Journal of Machine Learning Research 24 (2023) 1-33 Submitted 10/22; Revised 8/23; Published 9/23

Two Sample Testing in High Dimension via Maximum Mean Discrepancy

Hanjia Gao HANJIAG2@ILLINOIS.EDU
Department of Statistics
University of Illinois at Urbana-Champaign
Champaign, IL 61820-5711, USA

Xiaofeng Shao XSHAO@ILLINOIS.EDU
Department of Statistics
University of Illinois at Urbana-Champaign
Champaign, IL 61820-5711, USA

Editor: Aarti Singh

Abstract

Maximum Mean Discrepancy (MMD) has been widely used in the areas of machine learning and statistics to quantify the distance between two distributions in the p-dimensional Euclidean space. The asymptotic property of the sample MMD has been well studied when the dimension p is fixed using the theory of U-statistic. As motivated by the frequent use of MMD test for data of moderate/high dimension, we propose to investigate the behavior of the sample MMD in a high-dimensional environment and develop a new studentized test statistic. Specifically, we obtain the central limit theorems for the studentized sample MMD as both the dimension p and sample sizes n, m diverge to infinity. Our results hold for a wide range of kernels, including popular Gaussian and Laplacian kernels, and also cover energy distance as a special case. We also derive the explicit rate of convergence under mild assumptions and our results suggest that the accuracy of normal approximation can improve with dimensionality. Additionally, we provide a general theory on the power analysis under the alternative hypothesis and show that our proposed test can detect difference between two distributions in the moderately high dimensional regime. Numerical simulations demonstrate the effectiveness of our proposed test statistic and normal approximation.

Keywords: Berry-Esseen Bound, Distance Covariance, Energy Distance, Hilbert-Schmidt Independence Criterion, Kernel Method

1. Introduction

Testing whether two samples are drawn from the same distribution is a classical problem in statistics. Mathematically speaking, given independent and identically distributed (iid) p-dimensional samples $X_1, \ldots, X_n$ from the distribution F_X and $Y_1, \ldots, Y_m$ from the distribution F_Y, we aim to test the hypothesis $H_0 : F_X = F_Y$ versus $H_A : F_X \neq F_Y$. There is a rich literature for the two-sample testing and well-known tests include Kolmogorov-Smirnov test [Kolmogorov (1933), Smirnov (1939)], Cramer von-Mises test [Cramér (1928)] and Anderson-Darling test [Anderson and Darling (1952)]. Other notable ones include Wald-Wolfowitz runs test [Wald and Wolfowitz (1940)], Mann-Whitney test [Mann and

Whitney (1947)] for univariate distributions and their multivariate generalizations [Friedman and Rafsky (1979)], among others.

In this article, we focus on the test based on maximum mean discrepancy (MMD, hereafter) [Gretton et al. (2012)], which is defined as the largest difference in expectations over functions in the unit ball of a reproducing kernel Hilbert space (RKHS). Since its introduction in the machine learning literature, it has gained growing popularity in both statistics and machine learning and found numerous real-world applications, ranging from biological data integration [Borgwardt et al. (2006)], to neural networks training [Dziugaite et al. (2015)], to the evaluation of a generative model in generative adversarial networks (GAN) [Arbel et al. (2018), Bińkowski et al. (2018)].

As a distance metric that measures the closeness of two distributions, MMD belongs to the category of interpoint distance based metric. In this category, a notable member is energy distance (ED, hereafter) [Székely et al. (2004), Székely and Rizzo (2013b)], which can be viewed as a special case of MMD [Sejdinovic et al. (2013)]. ED has been applied to many statistical problems, including two sample testing [Székely et al. (2004), Zhu and Shao (2021)], change-point detection [Matteson and James (2014)], hierarchical clustering [Szekely et al. (2005)], assessment of the quality of probabilistic forecasts via new scoring rules [Gneiting and Raftery (2007)], and covariate balancing in causal inference [Huling and Mak (2020)].

Motivated by the increasing use of MMD test for data of moderate and high dimension [Borgwardt et al. (2006), Zhu et al. (2017), Zhao et al. (2019)], we propose to study the behavior of sample MMD in the high-dimensional setting, which seems relatively less explored. To the best of our knowledge, we are only aware of recent contributions from Zhu and Shao (2021) and Chakraborty and Zhang (2021). In Zhu and Shao (2021), they showed that under the setting $p \gg \max(n, m)$, the MMD permutation tests are inconsistent when the two high dimensional distributions correspond to the same marginal distributions but differ in other aspects of the distributions in that the ED and MMD tests mainly target the differences between marginal means and sum of componentwise variances; see Chakraborty and Zhang (2021) for similar findings. Note that the computational complexity of MMD permutation test is $O((n+m)^2 pB)$ with B being the number of permutations employed and the computational cost is expensive for large scale data, whereas that of our proposed method is $O((n+m)^2 p)$.

As close relatives of ED, distance covariance (dcov, hereafter) and its standardized version distance correlation (dcor, hereafter) were proposed by Székely et al. (2007) to measure the dependence between two random vectors $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$ of arbitrary dimensions. The high-dimensional behavior of sample dcov has been studied in Zhu et al. (2020) and Gao et al. (2021). In Zhu et al. (2020), they showed that under the setting $\min(p, q) \gg n$, the dcov is unable to capture full nonlinear dependence between X and Y and it is only capable of capturing componentwise cross-covariance, a phenomenon reminiscent of the one in Zhu and Shao (2021) and Chakraborty and Zhang (2021). Additionally, their results have been shown to hold for sample HSIC (Hilbert-Schmidt Independence Criterion), which can be viewed as a kernelized version of sample dcov; see Sejdinovic et al. (2013). On the other hand, Gao et al. (2021) showed that a rescaled sample dcor is capable of detecting full nonlinear dependence as long as $p = q = o(\sqrt{n})$ and other regularity conditions hold. Thus the results in Zhu et al. (2020) and Gao et al. (2021) complement each other and

suggest that there are several interesting regimes for the asymptotic behavior of sample dcov and sample dcor. Han and Shen (2021) derived the first non-null central limit theorem (CLT, hereafter) for the sample distance covariance, as well as the more general sample HSIC in high dimensions, and their results were obtained primarily in the Gaussian case.

Despite the aforementioned recent advances, the asymptotic theory for sample MMD under the null hypothesis H_0 in general case of n, m and p diverging in an arbitrary fashion remains unexplored. Our first main contribution is to obtain central limit theorems for a studentized sample MMD. We also obtain the explicit rates of convergence to the limiting standard normal distribution. As another important contribution, we provide a general theory for the power analysis for our studentized sample MMD and provide several non-overlapping cases to discuss when the power of our MMD test is asymptotically one. One of the main findings is that in the moderately high-dimensional regime, the proposed studentized test statistic is able to detect the difference between two distributions with high power. The difference can lie in the means, marginal variances, componentwise covariances, and higher-order features associated with two high-dimensional distributions. The theoretical results are new to the literature and can be considered as substantial extensions over those obtained in Zhu et al. (2020), Zhu and Shao (2021), and Gao et al. (2021). As compared to Zhu and Shao (2021), who focused on the behavior of MMD-based permutation test in both High-dimensional Low Sample Size (HDLSS) and High-dimensional Medium Sample Size (HDMSS) settings, we aim to derive a simple studentized test statistic with standard normal limiting null distribution, under less stringent restrictions on the growth rate of p as a function of n. Some detailed comparisons with their power results are deferred to Section 3.5.

