) \to \tau$. According to Theorem 1.1 of [16], $Q(x,y;x',y')$ have the following spectral decomposition

$$Q(x,y;x',y') = \sum_{k=1}^{\infty} \lambda_k f_k(x,y) f_k(x',y').$$

Thus,

$$\begin{aligned}
&\left(\frac{1}{n} + \frac{1}{m}\right)^{-1} \frac{2}{\binom{n}{2}\binom{m}{2}}\sum_{u<u'}^{n}\sum_{v<v'}^{m} Q(X_u, X_{u'}; Y_v, Y_{v'}) \\
\xrightarrow[n\to\infty]{d}& \sum_{k=1}^{\infty} 2\lambda_k[(a_k(\tau)Z_{1k} + b_k(\tau)Z_{2k})^2 - (a_k^2(\tau) + b_k^2(\tau))].
\end{aligned}$$

where $a_k^2(\tau) = (1-\tau)E_X[E_Y f_k(X,Y)]^2$, $b_k^2(\tau) = \tau E_Y[E_X f_k(X,Y)]^2$ and Z_{1k}, Z_{2k} are $i.i.d. N(0,1)$, $k = 1, 2, \cdots$.

Moreover, since

$$\hat{U}_A^{(4,1)} \xrightarrow[n,m\to\infty]{a.s.} E\hat{U}_A^{(4,1)}, \qquad \hat{U}_A^{(3,2)} \xrightarrow[n,m\to\infty]{a.s.} E\hat{U}_A^{(3,2)},$$

$$\hat{U}_C^{(1,4)} \xrightarrow[n,m\to\infty]{a.s.} E\hat{U}_C^{(1,4)}, \qquad \hat{U}_C^{(2,3)} \xrightarrow[n,m\to\infty]{a.s.} E\hat{U}_C^{(2,3)}.$$

Under the null hypothesis,

$$\begin{aligned}&E\hat{U}_A^{(4,1)} = E\hat{U}_C^{(1,4)} = 6E\hat{U}_A^{(3,2)} = 6E\hat{U}_C^{(2,3)}\\&= E[E(\delta(X_1, X_2, X)(1 - (\delta(X_1, X_2, Y)))|X_1, X_2].\end{aligned}$$

Let $\theta = 2E[E(\delta(X_1, X_2, X)(1 - \delta(X_1, X_2, Y))|X_1, X_2]$. By use of Slutsky Theorem, we have

$$\frac{nm}{n+m} D_{n,m} \xrightarrow[n\to\infty]{d} \sum_{k=1}^{\infty} 2\lambda_k[(a_k(\tau)Z_{1k} + b_k(\tau)Z_{2k})^2 - (a_k^2(\tau) + b_k^2(\tau))] + \theta.$$

□

The proof of Theorem 5. Following the V-statistic decomposition (6.1) and H decomposition (6.2), we have

$$\begin{aligned}D_{n,m} =& A_{n,m} + C_{n,m}\\=& \frac{48}{n^4m^2}\binom{n}{4}\binom{m}{2}\hat{U}_A^{(4,2)} + \frac{48}{n^2m^4}\binom{n}{2}\binom{m}{4}\hat{U}_C^{(2,4)} + O_p(\frac{1}{n} + \frac{1}{m})\\=& \frac{48}{n^4m^2}\binom{n}{4}\binom{m}{2}\sum_{c=0}^{4}\sum_{d=0}^{2}\binom{4}{c}\binom{2}{d}H_A^{(c,d)}\\&+ \frac{48}{n^2m^4}\binom{n}{2}\binom{m}{4}\sum_{c=0}^{2}\sum_{d=0}^{4}\binom{2}{c}\binom{4}{d}H_C^{(c,d)} + O_p(\frac{1}{n} + \frac{1}{m}).\end{aligned}$$

