Submitted to Bernoulli
arXiv: arXiv:0000.0000

Interpoint Distance Based Two Sample Tests in High Dimension

CHANGBO ZHU[1,*] and XIAOFENG SHAO[1,**]

[1] *Department of Statistics, University of Illinois at Urbana-Champaign, Champaign, IL, 61820, USA. E-mail:* [*] changbo2@illinois.edu; [**] xshao@illinois.edu

In this paper, we study a class of two sample test statistics based on inter-point distances in the high dimensional and low/medium sample size setting. Our test statistics include the well-known energy distance and maximum mean discrepancy with Gaussian and Laplacian kernels, and the critical values are obtained via permutations. We show that all these tests are inconsistent when the two high dimensional distributions correspond to the same marginal distributions but differ in other aspects of the distributions. The tests based on energy distance and maximum mean discrepancy mainly target the differences between marginal means and variances, whereas the test based on L^1-distance can capture the difference in marginal distributions. Our theory sheds new light on the limitation of inter-point distance based tests, the impact of different distance metrics, and the behavior of permutation tests in high dimension. Some simulation results and a real data illustration are also presented to corroborate our theoretical findings.

Keywords: Two Sample Test, High Dimensionality, Permutation Test, Power Analysis.

1. Introduction

In many statistical and machine learning applications, we need inference about the two populations or distributions based on the data samples collected. For example, we need to compare the effectiveness of two newly developed drugs in clinical research, the higher educational level between two countries in a social study and the global warming effects on two regions in environmental science. Two sample hypothesis testing is a statistical procedure to deal with such problems. Formally speaking, having i.i.d. p-dimensional samples $X_1, \cdots, X_n =^d X \sim F$ and $Y_1, \cdots, Y_m =^d Y \sim G$, we are interested in knowing whether the underlining distributions F and G which generate the two samples are the same, i.e. to test the following hypothesis,

$$H_0 : F = G \text{ versus } H_A : F \neq G.$$

The study of two-sample testing has a long history and dates back to Kolmogorov-Smirnov's test [19, 27], where the empirical CDFs are compared using the sup-norm. Related work for univariate two-sample tests includes Cramer von-Mises criterion [9, 29] and Anderson-Darling test [3]. Extensions to comparison of multivariate distributions and also the k-sample problem can be found in [5, 6, 12, 17, 26] among others. Some

other interesting work focusing on the "trimmed" comparison of distributions can be found in [1, 2, 11, 23].

However, all the afore-mentioned work focuses on the fixed dimensional case. If the dimension exceeds the sample size or is allowed to grow, some of the above methods are expected to fail. For example, the density-based methods suffer from the curse of high dimensionality in particular. In this paper, we study the two sample tests based on certain dissimilarity metrics that can be expressed as functions of the interpoint distances. Two of the most popular high dimensional two-sample tests that fall into this category are based on the Energy Distance (ED) [28] and the Maximum Mean Discrepancy (MMD) [13]. The former is based on the Euclidean distance between sample elements; while the latter is a kernel based method and is basically a variant of ED with a user-specified kernel as distance metric. To be more specific, both ED and MMD take the following form

$$\mathrm{ED}^k(F,G) = 2E[k(X,Y)] - E[k(X,X')] - E[k(Y,Y')], \tag{1}$$

where k is a user-specified kernel, X', Y' are i.i.d copies of X, Y respectively. For instance, k can be chosen as

$$\begin{aligned}
&L^2\text{-norm (Euclidean distance)}: && k(X,Y) = \|X-Y\|_2 = \sqrt{\textstyle\sum_{u=1}^p (x_u - y_u)^2},\\
&\text{Gaussian kernel}: && k(X,Y) = \exp\left(-\frac{\|X-Y\|_2^2}{2\gamma_p^2}\right),\\
&\text{Laplacian kernel}: && k(X,Y) = \exp\left(-\frac{\|X-Y\|_2}{\gamma_p}\right),\\
&L^1\text{-norm}: && k(X,Y) = \|X-Y\|_1 = \textstyle\sum_{u=1}^p |x_u - y_u|,
\end{aligned}$$

where $X = (x_1, \cdots, x_p)^T$, $Y = (y_1, \cdots, y_p)^T$ and γ_p is a user-specified bandwidth parameter. Then, the population version of ED is given by Equation (1) with k being the L^2-norm and the population version of MMD multiplied by -1 is given by Equation (1) with k being Gaussian or Laplacian kernel. When k is L^2-norm, Gaussian or Laplacian kernel, $\mathrm{ED}^k(F,G)$ enjoys the property that $\mathrm{ED}^k(F,G) = 0 \Leftrightarrow F = G$. In fact, $\mathrm{ED}^k(F,G) = 0 \Leftrightarrow F = G$ holds as long as k is a strongly negative definite kernel [18]. ED and MMD based tests are both nonparametric without any assumption on the underlying distributions and can be implemented conveniently in practice using permutations. In this work, we aim to address the following questions:

1, Can ED^k based permutation test maintain its power against all kinds of alternatives in the high dimensional setting?
2, What are the impact of different distance metrics?

To answer the above questions, we conduct rigorous theoretical analysis on the power of $\mathrm{ED}^k(F,G)$ based permutation test in the high dimensional low sample setting (HDLSS) [16] as well as high dimensional medium sample size setting (HDMSS) [4]. Naturally, we say a test is *consistent if its power goes to 1 under either HDLSS or HDMSS regime*. Here, we study the power property of the permutation based tests because they are frequently implemented for Energy Distance and its variants in real life applications.

Let $\mathbf{X} = (X_1, X_2, \cdots, X_n)^T$, $\mathbf{Y} = (Y_1, Y_2, \cdots, Y_m)^T$, $\mathbf{Z} = (\mathbf{X}^T, \mathbf{Y}^T)^T$ denote the sample matrices and $\mathrm{ED}_n^k(\mathbf{Z})$ be a U-statistic based unbiased estimator of $\mathrm{ED}^k(F,G)$.

Our main results include: (i) Derivation of the limiting distribution of $\mathrm{ED}_n^k(\mathbf{\Gamma Z})$ under both low and medium sample size setting, where $\mathbf{\Gamma} \sim \mathrm{Uniform}(\mathbb{P}_{n+m})$ and $\mathbb{P}_{n+m}$ is the set of permutation matrices of dimension $(n+m)\times(n+m)$. (ii) Based on the asymptotic results, we formulate different local alternatives, under which the power behavior of ED_n^k based permutation tests are discussed in detail. (iii) Our theories are applied to existing kernels and statistics, for example

1, Under both HDLSS and HDMSS, ED^k based permutation test w.r.t. L^2-norm, Gaussian and Laplacian kernel are consistent if the sum of component-wise mean or variance differences are not so small, i.e, $\lim_{p\to\infty}\sum_{u=1}^p(E(x_u)-E(y_u))^2/p \neq 0$ or $\lim_{p\to\infty}|\sum_{u=1}^p(var(x_u)-var(y_u))/p| \neq 0$. In addition, if the sum of component-wise mean and variance differences are both of order $o(\sqrt{p}/\sqrt{nm})$, i.e.,

$$\sum_{u=1}^p(E(x_u)-E(y_u))^2 = o\left(\frac{\sqrt{p}}{\sqrt{nm}}\right) \text{ and } \left|\sum_{u=1}^p(var(x_u)-var(y_u))\right| = o\left(\frac{\sqrt{p}}{\sqrt{nm}}\right),$$

these tests suffer substantial power loss (the limits of their power are derived) under HDLSS and have trivial power (power no larger than the significance level) under HDMSS. Furthermore, under HDLSS, the afore-mentioned tests have trivial power if additionally we have $\sum_{u,v=1}^p(\mathrm{cov}(x_u,x_v)-\mathrm{cov}(y_u,y_v))^2 = o(p)$.

2, When k is chosen as L^1-norm, ED^k based permutation test experiences a power drop under HDLSS and trivial power under HDMSS if X, Y have the same univariate marginal distribution, i.e. $x_u =^d y_u$ for $u = 1, 2, \cdots, p$. This phenomenon is consistent with the fact that ED^k with L^1-norm can characterise the discrepancies between the marginal univariate distributions. In addition, Under HDLSS, we show that the L^1-norm based test has trivial power when X and Y have the same bivariate marginal distribution, i.e., $(x_u, x_v) =^d (y_u, y_v), u, v = 1, \cdots, p$.

These findings are further corroborated in our simulation study. It is worth mentioning that Chakraborty and Zhang [8] investigate the energy distance, maximum mean discrepancy, distance covariance and Hilbert-Schmidt Independence Criterion in the high dimensional setting. They propose a new class of metrics which can detect/measure the equality of low-dimensional marginal distributions and a computational efficient t-test is further proposed based on the new metric. By contrast, our focus is on kernel-based permutation test and their asymptotic power properties in the high dimensional setting. In the following we introduce some notation and define some frequently used operators for later convenience.

1.1. Notation

Here, random data samples are denoted as, for each $i = 1, 2, \cdots, n$, $X_i =^d X = (x_1, \cdots, x_p)^T \in \mathbb{R}^p$ and for each $j = 1, 2, \cdots, m$, $Y_j =^d Y = (y_1, \cdots, y_p)^T \in \mathbb{R}^p$. Next, let $\mathbf{X} = (X_1, X_2, \cdots, X_n)^T$, $\mathbf{Y} = (Y_1, Y_2, \cdots, Y_m)^T$ and $\mathbf{Z} = (\mathbf{X}^T, \mathbf{Y}^T)^T = (Z_1, Z_2, \cdots, Z_{n+m})^T$ denote the random sample matrices. Furthermore, let $\mathbb{P}_{n+m} =$

$\{\Gamma_1, \Gamma_2, \cdots, \Gamma_{(n+m)!}\}$ be the group containing all permutation matrices of dimension $(n+m)\times(n+m)$ and for each i, let π_i be the permutation that corresponds to Γ_i via

$$\Gamma_i \mathbf{Z} = \left(Z_{\pi_i(1)}, Z_{\pi_i(2)}, \cdots, Z_{\pi_i(n+m)}\right)^T.$$

For a random permutation matrix $\mathbf{\Gamma} \sim \text{uniform}(\mathbb{P}_{n+m})$, we use $\boldsymbol{\pi}$ to represent its corresponding permutation. Next, given any function φ, $\varphi^{(i)}$ is used to denote its i-th order derivative. Calligraphic letters $(\mathcal{K}, \mathcal{L}, \mathcal{R}, \mathcal{W}, \mathcal{G})$ are used to denote self-defined operators that act on random variables to produce random variables. For two random variables W_1, W_2, the notation $W_1 =^d W_2$ means that they have the same distribution. Under HDLSS or HDMSS setting, we use $\xrightarrow{p}$, $\xrightarrow{d}$ to denote convergence in probability, in distribution respectively and utilize the order in probability notations such as stochastic boundedness O_p (big O in probability) and convergence in probability o_p (small o in probability).

2. Interpoint Distance Based Two Sample Tests

In this paper, we limit our attention to $\text{ED}^k(F,G)$, where k is a user specified dissimilarity metric [24] of the following form

$$k(X,Y) = \varphi\left\{\frac{1}{p}\sum_{u=1}^{p}\psi(x_u, y_u)\right\}, \tag{2}$$

where $\psi \geq 0$ and φ has continuous second order derivative on $(0, +\infty)$. The reason we focus on $\text{ED}^k(F,G)$ of the above form is that the metric k encompasses many well-known distance metrics such as L^2-norm, L^1-norm, Gaussian and Laplacian kernel. Consequently, Energy Distance (ED) and Maximum Mean Discrepancy (MMD) are just special cases of $\text{ED}^k(F,G)$. We summarize the commonly used distance metrics in Table 1. Following the literatures [13, 14], we consider the bandwidth parameter γ in Gaussian and Laplacian kernel as a fixed constant and note that its relationship with γ_p (introduced in Gaussian or Laplacian kernel on page 2) is given by $\gamma_p = \sqrt{p}\gamma$. Notice that if k is some well-known distance metrics such as L^2-norm, Gaussian kernel (multiplied by -1) and Laplacian kernel (multiplied by -1), a nice property for ED^k is that

$$\text{ED}^k(F,G) \geq 0 \text{ and } \text{ED}^k(F,G) = 0 \Leftrightarrow F = G. \tag{3}$$

Here, it is just for the ease of presentation and notational simplicity that k is set to be Gaussian or Laplacian kernel multiplied by -1. In fact, if k is a universal kernel (see Theorem 5 and Lemma 1 of [13]) or k is a strongly negative definite kernel (see Theorem 1.9 [18]), Property (3) still holds. On the other hand, using $\text{ED}^1(F,G)$ to denote $\text{ED}^k(F,G)$ when k is the L^1-distance, we observe that $\text{ED}^1(F,G) = \sum_{u=1}^{p}\text{ED}(F_u, G_u)$, from which it easily follows that

$$\text{ED}^1(F,G) \geq 0 \text{ and } \text{ED}^1(F,G) = 0 \Leftrightarrow F_u = G_u \text{ for all } u = 1, 2, \cdots, p.$$

$\psi(x,y)$	$\varphi(x)$	k	$\mathrm{ED}^k(F,G)$
$(x-y)^2$	$\sqrt{x}$	L^2-norm	Energy distance (ED) Székely and Rizzo [28]
	$-e^{-\frac{x}{2\gamma^2}}$	Gaussian kernel (multiplied by -1)	Maximum Mean Discrepancy (MMD) Gretton et al. [13]
	$-e^{-\frac{\sqrt{x}}{\gamma}}$	Laplacian kernel (multiplied by -1)	
$\lvert x-y\rvert$	x	L^1-norm	Used for some graph-based tests Sarkar et al. [24]

Table 1. Correspondence between different choices of ψ, φ and existing distance metrics as well as two sample test statistics in the literature.

Notice that it is possible to have $F_u = G_u$ for all $u = 1, 2, \cdots, p$ but $F \neq G$, under which we have $\mathrm{ED}^1(F,G) = 0$ while $\mathrm{ED}^k > 0$ if k is L^2-norm, Gaussian kernel (multiplied by -1) or Laplacian kernel (multiplied by -1). Thus, L^2-norm, Gaussian kernel or Laplacian kernel based test statistics have advantage over L^1-norm based test statistic in the low dimensional setting, but we will see later that the story is in a sense reversed under the high dimensional setting. Next, an unbiased estimator of ED^k is given as

$$\mathrm{ED}_n^k(\mathbf{Z}) = \frac{2}{mn}\sum_{i=1}^{n}\sum_{j=1}^{m} k(X_i, Y_j) - \frac{2}{n(n-1)}\sum_{1\le i<j\le n} k(X_i, X_j) - \frac{2}{m(m-1)}\sum_{1\le i<j\le m} k(Y_i, Y_j).$$

3. Power Analysis for Permutation Test

As permutation tests are commonly used for Energy Distance and kernel variants in practice due to their implementational convenience and accurate size, we study their asymptotic behavior under the high dimensional setting in this subsection. Since we have i.i.d samples, after we permute the data, i.e., shuffle the rows of $\mathbf{Z}$ as $\Gamma_i\mathbf{Z}$ by some permutation matrix Γ_i, what really matters to the distribution of $\mathrm{ED}_n^k(\Gamma_i\mathbf{Z})$ is how many X samples stay in the first n rows. Formally, let $|\mathbb{A}|$ be the cardinality of the set $\mathbb{A}$ and given a permutation matrix Γ_i with the corresponding permutation π_i, set

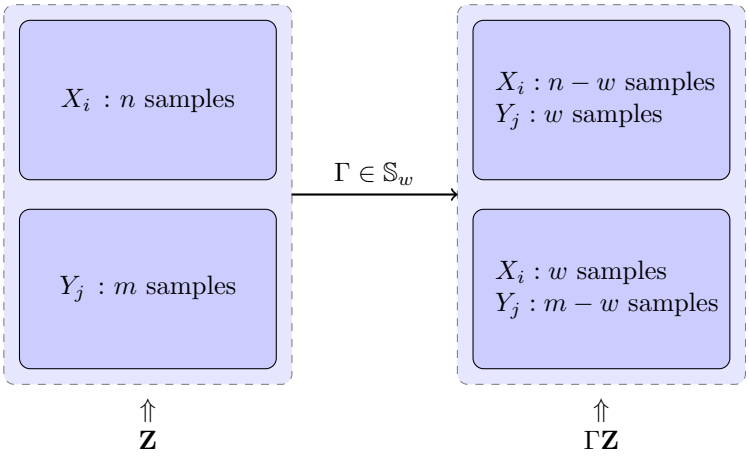
$$N(\Gamma_i) = \left|\{j \in \{1, 2, \cdots, n\} : 1 \le j \le n, n+1 \le \pi_i(j) \le n+m\}\right|.$$

The integer $n - N(\Gamma_i)$ actually counts the number of samples which belong to the first n rows of $\mathbf{Z}$ both before and after the permutation Γ_i. Notice that it is possible that $N(\Gamma_i) = N(\Gamma_j)$ for different permutations Γ_i and Γ_j. The set $\mathbb{S}_w$ collects all the

permutations Γ_i such that $N(\Gamma_i) = w$. Mathematically, fix $0 \le w \le \min\{n, m\}$, set $\mathbb{S}_w = \{\Gamma_i : N(\Gamma_i) = w,\ i = 1, 2, \cdots, (n+m)!\}$, then

$$|\mathbb{S}_w| = \binom{m}{w}\binom{n}{n-w} n!m!.$$

To differentiate from Γ_i, we use italic symbol $\mathit{\Gamma}_w$ to represent an element in $\mathbb{S}_w$. Intuitively, $|\mathbb{S}_w|$ is the number of permutations that would have $n - w$ samples stay in the first n rows of $\mathbf{Z}$ after we apply the corresponding permutation. The above process is further illustrated in the following diagram.