As two sample testing and independence testing are very much related, our work is also inspired by the dcov-based testing in high dimensional setting in Zhu et al. (2020) and Gao et al. (2021). In particular, since our work and Gao et al. (2021) share some technical arguments (say, Berry-Esseen bound for martingale), it pays to highlight the main difference between these two papers. First, the main U-statistic (that is, sample dcov) in Gao et al. (2021) is based on a one-sample kernel of order four, whereas we need to deal with a two-sample kernel of order $(2, 2)$. Consequently, some new theoretical tools need to be developed, such as the moment inequality for the two-sample U-statistic. Second, to form the studentized test statistic, we estimate the variance of sample MMD under the null using the pooled sample. The asymptotic behavior of this variance estimate is studied under both the null and the alternative. In particular, we have shown that it is a ratio-consistent estimator of HSIC of a mixture distribution with itself under some mild conditions. Lastly, our asymptotic theory is developed for a large class of kernels, including the L_2 norm as well as the Gaussian kernel, the Laplacian kernel, and many other kernels used in the machine learning literature. This generality is achieved by substantial new technical developments and very involved asymptotic analysis.

Recently, Yan and Zhang (2023) have obtained some related results for MMD-based test in high dimension. Specifically, they propose a studentized MMD-based test statistic under a specific model structure and establish the null CLT as well as the non-null CLTs under fixed and local alternatives for an (infeasible) standardized statistic. Though both papers consider the two-sample MMD-based testing problem when both (n, m) and p diverge and propose a studentized statistic, there are significant differences in terms of settings, technical

tools and theoretical results. Firstly, the problem set-ups are different. Yan and Zhang (2023) consider a special factor-like model which has been adopted in high-dimensional two sample mean testing [Chen and Qin (2010)]. All of our theory, including the CLT, the general Berry-Esseen bound and the power results, are established with no specific model constraints, and are thus applicable to a broader set of data generating processes. Secondly, the technical tools and primary results established in the two papers are very different. The most striking contribution in Yan and Zhang (2023) is the non-null CLTs for the standardized statistic, which are very interesting and seem only achievable under the specific model assumption, whereas we only present the null CLT for our studentized test statistic but additionally derive a Berry-Esseen bound under the null. Thirdly, the power results and the regimes under which the power approaches one are very different, and more discussion can be found in Section 3.4. Overall, we view the results in Yan and Zhang (2023) and our paper complementary to each other. Together they provide a more complete portrayal of the high-dimensional behavior of MMD-based statistics.

The rest of this paper is organized as follows. Section 2 introduces the maximum mean discrepancy, its sample version as a two sample U-statistic and its Hoeffeding decomposition. The distributional properties when the dimension p is fixed is also described. We propose a studentized test statistic and present the main theorems in Section 3. To be specific, we present the CLT for the studentized MMD and obtain the rates of convergence under the null. We also provide a general theory for the power under the alternative in this section. Finite sample performance is examined via simulations in Section 4. In Section 5, we summarize our results and discuss some potential extensions. Some illustrative examples, all the technical details, and some additional simulation results are presented in the online appendices; see `https://arxiv.org/abs/2109.14913`.

Let c, d be any positive integers and $\phi(x_1, \ldots, x_c, y_1, \ldots, y_d)$ denote a two-sample kernel function. For any $0 \le c' \le c$, $0 \le d' \le d$, and subsets $\{i_1, \ldots, i_{c'}\}$, $\{j_1, \ldots, j_{d'}\}$, define $\mathbb{E}_{X_{i_1},\ldots,X_{i_{c'}},Y_{j_1},\ldots,Y_{j_{d'}}}[\phi(X_1,\ldots,X_c,Y_1,\ldots,Y_d)] = \int\cdots\int \phi(X_1,\ldots,X_c,Y_1,\ldots,Y_d)\prod_{s=1}^{c'} dF_{X_{i_s}}\prod_{r=1}^{d'} dF_{Y_{j_r}}$. For simplicity, we write

$$\mathbb{E}_{X_1,\ldots,X_c,Y_1,\ldots,Y_d}[\phi(X_1,\ldots,X_c,Y_1,\ldots,Y_d)] = \mathbb{E}[\phi(X_1,\ldots,X_c,Y_1,\ldots,Y_d)].$$

For two random vectors V_1, V_2, the notation $V_1 =^d V_2$ means that they are identically distributed. We use $\to^d$ and $\to^p$ to denote convergence in distribution and in probability respectively. For two real-valued sequences a_n, b_n, we say $a_n = O(b_n)$ or $a_n \lesssim b_n$ if there exist $M, C > 0$, such that $a_n \le Cb_n$ for $n > M$. If there exist $M, C_1, C_2 > 0$, such that $C_1 b_n \le a_n \le C_2 b_n$ for $n > M$, then we say $a_n = O_s(b_n)$. In addition, we say $a_n = o(b_n)$ or $a_n \prec b_n$ if $a_n/b_n \to 0$ as $n \to \infty$. For any p-dimensional vectors a, b, we use $|a - b|$ to denote the Euclidean distance between a and b. For a function f, we use f_i to denote its i-th order derivative, and $\tilde{f}$ to denote its centered version, that is, $\tilde{f}(V_1, \ldots, V_k) = f(V_1, \ldots, V_k) - \mathbb{E}\big[f(V_1, \ldots, V_k)\big]$. We use $C(u_1, \ldots, u_k)$ to denote a positive and finite constant that depends only on the parameters $u_1, \ldots, u_k$ and the values of $C(u_1, \ldots, u_k)$ may vary from line to line. Additionally, we use $\mathrm{cum}(x_1, \cdots, x_k)$ to denote the joint cumulant of the random variables $x_1, \cdots, x_k$.

2. Maximum Mean Discrepancy and Its Properties

2.1 The Definition of Maximum Mean Discrepancy

We follow Definition 2 in Gretton et al. (2012) to provide a formal definition of MMD.

Definition 1 *Let $X \sim P_1$ and $Y \sim P_2$ be independent random vectors in $\mathbb{R}^p$ and let $\mathcal{F}_0$ be a class of functions $f : \mathbb{R}^p \to \mathbb{R}$. We define the maximum mean discrepancy (MMD) as*

$$MMD(P_1, P_2) = \sup_{f \in \mathcal{F}_0} \left\{ \mathbb{E}\big[f(X)\big] - \mathbb{E}\big[f(Y)\big] \right\}. \tag{1}$$

With properly selected function class $\mathcal{F}_0$, $\mathrm{MMD}(P_1, P_2)$ has some special properties. To facilitate the subsequent discussion, we follow the introduction in Section 2.2 of Gretton et al. (2012) to provide some basic properties of the reproducing kernel Hilbert space (RKHS).