Further, we obtain that

(6.3)

$$\begin{aligned}D_{n,m} =& \sum_{c=0}^{4}\sum_{d=0}^{2}\binom{4}{c}\binom{2}{d}H_A^{(c,d)} + \sum_{c=0}^{2}\sum_{d=0}^{4}\binom{2}{c}\binom{4}{d}H_C^{(c,d)}\\&- (\frac{6}{n} + \frac{1}{m})\sum_{c=0}^{4}\sum_{d=0}^{2}\binom{4}{c}\binom{2}{d}H_A^{(c,d)} - (\frac{1}{n} + \frac{6}{m})\sum_{c=0}^{2}\sum_{d=0}^{4}\binom{2}{c}\binom{4}{d}H_C^{(c,d)}\\&+ O_p(\frac{1}{n} + \frac{1}{m})\\=& \sum_{c=0}^{4}\sum_{d=0}^{2}\binom{4}{c}\binom{2}{d}H_A^{(c,d)} + \sum_{c=0}^{2}\sum_{d=0}^{4}\binom{2}{c}\binom{4}{d}H_C^{(c,d)} + O_p(\frac{1}{n} + \frac{1}{m}).\end{aligned}$$

Since the order of $\sum_{c+d\geq 2} C_4^c C_2^d H_A^{(c,d)}$ and $\sum_{c+d\geq 2} C_2^c C_4^d H_C^{(c,d)}$ are higher

than $O_p(\frac{1}{n}+\frac{1}{m})$. We can simplify (6.3) as

(6.4)

$$\begin{aligned}
D_{n,m} =& A + n^{-1}\sum_{u=1}^{n}(h_{A,1}^{(1,0)}(X_u)+h_{A,2}^{(1,0)}(X_u)+h_{A,3}^{(1,0)}(X_u)+h_{A,4}^{(1,0)}(X_u))\\
&+ m^{-1}\sum_{v=1}^{m}(h_{A,1}^{(0,1)}(Y_v)+h_{A,2}^{(0,1)}(Y_v)) + \sum_{c+d\geq 2} C_4^c C_2^d H_A^{(c,d)}\\
&+ C + n^{-1}\sum_{u=1}^{n}(h_{C,1}^{(1,0)}(X_u)+h_{C,2}^{(1,0)}(X_u))\\
&+ m^{-1}\sum_{v=1}^{m}(h_{C,1}^{(0,1)}(Y_v)+h_{C,2}^{(0,1)}(Y_v)+h_{C,3}^{(0,1)}(Y_v)+h_{C,4}^{(0,1)}(Y_v))\\
&+ \sum_{c+d\geq 2} C_2^c C_4^d H_C^{(c,d)} + O_p(\frac{1}{n}+\frac{1}{m})\\
=& A + C + n^{-1}\sum_{u=1}^{n}(h_{A,1}^{(1,0)}(X_u)+h_{A,2}^{(1,0)}(X_u)+h_{A,3}^{(1,0)}(X_u)+h_{A,4}^{(1,0)}(X_u)\\
&+ h_{C,1}^{(1,0)}(X_u)+h_{C,2}^{(1,0)}(X_u)) + m^{-1}\sum_{v=1}^{m}(h_{A,1}^{(0,1)}(Y_v)+h_{A,2}^{(0,1)}(Y_v)+h_{C,1}^{(0,1)}(Y_v)\\
&+ h_{C,2}^{(0,1)}(Y_v)+h_{C,3}^{(0,1)}(Y_v)+h_{C,4}^{(0,1)}(Y_v)) + O_p(\frac{1}{n}+\frac{1}{m}),
\end{aligned}$$

where

$$\begin{aligned}
h_{A,i}^{(1,0)}(x) =& E(\psi_A(X_1,X_2,X_3,X_4;Y_1,Y_2)|X_i=x) - A, i=1,\cdots,4,\\
h_{A,i}^{(0,1)}(y) =& E(\psi_A(X_1,X_2,X_3,X_4;Y_1,Y_2)|Y_i=y) - A, i=1,2,\\
h_{C,i}^{(1,0)}(x) =& E(\psi_C(X_1,X_2;Y_1,Y_2,Y_3,Y_4)|X_i=x) - C, i=1,2,\\
h_{C,i}^{(0,1)}(y) =& E(\psi_C(X_1,X_2;Y_1,Y_2,Y_3,Y_4)|Y_i=y) - C, i=1,\cdots,4.
\end{aligned}$$