For the inter-point distance based two sample tests, we can equivalently permute the weights on the pair-wise distances instead of permuting data points, i.e., for a fixed permutation matrix $\Gamma_s \in \mathbb{P}_{n+m}$ that corresponds to π_s, we can write $\mathrm{ED}_n^k(\Gamma_s\mathbf{Z})$ as

$$\mathrm{ED}_n^k(\Gamma_s\mathbf{Z}) = \sum_{i=2}^{n+m}\sum_{j=1}^{i-1} \Pi_{s,ij} k(Z_i, Z_j), \tag{4}$$

where $\Pi_{s,ij}$ is defined as

$$\Pi_{s,ij} = \begin{cases} -\frac{2}{n(n-1)}, & 1 \le \pi_s(i), \pi_s(j) \le n, \\ -\frac{2}{m(m-1)}, & n+1 \le \pi_s(i), \pi_s(j) \le n+m, \\ \frac{2}{mn}, & 1 \le \pi_s(i) \le n, n+1 \le \pi_s(j) \le n+m, \\ \frac{2}{mn}, & n+1 \le \pi_s(i) \le n+m, 1 \le \pi_s(j) \le m. \end{cases}$$

To formally define the permutation test for $\mathrm{ED}_n^k(\mathbf{Z})$, let $\widehat{R}$ denote the randomization distribution of $\mathrm{ED}_n^k(\mathbf{Z})$, which is defined by

$$\widehat{R}(t) = \frac{1}{(n+m)!}\sum_{i=1}^{(n+m)!} \mathbb{I}_{\left\{\mathrm{ED}_n^k(\Gamma_i\mathbf{Z}) \le t\right\}} = \frac{1}{(m+n)!}\sum_{w=0}^{\min\{n,m\}}\sum_{\Gamma\in\mathbb{S}_w} \mathbb{I}_{\left\{\mathrm{ED}_n^k(\Gamma\mathbf{Z}) \le t\right\}}.$$

For any distribution F, let the $(1-\alpha)$-th quantile of F be denoted by $Q_{F,1-\alpha}$. In particular, the $(1-\alpha)$th quantile of $\widehat{R}$ is $Q_{\widehat{R},1-\alpha}$, i.e.

$$Q_{\widehat{R},1-\alpha} = \widehat{R}^{-1}(1-\alpha) = \inf\left\{t : \widehat{R}(t) \geq 1-\alpha\right\}. \tag{5}$$

Then, the level-α permutation test w.r.t. $\mathrm{ED}_n^k(\mathbf{Z})$ is defined as

$$\text{Reject } H_0 \text{ if } \mathrm{ED}_n^k(\mathbf{Z}) > Q_{\widehat{R},1-\alpha}.$$

In real life applications, $(n+m)!$ might be large, we thus resort to an approximation of $Q_{\widehat{R},1-\alpha}$. Let $\mathbf{\Gamma}_1, \cdots, \mathbf{\Gamma}_{S-1}$ be i.i.d and uniformly sampled from $\mathbb{P}_{n+m}$ and we approximate the critical value by $Q_{\widetilde{R},1-\alpha}$, where

$$\widetilde{R}(t) := \frac{1}{S}\left(\mathbb{I}_{\left\{\mathrm{ED}_n^k(\mathbf{Z})\leq t\right\}} + \sum_{i=1}^{S-1}\mathbb{I}_{\left\{\mathrm{ED}_n^k(\mathbf{\Gamma}_i\mathbf{Z})\leq t\right\}}\right).$$

Under the null hypothesis, let the critical value c be chosen as $c = Q_{\widehat{R},1-\alpha}$ or $c = Q_{\widetilde{R},1-\alpha}$, then we have $P(\mathrm{ED}_n^k(\mathbf{Z}) > c) \leq \alpha$, where the equality may not hold (i.e., the size may not be equal to α) since our test is non-randomized. A randomization based test can be formulated but should make little difference in practice; see [21].

3.1. Local Alternatives

In this subsection, we define different local alternatives, under which the $\mathrm{ED}_n^k(\mathbf{Z})$ based permutation test will be consistent, have a nontrivial power limit and exhibit trivial power (power no larger than the significance level α) in the limit. To formally define the local alternative hypothesis, let the operator $\mathcal{K}$ be defined as

$$\begin{aligned}\mathcal{K}(Z_i, Z_j) = \frac{1}{\sqrt{p}}\sum_{u=1}^{p}\big\{&\psi(z_{iu}, z_{ju}) - E\left[\psi(z_{iu}, z_{ju})|z_{iu}\right] \\ &- E\left[\psi(z_{iu}, z_{ju})|z_{ju}\right] + E\left[\psi(z_{iu}, z_{ju})\right]\big\},\end{aligned} \tag{6}$$

It follows from Proposition 2.2.1 of [30] that $E[\mathcal{K}(Z_i, Z_j)\mathcal{K}(Z_{i'}, Z_{j'})] = 0$ if $\{i,j\} \neq \{i',j'\}$. Next, denote the average distance over components as

$$\overline{\psi}(Z_i, Z_j) = \frac{1}{p}\sum_{u=1}^{p}\psi(z_{iu}, z_{ju}).$$

In addition, we need to assume the existence of some constants to properly define the local alternatives. These constants will also appear in the limiting distribution of our test statistics.

Assumption 1. *As $p \to \infty$, assume the existence of the limiting mean*

$$e_x = \lim_{p\to\infty} E\left[\overline{\psi}(X, X')\right], e_y = \lim_{p\to\infty} E\left[\overline{\psi}(Y, Y')\right] \text{ and } e_{xy} = \lim_{p\to\infty} E\left[\overline{\psi}(X, Y)\right]$$

and also the limiting variances

$$v_x = \lim_{p\to\infty} var\left[\mathcal{K}(X, X')\right], v_y = \lim_{p\to\infty} var\left[\mathcal{K}(Y, Y')\right] \text{ and } v_{xy} = \lim_{p\to\infty} var\left[\mathcal{K}(X, Y)\right].$$

Then, we are ready to define the consistency space H_{A_c}, under which the $\mathrm{ED}_n^k(\mathbf{Z})$ implemented as permutation test can be shown to be consistent under both HDLSS and HDMSS settings.

$$H_{A_c} := \{(F, G) \mid 2\varphi(e_{xy}) \neq \varphi(e_x) + \varphi(e_y)\}.$$

We use $\mathbb{A}^c$ to denote the complement of any given set $\mathbb{A}$ and denote $F = (F_1, F_2, \cdots, F_p)$ and $G = (G_1, G_2, \cdots, G_p)$, where $F_u, G_u, u = 1, 2, \cdots, p$ are the marginal univariate distributions. For commonly used kernels, we have the following table characterizing H_{A_c} and the proof is postponed to subsection A.1.

k	H_{A_c} Characterization
L^2-norm	$H_{A_c} = \left\{ (F,G) \;\middle|\; \begin{array}{l} \sum_{u=1}^p (E(x_u) - E(y_u))^2 = o(p) \text{ and} \\ \left\lvert \sum_{u=1}^p (var(x_u) - var(y_u)) \right\rvert = o(p) \end{array} \right\}^c$
Gaussian kernel	
Laplacian kernel	
L^1-norm	$H_{A_c} = \{(F,G) \,\lvert \sum_{u=1}^p \mathrm{ED}(F_u, G_u) = o(p)\}^c$

Then, we present the space H_{A_l}, under which the normal limit of $\mathrm{ED}_n^k(\mathbf{Z})$ can be derived under both HDLSS and HDMSS.

$$H_{A_l} := \left\{ (F,G) \;\middle|\; \begin{array}{l} e_{xy} = e_x = e_y, \\ \left\lvert 2E\left[\overline{\psi}(X,Y)\right] - E\left[\overline{\psi}(X,X')\right] - E\left[\overline{\psi}(Y,Y')\right] \right\rvert = o(\sqrt{\frac{1}{nmp}}), \\ E\left[\left\lvert E\left[\overline{\psi}(X,Y)|X\right] - E\left[\overline{\psi}(X,X')|X\right] \right\rvert\right] = o(\sqrt{\frac{1}{nmp}}) \text{ and} \\ E\left[\left\lvert E\left[\overline{\psi}(X,Y)|Y\right] - E\left[\overline{\psi}(Y,Y')|Y\right] \right\rvert\right] = o(\sqrt{\frac{1}{nmp}}). \end{array} \right\}.$$

Under H_{A_l}, a limit for the power of $\mathrm{ED}_n^k(\mathbf{Z})$ (implemented as permutation test) is derived under HDLSS. On the other hand, its power is shown to be trivial (no larger than the significance level α) under HDMSS and H_{A_l}. Next, we provide sufficient conditions for

$(F,G) \in H_{A_l}$ with respect to the following well known kernels.

k	Sufficient conditions for H_{A_l}
L^2-norm	$\left\{ (F,G) \left\| \begin{array}{l} \sum_{u=1}^{p}(E(x_u)-E(y_u))^2 = o(\sqrt{\frac{p}{nm}}) \text{ and} \\ \left\|\sum_{u=1}^{p}(var(x_u)-var(y_u))\right\| = o(\sqrt{\frac{p}{nm}}) \end{array} \right. \right\} \subseteq H_{A_l}$
Gaussian kernel	
Laplacian kernel	
L^1-norm	$\{(F,G)\,\|F_u = G_u, u = 1,2,\cdots,p\} \subseteq H_{A_l}$

Then, the set of distributions H_{A_t} is defined as

$$H_{A_t} := \{(F,G)\,|(F,G) \in H_{A_l}, v_{xy} = v_x = v_y\}.$$

It can be shown that under H_{A_t}, the $\mathrm{ED}_n^k(\mathbf{Z})$ based permutation test has power no larger than the significance level α for both HDLSS and HDMSS settings. Sufficient conditions of being in H_{A_t} are provided in the following table.

k	Sufficient conditions for H_{A_t}
L^2-norm	$\left\{ (F,G) \left\| \begin{array}{l} \sum_{u=1}^{p}(E(x_u)-E(y_u))^2 = o(\sqrt{\frac{p}{nm}}), \\ \left\|\sum_{u=1}^{p}(var(x_u)-var(y_u))\right\| = o(\sqrt{\frac{p}{nm}}) \text{ and} \\ \sum_{u,v=1}^{p}(\mathrm{cov}(x_u,x_v)-\mathrm{cov}(y_u,y_v))^2 = o(p) \end{array} \right. \right\} \subseteq H_{A_t}$
Gaussian kernel	
Laplacian kernel	
L^1-norm	$\left\{(F,G)\,\middle\|(x_u,x_v) =^d (y_u,y_v), u,v = 1,\cdots,p\right\} \subseteq H_{A_t}$

Comparing the three local alternatives, it follows from the definition of $H_{A_c}, H_{A_l}, H_{A_t}$ that $H_{A_c}^c \supseteq H_{A_l} \supseteq H_{A_t}$. We also want to remark that it holds for arbitrary function φ and ψ that

$$\{(F,G)\,|F_u = G_u, u = 1,2,\cdots,p\} \subseteq H_{A_l},$$
$$\left\{(F,G)\,\middle|(x_u,x_v) =^d (y_u,y_v), u,v = 1,\cdots,p\right\} \subseteq H_{A_t}.$$

3.2. High Dimensional Low Sample Size (HDLSS)

The analysis in this subsection is conducted under the high dimensional low sample size setting (HDLSS), i.e, n, m are fixed constants and we let $p \to \infty$. Our final goal is to study the power of $\mathrm{ED}_n^k(\mathbf{Z})$ based permutation test under various local alternatives. To this end, we need the following assumption. Recall the operator $\mathcal{K}$ is defined in (6).

Assumption 2. *For fixed n, m, as $p \to \infty$,*

$$\begin{pmatrix} \mathcal{K}(X_i,Y_j) \\ \mathcal{K}(X_{i_1},X_{i_2}) \\ \mathcal{K}(Y_{j_1},Y_{j_2}) \end{pmatrix}_{i,j,i_1<i_2,j_1<j_2} \xrightarrow{d} \begin{pmatrix} b_{ij} \\ c_{i_1i_2} \\ d_{j_1j_2} \end{pmatrix}_{i,j,i_1<i_2,j_1<j_2},$$

where $\{b_{ij}, c_{i_1i_2}, d_{j_1j_2}\}_{i,j,i_1<i_2,j_1<j_2}$ *are uncorrelated and jointly Gaussian with mean 0 and variances* $var(b_{ij}) = v_{xy}$, $var(c_{i_1i_2}) = v_x$, $var(d_{j_1j_2}) = v_y$.

Remark 3.1. *The above multi-dimensional CLT result is classical and can be derived under suitable moment and weak dependence assumptions on the components of* X *and* Y.

In the above assumption, it is due to the use of double centered distance $\mathcal{K}(Z_i, Z_j)$ that the asymptotic covariance matrix is diagonal. Then, to provide some insights, the first step of our power analysis is the Taylor expansion w.r.t φ up to the second order, i.e., for $i \neq j$

$$k(Z_i, Z_j) = \varphi(e_{ij}) + \varphi^{(1)}(e_{ij})\mathcal{L}(Z_i, Z_j) + \mathcal{R}_2(Z_i, Z_j),$$

where $\mathcal{L}(Z_i, Z_j) := \overline{\psi}(Z_i, Z_j) - e_{ij}$ is an operator that acts on random variables,

$$e_{ij} = \begin{cases} e_x, & \text{if } 1 \leq i, j \leq n, \\ e_y, & \text{if } n+1 \leq i, j \leq n+m, \\ e_{xy}, & \text{otherwise.} \end{cases}$$

and $\mathcal{R}_2(Z_i, Z_j)$ is the remainder given by

$$\mathcal{R}_2(Z_i, Z_j) = \mathcal{L}^2(Z_i, Z_j) \int_0^1 \int_0^1 u\varphi^{(2)}\left(e_{ij} + uv\mathcal{L}(Z_i, Z_j)\right) dvdu.$$

In order to control the remainder term, we need assumptions about the decay rate of $E[\mathcal{L}^2(Z_i, Z_j)]$. Thus, we set

$$\alpha_x^2 = E[\mathcal{L}^2(X, X')],\ \alpha_y^2 = E[\mathcal{L}^2(Y, Y')] \text{ and } \alpha_{xy}^2 = E[\mathcal{L}^2(X, Y)].$$

It then follows from Markov's inequality that $\mathcal{L}(X, Y) = O_p(\alpha_{xy})$, $\mathcal{L}(X, X') = O_p(\alpha_x)$ and $\mathcal{L}(Y, Y') = O_p(\alpha_y)$. Then, our next two assumptions are used to control the remainder terms induced by taking the Taylor expansion.

Assumption 3. $\alpha_{xy}^2 = o(1), \alpha_x^2 = o(1)$ *and* $\alpha_y^2 = o(1)$.

Assumption 4. $\sqrt{p}\alpha_{xy}^2 = o(1), \sqrt{p}\alpha_x^2 = o(1)$ *and* $\sqrt{p}\alpha_y^2 = o(1)$.

Remark 3.2. *To gain some insights into the above assumptions, a straightforward calculation yields*

$$\alpha_{xy}^2 = \frac{1}{p^2} \sum_{u,v=1}^{p} cov(\psi(x_u, y_u), \psi(x_v, y_v)) + \left(\frac{\sum_{u=1}^{p} E[\psi(x_u, y_u)]}{p} - e_{xy} \right)^2.$$

Therefore, we have $\sqrt{p}\alpha_{xy}^2 = o(1)$ *if the component-wise dependencies of both* X *and* Y *are not so strong. For illustration purpose, suppose* X *and* Y *are* κ*-dependent weak stationary time series, i.e.,* $x_u \perp x_v$ *and* $y_u \perp y_v$ *if* $|u-v| > \kappa$. *Then, if* $\max_u E[\psi^2(x_u, y_u)] <$

∞, it is easy to see that $\alpha_{xy}^2 = O(\kappa/p)$ and as a consequence, Assumption 4 is satisfied as long as $\kappa/\sqrt{p} = o(1)$. In addition, it is indeed fairly straightforward to verify the above result when the sequence $\{(x_u, y_u)\}_{u=1}^p$ is α-mixing with geometrically decaying coefficients.

Remark 3.3. *When $\psi(x,y) = (x-y)^2$, some algebra shows that*

$$\sum_{u,v=1}^{p} cov(\psi(x_u, y_u), \psi(x_v, y_v)) = var(X^T X) + var(Y^T Y) + 4var(X^T Y) - 4cov(X^T X, X^T E[Y]) - 4cov(Y^T Y, Y^T E[X]).$$

Thus, suppose $\sum_{u=1}^p E[\psi(x_u, y_u)]/p - e_{xy} = o(p^{-1/4})$ and if $var(X^T X)$, $var(Y^T Y)$, $var(X^T Y)$, $var(X^T E[Y])$, $var(E[X]^T Y)$ all have order $o(p^{1.5})$, we have $\sqrt{p}\alpha_{xy}^2 = o(1)$.

In the next theorem, we state the asymptotic behavior of $\mathrm{ED}_n^k(\Gamma_w \mathbf{Z})$ for each fixed permutation matrix $\Gamma_w \in \mathbb{S}_w$. Here, we use the italic gamma $\Gamma_w \in \mathbb{S}_w$ to differentiate from $\Gamma_i \in \mathbb{P}_{n+m}$.

Theorem 3.1. *For fixed $\Gamma_w \in \mathbb{S}_w$,*

(i) Under Assumptions 1 and 3,

$$\mathrm{ED}_n^k(\Gamma_w \mathbf{Z}) \xrightarrow{p} \mu_{n,w},$$

where $\mu_{n,w}$ is defined as

$$\mu_{n,w} := \mu_n(\Gamma_w \mathbf{Z}) = (2\varphi(e_{xy}) - \varphi(e_x) - \varphi(e_y)) \times \left\{ 1 - \left(\frac{2m-1}{m(m-1)} + \frac{2n-1}{n(n-1)} \right) w + \left(\frac{2}{mn} + \frac{1}{n(n-1)} + \frac{1}{m(m-1)} \right) w^2 \right\}.$$

(ii) Under Assumptions 1, 2, 4 and local alternative H_{A_l},

$$\sqrt{p} \left(\mathrm{ED}_n^k(\Gamma_w \mathbf{Z}) - \mu_n(\Gamma_w \mathbf{Z}) \right) \xrightarrow{d} N\left(0, \sigma_{n,w}^2\right).$$

where $\sigma_{n,w}^2$ *is given as*

$$\begin{aligned}\sigma_{n,w}^2 :=& \sigma_n(\Gamma_w \mathbf{Z}) \\ =& \left\{ \frac{4}{nm} - 4\left(\frac{n+m}{n^2m^2} - \frac{n}{n^2(n-1)^2} - \frac{m}{m^2(m-1)^2} \right) w \right. \\ & \left. + 4\left(\frac{2}{n^2m^2} - \frac{1}{n^2(n-1)^2} - \frac{1}{m^2(m-1)^2} \right) w^2 \right\} v_{xy}[\varphi^{(1)}(e_{xy})]^2 \\ & + \left\{ \frac{2}{n(n-1)} + 2\left(\frac{2n}{n^2m^2} - \frac{2n-1}{n^2(n-1)^2} - \frac{1}{m^2(m-1)^2} \right) w \right. \\ & \left. - 2\left(\frac{2}{m^2n^2} - \frac{1}{n^2(n-1)^2} - \frac{1}{m^2(m-1)^2} \right) w^2 \right\} v_x[\varphi^{(1)}(e_x)]^2 \\ & + \left\{ \frac{2}{m(m-1)} + 2\left(\frac{2m}{n^2m^2} - \frac{1}{n^2(n-1)^2} - \frac{2m-1}{m^2(m-1)^2} \right) w \right. \\ & \left. - 2\left(\frac{2}{n^2m^2} - \frac{1}{m^2(m-1)^2} - \frac{1}{n^2(n-1)^2} \right) w^2 \right\} v_y[\varphi^{(1)}(e_y)]^2. \end{aligned}$$

We use $W \sim \text{Hypergeometric}(n+m, m, n)$ to denote that W follows the hypergeometric distribution, which describes the probability of n draws from a union of two groups (one group has m elements, the other has n elements) such that w of them are chosen from the group of size m. To be precise, W has probability mass function

$$P(W = w) = \frac{\binom{m}{w}\binom{n}{m-w}}{\binom{n+m}{n}} \text{ for } w \in \{0, 1, \cdots, \min\{n, m\}\}.$$

Then, the limiting distribution of $\text{ED}_n^k(\mathbf{\Gamma Z})$ is derived in the following proposition.