Specifically, let $\mathcal{F}$ be an RKHS on the separable metric space $(\mathbb{R}^p, \mathcal{P})$, where $\mathcal{P}$ denotes the set of Borel probability measures on $\mathbb{R}^p$. By the property of the RKHS and the Riesz representation theorem, there is a feature mapping $\phi(x) : \mathbb{R}^p \to \mathbb{R}$ such that $f(x) = \langle f, \phi(x) \rangle_{\mathcal{F}}$ for any $f \in \mathcal{F}$. Furthermore, there exists a symmetric and positive definite kernel $\bar{k}$ associated with $\mathcal{F}$ such that ϕ takes the canonical form $\phi(x) = \bar{k}(x, \cdot)$. It follows that $\bar{k}(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{F}}$ for any $x, x' \in \mathbb{R}^p$.

For any distribution $P \in \mathcal{P}$, we define the mean embedding of $\mu_P \in \mathcal{F}$ as the function satisfying that $\mathbb{E}[f(x)] = \langle f, \mu_P \rangle_F$ for any $f \in \mathcal{F}$. It is shown in Lemma 3 and Lemma 4 of Gretton et al. (2012) that, when the aforementioned kernel $\bar{k}$ is measurable and satisfies $\mathbb{E}[\sqrt{\bar{k}(X, X)}] < \infty$, $\mathbb{E}[\sqrt{\bar{k}(Y, Y)}] < \infty$, then MMD can be expressed as the distance in $\mathcal{F}$ between mean embeddings, that is, $\mathrm{MMD}^2(P_1, P_2) := \|\mu_{P_1} - \mu_{P_2}\|^2_{\mathcal{F}}$. Equivalently, as stated in Lemma 6 of Gretton et al. (2012), MMD can be expressed through the kernel as

$$\mathrm{MMD}(P_1, P_2) = \mathrm{MMD}(X, Y | \bar{k}) := \left(-2\mathbb{E}\left[\bar{k}(X, Y)\right] + \mathbb{E}\left[\bar{k}(X, X')\right] + \mathbb{E}\left[\bar{k}(Y, Y')\right] \right)^{1/2}, \tag{2}$$

where X', Y' are independent and identical copies of $X \sim P_1$ and $Y \sim P_2$, respectively.

When $\mathcal{F}_0$ in Equation (1) is the unit ball in the RKHS $(\mathcal{F}, \bar{k})$, Gretton et al. (2012) has shown that $\mathrm{MMD}(P_1, P_2)$ is a nonnegative metric and $\mathrm{MMD}(P_1, P_2) = 0$ if and only if $P_1 =^d P_2$. Similar results have been generalized by using the equivalent definition of MMD. In particular, if $\bar{k}$ in Equation (2) is characteristic on $\mathbb{R}^p$ (i.e., the corresponding mean map μ_P is injective), then the associated MMD is a metric on $\mathcal{P}$, which satisfies $\mathrm{MMD}(P_1, P_2) = 0$ if and only if $P_1 = P_2$ [Fukumizu et al. (2007), Sejdinovic et al. (2013)]. Many commonly used kernels are shown to be characteristic kernels on $\mathbb{R}^p$, including the Gaussian kernel and Laplacian kernel [Fukumizu et al. (2007)].

We note that when $k(x, y) = |x - y|$, Equation (2) coincides with the formulation of ED [Székely et al. (2004)].

Definition 2 *Let X, X', Y, Y' be independence random vectors in $\mathbb{R}^p$ that satisfies $X, X' \sim P_1$ and $Y, Y' \sim P_2$, we define the energy distance (ED) as*

$$ED(P_1, P_2) = ED(X, Y) = \left(2\mathbb{E}\big[|X - Y|\big] - \mathbb{E}\big[|X - X'|\big] - \mathbb{E}\big[|Y - Y'|\big] \right)^{1/2}. \tag{3}$$

ED is a nonnegative metric and it holds that $ED(P_1, P_2) = 0$ if any only if $P_1 = P_2$.

In this paper, we aim to provide a unified treatment of ED and MMD, so we follow the approach in Zhu and Shao (2021) and mimic the definition of energy distance in Székely and Rizzo (2013b) and Huang and Huo (2017) to give the definition of MMD with a general kernel k. Though kernel is commonly used to measure similarity in the machine learning literature, we use kernel throughout this article to refer to a broader range of metrics of dissimilarities, which include both a semimetric of strong negative type on $\mathbb{R}^p$ (Definition 1 and Definition 28 of Sejdinovic et al. (2013)) and a characteristic kernel multiplied by -1, as formally stated in Definition 3 below. For notational simplicity, we shall use MMD (ED) instead of MMD^2 (ED^2) as in their original definitions (2) and (3), and the same is done for dcov later.

Definition 3 *Define $k : \mathbb{R}^p \times \mathbb{R}^p \to [0, \infty)$ to be a kernel that satisfies either of the following conditions:*

(i) for any $x, y \in \mathbb{R}^p$, it holds that $k(x, y) = k(y, x)$ and $k(x, y) = 0$ if and only if $x = y$, and additionally, for any Borel probability measures P, Q on $\mathbb{R}^p$ satisfying $\int k(z, z)dP(z) < \infty$ and $\int k(z, z)dQ(z) < \infty$, $P \neq Q$ implies that $\int kd([P-Q] \times [P-Q]) < 0$.

(ii) $(\mathcal{F}, -k)$ is an RKHS on $(\mathbb{R}^p, \mathcal{P})$ and the kernel $-k$ is characteristic.

To ease the reading, we mimic Table 1 of Zhu and Shao (2021) to summarize a few kernels covered by Definition 3 in the following table.

Kernel k	Expression of k	Condition satisfied
Euclidean distance	$k(x, y) = \lvert x - y \rvert$	k satisfies (i)
Gaussian kernel (multiplied by -1)	$k(x, y) = -\exp\left(-\lvert x - y \rvert^2/(2\gamma^2)\right)$	k satisfies (ii)
Laplacian kernel (multiplied by -1)	$k(x, y) = -\exp\left(-\lvert x - y \rvert/\gamma\right)$	k satisfies (ii)

Table 1: Examples of kernel k covered by Definition 3.

Then we are ready to propose the unified definition of ED and MMD.

Definition 4 *Let k denote a kernel defined as Definition 3. Suppose that $X, Y \in \mathbb{R}^p$ are two independent random vectors satisfying that $\mathbb{E}\left[\lvert k(X, X') \rvert\right] + \mathbb{E}\left[\lvert k(X, Y) \rvert\right] + \mathbb{E}\left[\lvert k(Y, Y') \rvert\right] < \infty$, then we define*

$$\mathcal{E}^k(X, Y) = 2\mathbb{E}\left[k(X, Y)\right] - \mathbb{E}\left[k(X, X')\right] - \mathbb{E}\left[k(Y, Y')\right], \tag{4}$$

where X', Y' are independent and identical copies of X and Y, respectively.