Denote

$$\begin{aligned}
g^{(1,0)}(X_u) =& h_{A,1}^{(1,0)}(X_u)+h_{A,2}^{(1,0)}(X_u)+h_{A,3}^{(1,0)}(X_u)+h_{A,4}^{(1,0)}(X_u)\\
&+ h_{C,1}^{(1,0)}(X_u)+h_{C,2}^{(1,0)}(X_u),\\
g^{(0,1)}(Y_v) =& h_{A,1}^{(0,1)}(Y_v)+h_{A,2}^{(0,1)}(Y_v)+h_{C,1}^{(0,1)}(Y_v)+h_{C,2}^{(0,1)}(Y_u)\\
&+ h_{C,3}^{(0,1)}(Y_v)+h_{C,4}^{(0,1)}(Y_v),
\end{aligned}$$

and

$$\delta_{1,0}^2 = Var(g^{(1,0)}(X_u)) \quad and \quad \delta_{0,1}^2 = Var(g^{(0,1)}(Y_v)).$$

Then both $n^{-\frac{1}{2}}\sum_{u=1}^{n} g^{(1,0)}(X_u)$ and $m^{-\frac{1}{2}}\sum_{v=1}^{m} g^{(0,1)}(Y_v)$ converge to normal distributions with mean zero and variances $\delta_{1,0}^2$ and $\delta_{0,1}^2$, respectively. Suppose that n and $m \to \infty$ in such a way that $n/(n+m) \to \tau$. We have

$$\begin{aligned}&\sqrt{\frac{nm}{n+m}}(D_{n,m} - D(\mu,\nu))\\ =&\sqrt{\frac{m}{(n+m)n}}\sum_{u=1}^{n} g^{(1,0)}(X_u) + \sqrt{\frac{n}{(n+m)m}}\sum_{v=1}^{m} g^{(0,1)}(Y_v) + o_p(1).\end{aligned}$$

Since $\sum_{u=1}^{n} g^{(1,0)}(X_u)$ and $\sum_{v=1}^{m} g^{(0,1)}(Y_v)$ are independent, it follows that

$$\sqrt{\frac{nm}{n+m}}(D_{n,m} - D(\mu,\nu)) \xrightarrow[n\to\infty]{d} N(0, (1-\tau)\delta_{1,0}^2 + \tau\delta_{0,1}^2).$$

□

The proof of Theorem 6. According to the V-statistics decomposition (6.1), we have

$$\begin{aligned}D_{n,m} &= A_{n,m} + C_{n,m}\\ &= \frac{48}{n^4m^2}\binom{n}{4}\binom{m}{2}\hat{U}_A^{(4,2)} + \frac{48}{n^2m^4}\binom{n}{2}\binom{m}{4}\hat{U}_C^{(2,4)} + O_p(\frac{1}{n}+\frac{1}{m}),\end{aligned}$$

where $\hat{U}_A^{(4,2)}$ and $\hat{U}_C^{(2,4)}$ are the U-statistics with kernels ψ_A^s and ψ_C^s, respectively. That is,

$$\hat{U}_A^{(4,2)} = \frac{1}{\binom{n}{4}\binom{m}{2}}\sum_{i<j<u<u'}^{n}\sum_{v<v'}^{m}\psi_A^s(X_i, X_j, X_u, X_{u'}; Y_v, Y_{v'}),$$

and

$$\hat{U}_C^{(2,4)} = \frac{1}{\binom{n}{2}\binom{m}{4}}\sum_{u<u'}^{n}\sum_{i<j<v<v'}^{m}\psi_C^s(X_u, X_{u'}; Y_i, Y_j, Y_v, Y_{v'}).$$

According to (6.3) and (6.4), we have

$$Var(D_{n,m}) = \frac{\delta_{1,0}^2}{n} + \frac{\delta_{0,1}^2}{m} + O(\frac{1}{n}+\frac{1}{m}).$$

Thus, we can obtain that $\lim_{n\to\infty} Var_{H_1}(D_{n,m}) = 0$.