Proposition 3.1. *For* $\mathbf{\Gamma} \sim$ *Uniform*$(\mathbb{P}_{n+m})$*, which is independent of the data, let*

$$W := N(\mathbf{\Gamma}) \sim \textit{Hypergeometric}(n+m, m, n).$$

(i) Under Assumptions 1 and 3,

$$\text{ED}_n^k(\mathbf{\Gamma Z}) \xrightarrow{p} \mu_{n,W}.$$

(ii) Under Assumptions 1, 2, 4 and local alternative H_{A_l}*,*

$$\sqrt{p}\left(\text{ED}_n^k(\mathbf{\Gamma Z}) - \mu_{n,W} \right) \xrightarrow{d} N\left(0, \sigma_{n,W}^2\right).$$

In the above proposition, $N\left(0, \sigma_{n,W}^2\right)$ should be understood as a mixture of Gaussian with probability distribution

$$P\left(N\left(0, \sigma_{n,W}^2\right) \le a\right) = \sum_{w=1}^{\min\{n,m\}} P(W = w) P(N\left(0, \sigma_{n,w}^2\right) \le a).$$

Next, let Γ_0 corresponds to the identity permutation map, we present the power behavior of $\mathrm{ED}^k(\mathbf{Z})$ when the critical values are obtained via permutations.

Theorem 3.2. *Under Assumption 1 and assume that* $2\varphi(e_{xy}) \geq \varphi(e_x) + \varphi(e_y)$.

1, [**Consistency**] *Suppose Assumption 3 hold.*

(i) If the critical value is chosen as $Q_{\widehat{R},1-\alpha}$. *Let* n, m *be large enough such that* $n!m!/(n+m)! < 1-\alpha$ *if* $m \neq n$ *and* $2(n!)^2/(2n)! < 1-\alpha$ *if* $m = n$. *Then, we have*

$$\lim_{p\to\infty} P_{H_{A_c}}\left(\mathrm{ED}_n^k(\mathbf{Z}) > Q_{\widehat{R},1-\alpha}\right) = 1,$$

which means that the asymptotic power of ED^k *based permutation test is 1 as* p *goes to infinity.*

(ii) If the critical value is chosen as $Q_{\widetilde{R},1-\alpha}$. *Then, we have*

$$\lim_{p\to\infty} P_{H_{A_c}}\left(\mathrm{ED}_n^k(\mathbf{Z}) > Q_{\widetilde{R},1-\alpha}\right) \geq \begin{cases} 1 - \frac{S-1}{\lfloor \alpha S \rfloor}\frac{n!m!}{(n+m)!}, & \text{if } n \neq m, \\ 1 - \frac{S-1}{\lfloor \alpha S \rfloor}\frac{2(n!)^2}{(n+m)!}, & \text{if } n = m. \end{cases}$$

2, [**Power Limit**] *Suppose Assumptions 2 and 4 hold.*

(i) If the critical value is chosen as $Q_{\widehat{R},1-\alpha}$. *Then, we have*

$$\lim_{p\to\infty} P_{H_{A_l}}\left(\mathrm{ED}_n^k(\mathbf{Z}) > Q_{\widehat{R},1-\alpha}\right) = P(V(\Gamma_0) > Q_{\widehat{T},1-\alpha}),$$

where

$$\widehat{T}(t) := \frac{1}{(n+m)!}\sum_{i=1}^{(n+m)!} \mathbb{I}_{\{V(\Gamma_i)\leq t\}}$$

and

$$V(\Gamma_s) = \sum_{i=1}^{n}\sum_{j=1}^{m} \Pi_{s,ij} b_{ij} - \sum_{1\leq i<j\leq n} \Pi_{s,ij} c_{ij} - \sum_{1\leq i<j\leq m} \Pi_{s,ij} d_{ij}.$$

(ii) If the critical value is chosen as $Q_{\widetilde{R},1-\alpha}$.

$$\lim_{p\to\infty} P_{H_{A_l}}\left(\mathrm{ED}_n^k(\mathbf{Z}) > Q_{\widetilde{R},1-\alpha}\right) = P\left(V(\Gamma_0) > Q_{\widetilde{T},1-\alpha}\right),$$

where

$$\widetilde{T} := \frac{1}{S}\left(\mathbb{I}_{\{V(\Gamma_0)\leq t\}} + \sum_{i=1}^{S-1} \mathbb{I}_{\{V(\mathbf{\Gamma}_i)\leq t\}}\right).$$

3, [**Trivial Power**] *Suppose Assumptions 2 and 4 hold. Then, we have*

$$\lim_{p\to\infty} P_{H_{A_t}}\left(\mathrm{ED}_n^k(\mathbf{Z}) > c\right) \le \alpha \text{ where } c = Q_{\widehat{R},1-\alpha} \text{ or } c = Q_{\widetilde{R},1-\alpha},$$

which means that the asymptotic power of ED^k *based permutation test is no more than the level* α *when* p *goes to infinity.*

Remark 3.4. *The above theorem and discussions in subsection 3.1 indicate that*

1. *L^1-norm can be more advantageous than L^2-norm, Gaussian kernel and Laplacian kernel when the dimension is high, since L^1-distance leads to high power provided that the summation of discrepancies between marginal univariate distributions is not so small, while L^2-norm, Gaussian kernel and Laplacian kernel would result in power loss when the total of marginal univariate mean and variance differences between X and Y is of order $o(\sqrt{p})$. Notice that the distributions of X and Y can differ in other aspects of the marginal distribution even if they have the same marginal univariate mean and variance.*
2. *All the tests under examination are only capable of detecting the discrepancies of marginal distributions. If the two high dimensional distributions $F \neq G$, but $F_u = G_u$ for $u = 1, 2, \cdots, p$, then none of them have consistent power.*

3.3. High Dimensional Medium Sample Size (HDMSS)

In this subsection, the theories are developed under the high dimensional medium sample size setting (HDMSS), i.e., as $p \to \infty$, $n := n(p) \to \infty$ at a slower rate compared to p and $n/m = \rho$, where $\rho \in (0, \infty)$ is a fixed constant. Though the proofs are quite different, most results and phenomena under the HDLSS setting have their similar counterparts under the HDMSS setting. Now that we have n, m growing to infinity, we need a stronger version of Assumptions 3 and 4.

Assumption 5. $nm\alpha_{xy}^2 = o(1), n^2\alpha_x^2 = o(1)$ *and* $m^2\alpha_y^2 = o(1)$.

Assumption 6. $\sqrt{nmp}\alpha_{xy}^2 = o(1), n\sqrt{p}\alpha_x^2 = o(1)$ *and* $m\sqrt{p}\alpha_y^2 = o(1)$.

Remark 3.5. *Following Remark 3.2, for κ-dependent stationary time series, $\alpha_{xy}^2 = O(\kappa/p)$. Thus, Assumptions 5 and 6 both require that $nm\kappa = o(p)$.*

To derive the asymptotic distribution under the HDMSS, we note that the leading term of $\mathrm{ED}_n^k(\mathbf{Z})$ is a martingale, the following assumption is used to ensure the conditional Lindeberg condition and the requirements on the conditional variance in classic martingale central limit theorem.

Assumption 7. *For any $\Lambda_1, \Lambda_2, \Lambda_3, \Lambda_4 \in \{X, Y\}$, suppose*

$$E\left[\mathcal{K}^4(\Lambda_1, \Lambda_2')\right] = o\left(n^2\right),$$
$$E\left[\mathcal{K}^2(\Lambda_1, \Lambda_3'')\mathcal{K}^2(\Lambda_2', \Lambda_3'')\right] = o(n),$$
$$E\left[\mathcal{K}(\Lambda_1, \Lambda_3'')\mathcal{K}(\Lambda_1, \Lambda_4''')\mathcal{K}(\Lambda_2', \Lambda_4''')\mathcal{K}(\Lambda_2', \Lambda_3'')\right] = o(1),$$

where $(\Lambda_1', \Lambda_2', \Lambda_3', \Lambda_4')$, $(\Lambda_1'', \Lambda_2'', \Lambda_3'', \Lambda_4'')$ and $(\Lambda_1''', \Lambda_2''', \Lambda_3''', \Lambda_4''')$ are independent copies of $(\Lambda_1, \Lambda_2, \Lambda_3, \Lambda_4)$.

Remark 3.6. *For any function φ and ψ, suppose X and Y are κ dependent sequences, i.e., $x_u \perp x_v$ and $y_u \perp y_v$ if $|u - v| > \kappa$. If there exists some constant $C > 0$ such that*

$$\max\left\{\sup_u E\left[\psi^4(x_u, y_u)\right], \sup_u E\left[\psi^4(x_u, x_u')\right], \sup_u E\left[\psi^4(y_u, y_u')\right]\right\} \leq C.$$

Then, for notational convenience, let

$$\phi_{xy,u} = \psi(x_u, y_u) - E\left[\psi(x_u, y_{ju})|x_u\right] - E\left[\psi(x_u, y_u)|y_u\right] + E\left[\psi(x_u, y_u)\right],$$

we see that $\sup_u E[\phi_{xy,u}^4] \leq 4^4 C$ and thus $E[\mathcal{K}^4(X, Y)]$ can be bounded as following

$$E\left[\mathcal{K}^4(X, Y)\right] = \frac{1}{p^2}\sum_{s=1}^{p}\sum_{t,u,v=s-3\kappa}^{s+3\kappa} E[\phi_{xy,s}\phi_{xy,t}\phi_{xy,u}\phi_{xy,v}] = O\left(\frac{\kappa^3}{p}\right).$$

and similar results can be shown for $E[\mathcal{K}^4(X, X')]$ and $E[\mathcal{K}^4(Y, Y')]$. Thus, Assumption 7 is satisfied if $\kappa^3/p = o(1)$.

Let Φ denote the cdf of $N(0, 1)$. We shall show that $\mathrm{ED}^k(\Gamma_w \mathbf{Z})$ converges uniformly with respect to w under the HDMSS setting.

Theorem 3.3. *For $w = 0, 1, 2, \cdots, \min\{n, m\}$, fix $\Gamma_w \in \mathbb{S}_w$,*

(i) Under Assumptions 1 and 5,

$$\sup_w \left|\mathrm{ED}_n^k(\Gamma_w \mathbf{Z}) - \mu_{n,w}\right| = o_p(1).$$

where $\mu_{n,w}$ is the same as that in Theorem 3.1.

(ii) Under Assumptions 1, 5, 6, 7 and local alternative H_{A_l},

$$\sup_w \left| P\left(\sqrt{nmp}\left(\mathrm{ED}_n^k(\Gamma_w \mathbf{Z}) - \mu_{n,w}\right) \leq a\right) - \Phi\left(\frac{a}{\sqrt{nm\sigma_{n,w}^2}}\right)\right| = o(1)$$

where a is a fixed constant and $\sigma_{n,w}^2$ is the same as in Theorem 3.1.

Then, the following theorem states the asymptotic distribution of $\mathrm{ED}_n^k(\mathbf{\Gamma Z})$.

Theorem 3.4. *For $\mathbf{\Gamma} \sim$ Uniform$(\mathbb{P}_{n+m})$, which is independent of the data,*

(i) Under Assumptions 1 and 5,

$$\mathrm{ED}_n^k(\mathbf{\Gamma Z}) \xrightarrow{p} 0.$$

(ii) Under Assumptions 1, 5, 6, 7 and local alternative H_{A_l},

$$\sqrt{nmp}\begin{pmatrix} \mathrm{ED}_n^k(\mathbf{\Gamma Z}) \\ \mathrm{ED}_n^k(\mathbf{\Gamma' Z}) \end{pmatrix} \xrightarrow{d} N\left(0, \begin{pmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{pmatrix}\right)$$

where $\mathbf{\Gamma}'$ is an independent copy of $\mathbf{\Gamma}$ and σ^2 is the asymptotic variance defined as

$$\sigma^2 := 4v_{xy}[\varphi^{(1)}(e_{xy})]^2 + 2\rho v_x[\varphi^{(1)}(e_x)]^2 + \frac{2}{\rho}v_y[\varphi^{(1)}(e_y)]^2.$$

We need the limiting distribution of $(\mathrm{ED}_n^k(\mathbf{\Gamma Z}), \mathrm{ED}_n^k(\mathbf{\Gamma' Z}))$ to show that the variance of randomization distribution go to 0, from which it follows that the randomization distribution converges in probability to the limit of its mean. Furthermore, we can show that the critical values are concentrating on some constants.

Corollary 3.1. *Let $\mathbf{\Gamma}_1, \cdots, \mathbf{\Gamma}_S$ be i.i.d and uniformly sampled from $\mathbb{P}_{n+m}$.*

(i) Under Assumptions 1 and 5, as $n \wedge m \wedge p \wedge S \to \infty$,

$$\widehat{R}(t) \xrightarrow{p} \mathbb{I}_{\{t \geq 0\}} \text{ and } \widetilde{R}(t) \xrightarrow{p} \mathbb{I}_{\{t \geq 0\}}.$$

Consequently, we have $Q_{\widehat{R},1-\alpha} \xrightarrow{p} 0$ and $Q_{\widetilde{R},1-\alpha} \xrightarrow{p} 0$.

(ii) Under Assumptions 1, 5, 6, 7 and local alternative H_{A_l}, as $n \wedge m \wedge p \wedge S \to \infty$,

$$\frac{1}{(n+m)!}\sum_{i=1}^{(n+m)!} \mathbb{I}_{\left\{\sqrt{nmp}\mathrm{ED}_n^k(\Gamma_i \mathbf{Z}) \leq t\right\}} \xrightarrow{p} \Phi(t/\sigma),$$

$$\frac{1}{S}\sum_{i=1}^{S} \mathbb{I}_{\left\{\sqrt{nmp}\mathrm{ED}_n^k(\mathbf{\Gamma}_i \mathbf{Z}) \leq t\right\}} \xrightarrow{p} \Phi(t/\sigma)$$

Consequently, we have $\sqrt{nmp}Q_{\widehat{R},1-\alpha} \xrightarrow{p} \sigma Q_{\Phi,1-\alpha}$ and $\sqrt{nmp}Q_{\widetilde{R},1-\alpha} \xrightarrow{p} \sigma Q_{\Phi,1-\alpha}$, where σ^2 is defined in Theorem 3.4.

The power behavior of $\mathrm{ED}_n^k(\mathbf{Z})$ w.r.t permutation test under the HDMSS is stated in the following theorem.

Theorem 3.5. *Suppose Assumption 1 is true and assume that $2\varphi(e_{xy}) \geq \varphi(e_x)+\varphi(e_y)$. For any $c \in \{Q_{\widehat{R},1-\alpha}, Q_{\widetilde{R},1-\alpha}\}$, the following holds.*

1, [**Consistency**] *Under Assumption 5, we have*

$$\lim_{p\to\infty} P_{H_{A_c}}\left(\mathrm{ED}_n^k(\mathbf{Z}) > c\right) = 1,$$

which means that the asymptotic power of ED^k *based permutation test is 1 as* $p \wedge n \wedge m \to \infty$.

2, [**Trivial Power**] *Under Assumptions 5, 6 and 7, we have*

$$\lim_{p\to\infty} P_{H_{A_l}}\left(\mathrm{ED}_n^k(\mathbf{Z}) > c\right) \leq \alpha,$$

Thus, we have the asymptotic power of ED^k *based permutation test is no more than the level* α *when* $p \wedge n \wedge m \to \infty$.

Comparing with Theorem 3.2, the $\mathrm{ED}_n^k(\mathbf{Z})$ based permutation test have trivial power under H_{A_l} and the HDMSS setting. This is due to the interesting facts that $nm\sigma_{n,W}^2$ converges in probability to σ^2, which is also the limit of $nm\sigma_{n,0}^2$ as $n \to \infty$ and

$$\mathrm{cov}\left(\mathrm{ED}_n^k(\mathbf{\Gamma Z}), \mathrm{ED}_n^k(\mathbf{\Gamma' Z})\right) \to 0 \text{ as } n \to \infty,$$

which ensures that the randomization distribution converges in probability to its mean limit.

4. Numerical Studies

In this section, we consider several examples to demonstrate the finite sample performance of ED^k based permutation test for different distance metrics. In our numerical comparison, we include the tests of Li [22] (denoted as JL) and Biswas and Ghosh [7] (denoted as BG) as these two were shown to have higher power over others in Li [22]. The critical values of JL test are determined by its asymptotic distribution, whereas BG test is also implemented as a permutation test.

4.1. Performance on simulated data

In all our simulations, we set $\alpha = 0.05$ and perform 1000 Monte Carlo replications with 300 permutations for each test. The first example is adopted from the simulation setting of [22] to study the size accuracy.

Example 4.1. *Generate samples as*

$$X = (V^{1/2}RV^{1/2})^{1/2}Z_1,$$
$$Y = (V^{1/2}RV^{1/2})^{1/2}Z_2,$$

where $R = (r_{ij})_{i,j=1}^{p}$, $r_{ij} = \rho^{|i-j|}$ *and* $\rho = 0.5$ *or* 0.8*;* V *is a diagonal matrix with* $V_{ii}^{1/2} = 1$ *or uniformly drawn from (1,5).* Z_1, Z_2 *are i.i.d copies of* Z *with*

$$Z = (\underbrace{z_1, z_2, \cdots, z_p}_{\overset{iid}{\sim} N(0,1)}) \text{ or } Z = (\underbrace{z_1, z_2, \cdots, z_p}_{\overset{iid}{\sim} Exponential(1)}) - \mathbf{1}_p.$$

In Example 4.1, X and Y follow the same distribution and we consider cases that $n = m = 50$ or $n = 70, m = 30$. From Table 2, we can see that all the tests have quite accurate size. To compare the power, we first use an example from [22], which include

Table 2. Size comparison from Example 4.1 for $p = 500$

	ρ	$V_{ii}^{1/2}$	n	m	ED$^{L^2\text{-norm}}$	ED$^{\text{Gaussian}}$	ED$^{\text{Laplacian}}$	ED$^{L^1\text{-norm}}$	BG	JL
Normal	0.5	1	50	50	0.06	0.06	0.058	0.059	0.053	0.053
	0.5	1	70	30	0.07	0.07	0.068	0.073	0.047	0.057
	0.5	Un(1,5)	50	50	0.052	0.052	0.05	0.051	0.056	0.057
	0.5	Un(1,5)	70	30	0.059	0.059	0.061	0.05	0.049	0.045
	0.8	1	50	50	0.053	0.053	0.052	0.059	0.054	0.055
	0.8	1	70	30	0.045	0.046	0.046	0.05	0.052	0.055
	0.8	Un(1,5)	50	50	0.045	0.045	0.049	0.048	0.054	0.054
	0.8	Un(1,5)	70	30	0.05	0.05	0.049	0.046	0.051	0.051
Exponential	0.5	1	50	50	0.06	0.06	0.058	0.059	0.053	0.053
	0.5	1	70	30	0.063	0.063	0.063	0.058	0.048	0.053
	0.5	Un(1,5)	50	50	0.057	0.057	0.058	0.055	0.049	0.06
	0.5	Un(1,5)	70	30	0.056	0.056	0.06	0.058	0.059	0.058
	0.8	1	50	50	0.054	0.054	0.051	0.047	0.065	0.062
	0.8	1	70	30	0.061	0.061	0.062	0.065	0.057	0.06
	0.8	Un(1,5)	50	50	0.051	0.05	0.052	0.046	0.045	0.057
	0.8	Un(1,5)	70	30	0.062	0.062	0.062	0.062	0.06	0.064

the situation when X and Y only differ in their means or only differ in their covariance matrices or differ in both, where $\beta \in [0, 1]$ is the percentage of the p components that differ in their distributions.