As shown in Sejdinovic et al. (2013), $\mathcal{E}^k(X, Y)$ is always non-negative and is zero if and only if $X =^d Y$. Similar to Huang and Huo (2017), who expressed sample ED as a U-statistic, we can find an unbiased estimator of $\mathcal{E}^k(X, Y)$ via a U-statistic with a two-sample kernel.

Proposition 5 *Define the two-sample kernel*

$$h^k(X_1, X_2, Y_1, Y_2) = \frac{1}{2}\sum_{i=1}^{2}\sum_{j=1}^{2} k(X_i, Y_j) - k(X_1, X_2) - k(Y_1, Y_2), \tag{5}$$

which satisfies $\mathbb{E}\left[h^k(X_1, X_2, Y_1, Y_2)\right] = \mathcal{E}^k(X, Y)$. *Then an unbiased estimator of* $\mathcal{E}^k(X, Y)$ *can be defined as*

$$\begin{aligned}\mathcal{E}^k_{n,m}(X,Y) &= \binom{n}{2}^{-1}\binom{m}{2}^{-1}\sum_{1\le i_1<i_2\le n}\sum_{1\le j_1<j_2\le m} h^k(X_{i_1}, X_{i_2}, Y_{j_1}, Y_{j_2}) \\ &= \frac{2}{mn}\sum_{i=1}^{n}\sum_{j=1}^{m} k(X_i, Y_j) - \binom{n}{2}^{-1}\sum_{1\le i<j\le n} k(X_i, X_j) - \binom{m}{2}^{-1}\sum_{1\le i<j\le m} k(Y_i, Y_j).\end{aligned}$$

It follows from the Hoeffding decomposition that two-sample U-statistic $\mathcal{E}^k_{n,m}(X, Y)$ can be decomposed into the sum of a leading term and a remainder term. In particular, let G_x denote the distribution function of a single point mass at x, and for $0 \le c, d \le 2$, define

$$\begin{aligned}h^{(c,d)}(X_1,\ldots,X_c;Y_1,\ldots,Y_d) = \int\cdots\int h^k(u_1,u_2,v_1,v_2)\prod_{i=1}^{c}\left(dG_{X_i}(u_i) - dF_X(u_i)\right)\prod_{i=c+1}^{2} dF_X(u_i) \\ \times\prod_{j=1}^{d}\left(dG_{Y_j}(v_j) - dF_Y(v_j)\right)\prod_{j=d+1}^{2} dF_Y(v_j).\end{aligned}$$

Then it holds that $\mathcal{E}^k_{n,m}(X, Y) = L^k_{n,m}(X, Y) + R^k_{n,m}(X, Y)$, where

$$\begin{aligned}L^k_{n,m}(X,Y) =& \mathcal{E}^k(X,Y) + \frac{2}{n}\sum_{i=1}^{n} h^{(1,0)}(X_i) + \frac{2}{m}\sum_{j=1}^{m} h^{(0,1)}(Y_j) \\ &+ \frac{2}{n(n-1)}\sum_{1\le i_1<i_2\le n} h^{(2,0)}(X_{i_1}, X_{i_2}) + \frac{4}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} h^{(1,1)}(X_i, Y_j) \\ &+ \frac{2}{m(m-1)}\sum_{1\le j_1<j_2\le m} h^{(0,2)}(Y_{j_1}, Y_{j_2}), \\ R^k_{n,m}(X,Y) =& \frac{4}{n(n-1)m}\sum_{1\le i_1<i_2\le n}\sum_{j=1}^{m} h^{(2,1)}(X_{i_1}, X_{i_2}, Y_j) \\ &+ \frac{4}{nm(m-1)}\sum_{i=1}^{n}\sum_{1\le j_1<j_2\le m} h^{(1,2)}(X_i, Y_{j_1}, Y_{j_2}) \\ &+ \frac{4}{n(n-1)m(m-1)}\sum_{1\le i_1<i_2\le n}\sum_{1\le j_1<j_2\le m} h^{(2,2)}(X_{i_1}, X_{i_2}, Y_{j_1}, Y_{j_2}),\end{aligned}$$

By generalizing some results established in Huang and Huo (2017), the expressions of $L^k_{n,m}(X, Y)$ and $R^k_{n,m}(X, Y)$ can be greatly simplified, which are stated in Proposition 6.

Proposition 6 *Assume that* $\mathbb{E}\left[|k(X, X')|\right] + \mathbb{E}\left[|k(X, Y)|\right] + \mathbb{E}\left[|k(Y, Y')|\right] < \infty$, *then it holds that* $R^k_{n,m}(X, Y) = 0$, *and*

$$\begin{aligned}L^k_{n,m}(X,Y) &= 3\mathcal{E}^k(X,Y) - \tfrac{4}{n}\sum_{i=1}^{n} h_{10}(X_i) - \tfrac{4}{m}\sum_{j=1}^{m} h_{01}(Y_j) + \tfrac{4}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} h_{11}(X_i, Y_j) \\ &\quad + \binom{n}{2}^{-1}\sum_{1\le i_1<i_2\le n} h_{20}(X_{i_1}, X_{i_2}) + \binom{m}{2}^{-1}\sum_{1\le j_1<j_2\le m} h_{02}(Y_{j_1}, Y_{j_2}),\end{aligned}$$

where for $1 \le i_1 < i_2 \le n$ *and* $1 \le j_1 < j_2 \le m$, *we define*

$$h_{10}(X_{i_1}) = \mathbb{E}_{X_{i_2},Y_{j_1},Y_{j_2}}[h^k(X_{i_1},X_{i_2},Y_{j_1},Y_{j_2})], \quad h_{01}(Y_{j_1}) = \mathbb{E}_{X_{i_1},X_{i_2},Y_{j_2}}[h^k(X_{i_1},X_{i_2},Y_{j_1},Y_{j_2})],$$
$$h_{20}(X_{i_1},X_{i_2}) = \mathbb{E}_{Y_{j_1},Y_{j_2}}[h^k(X_{i_1},X_{i_2},Y_{j_1},Y_{j_2})], \quad h_{02}(Y_{j_1},Y_{j_2}) = \mathbb{E}_{X_{i_1},X_{i_2}}[h^k(X_{i_1},X_{i_2},Y_{j_1},Y_{j_2})],$$

and $h_{11}(X_{i_1},Y_{j_1}) = \mathbb{E}_{X_{i_2},Y_{j_2}}[h^k(X_{i_1},X_{i_2},Y_{j_1},Y_{j_2})]$.