Since the kernel ψ_A^s is the symmetrized form of ψ_A, they actually share the same expectation and we thus can calculate the expectation of $\hat{U}_A^{(4,2)}$ with

$$\begin{aligned}
E\hat{U}_A^{(4,2)} =& \frac{1}{\binom{n}{4}\binom{m}{2}} \sum_{i<j<u<u'}^{n} \sum_{v<v'}^{m} E\psi_A^s(X_1, X_2, X_3, X; Y_3, Y) \\
=& E\psi_A^s(X_1, X_2, X_3, X; Y_3, Y) \\
=& E\psi_A(X_1, X_2, X_3, X; Y_3, Y) \\
=& E\big[E(\psi_A(X_1, X_2, X_3, X; Y_3, Y)|X_1, X_2)\big] \\
=& E\big[E(\delta(X_1, X_2, X)|X_1, X_2) - E(\delta(X_1, X_2, Y)|X_1, X_2)\big]^2 \\
=& A.
\end{aligned}$$

Similarly, we can also figure out the expectation of $\hat{U}_C^{(2,4)}$ by

$$\begin{aligned}
E\hat{U}_C^{(2,4)} =& \frac{1}{\binom{n}{2}\binom{m}{4}} \sum_{u<u'}^{n} \sum_{i<j<v<v'}^{m} E\psi_C^s(X_3, X; Y_1, Y_2, Y_3, Y) \\
=& E\big[E(\psi_C(X_3, X; Y_1, Y_2, Y_3, Y)|Y_1, Y_2)\big] \\
=& E\big[E(\delta(Y_1, Y_2, X)|Y_1, Y_2) - E(\delta(Y_1, Y_2, Y)|Y_1, Y_2)\big]^2 \\
=& C.
\end{aligned}$$

With the above equations, the expectation of $D_{n,m}$ can be expressed as

$$\begin{aligned}
ED_{n,m} =& \frac{48}{n^4m^2}\binom{n}{4}\binom{m}{2}E\hat{U}_A^{(4,2)} + \frac{48}{n^2m^4}\binom{n}{2}\binom{m}{4}E\hat{U}_C^{(2,4)} + O_p(\frac{1}{n} + \frac{1}{m}) \\
=& \frac{(n-1)(n-2)(n-3)(m-1)}{n^3m}A + \frac{(n-1)(m-1)(m-2)(m-3)}{nm^3}C \\
&+ O(\frac{1}{n} + \frac{1}{m}).
\end{aligned}$$

As $n \to \infty$ and $n \le m$, we therefore have

$$\begin{aligned}
\liminf_{n\to\infty} ED_{n,m} =& \liminf_{n\to\infty}(\frac{(n-1)(n-2)(n-3)(m-1)}{n^3m})A \\
&+ \liminf_{n\to\infty}(\frac{(n-1)(m-1)(m-2)(m-3)}{nm^3})C \\
=& A + C \\
=& D(\mu, \nu).
\end{aligned}$$

According to Theorem 1, we know that $D(\mu,\nu) \geq 0$ always holds and the equality will be reached if and only if $\mu = \nu$. That is, $D(\mu,\nu)$ will equal to 0 only when null hypothesis H_0 holds while under alternative hypothesis H_1, we always have $D(\mu,\nu) > 0$. Hence, it can be seen that

$$\begin{aligned}\Delta(\eta) =& \liminf_{n\to\infty}(E_{H_1}D_{n,m} - E_{H_0}D_{n,m})\\ =& D(\mu,\nu)|_{H_1} - D(\mu,\nu)|_{H_0}\\ =& D(\mu,\nu)|_{H_1} > 0,\end{aligned}$$

where the value of $\Delta(\eta)$ will be only affected by the alternative distributions while independent of sample size ratio η. □