Example 4.2. *Let* R, V, Z_1, Z_2 *be defined the same as in Example 4.1 and we choose* $\rho = 0.5$ *here. Generate samples as*

(i)

$$X = (V^{1/2} R V^{1/2})^{1/2} Z_1,$$
$$Y = (0.125 \times \mathbf{1}_{\beta p}, \mathbf{0}_{(1-\beta)p}) + (V^{1/2} R V^{1/2})^{1/2} Z_2.$$

(ii) Let V^* *be a diagonal matrix with* $V_{ii}^{*1/2} = 1.05$ *for* $i = 1, 2, \cdots, \beta p$ *and* $V_{ii}^{*1/2} = 1$ *for* $i = \beta p + 1, \cdots, \beta p$.

$$X = (V^{1/2} R V^{1/2})^{1/2} Z_1,$$
$$Y = (V^{*1/2} R V^{*1/2})^{1/2} Z_2.$$

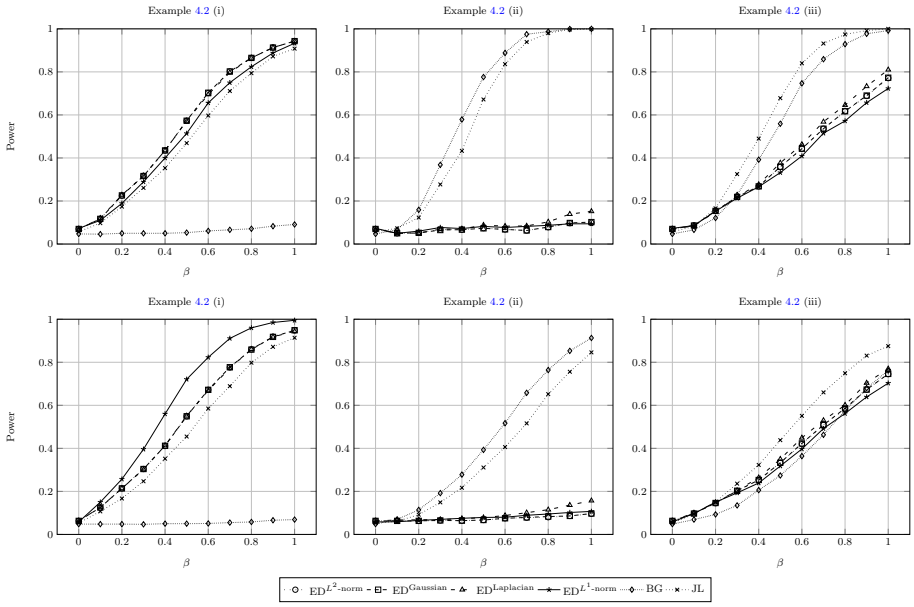


Figure 1: Power comparison for example 4.2 and $n = 70$, $m = 30$, $p = 500$, where in the top 3 figures Z_1, Z_2 are generated from normal distribution and in the bottom 3 figures, Z_1, Z_2 are generated from exponential distribution.

(iii) Let $V_{ii}^{*1/2} = 1.04$ *for* $i = 1, 2, \cdots, \beta p$ *and* $V_{ii}^{*1/2} = 1$ *for* $i = \beta p + 1, \cdots, \beta p$.

$$X = (V^{1/2} R V^{1/2})^{1/2} Z_1,$$
$$Y = (0.1 \times \mathbf{1}_{\beta p}, \mathbf{0}_{(1-\beta)p}) + (V^{*1/2} R V^{*1/2})^{1/2} Z_2.$$

From Figure 1, we can see that (1) when there is a small difference in the means, ED^k-based tests and JL perform similarly, while BG barely show any power. (2) when there is a small difference in the scales, JL and BG are consistent and ED^k-based tests have very little power. Similar phenomenon by Li [22] were also observed, i.e., ED^k based permutation test is not sensitive to small scale differences and the method proposed by Li [22] and Biswas and Ghosh [7] have dominant power in this case. Note that there is a tuning parameter involved in JL test and its choice could have a big impact on the size and power; results not shown. (3) when there are differences for both the means and scales, all the tests performs comparably.

Next, Example 4.3 examines the situation when X and Y have the same marginal univariate mean and variance, but different marginal univariate distributions.

Example 4.3. *Generate samples as*

(i) Let Rademacher(0.5) be the Rademacher distribution with success probability 0.5, e.g. $P(y_{iu} = -1) = P(y_{iu} = 1) = 0.5$.

$$X = (x_1, ..., x_p) \overset{iid}{\sim} N(0,1),$$
$$Y = (\underbrace{y_1, y_2, \cdots, y_{\beta p}}_{\overset{iid}{\sim} Rademacher(0.5)}, \underbrace{y_{\beta p+1}, y_{\beta p+2} \cdots, y_p}_{\overset{iid}{\sim} N(0,1)}).$$

(ii)

$$X = (x_1, ..., x_p) \overset{iid}{\sim} N(0,1),$$
$$Y = (\underbrace{y_1, y_2, \cdots, y_{\beta p}}_{\overset{iid}{\sim} Uniform(-\sqrt{3},\sqrt{3})}, \underbrace{y_{\beta p+1}, y_{\beta p+2} \cdots, y_p}_{\overset{iid}{\sim} N(0,1)}).$$

From Figure 2, we see that only ED$^{L^1\text{-norm}}$ based permutation test has power growing as β elevates (p fixed) or p increases (β fixed). This phenomenon matches with our theories, which indicate that L^2-norm, Gaussian and Laplacian kernel can detect only marginal mean and variance differences. For ED$^{L^1\text{-norm}}$ based permutation test, the power is growing more rapidly for Example 4.3 (i) than Example 4.3 (ii), which might suggest that L^1-distance is more sensitive for the difference between continuous and discrete distributions. It is also apparent that the JL and BG tests show little power in this example. The next example examines the case where X and Y have the same marginal univariate distributions.

Example 4.4. *Generate samples as*

(i) Let $(y'_1, y'_2, \cdots, y'_{\beta p/2}) \overset{iid}{\sim} Bernoulli(0.5)$

$$X = (x_1, ..., x_p) \overset{iid}{\sim} Bernoulli(0.5),$$
$$Y = \big(y'_1, \mathbb{I}_{\{y'_1=1\}}, y'_2, \mathbb{I}_{\{y'_2=1\}} \cdots, y'_{\beta p/2}, \mathbb{I}_{\{y'_{\beta p/2}=1\}}, \underbrace{y_1, y_2, \cdots, y_{(1-\beta)p}}_{\overset{iid}{\sim} Bernoulli(0.5)}\big).$$

(ii) Let $(y'_1, y'_2, \cdots, y'_{\beta p/3}) \overset{iid}{\sim} Bernoulli(0.5)$ *and* $(y''_1, y''_2, \cdots, y''_{\beta p/3}) \overset{iid}{\sim} Bernoulli(0.5)$

$$X = (x_1, ..., x_p) \overset{iid}{\sim} Bernoulli(0.5),$$
$$Y = \big(y'_1, y''_1, \mathbb{I}_{\{y'_1=y''_1\}}, \cdots, y'_{\beta p/3}, y''_{\beta p/3}, \mathbb{I}_{\{y'_{\beta p/3}=y''_{\beta p/3}\}}, \underbrace{y_1, y_2, \cdots, y_{(1-\beta)p}}_{\overset{iid}{\sim} Bernoulli(0.5)}\big).$$

Notice that in Example 4.4 (i) X, Y have the same marginal univariate distribution, but different marginal bivariate distributions and in Example 4.4 (ii) X, Y have the same

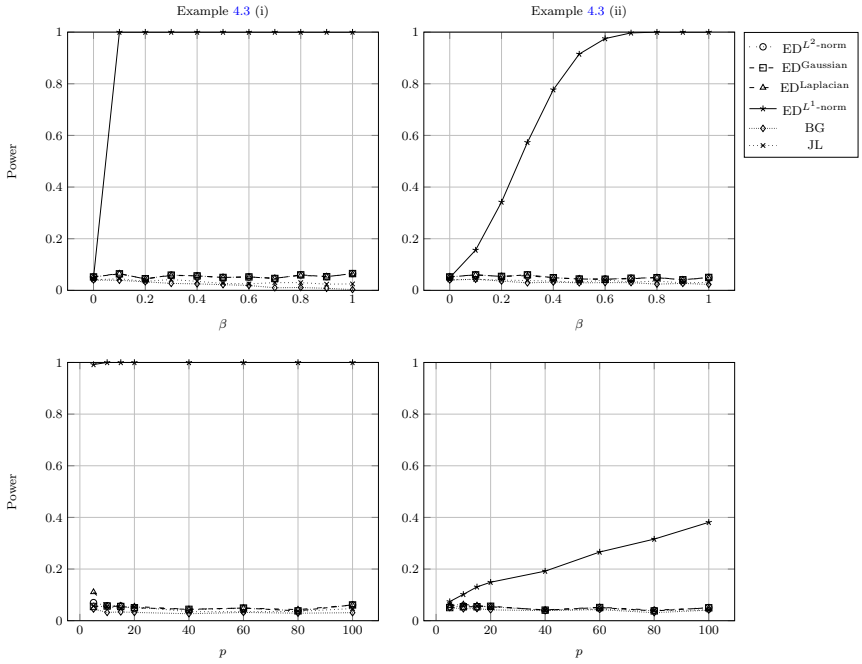


Figure 2: Power comparison for Example 4.3 and $n = 70$, $m = 30$. For the top two figures, the dimension p is equal to 500 and we plot the power as β ranges from 0 to 1. For the bottom two figures, β is fixed to be 0.5 and the power is plotted with respect to p.

marginal bivariate distribution, but different joint distribution. Theorem 3.2 (ii) and Theorem 3.5 (ii) both provide insights that L^2-norm, L^1-norm, Gaussian or Laplacian kernel based tests all suffer substantial power loss under Example 4.4 (i). On the other hand, Theorem 3.2 (iii) suggests us that since Example 4.4 (ii) belong to class H_{A_t}, all these tests have trivial power. The simulation results of Example 4.4 are in Figure 3 and they again corroborate our theoretical findings.

4.2. Performance on real data

We also compare the power of the above tests on the following real data sets.

- Strawberry data: this data set contains the spectrographs of fruit purees. There are totally two classes: one is strawberry purees (authentic samples) and the other one is non-strawberry purees (adulterated strawberries and other fruits). Each data point is of length 235.
- SmallKitchenAppliances data: this data sets contains records of the electricity usage of some kitchen appliances. We only use classes Kettle and Microwave. Each data

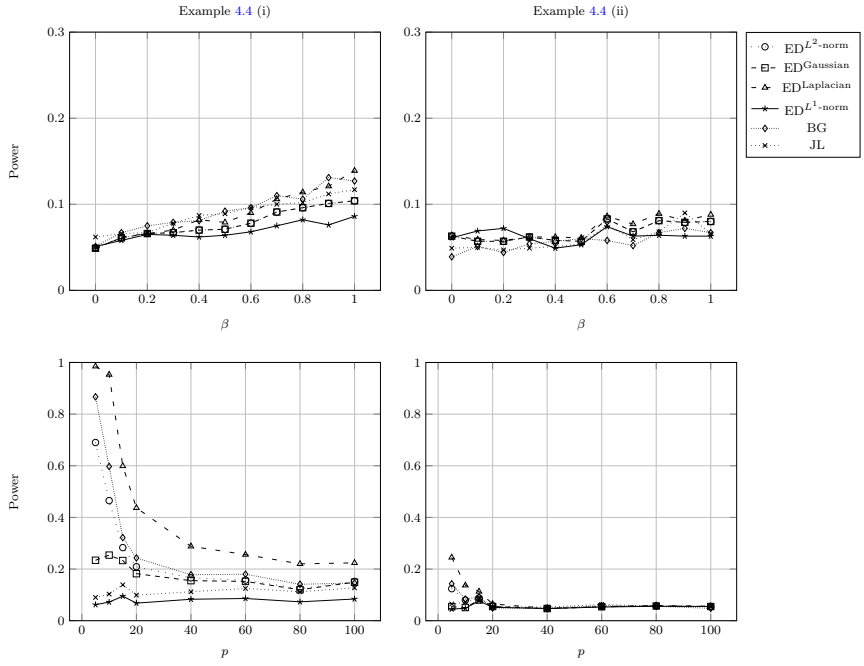


Figure 3: Power comparison for Example 4.4 and $n = 70$, $m = 30$. For the top two figures, the dimension p is equal to 500 and we plot the power as β ranges from 0 to 1. For the bottom two figures, β is fixed to be 1 and the power is plotted with respect to p.

point has readings taken every 2 minutes over 24 hours.

- Earthquakes data: this data set is from Northern California Earthquake Data Center and has classes of positive and negative major earthquake events. There are 368 negative and 93 positive cases and each data point is of length 512.

All the above data sets are downloaded from UCR Time Series Classification Archive [10] (https://www.cs.ucr.edu/~eamonn/time_series_data_2018/) and a glance of these data sets is provided in Figure 4. For each of the three data sets, the data points have two classes and we want to compare the underlining distributions of the two classes. Following the procedures of [7] and [24], for each $m = n \in \{10, 20, 30, 40, 50, 60\}$, we randomly sample n points from each class and test whether the two distributions are the same using the afore-mentioned tests. The same procedure is repeated 1000 times to calculate the power.

The experimental results for these data sets are shown in Figure 5, from which we see that all the tests have very high power for the Strawberry data with relatively low sample size. As for the SmallKitchenAppliances and Earthquakes data sets, the L^1-norm based test demonstrates superior power compared to other tests. It is also worth noting that BG and JL barely exhibit any power for the Earthquakes data.

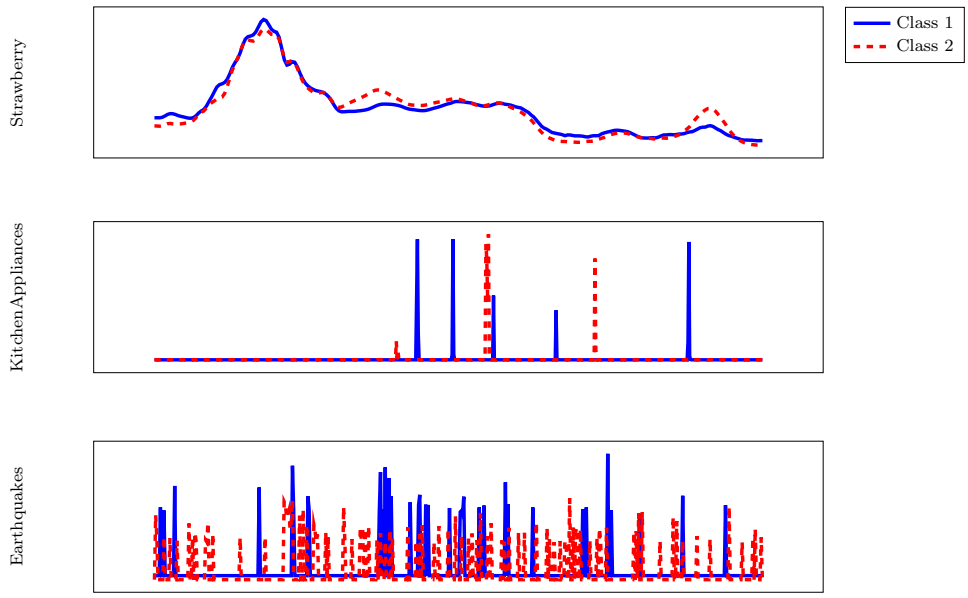


Figure 4: A glance of the data in Section 4.2, where we plot one point from each of the two classes for each data set.

5. Discussions&Conclusion

In this paper, we study the two-sample hypothesis testing problem in a high dimension and low/medium sample size setting. Our focus is on the interpoint distance based permutation tests, such as those based on Energy Distance (ED) and Maximum Mean Discrepancy (MMD). Our theory demonstrates that all these tests under examination are unable to detect the difference between two high dimensional distributions beyond univariate marginal distributions. In particular, the ED test with L^2-norm and MMD with Gaussian or Laplacian kernels suffer substantial power loss under the HDLSS and have trivial power under the HDMSS when the average of component-wise mean and variance discrepancies between two distributions are both asymptotically zero at the rate of $o(1/\sqrt{nmp})$. Thus these tests mainly target mean and variance differences in marginal distributions. By contrast, if we use L^1-norm in ED test, then the non-negligible difference in marginal univariate distributions, as quantified by cumulative energy distance of marginal distributions, can be detected with high power. Thus the theory suggests that

1), The ED with L^2-norm, and MMD with Gaussian and Laplacian kernels are of the same category, as they all depend on the interpoint distance as measured by Euclidean distance, which leads to undesirable power limitation.

2), Although in a low dimensional setting the use of L^1-norm in ED is not preferred due to the fact that it does not completely characterize the difference between two distributions since $\mathrm{ED}^1(F,G) = 0$ does not necessarily imply $F = G$, it seems to have some

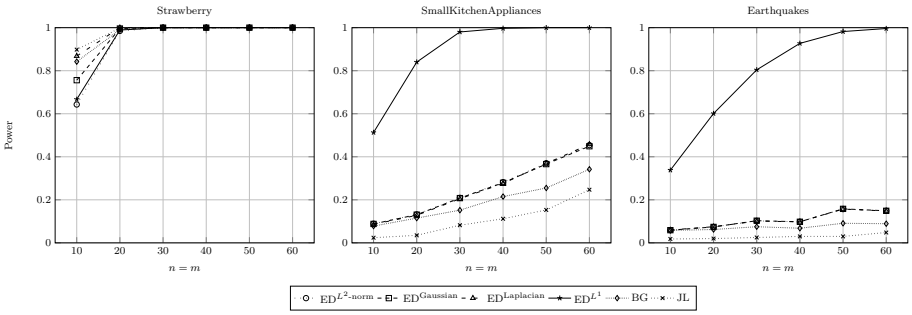


Figure 5: Power comparison for real data examples in Section 4.2.

advantage over the ED with L^2-norm and MMD with Gaussian and Laplacian kernels in the high dimensional setting, as shown in both theory and numerical studies. Interestingly, for each fixed p, it is shown in [25] that a class of methods based on kernel embedding have better power if using L_1-norm to differentiate the expectations of analytic kernels evaluated at some well-chosen distribution locations instead of L^2-norm, since these features are dense due to the use of analytic functions.