If additionally $X =^d Y$, *then* $\mathcal{E}^k_{n,m}(X,Y) = L^k_{n,m}(X,Y)$ *with the simplified expression*

$$L^k_{n,m}(X,Y) = \binom{n}{2}^{-1} \sum_{1\le i_1<i_2\le n} h_{20}(X_{i_1},X_{i_2}) + \tfrac{4}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} h_{11}(X_i,Y_j) + \binom{m}{2}^{-1} \sum_{1\le j_1<j_2\le m} h_{02}(Y_{j_1},Y_{j_2}).$$

Proposition 6 simplifies $\mathcal{E}^k_{n,m}(X,Y)$ and facilitates the subsequent analysis.

2.2 Distributional Properties of $\mathcal{E}^k_{n,m}(X,Y)$ with Fixed p

The asymptotic behavior of sample MMD (and sample ED) has been well studied when the dimension p is fixed; see Székely et al. (2004), Gretton et al. (2009) and Gretton et al. (2012). In particular, the asymptotic distribution of $\mathcal{E}^k_{n,m}(X,Y)$ (in the case of sample MMD) under the null is established in Theorem 12 of Gretton et al. (2012). Define

$$d^k(X_1,X_2) = k(X_1,X_2) - \mathbb{E}_{X_1}\left[k(X_1,X_2)\right] - \mathbb{E}_{X_2}\left[k(X_1,X_2)\right] + \mathbb{E}\left[k(X_1,X_2)\right]. \tag{6}$$

Assume that $\mathbb{E}\left[k^2(X,X')\right] < \infty$, and that $\lim_{n,m\to\infty} \frac{n}{n+m} = \rho$ for some fixed $0 < \rho < 1$, then under the null, $\mathcal{E}^k_{n,m}(X,Y)$ converges in distribution according to

$$(n+m)\mathcal{E}^k_{n,m}(X,Y) \to^d \sum_{\ell=1}^{\infty} \lambda_\ell \left((\rho(1-\rho))^{-1} - \left(\rho^{-1/2}a_\ell - (1-\rho)^{-1/2}b_\ell\right)^2 \right),$$

where $\{a_\ell\}_{\ell\ge1}, \{b_\ell\}_{\ell\ge1} \sim \mathcal{N}(0,1)$ are two independent sequences of iid Gaussian random variables, and $\{\lambda_i\}_{i=1}^{\infty}$ and $\{\Psi_i(x)\}_{i=1}^{\infty}$ are respectively the eigenvalues and the eigenfunctions of the equation $\mathbb{E}_X\left[d^k(X,X')\Psi_i(X)\right] = \lambda_i\Psi_i(X')$. Note that the above limiting null distribution is not pivotal, so critical values are not directly available. Several approximation methods have been developed in the special case of $m = n$ in Gretton et al. (2009).

3. Studentized Statistic and Asymptotic Theory

3.1 Studentized Statistic

To develop our studentized statistic, we need to find the variance of $L^k_{n,m}$ under the null, which can be shown to have a strong connection with the Hilbert-Schmidt independence criterion (HSIC) [Gretton et al. (2007)]. We thus start from the definition and some basic results of HSIC, and then move on to deriving the variance of $L^k_{n,m}$ under the null before proposing a studentized test statistic.

Definition 7 *Suppose* $X \in \mathbb{R}^p$ *and* $Y \in \mathbb{R}^q$ *with integers* $p, q \ge 1$. *Let* k *be a kernel defined as Definition 3, then a generalized Hilbert-Schmidt independence criterion (HSIC)* $\mathcal{V}^2_k(X,Y)$ *between the distributions* X *and* Y *is defined as*

$$\mathcal{V}^2_k(X,Y) = \mathbb{E}\big[k(X_1,X_2)k(Y_1,Y_2)\big] - 2\mathbb{E}\big[k(X_1,X_2)k(Y_1,Y_3)\big] + \mathbb{E}\big[k(X_1,X_2)\big]\mathbb{E}\big[k(Y_1,Y_2)\big], \tag{7}$$

where (X_1,Y_1), (X_2,Y_2) *and* (X_3,Y_3) *are independent copies of* (X,Y).

Remark 8 *This definition directly follows from Lemma 1 of Gretton et al. (2005) when $-k(\cdot,\cdot)$ is a characteristic kernel w.r.t. $(\mathbb{R}^p, \mathcal{P})$. When k is a semimetric of strong negative type according to Definition 3, the resulting metric also fully quantifies nonlinear dependence in the sense that $\mathcal{V}_k^2(X,Y) \geq 0$ and equals to zero if and only if X, Y are independent; see Sejdinovic et al. (2013). In fact, $\mathcal{V}_k^2(X,Y)$ can be viewed as a generalization of distance covariance (dcov), as the expression of $\mathcal{V}_k^2(X,Y)$ with $k(x,y) = |x-y|$ coincides with the well-known dcov [Székely et al. (2007)]. For simplicity, we call $\mathcal{V}_k^2(X,Y)$ the HSIC in both cases.*

Next, we introduce a mixture distribution of X and Y, defined by

$$Z = \begin{cases} X, & \text{with probability } \rho, \\ Y, & \text{with probability } 1-\rho, \end{cases} \tag{8}$$

where ρ is the limit of the size proportion $n/(n+m)$. By introducing the mixture distribution Z, we are able to aggregate X and Y with respect to their occurrence frequencies in the pooled sample. Consequently, we can directly establish some unified results in terms of Z, which are more succinct than establishing the counterparts in terms of X and Y respectively.

Let Z_1, Z_2, Z_3 be three independent copies of Z, then HSIC of Z with itself is given by

$$\mathcal{V}_k^2(Z) = \mathbb{E}\left[k^2(Z_1,Z_2)\right] - 2\mathbb{E}\left[k(Z_1,Z_2)k(Z_1,Z_3)\right] + \left(\mathbb{E}\left[k(Z_1,Z_2)\right]\right)^2. \tag{9}$$

With $\mathcal{V}_k^2(X,Y)$ being a generalization of distance covariance, we can also view $\mathcal{V}_k^2(Z)$ as a generalization of distance variance of Z. It is trivial that $X =^d Y =^d Z$ under the null. Furthermore, the variance of $L_{n,m}^k$ under the null can be written in terms of $\mathcal{V}_k^2(Z)$.

Proposition 9 *If X and Y are identically distributed, and $\mathbb{E}\left[k^2(X_1,X_2)\right] < \infty$, then it holds that $Var(L_{n,m}^k(X,Y)) = c_{n,m}\mathcal{V}_k^2(Z)$, where $c_{n,m} = \frac{2}{n(n-1)} + \frac{4}{nm} + \frac{2}{m(m-1)}$.*

If given n independent and identically distributed observations $\mathbf{Z} = (Z_1, \ldots, Z_n)$ from the mixture distribution Z, an unbiased estimator of $\mathcal{V}_k^2(Z)$ with $k(x,y) = |x-y|$ can be obtained using the U-centering approach in Székely and Rizzo (2013a) and Székely and Rizzo (2014). However, the mixture distribution Z is unobserved, and we only have two independent random samples $\mathbf{X} = (X_1, \ldots, X_n)$ from the distribution of X and $\mathbf{Y} = (Y_1, \ldots, Y_m)$ from the distribution of Y. Let $N = n + m$ denote the total sample size. Throughout, we assume that there exists some constant $0 < \rho < 1$, such that $n/N \to \rho$ as $\min\{n,m\} \to \infty$. We propose to use the pooled sample to estimate $\mathcal{V}_k^2(Z)$ as follows.