Computational Complexity. The computational complexity of BD is $O(n^3 + m^3)$ if we compute it exactly from the definition of BD:

$$D_{n,m} = A_{n,m} + C_{n,m},$$

where

$$\begin{aligned}A_{n,m} &= \frac{1}{n^2}\sum_{i,j=1}^{n}(A_{ij}^X - A_{ij}^Y)^2,\ C_{n,m} = \frac{1}{m^2}\sum_{k,l=1}^{m}(C_{kl}^X - C_{kl}^Y)^2,\\ A_{ij}^X &= \frac{1}{n}\sum_{u=1}^{n}\delta(X_i, X_j, X_u),\ A_{ij}^Y = \frac{1}{m}\sum_{v=1}^{m}\delta(X_i, X_j, Y_v),\\ C_{kl}^X &= \frac{1}{n}\sum_{u=1}^{n}\delta(Y_k, Y_l, X_u),\ C_{kl}^Y = \frac{1}{m}\sum_{v=1}^{m}\delta(Y_k, Y_l, Y_v).\end{aligned}$$

However, we could reduce its computational complexity by some fast sorting algorithms because $nA_{ij}^X, mA_{ij}^Y, nC_{kl}^X$ and mC_{kl}^Y are some ranks. For example, nA_{ij}^X is the rank of $\rho(X_i, X_j)$ among $\{\rho(X_i, X_u), u = 1, \cdots, n\}$ and $nA_{ij}^X + mA_{ij}^Y$ is the rank of $\rho(X_i, X_j)$ among $\{\rho(X_i, X_u), u = 1, \cdots, n\} \cup \{\rho(X_i, Y_v), v = 1, \cdots, m\}$.

Given two independent samples $\mathcal{X}_n = \{X_1, \ldots, X_n\}$ and $\mathcal{Y}_m = \{Y_1, \ldots, Y_m\}$, let $\mathcal{Z} = \mathcal{X}_n \cup \mathcal{Y}_m$. The following algorithm reduces the computational complexity to $O(n^2 \log n + m^2 \log m)$.

Step 1) Calculate the pairwise distances between the points of $\mathcal{Z}$ to get the distance matrix

$$D_{\mathcal{Z}\mathcal{Z}} = \begin{pmatrix} D_{\mathcal{X}\mathcal{X}} & D_{\mathcal{X}\mathcal{Y}} \\ D_{\mathcal{Y}\mathcal{X}} & D_{\mathcal{Y}\mathcal{Y}} \end{pmatrix}.$$

Step 2) Rank $D_{\mathcal{X}\mathcal{X}}$ row by row to get the modified competition ranking matrix $R_{\mathcal{X}}$ in which the (i,j) entry is nA_{ij}^X; Get $R_{\mathcal{Y}}$ similarly.

Step 3) Rank $D_{\mathcal{Z}\mathcal{Z}}$ row by row to get the modified competition ranking matrix

$$R_{\mathcal{Z}\mathcal{Z}} = \begin{pmatrix} R_{\mathcal{X}\mathcal{X}} & R_{\mathcal{X}\mathcal{Y}} \\ R_{\mathcal{Y}\mathcal{X}} & R_{\mathcal{Y}\mathcal{Y}} \end{pmatrix},$$

in which, the (i, j) entry of $R_{\mathcal{X}\mathcal{X}}$ is $nA^X_{ij} + mA^Y_{ij}$ and the (k, l) entry of $R_{\mathcal{Y}\mathcal{Y}}$ is $nC^X_{kl} + mC^Y_{kl}$.

Step 4) Calculate $A_{n,m}$ and $C_{n,m}$, and then $D_{n,m}$.

References.

[1] Andersen, L., Friis, S., Hallas, J., Ravn, P., Schrøder, H. D. and Gaist, D. (2014). Hormone Replacement Therapy Increases The Risk Of Cranial Meningioma (P3. 325). *Neurology* **82** P3–325.

[2] Bai, Z. D. and Saranadasa, H. (1996). Effect of high dimension: by an example of a two sample problem. *Statistica Sinica* **6** 311–329.

[3] Bogachev, V. I. (2007). *Measure Theory Volume I*. Springer.

[4] Chen, L., Dou, W. W. and Qiao, Z. (2013). Ensemble Subsampling for Imbalanced Multivariate Two-Sample Tests. *Journal of the American Statistical Association* just-accepted.

[5] Chen, S. X. and Qin, Y.-L. (2010). A two-sample test for high-dimensional data with applications to gene-set testing. *The Annals of Statistics* **38** 808–835.