3), As shown in our simulations and data illustration, the existing interpoint distance test by [22] and [7] also suffer from low power when the two distributions have the same marginal mean and variances but different marginal distributions. So in this sense, they are also inferior to the ED test with L^1-norm.

4), The difference in marginal distributions of two high dimensional distributions can be interpreted as the main effect of the distribution differences. It is a standard statistical practice to test for the nullity of main effects first, before proceeding to the higher-order interactions. Thus we advocate the use of L^1-norm based test to test for the presence of main differences in two high dimensional distributions.

To conclude the paper, we shall mention a few future directions. First, we are holding the bandwidth parameter in Gaussian and Laplacian kernels fixed for theoretical convenience, and it would be interesting to relax this restriction by allowing it to be data-dependent. Second, there might be some intrinsic difficulty of capturing all kinds of differences in two high dimensional distributions with limited sample sizes, so it seems natural to ask whether it is possible to detect any difference beyond marginal univariate distributions. If possible, what would be the form of the new tests? We leave these topics for future investigation.

Appendix A: Technical Details

A.1. Proof of Sufficient Conditions for Local Alternatives

When $\psi(x,y) = (x-y)^2$, φ is strictly concave, strictly increasing on $(0,+\infty)$ (e.g. L^2-norm, Gaussian kernel myltiplied by -1 and Laplacian kernel multiplied by -1), we first

note that $2e_{xy} - e_x - e_y = 2\lim_{p\to\infty}\sum_{u=1}^{p}(E(x_u) - E(y_u))^2/p \geq 0$ and

$$\varphi(e_{xy}) - \frac{\varphi(e_x) + \varphi(e_y)}{2} \geq \varphi(e_{xy}) - \varphi\left(\frac{e_x + e_y}{2}\right) \geq 0,$$

where the equality holds iff $e_{xy} = e_x = e_y$. Also, some algebra shows that

$$\begin{aligned} e_{xy} &= e_x + \lim_{p\to\infty}\frac{1}{p}\sum_{u=1}^{p}(E(x_u) - E(y_u))^2 + \lim_{p\to\infty}\frac{1}{p}\sum_{u=1}^{p}(var(y_u) - var(x_u)) \\ &= e_y + \lim_{p\to\infty}\frac{1}{p}\sum_{u=1}^{p}(E(x_u) - E(y_u))^2 + \lim_{p\to\infty}\frac{1}{p}\sum_{u=1}^{p}(var(x_u) - var(y_u)). \end{aligned}$$

Thus, in summary we have

$$\begin{aligned} 2\varphi(e_{xy}) = \varphi(e_x) + \varphi(e_y) &\Leftrightarrow e_{xy} = e_x = e_y \\ &\Leftrightarrow \sum_{u=1}^{p}(E(x_u) - E(y_u))^2 = o(p) \text{ and } \left|\sum_{u=1}^{p}(var(x_u) - var(y_u))\right| = o(p). \end{aligned}$$

This proves the result for H_{A_c} characterization. Next, for sufficient conditions of H_{A_l}, if we have

$$\sum_{u=1}^{p}(E(x_u) - E(y_u))^2 = o\left(\tfrac{\sqrt{p}}{\sqrt{nm}}\right) \text{ and } \left|\sum_{u=1}^{p}(var(x_u) - var(y_u))\right| = o\left(\tfrac{\sqrt{p}}{\sqrt{nm}}\right),$$

then it holds that $e_{xy} = e_x = e_y$ and

$$\begin{aligned} &E\left[\left|E\left[\overline{\psi}(X,Y)|X\right] - E\left[\overline{\psi}(X,X')|X\right]\right|\right] \\ &\quad\leq \frac{2}{p}\sqrt{\sum_{u=1}^{p}E(x_u^2)\sum_{u=1}^{p}(E(x_u) - E(y_u))^2} + \frac{1}{p}\left|\sum_{u=1}^{p}(\mathrm{var}(y_u) - \mathrm{var}(x_u))\right| \\ &\qquad + \frac{1}{p}\sqrt{\sum_{u=1}^{p}(E(x_u) + E(y_u))^2\sum_{u=1}^{p}(E(x_u) - E(y_u))^2} = o\left(\frac{1}{\sqrt{nmp}}\right). \end{aligned}$$

For H_{A_t}, a straight forward calculation shows that

$$\psi(x,y) - E\left[\psi(x,y)|x\right] - E\left[\psi(x,y)|y\right] + E\left[\psi(x,y)\right] = -2(x - E(x))(y - E(y))$$

and $v_{xy} = \sum_{u,v=1}^{p} 4\mathrm{cov}(x_u, x_v)\mathrm{cov}(y_u, y_v)/p$. Thus, from Cauchy–Schwarz inequality, we have

$$\sum_{u,v=1}^{p}(\mathrm{cov}(x_u, x_v) - \mathrm{cov}(y_u, y_v))^2 = o(p) \Rightarrow v_{xy} = v_x = v_y.$$

When $\psi(x,y)=|x-y|$, $\varphi(x)=x$, the results follow from the following equality.

$$\begin{aligned}
&2\varphi(e_{xy})-\varphi(e_x)-\varphi(e_y)\\
=&2\lim_{p\to\infty}\frac{1}{p}\sum_{u=1}^{p}E\left[|x_{1u}-y_{1u}|\right]-\lim_{p\to\infty}\frac{1}{p}\sum_{u=1}^{p}E\left[|x_{1u}-x_{2u}|\right]-\lim_{p\to\infty}\frac{1}{p}\sum_{u=1}^{p}E\left[|y_{1u}-y_{2u}|\right]\\
=&\lim_{p\to\infty}\frac{1}{p}\sum_{u=1}^{p}\left(2E\left[|x_{1u}-y_{1u}|\right]-E\left[|x_{1u}-x_{2u}|\right]-E\left[|y_{1u}-y_{2u}|\right]\right)\\
=&\lim_{p\to\infty}\frac{1}{p}\sum_{u=1}^{p}\mathrm{ED}(F_u,G_u).
\end{aligned}$$

A.2. Proof of Theorem 3.1

Proof. (i) Taking a first order Taylor expansion w.r.t φ gives

$$k(Z_i,Z_j)=\varphi\left(e_{ij}\right)+\mathcal{R}_1(Z_i,Z_j),$$

where $\mathcal{R}_1(Z_i,Z_j)$ is an operator that acts on random variables

$$\mathcal{R}_1(Z_i,Z_j)=\mathcal{L}(Z_i,Z_j)\int_0^1\varphi^{(1)}\left(e_{ij}+v\mathcal{L}(Z_i,Z_j)\right)dv$$

For each fixed permutation matrix $\Gamma_w\in\mathbb{S}_w$

$$\mathrm{ED}_n^k(\Gamma_w\mathbf{Z})=\underbrace{\sum_{i=2}^{n+m}\sum_{j=1}^{i-1}\Pi_{w,ij}\varphi\left(e_{ij}\right)}_{:=\mu_n(\Gamma_w\mathbf{Z})}+\underbrace{\sum_{i=2}^{n+m}\sum_{j=1}^{i-1}\Pi_{w,ij}\mathcal{R}_1(Z_i,Z_j)}_{:=R_1(\Gamma_w\mathbf{Z})},\tag{7}$$

where $\mu_n(\Gamma_w\mathbf{Z})$ is the asymptotic mean for the permuted data and equals

$$\begin{aligned}
&\frac{2}{mn}\left(\left(w^2+(n-w)(m-w)\right)\varphi(e_{xy})+(n-w)w\varphi(e_x)+(m-w)w\varphi(e_y)\right)\\
&\quad-\frac{1}{n(n-1)}\left(2w(n-w)\varphi(e_{xy})+(n-w)(n-w-1)\varphi(e_x)+w(w-1)\varphi(e_y)\right)\\
&\quad-\frac{1}{m(m-1)}\left(2w(m-w)\varphi(e_{xy})+w(w-1)\varphi(e_x)+(m-w)(m-w-1)\varphi(e_y)\right).
\end{aligned}$$

Then, after re-arranging the terms according to the powers of w, we have

$$\mu_n(\Gamma_w\mathbf{Z})=\left(2\varphi(e_{xy})-\varphi(e_x)-\varphi(e_y)\right)f(w),$$

where $f(w)$ is a second order polynomial with respect to w

$$f(w)=1-\left(\frac{2m-1}{m(m-1)}+\frac{2n-1}{n(n-1)}\right)w+\left(\frac{2}{mn}+\frac{1}{n(n-1)}+\frac{1}{m(m-1)}\right)w^2.$$

For the remainder term $R_1(\Gamma_w\mathbf{Z})$, notice that $\mathcal{L}(Z_i, Z_j) \xrightarrow{p} 0$ for any $1 \le i < j < n+m$. By the continuous mapping theorem, we know

$$\int_0^1 \varphi^{(1)}(e_{ij} + v\mathcal{L}(Z_i, Z_j))dv \xrightarrow{p} \int_0^1 \varphi^{(1)}(e_{ij})dv.$$

Thus, it holds that $\mathcal{R}_1(Z_i, Z_j) \asymp_p \mathcal{L}(Z_i, Z_j)$ and $R_1(\Gamma_w\mathbf{Z}) = O_p(\alpha_{xy} + \alpha_x + \alpha_y) = o_p(1)$.

(ii) Taking a second order Taylor expansion w.r.t φ gives

$$k(Z_i, Z_j) = \varphi\left(e_{ij}\right) + \varphi^{(1)}\left(e_{ij}\right)\frac{\mathcal{K}(Z_i, Z_j) + \mathcal{W}(Z_i, Z_j)}{\sqrt{p}} + \mathcal{R}_2(Z_i, Z_j),$$

where $\mathcal{R}_2$ and $\mathcal{W}$ are defined as

$$\mathcal{R}_2(Z_i, Z_j) = \mathcal{L}^2(Z_i, Z_j)\int_0^1\int_0^1 u\varphi^{(2)}\left(e_{ij} + uv\mathcal{L}(Z_i, Z_j)\right)dvdu,$$

$$\mathcal{W}(Z_i, Z_j) = \frac{1}{\sqrt{p}}\sum_{u=1}^{p}\left(E[\psi(z_{iu}, z_{ju})|z_{iu}] + E[\psi(z_{iu}, z_{ju})|z_{ju}] - E[\psi(z_{iu}, z_{ju})] - e_{ij}\right).$$

Accordingly, we can decompose the sample energy distance as

$$\begin{aligned}\sqrt{p}\left(\mathrm{ED}_n^k(\Gamma_w\mathbf{Z}) - \mu_n(\Gamma_w\mathbf{Z})\right) = &\underbrace{\sum_{i=2}^{n+m}\sum_{j=1}^{i-1}\Pi_{w,ij}\varphi^{(1)}\left(e_{ij}\right)\mathcal{K}(Z_i, Z_j)}_{:=L(\Gamma_w\mathbf{Z})} \\ &+ \underbrace{\sum_{i=2}^{n+m}\sum_{j=1}^{i-1}\Pi_{w,ij}\varphi^{(1)}(e_{ij})\mathcal{W}(Z_i, Z_j)}_{:=R_l(\Gamma_w\mathbf{Z})} + \underbrace{\sqrt{p}\sum_{i=2}^{n+m}\sum_{j=1}^{i-1}\Pi_{w,ij}\mathcal{R}_2(Z_i, Z_j)}_{:=R_2(\Gamma_w\mathbf{Z})}. \quad (8)\end{aligned}$$

For the leading term $L(\Gamma_w\mathbf{Z})$, notice that under Assumption 2, $(\mathcal{K}(Z_i, Z_j))_{i<j}$ converges jointly to a multivariate normal with mean 0 and a diagonal covariance matrix. Thus,

given a permutation matrix Γ_w, we are able to obtain $L(\Gamma_w\mathbf{Z}) \xrightarrow{d} N(0, \sigma_n^2(\Gamma_w\mathbf{Z}))$, where

$$\begin{aligned}\sigma_n^2(\Gamma_w\mathbf{Z}) =& \frac{4}{n^2(n-1)^2}\Bigg\{\frac{(n-w)(n-w-1)}{2}v_x[\varphi^{(1)}(e_x)]^2 \\ &+ \frac{w(w-1)}{2}v_y[\varphi^{(1)}(e_y)]^2 + (n-w)wv_{xy}[\varphi^{(1)}(e_{xy})]^2\Bigg\} \\ &+\frac{4}{m^2(m-1)^2}\Bigg\{\frac{w(w-1)}{2}v_x[\varphi^{(1)}(e_x)]^2 \\ &+ \frac{(m-w)(m-w-1)}{2}v_y[\varphi^{(1)}(e_y)]^2 + w(m-w)v_{xy}[\varphi^{(1)}(e_{xy})]^2\Bigg\} \\ &+\frac{4}{n^2m^2}\Bigg\{(n-w)wv_x[\varphi^{(1)}(e_x)]^2 \\ &+ w(m-w)v_y[\varphi^{(1)}(e_y)]^2 + ((n-w)(m-w)+w^2)v_{xy}[\varphi^{(1)}(e_{xy})]^2\Bigg\}.\end{aligned}$$

By collecting terms with respect to v_{xy}, v_x, v_y, we obtain

$$\begin{aligned}\sigma_{n,w}^2 =& \Bigg\{\frac{4}{nm} - 4\left(\frac{n+m}{n^2m^2} - \frac{n}{n^2(n-1)^2} - \frac{m}{m^2(m-1)^2}\right)w \\ &+ 4\left(\frac{2}{n^2m^2} - \frac{1}{n^2(n-1)^2} - \frac{1}{m^2(m-1)^2}\right)w^2\Bigg\}v_{xy}[\varphi^{(1)}(e_{xy})]^2 \\ &+\Bigg\{\frac{2}{n(n-1)} + 2\left(\frac{2n}{n^2m^2} - \frac{2n-1}{n^2(n-1)^2} - \frac{1}{m^2(m-1)^2}\right)w \\ &- 2\left(\frac{2}{m^2n^2} - \frac{1}{n^2(n-1)^2} - \frac{1}{m^2(m-1)^2}\right)w^2\Bigg\}v_x[\varphi^{(1)}(e_x)]^2 \\ &+\Bigg\{\frac{2}{m(m-1)} + 2\left(\frac{2m}{n^2m^2} - \frac{1}{n^2(n-1)^2} - \frac{2m-1}{m^2(m-1)^2}\right)w \\ &- 2\left(\frac{2}{n^2m^2} - \frac{1}{m^2(m-1)^2} - \frac{1}{n^2(n-1)^2}\right)w^2\Bigg\}v_y[\varphi^{(1)}(e_y)]^2.\end{aligned}$$

We then conclude the result by showing that the remainder terms are negligible. $R_l(\Gamma_w\mathbf{Z}) = o_p(1)$ is proved in lemma A.3. For the $R_2(\Gamma_w\mathbf{Z})$ term, it can be shown similarly that $\mathcal{R}_2(Z_i, Z_j) \asymp_p \mathcal{L}^2(Z_i, Z_j) = O_p(\alpha_{xy}^2+\alpha_x^2+\alpha_y^2)$, which implies that $R_2(\Gamma_w\mathbf{Z}) = O_p(\sqrt{p}(\alpha_{xy}^2+\alpha_x^2+\alpha_y^2)) = o_p(1)$ under Assumption 4. □

A.3. Proof of Proposition 3.1

Proof. (i) From Equation 7, we obtain

$$\mathrm{ED}_n^k(\mathbf{\Gamma Z}) = \underbrace{\sum_{i=2}^{n+m}\sum_{j=1}^{i-1}\mathbf{\Pi}_{ij}\varphi\left(e_{ij}\right)}_{:=\mu_{n,W}} + o_p(1),$$

where $\mathbf{\Pi}_{ij}$ corresponds to $\mathbf{\Gamma}$ and $W = N(\mathbf{\Gamma}) \sim \text{Hypergeometric}(m+n, m, n)$.

(ii) It follows from Equation 8, Lemma A.3 and the proof of Theorem 3.1 that

$$\sqrt{p}\left(\mathrm{ED}_n^k(\mathbf{\Gamma Z}) - \mu_n(\mathbf{\Gamma Z})\right) = \underbrace{\sum_{i=2}^{n+m}\sum_{j=1}^{i-1}\mathbf{\Pi}_{ij}\varphi^{(1)}\left(e_{ij}\right)\mathcal{K}(Z_i, Z_j)}_{:=L(\mathbf{\Gamma Z})} + o_p(1).$$

Then, under Assumption 2, it is not hard to see that $L(\mathbf{\Gamma Z}) \xrightarrow{d} N(0, \sigma^2_{n,W})$, where $W = N(\mathbf{\Gamma}) \sim \text{Hypergeometric}(m+n, m, n)$. This concludes the proposition. □

A.4. Proof of Theorem 3.2

1, For any $\mathbf{a} \in \mathbb{R}^{(n+m)!}$, we define the α-th quantile of the set $\{a_1, \cdots, a_{(n+m)!}\}$ as

$$Q_{1-\alpha}\left\{a_1, \cdots, a_{(n+m)!}\right\} = \min\left\{a_i : \frac{1}{(n+m)!}\sum_{i=1}^{(n+m)!}\mathbb{I}_{\{a_i \le t\}} \ge 1-\alpha\right\}.$$

Then, we can view $Q_{1-\alpha}$ as a continuous function on $\mathbb{R}^{(n+m)!}$.