Proposition 10 *For any fixed p and kernel $k : \mathbb{R}^p \times \mathbb{R}^p \mapsto \mathbb{R}$, assume that $\mathbb{E}\left[k^2(X_1,X_2)\right]$, $\mathbb{E}\left[k^2(X_1,Y_1)\right]$ and $\mathbb{E}\left[k^2(Y_1,Y_2)\right]$ are all finite, and $k(X,X) = a_0^k$ is a finite constant independent of X. For $1 \leq s,t \leq N$, define*

$$a_{s,t}^k = \begin{cases} k(X_s,X_t), & 1 \leq s,t \leq n \\ k(X_s,Y_{t-n}), & 1 \leq s \leq n < t \leq N \\ k(X_t,Y_{s-n}), & 1 \leq t \leq n < s \leq N \\ k(Y_{s-n},Y_{t-n}), & n+1 \leq s,t \leq N \end{cases} \tag{10}$$

Define the $\mathcal{U}$-centered distances with kernel k as $A^{k*}_{s,t} = a^k_{s,t} - \tilde{a}^k_{\cdot t} - \tilde{a}^k_{s\cdot} + \tilde{a}^k_{\cdot\cdot}$, where

$$\tilde{a}^k_{\cdot t} = \frac{1}{N-2}\sum_{i=1}^{N} a^k_{i,t}, \quad \tilde{a}^k_{s\cdot} = \frac{1}{N-2}\sum_{j=1}^{N} a^k_{s,j}, \quad \tilde{a}^k_{\cdot\cdot} = \frac{1}{(N-1)(N-2)}\sum_{i,j=1}^{N} a^k_{i,j}.$$

Then under the null, it holds for any fixed p and kernel k that,

$$\mathcal{V}^{k*}_{n,m}(X,Y) = \frac{1}{N(N-3)}\sum_{s\neq t}\left(A^{k*}_{s,t}\right)^2 - \frac{(a^k_0)^2}{(N-1)(N-3)} \tag{11}$$

is an unbiased estimator of $\mathcal{V}^2_k(Z)$. Furthermore, under the alternative, $\mathcal{V}^{k*}_{n,m}(X,Y)$ is asymptotically unbiased of $\mathcal{V}^2_k(Z)$ for any fixed p and kernel k, that is, $\mathbb{E}\left[\mathcal{V}^{k*}_{n,m}(X,Y)\right] \to \mathcal{V}^2_k(Z)$ as $n, m \to \infty$.

Remark 11 *If the kernel k is chosen to be the L_2 norm, we have $a^k_0 = 0$ and the estimate $\mathcal{V}^{k*}_{n,m}(X,Y)$ reduces to the traditional $\mathcal{U}$-centering based sample distance variance based on the pooled sample. However, for a general kernel k, a^k_0 may be nonzero, and the correction term $-\frac{(a^k_0)^2}{(N-1)(N-3)}$ is necessary to obtain the unbiasedness. This bias correction is important for Gaussian and Laplacian kernels as the use of biased variance estimate leads to noticeable size distortion in the small sample in our (unreported) simulations.*

To our best knowledge, the proposed estimate of $\mathcal{V}^2_k(Z)$ based on the pooled sample is a new addition to the literature, and it is different from the studentizers proposed in Chakraborty and Zhang (2021) and Yan and Zhang (2023); see Remark 12 and Remark 13.

In terms of computational complexity, the computation of all $a^k_{s,t}$'s is of order $O((n+m)^2p)$. Since the $\mathcal{U}$-centering only requires $O((n+m)^2)$ computation, the computational complexity of $\mathcal{V}^{k*}_{n,m}(X,Y)$ and the studentized statistic $T^k_{n,m,p}$ defined below is of order $O((n+m)^2p)$. By contrast, the computational complexity for the permutation based test in Zhu and Shao (2021) is of order $O((n+m)^2pB)$, where B is the number of permutations.

To test $H_0 : X =^d Y$ against $H_a : X \neq^d Y$, it is natural to use the following studentized test statistic:

$$T^k_{n,m,p} = \frac{\mathcal{E}^k_{n,m}(X,Y)}{\sqrt{c_{n,m}\mathcal{V}^{k*}_{n,m}(X,Y)}}, \tag{12}$$

where $c_{n,m}$ is defined in Proposition 9.

Similar test statistics for the two-sample problem have been previously discussed in other existing papers; see Chakraborty and Zhang (2021) and Yan and Zhang (2023). We conjecture that all three studentizers are asymptotically equivalent. Some additional discussions can be found in the following remarks.

Remark 12 *In Chakraborty and Zhang (2021), a different studentized test statistic is proposed in the form of $\tilde{T}^k_{n,m,p} = \frac{\mathcal{E}^k_{n,m}(X,Y)}{\sqrt{c_{n,m}S_{n,m}/2}}$. The major difference between $T^k_{n,m,p}$ and $\tilde{T}^k_{n,m,p}$ is the variance estimator of $\mathcal{E}^k_{n,m}(X,Y)$ in the denominator. Specifically,*

$$S_{n,m} = \frac{4(n-1)(m-1)cdCov^2_{n,m}(X,Y) + 4v_n\mathcal{V}^{k*}_n(X) + 4v_m\mathcal{V}^{k*}_m(Y)}{(n-1)(m-1) + n(n-3)/2 + m(m-3)/2},$$

where $\mathcal{V}_n^{k*}(X), \mathcal{V}_m^{k*}(Y)$ *are respectively the* $\mathcal{U}$*-centering based unbiased estimators of* $\mathcal{V}_k^2(X), \mathcal{V}_k^2(Y)$*, and* $cdCov_{n,m}^2(X,Y)$ *is the cross distance covariance between* X *and* Y*, given by*

$$cdCov_{n,m}^2(X,Y) = \frac{1}{(n-1)(m-1)} \sum_{k=1}^{n} \sum_{l=1}^{m} \hat{k}(X_k, Y_l)^2$$

with $\hat{k}(X_k, Y_l) = k(X_k, Y_l) - \frac{1}{n}\sum_{i=1}^{n} k(X_i, Y_l) - \frac{1}{m}\sum_{j=1}^{m} k(X_k, Y_j) + \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} k(X_i, Y_j)$*. In Theorem 4.2 of Chakraborty and Zhang (2021), they derived the limiting distribution of* $\tilde{T}_{n,m,p}^k$ *under both the null and alternatives when* $p \to \infty$ *whereas* (n, m) *are fixed.*

Remark 13 *In a very recent paper by Yan and Zhang (2023), they also proposed a studentized MMD test, and their studentizer is based on a linearization argument and differs from ours and the one in Chakraborty and Zhang (2021). However, the CLT results in Yan and Zhang (2023) are established for the standardized statistic instead of the studentized statistic under both the null and alternative when both the dimension and sample size diverge, and the standardizer is actually infeasible.*

As we present below, we will be investigating the asymptotic behavior of our studentized statistic $T_{n,m,p}^k$ under the setting $\min(n, m, p) \to \infty$ using a different set of technical arguments and our results are complementary to those in theirs.