[6] Chiu, S. N. and Liu, K. I. (2009). Generalized Cramér-Von Mises goodness-of-fit tests for multivariate distributions. *Computational Statistics & Data Analysis* **53** 3817 - 3834.

[7] Denti, L. (2009). The hormone replacement therapy (HRT) of menopause: focus on cardiovascular implications. *Acta bio-medica: Atenei Parmensis* **81** 73–76.

[8] Dumeaux, V. (2006). Gene expression profiling of whole-blood samples from women exposed to hormone replacement therapy. *Molecular Cancer Therapeutics* **5** 868–876.

[9] Gehan, E. A. (1965). A generalized two-sample Wilcoxon test for doubly censored data. *Biometrika* 650–653.

[10] Gretton, A., Borgwardt, K. M., Rasch, M., Schölkopf, B. and Smola, A. J. (2006). A kernel method for the two-sample-problem. In *Advances in neural information processing systems* 513–520.

[11] Hou, N., Hong, S., Wang, W., Olopade, O. I., Dignam, J. J. and Huo, D. (2013). Hormone replacement therapy and breast cancer: heterogeneous risks by race, weight, and breast density. *Journal of the National Cancer Institute* **105** 1365–1372.

[12] Jackson, S. and Mauldin, R. D. (1999). On the sigma-class generated by open balls. *Proc. Cambridge Phil. Soc.* **127** 99-10.

[13] Justel, A., Peña, D. and Zamar, R. (1997). A multivariate Kolmogorov-Smirnov test of goodness of fit. *Statistics & Probability Letters* **35** 251–259.

[14] Kosorok, M. R. and Ma, S. (2007). Marginal asymptotics for the large p, small n paradigm: with applications to microarray data. *The Annals of Statistics* **35** 1456–1486.

[15] Lee, A. J. (1990). *U-Statistics: Theory and Practice. Statistics: Textbooks and Monographs*. M. Dekker.

[16] Neuhaus, G. (1977). Functional limit theorems for U-statistics in the degenerate case. *Journal of Multivariate Analysis* **7** 424 - 439.

[17] Preiss, D. and Tiser, J. (1991). Measures in Banach spaces are determined by their values on balls. *Mathematika* **38** 391-397.

[18] Reshef, D. N., Reshef, Y. A., Finucane, H. K., Grossman, S. R., McVean, G., Turnbaugh, P. J., Lander, E. S., Mitzenmacher, M. and Sabeti, P. C. (2011). Detecting novel associations in large data sets. *science* **334** 1518–1524.

[19] Schierz, A. (2009). Virtual screening of bioassay data. *Journal of Cheminformatics* **1**.

[20] Schoenberg, I. (1937). On Certain Metric Spaces Arising From Euclidean Spaces by a Change of Metric and Their Imbedding in Hilbert Space. *Annals of Mathematics* **38** 787–793.

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

[22] Sejdinovic, D., Sriperumbudur, B. K., Gretton, A. and Fukumizu, K. (2013). Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *Annals of Statistics* **41** 2263–2291.

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

[24] Van Der Laan, M. J. and Bryan, J. (2001). Gene expression analysis with the parametric bootstrap. *Biostatistics* **2** 445–461.

[25] Wang, X. et al. (2017). Multiple change point detection in time series by ball divergence. *Preprint*.

Southern China Research
Center of Statistical Science
School of Mathematical
and Computational Science
Sun Yat-sen University
Guangzhou, GD 510275, China
sysu.wenliang@gmail.com
tiany5@mail2.sysu.edu.cn

Southern China Research
Center of Statistical Science
School of Mathematical
and Computational Science
Sun Yat-sen University
Guangzhou, GD 510275, China
Zhongshan School of Medicine
Sun Yat-sen University
Guangzhou, GD 510080, China
wangxq88@mail.sysu.edu.cn

Southern China Research
Center of Statistical Science
School of Mathematical
and Computational Science
Sun Yat-sen University
Guangzhou, GD 510275, China
Department of Biostatistics
Yale University School of Public Health
New Haven, CT 06520-8034, USA
heping.zhang@yale.edu