(i) By Theorem 3.1, for any fixed $\Gamma_i \in \mathbb{P}_{n+m}$, we have $\mathrm{ED}_n^k(\Gamma_i\mathbf{Z}) \xrightarrow{p} \mu_n(\Gamma_i\mathbf{Z})$. The continuous mapping theorem implies

$$Q_{1-\alpha}\left\{\mathrm{ED}_n^k(\Gamma_1\mathbf{Z}), \cdots, \mathrm{ED}_n^k(\Gamma_{(n+m)!}\mathbf{Z})\right\} \xrightarrow{p} Q_{1-\alpha}\left\{\mu_n(\Gamma_1\mathbf{Z}), \cdots, \mu_n(\Gamma_{(n+m)!}\mathbf{Z})\right\}.$$

Then, it follows from the definition of $\mu_n(\cdot)$ in Theorem 3.1 that

$$\mu_n(\mathbf{Z}) = \mu_n(\Gamma_0\mathbf{Z}) = \max\left\{\mu_n(\Gamma_1\mathbf{Z}), \cdots, \mu_n(\Gamma_{(n+m)!}\mathbf{Z})\right\}.$$

Notice that

$$\begin{cases} \frac{n!m!}{(n+m)!} < 1-\alpha & \text{if } m \neq n, \\ \frac{2(n!)^2}{(2n)!} < 1-\alpha & \text{if } m = n, \end{cases} \quad \text{implies} \quad \begin{cases} \frac{|\mathbb{S}_0|}{(n+m)!} < 1-\alpha & \text{if } m \neq n, \\ \frac{|\mathbb{S}_0|+|\mathbb{S}_{\min\{n,m\}}|}{(n+m)!} < 1-\alpha & \text{if } m = n, \end{cases}$$

and so $\mu_n(\Gamma_0\mathbf{Z}) > Q_{1-\alpha}\left\{\mu_n(\Gamma_1\mathbf{Z}),\cdots,\mu_n(\Gamma_{(n+m)!}\mathbf{Z})\right\}$. Thus, as $p\to\infty$, we conclude

$$P\left(\mathrm{ED}_n^k(\mathbf{Z}) > Q_{1-\alpha}\left\{\mathrm{ED}_n^k(\Gamma_1\mathbf{Z}),\cdots,\mathrm{ED}_n^k(\Gamma_{(n+m)!}\mathbf{Z})\right\}\right) \\ \to P\left(\mu_n(\Gamma_0\mathbf{Z}) > Q_{1-\alpha}\left\{\mu_n(\Gamma_1\mathbf{Z}),\cdots,\mu_n(\Gamma_{(n+m)!}\mathbf{Z})\right\}\right) = 1.$$

(ii) For any random permutation matrix $\mathbf{\Gamma}_s$, by Proposition 3.1, we have $\mathrm{ED}_n^k(\mathbf{\Gamma}_s\mathbf{Z}) \xrightarrow{p} \mu_{n,W_s}$, where $W_s = N(\mathbf{\Gamma}_s) \sim \text{Hypergeometric}(n+m,m,n)$. Then, the continuous mapping theorem implies that

$$P\left(\mathrm{ED}_n^k(\mathbf{Z}) > Q_{1-\alpha}\left\{\mathrm{ED}_n^k(\mathbf{Z}),\mathrm{ED}_n^k(\mathbf{\Gamma}_1\mathbf{Z}),\cdots,\mathrm{ED}_n^k(\mathbf{\Gamma}_{S-1}\mathbf{Z})\right\}\right) \\ \to P\left(\mu_{n,0} > Q_{1-\alpha}\left\{\mu_{n,0},\mu_{n,W_1},\cdots,\mu_{n,W_{S-1}}\right\}\right).$$

Since $\mu_{n,0} = \max_w \mu_{n,w}$, in order to have $\mu_{n,0} = Q_{1-\alpha}\left\{\mu_{n,0},\mu_{n,W_1},\cdots,\mu_{n,W_{S-1}}\right\}$, at least $\lfloor \alpha S\rfloor + 1$ elements of $\left\{\mu_{n,0},\mu_{n,W_1},\cdots,\mu_{n,W_{S-1}}\right\}$ should be equal to $\mu_{n,0}$. Thus, we get

$$P\left(\mu_{n,0} > Q_{1-\alpha}\left\{\mu_{n,0},\mu_{n,W_1},\cdots,\mu_{n,W_{S-1}}\right\}\right) \\ = 1 - P\left(\mu_{n,0} = Q_{1-\alpha}\left\{\mu_{n,0},\mu_{n,W_1},\cdots,\mu_{n,W_{S-1}}\right\}\right) \\ \geq \begin{cases} 1 - \frac{S-1}{\lfloor \alpha S\rfloor}\frac{n!m!}{(n+m)!}, & \text{if } n\neq m, \\ 1 - \frac{S-1}{\lfloor \alpha S\rfloor}\frac{2(n!)^2}{(n+m)!}, & \text{if } n = m. \end{cases}$$

2, (i) Since $\mu_n(\Gamma_u\mathbf{Z}) = 0$ for all $u = 1,2,\cdots,(n+m)!$ under H_{A_l}, Assumption 2 implies that

$$\sqrt{p}\mathrm{ED}_n^k(\Gamma_u\mathbf{Z}) \xrightarrow{d} \sum_{i=1}^{n}\sum_{j=1}^{m}\Pi_{u,ij}b_{ij} - \sum_{1\leq i<j\leq n}\Pi_{u,ij}c_{ij} - \sum_{1\leq i<j\leq m}\Pi_{u,ij}d_{ij},$$

where $\Pi_{u,ij}$ corresponds to Γ_u. Then, the continuous mapping theorem entails

$$P\left(\mathrm{ED}_n^k(\mathbf{Z}) > Q_{1-\alpha}\left\{\mathrm{ED}_n^k(\Gamma_1\mathbf{Z}),\cdots,\mathrm{ED}_n^k(\Gamma_{(n+m)!}\mathbf{Z})\right\}\right) \\ = P\left(\sqrt{p}\mathrm{ED}_n^k(\mathbf{Z}) > Q_{1-\alpha}\left\{\sqrt{p}\mathrm{ED}_n^k(\Gamma_1\mathbf{Z}),\cdots,\sqrt{p}\mathrm{ED}_n^k(\Gamma_{(n+m)!}\mathbf{Z})\right\}\right) \\ \to P(V(\Gamma_0) > Q_{\widehat{T},1-\alpha}).$$

(ii) Conditioned on $\mathbf{\Gamma}_1,\mathbf{\Gamma}_2,\cdots,\mathbf{\Gamma}_{S-1}$, the result can be shown similarly with part(i). Then, since the number of permutations is fixed and finite, the unconditioned version follows straightforwardly.

3, (i) By construction, we have

$$\frac{1}{(n+m)!}\sum_{u=1}^{(n+m)!}\mathbb{I}_{\left\{V(\Gamma_u) > Q_{1-\alpha}\left\{V(\Gamma_1),\cdots,V(\Gamma_{(n+m)!})\right\}\right\}} \leq \alpha.$$

If $\varphi'(e_{xy}) = \varphi'(e_x) = \varphi'(e_y)$ and $v_{xy} = v_x = v_y$, then $V(\Gamma_u) =^d V(\Gamma_0)$ for any $u = 1, 2, \cdots, (n+m)!$ and so

$$\left(V(\Gamma_u), Q_{1-\alpha}\left\{V(\Gamma_1), \cdots, V(\Gamma_{(n+m)!})\right\}\right) =^d \left(V(\Gamma_0), Q_{1-\alpha}\left\{V(\Gamma_1), \cdots, V(\Gamma_{(n+m)!})\right\}\right).$$

Thus, we have

$$P\left(V(\Gamma_0) > Q_{1-\alpha}\left\{V(\Gamma_1), \cdots, V(\Gamma_{(n+m)!})\right\}\right) = E\left[\frac{1}{(n+m)!}\sum_{u=1}^{(n+m)!} \mathbb{I}_{\left\{V(\Gamma_u) > Q_{1-\alpha}\left\{V(\Gamma_1), \cdots, V(\Gamma_{(n+m)!})\right\}\right\}}\right] \leq \alpha.$$

(ii) The proof follows similarly from part (i) by observing that for any $s = 1, 2, \cdots, S$

$$\left(V(\mathbf{\Gamma}_s), Q_{1-\alpha}\left\{V(\Gamma_0), V(\mathbf{\Gamma}_1), \cdots, V(\mathbf{\Gamma}_{S-1})\right\}\right) =^d \left(V(\Gamma_0), Q_{1-\alpha}\left\{V(\Gamma_0), V(\mathbf{\Gamma}_1), \cdots, V(\mathbf{\Gamma}_{S-1})\right\}\right).$$

A.5. Proof of Theorem 3.3

(i) Recall that for a fixed permutation matrix $\mathit{\Gamma}_w \in \mathbb{S}_w$ that corresponds to $\mathit{\Pi}_{w,ij}$,

$$\mathrm{ED}_n^k(\mathit{\Gamma}_w\mathbf{Z}) = \underbrace{\sum_{i=2}^{n+m}\sum_{j=1}^{i-1}\mathit{\Pi}_{w,ij}\varphi\left(e_{ij}\right)}_{:=\mu_n(\mathit{\Gamma}_w\mathbf{Z})} + \underbrace{\sum_{i=2}^{n+m}\sum_{j=1}^{i-1}\mathit{\Pi}_{w,ij}\mathcal{R}_1(Z_i, Z_j)}_{:=R_1(\mathit{\Gamma}_w\mathbf{Z})}.$$

Under the HDMSS setting, part (i) follows from Lemma A.1.

Lemma A.1. *Under Assumption 5,* $\sup_\Gamma |R_1(\Gamma\mathbf{Z})| = o_p(1)$.

Proof. Consider the events $B_{\mathbf{XY}}, B_{\mathbf{X}}, B_{\mathbf{Y}}$ and their complements $B^c_{\mathbf{XY}}, B^c_{\mathbf{X}}, B^c_{\mathbf{Y}}$, where

$$B_{\mathbf{XY}} = \left\{\min_{1\leq s\leq n, 1\leq t\leq m}\mathcal{L}(X_s, Y_t) \leq -\frac{1}{2}e_{xy} \text{ or } \max_{1\leq s\leq n, 1\leq t\leq m}\mathcal{L}(X_s, Y_t) \geq \frac{1}{2}e_{xy}\right\},$$
$$B_{\mathbf{X}} = \left\{\min_{1\leq s\neq t\leq n}\mathcal{L}(X_s, X_t) \leq -\frac{1}{2}e_{x} \text{ or } \max_{1\leq s\neq t\leq n}\mathcal{L}(X_s, X_t) \geq \frac{1}{2}e_{x}\right\},$$
$$B_{\mathbf{Y}} = \left\{\min_{1\leq s\neq t\leq m}\mathcal{L}(Y_s, Y_t) \leq -\frac{1}{2}e_{y} \text{ or } \max_{1\leq s\neq t\leq m}\mathcal{L}(Y_s, Y_t) \geq \frac{1}{2}e_{y}\right\}.$$

Then, under assumption 5, as $n \wedge m \wedge p \to \infty$

$$\begin{aligned}
P(B_{\mathbf{XY}}) &= P\left(\bigcup_{1\le s\le n, 1\le t\le m}\left\{\mathcal{L}(X_s, Y_t) \le -\frac{1}{2}e_{xy} \text{ or } \mathcal{L}(X_s, Y_t) \ge \frac{1}{2}e_{xy}\right\}\right) \\
&\le \sum_{1\le s\le n, 1\le t\le m} P\left(|\mathcal{L}(X_s, Y_t)| \ge \frac{1}{2}e_{xy}\right) \\
&\le nmP\left(|\mathcal{L}(X, Y)| \ge \frac{1}{2}e_{xy}\right) \\
&\le \frac{4nmE\left[\mathcal{L}(X, Y)^2\right]}{e_{xy}^2} \\
&= o(1).
\end{aligned}$$

Similarly, we can show that $P(B_{\mathbf{X}}) = o(1)$ and $P(B_{\mathbf{Y}}) = o(1)$. Conditioned on event $B_{\mathbf{XY}}^c B_{\mathbf{X}}^c B_{\mathbf{Y}}^c$, we have $e_{ij} \le e_{ij} + v\mathcal{L}(Z_i, Z_j) \le 3e_{ij}/2$ for any $0 \le v \le 1$. Suppose $\varphi^{(1)}(\cdot)$ is a continuous function on $(0, +\infty)$, we know there exist a constant C such that $|\varphi^{(1)}(e_{ij} + v\mathcal{L}(Z_i, Z_j))| \le C$ and consequently, we have

$$|\mathcal{R}_1(Z_i, Z_j)| \le C'|\mathcal{L}(Z_i, Z_j)|,$$

where C' is a constant depends only on φ, e_{xy}, e_x and e_y. Let Π_{ij} corresponds to Γ,

$$\begin{aligned}
&\sup_{\Gamma}\left|\sum_{i=2}^{n+m}\sum_{j=1}^{i-1}\Pi_{ij}\mathcal{R}_1(Z_i, Z_j)\right| \\
\le &\sup_{\Gamma}\left|\sum_{i=2}^{n+m}\sum_{j=1}^{i-1}\Pi_{ij}\mathcal{R}_1(Z_i, Z_j)\left\{\mathbb{I}_{B_{\mathbf{XY}}^c}\mathbb{I}_{B_{\mathbf{X}}^c}\mathbb{I}_{B_{\mathbf{Y}}^c} + \mathbb{I}_{B_{\mathbf{XY}}} + \mathbb{I}_{B_{\mathbf{X}}} + \mathbb{I}_{B_{\mathbf{Y}}}\right\}\right| \\
\le &\sup_{\Gamma}\sum_{i=2}^{n+m}\sum_{j=1}^{i-1}|\Pi_{ij}\mathcal{R}_1(Z_i, Z_j)|\mathbb{I}_{B_{\mathbf{XY}}^c}\mathbb{I}_{B_{\mathbf{X}}^c}\mathbb{I}_{B_{\mathbf{Y}}^c} + o_p(1) \\
\le &\sum_{i=2}^{n+m}\sum_{j=1}^{i-1}\frac{C''}{nm}|\mathcal{R}_1(Z_i, Z_j)|\mathbb{I}_{B_{\mathbf{XY}}^c}\mathbb{I}_{B_{\mathbf{X}}^c}\mathbb{I}_{B_{\mathbf{Y}}^c} + o_p(1),
\end{aligned}$$

where C'' is a constant depends only on ρ. Then, for any $\epsilon > 0$, by Markov's inequality

$$\begin{aligned}
&P\left(\frac{1}{mn}\sum_{i=2}^{n+m}\sum_{j=1}^{i-1}|\mathcal{R}_1(Z_i, Z_j)|\mathbb{I}_{B_{\mathbf{XY}}^c}\mathbb{I}_{B_{\mathbf{X}}^c}\mathbb{I}_{B_{\mathbf{Y}}^c} > \epsilon\right) \\
&\qquad \le \frac{1}{\epsilon mn}\sum_{i=2}^{n+m}\sum_{j=1}^{i-1}E[|\mathcal{R}_1(Z_i, Z_j)|\mathbb{I}_{B_{\mathbf{XY}}^c}\mathbb{I}_{B_{\mathbf{X}}^c}\mathbb{I}_{B_{\mathbf{Y}}^c}] \le C'''\frac{(\alpha_{xy} + \alpha_x + \alpha_y)}{\epsilon},
\end{aligned}$$

where C''' is a constant depends only on ρ, φ, e_{xy}, e_x and e_y. □

(ii) Similar to the HDLSS setting, we consider the following decomposition

$$\sqrt{nmp}\mathrm{ED}_n^k(\Gamma_w\mathbf{Z}) = \sqrt{nmp}\mu_n(\Gamma_w\mathbf{Z}) + \underbrace{\sqrt{nm}\sum_{i=2}^{n+m}\sum_{j=1}^{i-1}\Pi_{w,ij}\varphi^{(1)}(e_{ij})\mathcal{K}(Z_i,Z_j)}_{:=\sqrt{nm}L(\Gamma_w\mathbf{Z})}$$
$$+\underbrace{\sqrt{nm}\sum_{i=2}^{n+m}\sum_{j=1}^{i-1}\Pi_{w,ij}\varphi^{(1)}(e_{ij})\mathcal{W}(Z_i,Z_j)}_{:=\sqrt{nm}R_l(\Gamma_w\mathbf{Z})} + \underbrace{\sqrt{nmp}\sum_{i=2}^{n+m}\sum_{j=1}^{i-1}\Pi_{w,ij}\mathcal{R}_2(Z_i,Z_j)}_{:=\sqrt{nm}R_2(\Gamma_w\mathbf{Z})}.$$

Next, for any $-\infty < a < \infty$ and $\epsilon > 0$, using the inequality $P(X \le a) \le P(Y \le a+\epsilon) + P(|X-Y| > \epsilon)$, we can show that

$$\begin{aligned}
&P(\sqrt{nm}L(\Gamma_w\mathbf{Z}) \le a-\epsilon) - P(|\sqrt{nm}R_l(\Gamma_w\mathbf{Z}) + \sqrt{nm}R_2(\Gamma_w\mathbf{Z})| > \epsilon)\\
&\qquad \le P(\sqrt{nmp}[\mathrm{ED}_n^k(\Gamma_w\mathbf{Z}) - \mu_n(\Gamma_w\mathbf{Z})] \le a)\\
&\qquad \le P(\sqrt{nm}L(\Gamma_w\mathbf{Z}) \le a+\epsilon) + P(|\sqrt{nm}R_l(\Gamma_w\mathbf{Z}) + \sqrt{nm}R_2(\Gamma_w\mathbf{Z})| > \epsilon).
\end{aligned}$$

Then, some algebra shows that

$$\begin{aligned}
&\sup_w \left|P(\sqrt{nmp}[\mathrm{ED}_n^k(\Gamma_w\mathbf{Z}) - \mu_n(\Gamma_w\mathbf{Z})] \le a) - \Phi\left(a/\sqrt{nm\sigma_n^2(\Gamma_w\mathbf{Z})}\right)\right|\\
\le &\sup_w \left|P(\sqrt{nm}L(\Gamma_w\mathbf{Z}) \le a-\epsilon) - \Phi\left((a-\epsilon)/\sqrt{nm\sigma_n^2(\Gamma_w\mathbf{Z})}\right)\right|\\
&\qquad + \sup_w \left|\Phi\left(a/\sqrt{nm\sigma_n^2(\Gamma_w\mathbf{Z})}\right) - \Phi\left((a-\epsilon)/\sqrt{nm\sigma_n^2(\Gamma_w\mathbf{Z})}\right)\right|\\
&+\sup_w \left|P(\sqrt{nm}L(\Gamma_w\mathbf{Z}) \le a+\epsilon) - \Phi\left((a+\epsilon)/\sqrt{nm\sigma_n^2(\Gamma_w\mathbf{Z})}\right)\right|\\
&\qquad + \sup_w \left|\Phi\left(a/\sqrt{nm\sigma_n^2(\Gamma_w\mathbf{Z})}\right) - \Phi\left((a+\epsilon)/\sqrt{nm\sigma_n^2(\Gamma_w\mathbf{Z})}\right)\right|\\
&+2\sup_w P(|\sqrt{nm}R_l(\Gamma_w\mathbf{Z}) + \sqrt{nm}R_2(\Gamma_w\mathbf{Z})| > \epsilon).
\end{aligned}$$

Next, by Lemma A.2, A.3 and A.4. the right hand side can be made arbitrarily small by first choose ϵ small enough, then n, m, p large enough.

Lemma A.2. *Under Assumption 5 and 6,* $\sup_\Gamma |\sqrt{nm}R_2(\Gamma\mathbf{Z})| = o_p(1)$.