3.2 Asymptotic Distributions

For each $p \in \mathbb{N}$, let $k^{(p)} : \mathbb{R}^p \times \mathbb{R}^p \mapsto \mathbb{R}$ be a kernel defined as in Definition 3 and $(k^{(p)} : p \in \mathbb{N})$ thus forms a sequence of kernels. Throughout the paper, we let $\mathcal{C} = \{(k^{(p)} : p \in \mathbb{N})\}$ be the set of all the kernel sequences of interest. Again, we drop the symbol (p) for simplicity when we are focusing on a specific kernel given a fixed p.

Let $\tilde{k}(Z_1, Z_2) = k(Z_1, Z_2) - \mathbb{E}\big[k(Z_1, Z_2)\big]$ denote the centered version of k. We have already shown that $\mathcal{V}_{n,m}^{k*}(X,Y)$ is an unbiased estimator of $\mathcal{V}_k^2(Z)$ under the null and is asymptotically unbiased under the alternative, then we are ready to state that $\mathcal{V}_{n,m}^{k*}(X,Y)$ is ratio-consistent for $\mathcal{V}_k^2(Z)$ under both the null and the alternative with some conditions.

Proposition 14 *Assume that* $\mathbb{E}\left[k^4(Z_1, Z_2)\right] < \infty$ *for each* $k = k^{(p)} \in \mathcal{C}$ *and* $n/N \to \rho$ *as* $n, m \to \infty$ *for some* $0 < \rho < 1$*. Under the null when* $X =^d Y =^d Z$*, if for some constant* $0 < \tau \leq 1$*, it is satisfied when* $N, p \to \infty$ *that*

$$\frac{\mathbb{E}\left[\left|\tilde{k}(Z_1, Z_2)\right|^{2+2\tau}\right]}{N^{\tau}\left(\mathcal{V}_k^2(Z)\right)^{1+\tau}} \to 0, \tag{13}$$

then we have $\dfrac{\mathcal{V}_{n,m}^{k*}(X,Y)}{\mathcal{V}_k^2(Z)} \longrightarrow^p 1$*. As a direct consequence, under the null we have that* $\dfrac{c_{n,m}\mathcal{V}_{n,m}^{k*}(X,Y)}{Var(L_{n,m}^k(X,Y))} \to^p 1$*.*

Condition (13) is directly from the use of Markov's inequality. In fact, to show that $\mathcal{V}^{k*}_{n,m}(X,Y)$ is ratio-consistent for $\mathcal{V}^2_k(Z)$, it suffices to find an upper bound of $\frac{\mathbb{E}\left[\left|\mathcal{V}^{k*}_{n,m}(X,Y)-\mathcal{V}^2_k(Z)\right|^{1+\tau}\right]}{\left(\mathcal{V}^2_k(Z)\right)^{1+\tau}}$, where $\tau \in (0,1]$. As shown in Proposition 10, $\mathcal{V}^{k*}_{n,m}(X,Y)$ is unbiased of $\mathcal{V}^2_k(Z)$ under the null. Furthermore, $\mathcal{V}^{k*}_{n,m}(X,Y) - \mathcal{V}^2_k(Z)$ can be decomposed as a summation of multiple U-statistics with mean zero. By applying a moment inequality for the U-statistics, we can show that the deviation $\mathbb{E}\left[|\mathcal{V}^{k*}_{n,m}(X,Y) - \mathcal{V}^2_k(Z)|^{1+\tau}\right]$ is bounded by $\mathbb{E}\left[|\tilde{k}(Z_1,Z_2)|^{2+2\tau}\right]/N^\tau$ from above. This leads to condition (13).

As a counterpart of Proposition 14, the ratio-consistency of the sample estimate $\mathcal{V}^{k*}_{n,m}(X,Y)$ under the alternative is established in Proposition 15.

Proposition 15 *Assume that* $\mathbb{E}\left[k^4(Z_1,Z_2)\right] < \infty$ *for each* $k = k^{(p)} \in \mathcal{C}$ *and* $n/N = \rho + O(1/N^s)$ *as* $n,m \to \infty$ *for some* $0 < \rho < 1$ *and* $s > 0$. *Under the alternative, if for some constant* $0 < \tau \le 1$, *it is satisfied that*

$$\frac{\left|\mathcal{E}^k(X,Y)\right|^{2+2\tau}}{N^\tau\left(\mathcal{V}^2_k(Z)\right)^{1+\tau}} \to 0, \qquad \frac{\mathbb{E}\left[k^2(Z_1,Z_2)\right]}{N^s\mathcal{V}^2_k(Z)} \to 0, \tag{14}$$

and

$$\frac{\mathbb{E}\left[\left|\tilde{k}(X_1,X_2)\right|^{2+2\tau} + \left|\tilde{k}(X_1,Y_1)\right|^{2+2\tau} + \left|\tilde{k}(Y_1,Y_2)\right|^{2+2\tau}\right]}{N^\tau\left(\mathcal{V}^2_k(Z)\right)^{1+\tau}} \to 0, \tag{15}$$

then it holds that $\dfrac{\mathcal{V}^{k*}_{n,m}(X,Y)}{\mathcal{V}^2_k(Z)} \longrightarrow^p 1$ *as* $N,p \to \infty$.

The argument to show Proposition 15 is quite similar to that of Proposition 14, with the main difference being attributed to the fact that $\mathcal{V}^{k*}_{n,m}(X,Y)$ is not an unbiased estimator of $\mathcal{V}^2_k(Z)$ under the alternative. In this case, to bound $\mathbb{E}\left[|\mathcal{V}^{k*}_{n,m}(X,Y) - \mathcal{V}^2_k(Z)|^{1+\tau}\right]$, we break it into two parts, namely, $\mathbb{E}\left[|\mathcal{V}^{k*}_{n,m}(X,Y) - \mathbb{E}[\mathcal{V}^{k*}_{n,m}(X,Y)]|^{1+\tau}\right]$ and $|\mathbb{E}[\mathcal{V}^{k*}_{n,m}(X,Y)] - \mathcal{V}^2_k(Z)|^{1+\tau}$. Note that $\mathcal{V}^{k*}_{n,m}(X,Y) - \mathbb{E}\left[\mathcal{V}^{k*}_{n,m}(X,Y)\right]$ can be decomposed as a combination of multiple U-statistics and its upper bound is obtained by a moment inequality. This is manifested in the first condition in (14) and condition (15). A major difference from the condition under the null is that, to make the pooled-sample estimate $\mathcal{V}^{k*}_{n,m}(X,Y)$ ratio-consistent for $\mathcal{V}^2_k(Z)$, the discrepancy between the distributions of X and Y, as quantified by $\mathcal{E}^k(X,Y)$, cannot be too large, as regulated by the first condition in (14).