Proof. The proof is similar with Lemma A.1 by observing that conditioned on event $B^c_{\mathbf{XY}}B^c_{\mathbf{X}}B^c_{\mathbf{Y}}$, it holds for some constant C that $|\mathcal{R}_2(Z_i,Z_j)| \le C\left|\mathcal{L}^2(Z_i,Z_j)\right|$. □

Lemma A.3. *Under* H_{A_l}, $\sup_{\Gamma\in\mathbb{P}_{n+m}} |\sqrt{nm}R_l(\Gamma\mathbf{Z})| = o_p(1)$.

Proof. For any fixed permutaiton marix $\Gamma \in \mathbb{P}_{n+m}$, we have

$$\sqrt{nm}R_l(\Gamma\mathbf{Z}) = \sqrt{nmp}\sum_{i=2}^{n+m}\sum_{j=1}^{i-1}\Pi_{ij}\varphi^{(1)}\left(e_{ij}\right)\left(E\left[\overline{\psi}(Z_i,Z_j)|Z_i\right]+E\left[\overline{\psi}(Z_i,Z_j)|Z_j\right]\right)$$
$$-\sqrt{nmp}\sum_{i=2}^{n+m}\sum_{j=1}^{i-1}\Pi_{ij}\varphi^{(1)}\left(e_{ij}\right)\left(E\left[\overline{\psi}(Z_i,Z_j)\right]+e_{ij}\right).$$

Let $w = N(\Gamma)$, similar to the computation of $\mu_{n,w}$, we obtain

$$\sqrt{nmp}\sum_{i=2}^{n+m}\sum_{j=1}^{i-1}\Pi_{ij}\varphi^{(1)}\left(e_{ij}\right)E\left[\overline{\psi}(Z_i,Z_j)\right] = \sqrt{nmp}\Big\{2\varphi^{(1)}\left(e_{xy}\right)E\left[\overline{\psi}(X,Y)\right]$$
$$-\varphi^{(1)}\left(e_x\right)E\left[\overline{\psi}(X,X')\right]-\varphi^{(1)}\left(e_y\right)E\left[\overline{\psi}(Y,Y')\right]\Big\}f(w),$$

where the right hand side is of order $o_p(1)$ under H_{A_l}. Let π corresponds to Γ, then for each $1 \le i \le n$ such that $1 \le \pi(i) \le n$, it follows from the definition of Π_{ij} that

$$\begin{aligned}
&\frac{1}{4}\sum_{1\le j\le n+m}^{j\neq i}\Pi_{ij}\varphi^{(1)}\left(e_{ij}\right)E\left[\overline{\psi}(X_i,Z_j)|X_i\right]\\
=&-\frac{(n-w-1)}{n(n-1)}\varphi^{(1)}\left(e_x\right)E\left[\overline{\psi}(X_i,X)|X_i\right]-\frac{w}{n(n-1)}\varphi^{(1)}\left(e_{xy}\right)E\left[\overline{\psi}(X_i,Y)|X_i\right]\\
&+\frac{w}{nm}\varphi^{(1)}\left(e_x\right)E\left[\overline{\psi}(X_i,X)|X_i\right]+\frac{m-w}{nm}\varphi^{(1)}\left(e_{xy}\right)E\left[\overline{\psi}(X_i,Y)|X_i\right]\\
=&\left(\frac{1}{n}-\frac{w}{nm}-\frac{w}{n(n-1)}\right)\left\{\varphi^{(1)}\left(e_{xy}\right)E\left[\overline{\psi}(X_i,Y)|X_i\right]-\varphi^{(1)}(e_x)E\left[\overline{\psi}(X_i,X)|X_i\right]\right\},
\end{aligned}$$

which entails

$$\sup_{\Gamma}\left|\frac{1}{4}\sum_{1\le j\le n+m}^{j\neq i}\Pi_{ij}\varphi^{(1)}\left(e_{ij}\right)E\left[\overline{\psi}(X_i,Z_j)|X_i\right]\right| \le$$
$$\frac{C}{\sqrt{nm}}\left|\varphi^{(1)}\left(e_{xy}\right)E\left[\overline{\psi}(X_i,Y)|X_i\right]-\varphi^{(1)}(e_x)E\left[\overline{\psi}(X_i,X)|X_i\right]\right|,$$

where C is a constant that only depends on ρ. Using the same approach, the above bound can be shown to hold for each $1 \le i \le n$ such that $n+1 \le \pi(i) \le n+m$. Similarly, we can show that for each $n+1 \le i \le n+m$,

$$\sup_{\Gamma}\left|\frac{1}{4}\sum_{1\le j\le n+m}^{j\neq i}\Pi_{ij}\varphi^{(1)}(e_{ij})E\left[\overline{\psi}(Y_i,Z_j)|Y_i\right]\right|$$
$$\le\frac{C}{\sqrt{nm}}\left|\varphi^{(1)}(e_{xy})E\left[\overline{\psi}(Y_i,X)|Y_i\right]-\varphi^{(1)}(e_y)E\left[\overline{\psi}(Y_i,Y)|Y_i\right]\right|.$$

Consequently, the following bound holds

$$\begin{aligned}
&\sup_{\Gamma}\left|\sqrt{nmp}\sum_{i=2}^{n+m}\sum_{j=1}^{i-1}\Pi_{ij}\varphi^{(1)}(e_{ij})\left(E\left[\overline{\psi}(Z_i,Z_j)|Z_i\right]+E\left[\overline{\psi}(Z_i,Z_j)|Z_j\right]\right)\right| \\
&\qquad\le C'\sqrt{p}\sum_{i=1}^{n}\left|\varphi^{(1)}(e_{xy})E\left[\overline{\psi}(X_i,Y)|X_i\right]-\varphi^{(1)}(e_x)E\left[\overline{\psi}(X_i,X)|X_i\right]\right| \\
&\qquad\qquad+C'\sqrt{p}\sum_{i=n+1}^{n+m}\left|\varphi^{(1)}(e_{xy})E\left[\overline{\psi}(Y_i,X)|Y_i\right]-\varphi^{(1)}(e_y)E\left[\overline{\psi}(Y_i,Y)|Y_i\right]\right|,
\end{aligned}$$

where C' is a constant. Finally, an application of Markov's inequality shows that the right hand side is of order $o_p(1)$ under H_{A_l}. □

Lemma A.4. *Under Assumptions 1. Let $\Gamma_1,\Gamma_2\in\mathbb{P}_{n+m}$. Then, for any constants a_1,a_2,b, we have*

$$\begin{aligned}
&\sup_{\Gamma_1,\Gamma_2}\left|P(a_1\sqrt{nm}L(\Gamma_1\mathbf{Z})+a_2\sqrt{nm}L(\Gamma_2\mathbf{Z})\le b)-\Phi\left(\frac{b}{\eta_{a_1,a_2}(\Gamma_1,\Gamma_2)}\right)\right| \\
&\quad\le C\Bigg\{\max_{\Lambda_1,\Lambda_2\in\{X,Y\}}E\left[\mathcal{K}^4(\Lambda_1,\Lambda_2')\right]/n^2 \\
&\qquad+\max_{\Lambda_1,\Lambda_2,\Lambda_3\in\{X,Y\}}E\left[\mathcal{K}^2(\Lambda_1,\Lambda_3'')\mathcal{K}^2(\Lambda_2',\Lambda_3'')\right]/n \\
&\qquad+\max_{\Lambda_1,\Lambda_2,\Lambda_3,\Lambda_4\in\{X,Y\}}E\left[\mathcal{K}(\Lambda_1,\Lambda_3'')\mathcal{K}(\Lambda_1,\Lambda_4''')\mathcal{K}(\Lambda_2',\Lambda_4''')\mathcal{K}(\Lambda_2',\Lambda_3'')\right] \\
&\qquad+(v_x-var[\mathcal{K}(X,X')])^2+(v_y-var[\mathcal{K}(Y,Y')])^2+(v_{xy}-var[\mathcal{K}(X,Y)])^2\Bigg\}^{1/5},
\end{aligned}$$

where C is a constant depend on φ,ρ,e_x,e_y and e_{xy} only; $\eta_{a_1,a_2}(\Gamma_1,\Gamma_2)$ is defined as

$$[\eta_{a_1,a_2}(\Gamma_1,\Gamma_2)]^2=nm\sum_{i=2}^{n+m}\sum_{j=1}^{i-1}(a_1\Pi_{1,ij}+a_2\Pi_{2,ij})^2[\varphi^{(1)}(e_{ij})]^2v_{ij},$$

where $\Pi_{1,ij},\Pi_{2,ij}$ correspond to Γ_1,Γ_2 respectively and

$$v_{ij}=\begin{cases}v_x, & \text{if } 1\le i,j\le n,\\ v_y, & \text{if } n+1\le i,j\le n+m,\\ v_{xy}, & \text{otherwise.}\end{cases}$$

Proof. Notice that

$$\sqrt{nm}\left(a_1L(\Gamma_1\mathbf{Z})+a_2L(\Gamma_2\mathbf{Z})\right)=\sqrt{nm}\sum_{i=2}^{n+m}\sum_{j=1}^{i-1}(a_1\Pi_{1,ij}+a_2\Pi_{2,ij})\varphi^{(1)}(e_{ij})\mathcal{K}(Z_i,Z_j).$$

Then, for notational convenience, set

$$\mathcal{H}(Z_i, Z_j) = \sqrt{nm}(a_1\Pi_{1,ij} + a_2\Pi_{2,ij})\varphi^{(1)}(e_{ij})\mathcal{K}(Z_i, Z_j),$$

and $S_{n+m,l} = \sum_{i=2}^{l} \xi_{n+m,i}$, where $\xi_{n+m,i} = \sum_{j=1}^{i-1} \mathcal{H}(Z_i, Z_j)$. Next, let

$$\mathcal{F}_{n+m,l} = \sigma(Z_1, Z_2 \cdots, Z_l)$$

be the σ-algebra generated by $Z_1, \cdots, Z_l$, we have $\{S_{n+m,l}, \mathcal{F}_{n+m,l}, 1 \le l \le n+m\}$ is a martingale array and thus we can apply the Berry-Esseen type bound for martingale sequences [Theorem 1 of [15]]. By setting $m = 0$ and $\delta = 1$ in Theorem 1 of [15], we compute

$$\sum_{i=2}^{n+m} E\left[\xi_{n+m,i}^2\right], var\left[\sum_{i=2}^{n+m} E\left[\xi_{n+m,i}^2 \middle| \mathcal{F}_{n+m,i-1}\right]\right] \text{ and } \sum_{i=2}^{n+m} E\left[\xi_{n+m,i}^4\right]$$

Firstly, due to the property of double centering, $\mathrm{E}[\mathcal{H}(Z_i, Z_j)\mathcal{H}(Z_{i'}, Z_{j'})] \neq 0$ only when $\{i, j\} = \{i', j'\}$. Then, let $\eta_{a_1,a_2}(\Gamma_1, \Gamma_2)$ be in Theorem 1 of [15]

$$\frac{\eta_{a_1,a_2}^2(\Gamma_1, \Gamma_2)}{nm} := \lim_{p\to\infty} \frac{\sum_{i=2}^{n+m} E\left[\xi_{n+m,i}^2\right]}{nm} = \sum_{i=2}^{n+m}\sum_{j=1}^{i-1}(a_1\Pi_{1,ij} + a_2\Pi_{2,ij})^2[\varphi^{(1)}(e_{ij})]^2 v_{ij}.$$

Then, to calculate the variance, notice that

$$var\left[\sum_{i=2}^{n+m} E\left[\xi_{n+m,i}^2 \middle| \mathcal{F}_{n+m,i-1}\right]\right] = \sum_{i_1,i_2=2}^{n+m}\sum_{j_1,j_2=1}^{i_1-1}\sum_{j_3,j_4=1}^{i_2-1} \Theta(i_1, i_2; j_1, j_2, j_3, j_4),$$

where $\Theta(i_1, i_2; j_1, j_2, j_3, j_4)$ is defined as

$$\Theta(i_1, i_2; j_1, j_2, j_3, j_4) = cov\left[E\left[\mathcal{H}(Z_{i_1}, Z_{j_1})\mathcal{H}(Z_{i_1}, Z_{j_2})|Z_{j_1}, Z_{j_2}\right], E\left[\mathcal{H}(Z_{i_2}, Z_{j_3})\mathcal{H}(Z_{i_2}, Z_{j_4})|Z_{j_3}, Z_{j_4}\right]\right].$$

Next, for any $1 \le j_1, j_2 \le n+m$, $\Lambda \in \{X, Y\}$, denote

$$\mathcal{G}_\Lambda(Z_{j_1}, Z_{j_2}) = E[\mathcal{K}(\Lambda, Z_{j_1})\mathcal{K}(\Lambda, Z_{j_2})|Z_{j_1}, Z_{j_2}].$$

To bound each $\Theta(i_1, i_2; j_1, j_2, j_3, j_4)$, we need to study the covariance between $\mathcal{G}_{\Lambda_1}(Z_{j_1}, Z_{j_2})$ and $\mathcal{G}_{\Lambda_2}(Z_{j_1'}, Z_{j_2'})$.

Lemma A.5. *Then, for any $1 \le j_1, j_2, j_1', j_2' \le n+m$, $\Lambda_1, \Lambda_2 \in \{X, Y\}$, we have*

$$cov\left[\mathcal{G}_{\Lambda_1}(Z_{j_1}, Z_{j_2}), \mathcal{G}_{\Lambda_2}(Z_{j_1'}, Z_{j_2'})\right]$$

$$= \begin{cases} E\left[\mathcal{K}^2(\Lambda_1, Z_{j_1})\mathcal{K}^2(\Lambda_2', Z_{j_1})\right] - E\left[\mathcal{K}^2(\Lambda_1, Z_{j_1})\right] E\left[\mathcal{K}^2(\Lambda_2, Z_{j_1})\right], & j_1 = j_2 = j_1' = j_2'; \\ E\left[\mathcal{K}(\Lambda_1, Z_{j_1})\mathcal{K}(\Lambda_1, Z_{j_2})\mathcal{K}(\Lambda_2', Z_{j_2})\mathcal{K}(\Lambda_2', Z_{j_1})\right], & j_1 = j_1' \neq j_2 = j_2'; \\ E\left[\mathcal{K}(\Lambda_1, Z_{j_1})\mathcal{K}(\Lambda_1, Z_{j_2})\mathcal{K}(\Lambda_2', Z_{j_2})\mathcal{K}(\Lambda_2', Z_{j_1})\right], & j_1 = j_2' \neq j_2 = j_1'; \\ 0, & \textit{otherwise.} \end{cases}$$

Proof. If $j_1 = j_2' \neq j_2 = j_1'$,

$$\begin{aligned}
&E\left[\mathcal{G}_{\Lambda_1}(Z_{j_1}, Z_{j_2})\mathcal{G}_{\Lambda_2}(Z_{j_1'}, Z_{j_2'})\right]\\
=&E\left[E\left[\mathcal{K}(\Lambda_1, Z_{j_1})\mathcal{K}(\Lambda_1, Z_{j_2})\middle| Z_{j_1}, Z_{j_2}\right] E\left[\mathcal{K}(\Lambda_2', Z_{j_1'})\mathcal{K}(\Lambda_2', Z_{j_2'})\middle| Z_{j_1}, Z_{j_2}\right]\right]\\
=&E\left[E\left[\mathcal{K}(\Lambda_1, Z_{j_1})\mathcal{K}(\Lambda_1, Z_{j_2})\mathcal{K}(\Lambda_2', Z_{j_2})\mathcal{K}(\Lambda_2', Z_{j_1})\middle| Z_{j_1}, Z_{j_2}\right]\right]\\
=&E\left[\mathcal{K}(\Lambda_1, Z_{j_1})\mathcal{K}(\Lambda_1, Z_{j_2})\mathcal{K}(\Lambda_2', Z_{j_2})\mathcal{K}(\Lambda_2', Z_{j_1})\right]
\end{aligned}$$

It can be shown similarly for cases $j_1 = j_2 = j_3 = j_4$ and $j_1 = j_1' \neq j_2 = j_2'$. Next, we show that for other cases, the covariance is 0. We take $j_1 = j_1'$, $j_1 \neq j_2$, $j_1 \neq j_2'$, $j_2 \neq j_2'$ as an example

$$\begin{aligned}
&E\left[\mathcal{G}_{\Lambda_1}(Z_{j_1}, Z_{j_2})\mathcal{G}_{\Lambda_2}(Z_{j_1'}, Z_{j_2'})\right]\\
=&E\left[E\left[\mathcal{K}(\Lambda_1, Z_{j_1})\mathcal{K}(\Lambda_1, Z_{j_2})\middle| Z_{j_1}, Z_{j_2}\right] E\left[\mathcal{K}(\Lambda_2', Z_{j_1})\mathcal{K}(\Lambda_2', Z_{j_2'})\middle| Z_{j_1}, Z_{j_2'}\right]\right]\\
=&E\left[E\left[\mathcal{K}(\Lambda_1, Z_{j_1})\mathcal{K}(\Lambda_1, Z_{j_2})\mathcal{K}(\Lambda_2', Z_{j_1})\mathcal{K}(\Lambda_2', Z_{j_2'})\middle| Z_{j_1}, Z_{j_2}, Z_{j_2'}\right]\right]\\
=&E\left[\mathcal{K}(\Lambda_1, Z_{j_1})\mathcal{K}(\Lambda_1, Z_{j_2})\mathcal{K}(\Lambda_2', Z_{j_1})\mathcal{K}(\Lambda_2', Z_{j_2'})\right]\\
=&E\left[\mathcal{K}(\Lambda_1, Z_{j_1})\mathcal{K}(\Lambda_2', Z_{j_1})E\left[\mathcal{K}(\Lambda_1, Z_{j_2})\middle| \Lambda_1, \Lambda_2', Z_{j_1}\right] E\left[\mathcal{K}(\Lambda_2', Z_{j_2'})\middle| \Lambda_1, \Lambda_2', Z_{j_1}\right]\right]\\
=&0.
\end{aligned}$$

□

Next, we can bound $var\left[\sum_{i=2}^{n+m} E\left[\xi_{n+m,i}^2\middle| \mathcal{F}_{n+m,i-1}\right]\right]$ as

$$\begin{aligned}
&\sum_{i_1,i_2=1}^{n+m}\sum_{j_1,j_2=1}^{i_1-1}\sum_{j_3,j_4=1}^{i_2-1}\Theta(i_1,i_2;j_1,j_2,j_3,j_4)\\
=&\sum_{i=1}^{n+m}\left\{\sum_{j=1}^{i-1}\Theta(i,i;j,j,j,j)+2\sum_{1\le j_1\neq j_2\le i-1}\Theta(i,i;j_1,j_2,j_1,j_2)\right\}\\
&\quad+2\sum_{1\le i_1<i_2\le n+m}\left\{\sum_{j=1}^{i_1-1}\Theta(i_1,i_2;j,j,j,j)+2\sum_{1\le j_1\neq j_2\le i_1-1}\Theta(i_1,i_2;j_1,j_2,j_1,j_2)\right\}\\
=&O\Big(\max_{\Lambda_1,\Lambda_2,\Lambda_3\in\{X,Y\}}E\left[\mathcal{K}^2(\Lambda_1,\Lambda_3'')\mathcal{K}^2(\Lambda_2',\Lambda_3'')\right]/n\\
&\quad+\max_{\Lambda_1,\Lambda_2,\Lambda_3,\Lambda_4\in\{X,Y\}}E\left[\mathcal{K}(\Lambda_1,\Lambda_3'')\mathcal{K}(\Lambda_1,\Lambda_4''')\mathcal{K}(\Lambda_2',\Lambda_4''')\mathcal{K}(\Lambda_2',\Lambda_3'')\right]\Big)
\end{aligned}$$

Finally, to find the upper bound of $\sum_{i=2}^{n+m} E\left(\xi_{n+m,i}^4\right)$,

$$\begin{aligned}
&\sum_{i=2}^{n+m} E\left(\xi_{n+m,i}^4\right) \\
=&\sum_{i=2}^{n+m} \sum_{j_1,j_2,j_3,j_4=1}^{i-1} E\left[\mathcal{H}(Z_i,Z_{j_1})\mathcal{H}(Z_i,Z_{j_2})\mathcal{H}(Z_i,Z_{j_3})\mathcal{H}(Z_i,Z_{j_4})\right] \\
=&\sum_{i=2}^{n+m}\sum_{j=1}^{i-1} E\left[\mathcal{H}^4(Z_i,Z_j)\right] + 6\sum_{i=2}^{n+m}\sum_{1\le j_1<j_2\le i-1} E\left[\mathcal{H}^2(Z_i,Z_{j_1})\mathcal{H}^2(Z_i,Z_{j_2})\right] \\
=&O\left(\max_{\Lambda_1,\Lambda_2\in\{X,Y\}} E\left[\mathcal{K}^4(\Lambda_1,\Lambda_2')\right]/n^2\right) \\
&\quad + O\left(\max_{\Lambda_1,\Lambda_2,\Lambda_3\in\{X,Y\}} E\left[\mathcal{K}^2(\Lambda_1,\Lambda_3'')\mathcal{K}^2(\Lambda_2',\Lambda_3'')\right]/n\right).
\end{aligned}$$

Combining the above bounds, the lemma is a consequence of Theorem 1 in [15]. □

A.6. Proof of Theorem 3.4

(i) For a random permutation matrix $\mathbf{\Gamma} \sim \text{Uniform}(\mathbb{P}_{n+m})$, it follows from Lemma A.1 that

$$\text{ED}_n^k(\mathbf{\Gamma Z}) = \mu_n(\mathbf{\Gamma Z}) + R_1(\mathbf{\Gamma Z}) = \left(2\varphi\left(e_{xy}\right) - \varphi\left(e_x\right) - \varphi\left(e_y\right)\right) f(W) + o_p(1)$$

where $W = N(\mathbf{\Gamma}) \sim \text{Hypergeometric}(m+n, m, n)$. From the normal limit of hypergeometric distribution [20], we know that

$$\frac{W}{\sqrt{nm}} \xrightarrow{p} \frac{\sqrt{\rho}}{1+\rho}.$$

Next, some algebra shows that $f(W) \xrightarrow{p} 0$ and so the result is proved.
(ii) Recall that we can decompose the sample energy distance as

$$\begin{aligned}
\sqrt{nmp}\text{ED}_n^k(\mathbf{\Gamma Z}) = \sqrt{nmp}\mu_n(\mathbf{\Gamma Z}) &+ \underbrace{\sqrt{nm}\sum_{i=2}^{n+m}\sum_{j=1}^{i-1}\mathbf{\Pi}_{ij}\varphi^{(1)}\left(e_{ij}\right)\mathcal{K}(Z_i,Z_j)}_{:=\sqrt{nm}L(\mathbf{\Gamma Z})} \\
&+ \underbrace{\sqrt{nm}\sum_{i=2}^{n+m}\sum_{j=1}^{i-1}\mathbf{\Pi}_{ij}\varphi^{(1)}(e_{ij})\mathcal{W}(Z_i,Z_j)}_{:=\sqrt{nm}R_l(\mathbf{\Gamma Z})} + \underbrace{\sqrt{nmp}\sum_{i=2}^{n+m}\sum_{j=1}^{i-1}\mathbf{\Pi}_{ij}\mathcal{R}_2(Z_i,Z_j)}_{:=\sqrt{nm}R_2(\mathbf{\Gamma Z})}.
\end{aligned}$$

The result is a consequence of Lemma A.3, A.2 and A.6.

Lemma A.6. *Under Assumptions 1 and 7,*

$$\sqrt{nm}\begin{pmatrix} L(\mathbf{\Gamma Z}) \\ L(\mathbf{\Gamma' Z}) \end{pmatrix} \xrightarrow{d} N\left(0, \begin{pmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{pmatrix}\right)$$

where $\mathbf{\Gamma}'$ *is independent copy of* $\mathbf{\Gamma}$ *and* σ^2 *is the asymptotic variance defined as*

$$\sigma^2 := 4v_{xy}[\varphi^{(1)}(e_{xy})]^2 + 2\rho v_x[\varphi^{(1)}(e_x)]^2 + \frac{2}{\rho}v_y[\varphi^{(1)}(e_y)]^2.$$

Proof. We apply the Cramér-Wold device. For any constants a_1, a_2, we have

$$\begin{aligned}\eta^2_{a_1,a_2}(\mathbf{\Gamma}, \mathbf{\Gamma}') &= nm\sum_{i=2}^{n+m}\sum_{j=1}^{i-1}(a_1\mathbf{\Pi}_{ij} + a_2\mathbf{\Pi}'_{ij})^2[\varphi^{(1)}(e_{ij})]^2 v_{ij} \\ &= nm\sum_{i=2}^{n+m}\sum_{j=1}^{i-1}\left(a_1^2\mathbf{\Pi}_{ij}^2 + a_2^2(\mathbf{\Pi}'_{ij})^2 + 2a_1a_2\mathbf{\Pi}_{ij}\mathbf{\Pi}'_{ij}\right)[\varphi^{(1)}(e_{ij})]^2 v_{ij}.\end{aligned}$$

Notice that for $\{i_1, j_1\} \cap \{i_2, j_2\} = \emptyset$, it can be shown that $E[\mathbf{\Pi}_{i_1j_1}\mathbf{\Pi}_{i_2j_2}] = O(1/n^5)$. Then, denote $c_{ij} = 2a_1a_2[\varphi^{(1)}(e_{ij})]^2 v_{ij}$, we have

$$\begin{aligned}E\left[\left(nm\sum_{i=2}^{n+m}\sum_{j=1}^{i-1}c_{ij}\mathbf{\Pi}_{ij}\mathbf{\Pi}'_{ij}\right)^2\right] =& n^2m^2\sum_{i=2}^{n+m}\sum_{j_1=1}^{i-1}\sum_{j_2=1}^{i-1}c_{ij_1}c_{ij_2}E^2[\mathbf{\Pi}_{ij_1}\mathbf{\Pi}_{ij_2}] \\ &+ 2n^2m^2\sum_{2\le i_1<i_2\le n+m}\sum_{j=1}^{i_1-1}c_{i_1j}c_{i_2j}E^2[\mathbf{\Pi}_{i_1j}\mathbf{\Pi}_{i_2j}] \\ &+ n^2m^2\sum_{2\le i_1\ne i_2\le n+m}\sum_{j_1\ne j_2}c_{i_1j_1}c_{i_2j_2}E^2[\mathbf{\Pi}_{i_1j_1}\mathbf{\Pi}_{i_2j_2}] \\ =& O(1/n).\end{aligned}$$

In addition, let $W = N(\mathbf{\Gamma})$, we obtain

$$\begin{aligned}
\sigma_n^2(\mathbf{\Gamma Z}) :=& \sum_{i=2}^{n+m}\sum_{j=1}^{i-1} \mathbf{\Pi}_{ij}^2[\varphi^{(1)}(e_{ij})]^2 v_{ij} \\
=& \left\{ \frac{4}{nm} - 4\left(\frac{n+m}{n^2m^2} - \frac{n}{n^2(n-1)^2} - \frac{m}{m^2(m-1)^2}\right) W \right. \\
&\left. + 4\left(\frac{2}{n^2m^2} - \frac{1}{n^2(n-1)^2} - \frac{1}{m^2(m-1)^2}\right) W^2 \right\} v_{xy}[\varphi^{(1)}(e_{xy})]^2 \\
&+ \left\{ \frac{2}{n(n-1)} + 2\left(\frac{2n}{n^2m^2} - \frac{2n-1}{n^2(n-1)^2} - \frac{1}{m^2(m-1)^2}\right) W \right. \\
&\left. - 2\left(\frac{2}{m^2n^2} - \frac{1}{n^2(n-1)^2} - \frac{1}{m^2(m-1)^2}\right) W^2 \right\} v_x[\varphi^{(1)}(e_x)]^2 \\
&+ \left\{ \frac{2}{m(m-1)} + 2\left(\frac{2m}{n^2m^2} - \frac{1}{n^2(n-1)^2} - \frac{2m-1}{m^2(m-1)^2}\right) W \right. \\
&\left. - 2\left(\frac{2}{n^2m^2} - \frac{1}{m^2(m-1)^2} - \frac{1}{n^2(n-1)^2}\right) W^2 \right\} v_y[\varphi^{(1)}(e_y)]^2.
\end{aligned}$$

Since $W/\sqrt{nm} \xrightarrow{p} \sqrt{\rho}/(1+\rho)$, some algebra shows that

$$\sigma_n^2(\mathbf{\Gamma Z}) := \sum_{i=2}^{n+m}\sum_{j=1}^{i-1} \mathbf{\Pi}_{ij}^2[\varphi^{(1)}(e_{ij})]^2 v_{ij} \xrightarrow{p} \sigma^2,$$

which entails that $\eta_{a_1,a_2}^2(\mathbf{\Gamma}, \mathbf{\Gamma}') \xrightarrow{p} a_1^2\sigma^2 + a_2^2\sigma^2$. Since $|\Phi(\cdot)| \le 1$, we have

$$E\left[\left|\Phi\left(\frac{b}{\eta_{a_1,a_2}(\mathbf{\Gamma}, \mathbf{\Gamma}')}\right) - \Phi\left(\frac{b}{\sqrt{a_1^2\sigma^2 + a_2^2\sigma^2}}\right)\right|\right] \to 0.$$

Next, by a simple triangle inequality

$$\begin{aligned}
&\left| P\left(a_1\sqrt{nm}L(\mathbf{\Gamma Z}) + a_2\sqrt{nm}L(\mathbf{\Gamma}'\mathbf{Z}) \le b\right) - \Phi\left(\frac{b}{\sqrt{a_1^2\sigma^2 + a_2^2\sigma^2}}\right)\right| \le \\
&\qquad \left| P\left(a_1\sqrt{nm}L(\mathbf{\Gamma Z}) + a_2\sqrt{nm}L(\mathbf{\Gamma}'\mathbf{Z}) \le b\right) - \Phi\left(\frac{b}{\sqrt{\eta_{a_1,a_2}^2(\mathbf{\Gamma}, \mathbf{\Gamma}')}}\right)\right| \\
&\qquad\qquad + \left|\Phi\left(\frac{b}{\sqrt{\eta_{a_1,a_2}^2(\mathbf{\Gamma}, \mathbf{\Gamma}')}}\right) - \Phi\left(\frac{b}{\sqrt{a_1^2\sigma^2 + a_2^2\sigma^2}}\right)\right|.
\end{aligned}$$

Taking expectation with respect to $\mathbf{\Gamma}, \mathbf{\Gamma}'$ on both sides, then it follows from Lemma A.4

and Assumption 7 that

$$\left| P\left(a_1\sqrt{nm}L(\mathbf{\Gamma Z}) + a_2\sqrt{nm}L(\mathbf{\Gamma' Z}) \le b\right) - \Phi\left(\frac{b}{\sqrt{a_1^2\sigma^2 + a_2^2\sigma^2}}\right)\right| \le$$
$$E\left[\left| P\left(a_1\sqrt{nm}L(\mathbf{\Gamma Z}) + a_2\sqrt{nm}L(\mathbf{\Gamma' Z}) \le b\right) - \Phi\left(\frac{b}{\sqrt{\eta^2_{a_1,a_2}(\mathbf{\Gamma},\mathbf{\Gamma}')}}\right)\right|\right]$$
$$+ E\left[\left|\Phi\left(\frac{b}{\sqrt{\eta^2_{a_1,a_2}(\mathbf{\Gamma},\mathbf{\Gamma}')}}\right) - \Phi\left(\frac{b}{\sqrt{a_1^2\sigma^2 + a_2^2\sigma^2}}\right)\right|\right] = o(1).$$

□

A.7. Proof of Corollary 3.1

By using Theorem 15.2.3 of [21], the result is a consequence of Theorem 3.4.

A.8. Proof of Theorem 3.5

(i) By Corollary 3.1 and Theorem 3.3

$$\text{Power} = P_{H_{A_c}}\left(\text{ED}_n^k(\mathbf{Z}) > c\right) \to P\left(2\varphi(e_{xy}) - \varphi(e_x) - \varphi(e_y) > 0\right) = 1.$$

(ii) By Corollary 3.1 and Theorem 3.3

$$\text{Power} = P_{H_{A_l}}\left(\text{ED}_n^k(\mathbf{Z}) > c\right) = P_{H_{A_l}}\left(\sqrt{nmp}\text{ED}_n^k(\mathbf{Z}) > \sqrt{nmp}c\right)$$
$$\to P\left(N(0,\sigma^2) > \sigma Q_{\Phi,1-\alpha}\right) = \alpha.$$

Acknowledgements

We would like to thank Dr. Jun Li for providing the code used in [22]. We are also grateful to the three reviewers for their very helpful comments. The partial support from a US NSF grant is gratefully acknowledged.

References

[1] Alvarez-Esteban, P. C., Del Barrio, E., Cuesta-Albertos, J. A. and Matran, C. [2008], 'Trimmed comparison of distributions', *Journal of the American Statistical Association* **103**(482), 697–704.

[2] Alvarez-Esteban, P. C., Del Barrio, E., Cuesta-Albertos, J. A., Matrán, C. et al. [2012], 'Similarity of samples and trimming', *Bernoulli* **18**(2), 606–634.
[3] Anderson, T. W. and Darling, D. A. [1952], 'Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes', *The Annals of Mathematical Statistics* **23**(2), 193–212.
[4] Aoshima, M., Shen, D., Shen, H., Yata, K., Zhou, Y.-H. and Marron, J. [2018], 'A survey of high dimension low sample size asymptotics', *Australian & New Zealand Journal of Statistics* **60**(1), 4–19.
[5] Bickel, P. J. [1969], 'A distribution free version of the Smirnov two sample test in the p-variate case', *The Annals of Mathematical Statistics* **40**(1), 1–23.
[6] Bickel, P. J. and Breiman, L. [1983], 'Sums of functions of nearest neighbor distances, moment bounds, limit theorems and a goodness of fit test', *The Annals of Probability* **11**(1), 185–214.
[7] Biswas, M. and Ghosh, A. K. [2014], 'A nonparametric two-sample test applicable to high dimensional data', *Journal of Multivariate Analysis* **123**, 160–171.
[8] Chakraborty, S. and Zhang, X. [2019], 'A new framework for distance and kernel-based metrics in high dimensions', *arXiv preprint arXiv:1909.13469* .
[9] Cramér, H. [1928], 'On the composition of elementary errors: First paper: Mathematical deductions', *Scandinavian Actuarial Journal* **1928**(1), 13–74.
[10] Dau, H. A., Keogh, E., Kamgar, K., Yeh, C.-C. M., Zhu, Y., Gharghabi, S., Ratanamahatana, C. A., Yanping, Hu, B., Begum, N., Bagnall, A., Mueen, A. and Batista, G. [2018], 'The ucr time series classification archive'. https://www.cs.ucr.edu/~eamonn/time_series_data_2018/.
[11] Freitag, G., Czado, C. and Munk, A. [2007], 'A nonparametric test for similarity of marginals with applications to the assessment of population bioequivalence', *Journal of Statistical Planning and Inference* **137**(3), 697–711.
[12] Friedman, J. H. and Rafsky, L. C. [1979], 'Multivariate generalizations of the wald-wolfowitz and smirnov two-sample tests', *The Annals of Statistics* **7**(4), 697–717.
[13] Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B. and Smola, A. [2012], 'A kernel two-sample test', *Journal of Machine Learning Research* **13**(Mar), 723–773.
[14] Gretton, A., Fukumizu, K., Teo, C. H., Song, L., Schölkopf, B. and Smola, A. J. [2008], A kernel statistical test of independence, *in* 'Advances in Neural Information Processing Systems', pp. 585–592.
[15] Hall, P. and Heyde, C. C. [1981], 'Rates of convergence in the martingale central limit theorem', *The Annals of Probability* **9**(3), 395–404.
[16] Hall, P., Marron, J. S. and Neeman, A. [2005], 'Geometric representation of high dimension, low sample size data', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67**(3), 427–444.
[17] 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.
[18] Klebanov, L. B., Beneš, V. and Saxl, I. [2005], *N-distances and Their Applications*, Charles University in Prague, the Karolinum Press.
[19] Kolmogorov, A. N. [1933], *Sulla determinazione empirica di una legge di distribuzione*, NA.

[20] Lahiri, S. N., Chatterjee, A. and Maiti, T. [2006], 'A sub-gaussian berry-esseen theorem for the hypergeometric distribution', *arXiv preprint math/0602276* .

[21] Lehmann, E. L. and Romano, J. P. [2006], *Testing Statistical Hypotheses*, Springer Science & Business Media.

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

[23] Munk, A. and Czado, C. [1998], 'Nonparametric validation of similar distributions and assessment of goodness of fit', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **60**(1), 223–241.

[24] Sarkar, S., Biswas, R. and Ghosh, A. K. [2018], 'On high-dimensional modifications of some graph-based two-sample tests', *arXiv preprint arXiv:1806.02138* .

[25] Scetbon, M. and Varoquaux, G. [2019], Comparing distributions: l_1 geometry improves kernel two-sample testing, *in* 'Advances in Neural Information Processing Systems 32', Curran Associates, Inc., pp. 12327–12337.

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

[27] Smirnov, N. [1948], 'Table for estimating the goodness of fit of empirical distributions', *The Annals of Mathematical Statistics* **19**(2), 279–281.

[28] Székely, G. J. and Rizzo, M. L. [2004], 'Testing for equal distributions in high dimension', *InterStat* **5**, 1–6.

[29] Von Mises, R. [1928], 'Statistik und wahrheit', *Julius Springer* .

[30] Zhu, C., Yao, S., Zhang, X. and Shao, X. [2019], 'Distance-based and rkhs-based dependence metrics in high dimension', *arXiv preprint arXiv:1902.03291* .