The upper bound of $|\mathbb{E}\left[\mathcal{V}^{k*}_{n,m}(X,Y)\right] - \mathcal{V}^2_k(Z)|^{1+\tau}$ corresponds to the second condition in (14). In fact, it follows from some simple calculations that, under the assumption $n/N = \rho + O\left(1/N^s\right)$, the bias $\mathbb{E}\left[\mathcal{V}^{k*}_{n,m}(X,Y)\right] - \mathcal{V}^2_k(Z)$ can be bounded by $\mathbb{E}\left[k^2(Z_1,Z_2)\right]/N^s$ up to a multiplicative constant, where the convergence rate of n/N is involved.

To establish the central limit theorem for the proposed test, we define the functionals

$$g^k(X_1,X_2,X_3,X_4) = d^k(X_1,X_2)d^k(X_1,X_3)d^k(X_2,X_4)d^k(X_3,X_4), \tag{16}$$

where d^k is defined as (6). We can obtain the following central limit theorem for the proposed test statistic under the null.

Theorem 16 *Assume that* $\mathbb{E}\left[k^4(Z_1, Z_2)\right] < \infty$ *for each* $k = k^{(p)} \in \mathcal{C}$ *and* $n/N \to \rho$ *as* $n, m \to \infty$ *for some* $0 < \rho < 1$. *Under the null when* $X =^d Y =^d Z$, *if for some constant* $0 < \tau \leq 1$, *it is satisfied for some* $\{(k^{(p)})\} \in \mathcal{C}$ *that*

$$\frac{\mathbb{E}\left[\left|\tilde{k}(Z_1, Z_2)\right|^{2+2\tau}\right]}{N^{\tau}\left(\mathcal{V}_k^2(Z)\right)^{1+\tau}} \to 0, \tag{17}$$

and

$$\frac{\left(\mathbb{E}\left[g^k(Z_1, Z_2, Z_3, Z_4)\right]\right)^{(1+\tau)/2}}{\left(\mathcal{V}_k^2(Z)\right)^{1+\tau}} \to 0, \tag{18}$$

when $N, p \to \infty$, *then it holds for this sequence* $\{k^{(p)}\}$ *that* $T^k_{n,m,p} \longrightarrow^d \mathcal{N}(0, 1)$.

This theorem can be viewed as a counterpart of Theorem 1 in Gao et al. (2021) but is stated for a general kernel. Under the null, it follows from Proposition 14 that $\mathcal{V}^{k*}_{n,m}(X, Y)$ is ratio-consistent for $\mathcal{V}_k^2(Z)$ under H_0 when (17) is satisfied, and it is derived in Proposition 6 that $\mathcal{E}^k_{n,m}(X, Y) = L^k_{n,m}(X, Y)$. To derive the central limit theorem of $T^k_{n,m,p}$, it suffices to investigate the asymptotic behavior of $\frac{L^k_{n,m}(X,Y)}{\sqrt{c_{n,m}\mathcal{V}_k^2(Z)}}$ using the martingale central limit theorem, since $L^k_{n,m}(X, Y)$ forms a martingale. Condition (18) is basically Lyapunov-type condition in the use of martingale central limit theorem.

Note that condition (18) only depends on p and is free of the sample size N, whereas condition (17) depend on both N and p and thus might impose some implicit constraints between the divergence rate of N and p. For any fixed τ, if the order of $\mathbb{E}\big[|\tilde{k}(Z_1, Z_2)|\big]^{2+2\tau}$ does not exceed that of $(\mathcal{V}_k^2(Z))^{1+\tau}$, then the quantity in (17) naturally goes to zero as long as N diverges without additional restrictions between N and p. As it turns out, it can be shown that the orders of $\mathbb{E}\big[|\tilde{k}(Z_1, Z_2)|\big]^{2+2\tau}$ and $(\mathcal{V}_k^2(Z))^{1+\tau}$ are the same for the Gaussian kernel, the Laplacian kernel as well as the L_2 norm, hence the first term is independent of p for these kernels. Furthermore, later in this paper, we show that this is true as long as the kernel satisfies some technical conditions.

In the literature, Zhu and Shao (2021) obtained the asymptotic distribution for the MMD permutation test statistics under the HDLSS (high-dimensional low sample size, where p grows to infinity and (n, m) is fixed) and HDMSS (high-dimensional medium sample size, where (p, n, m) all grow to infinity but p grows faster than N). The asymptotic results for the studentized test proposed in Chakraborty and Zhang (2021) are also limited to the HDLSS setting. Yan and Zhang (2023) obtained the CLT of a standardized MMD statistic for a factor-like model allowing (p, n, m) to diverge without constraints.

3.3 Rate of Convergence

We can further obtain the rate of convergence of the test statistic under the null using the Berry-Esseen bound for martingales, which has been used in Gao et al. (2021). Here we first follow their steps to find an upper bound of $\sup_{x\in\mathbb{R}} \left|\mathbb{P}\left(T^k_{n,m,p} \leq x\right) - \Phi(x)\right|$.

Note that under the null, for any $0 < \gamma < 1$, we have

$$\begin{aligned}
&\sup_{x\in\mathbb{R}}\left|\mathbb{P}\left(T^k_{n,m,p}\le x\right)-\Phi(x)\right| = \sup_{x\in\mathbb{R}}\left|\mathbb{P}\left(\frac{\mathcal{E}^k_{n,m}}{\sqrt{c_{n,m}\mathcal{V}^{k*}_{n,m}(X,Y)}}\le x\right)-\Phi(x)\right| \\
\le\ & 2\sup_{x\in\mathbb{R}}\left|\mathbb{P}\left(\frac{L^k_{n,m}}{\sqrt{\mathrm{Var}(L^k_{n,m})}}\le x\right)-\Phi(x)\right| + \sup_{x\in\mathbb{R}}\left|\Phi(x)-\Phi(x\sqrt{1+\gamma})\right| \\
&+\sup_{x\in\mathbb{R}}\left|\Phi(x)-\Phi(x\sqrt{1-\gamma})\right| + 2\mathbb{P}\left(\left|\frac{c_{n,m}\mathcal{V}^{k*}_{n,m}(X,Y)}{\mathrm{Var}(L^k_{n,m})}-1\right|>\gamma\right) =: 2P_1+P_2+P_3+2P_4.
\end{aligned}$$

By upper bounding each P_i, we obtain the following theorem.

Theorem 17 *Let Z denote the mixture distribution of X and Y defined as (8). Assume that $\mathbb{E}\left[k^4(Z_1,Z_2)\right] < \infty$ for each $k = k^{(p)} \in \mathcal{C}$ and $n/N \to \rho$ as $n, m \to \infty$ for some $0 < \rho < 1$. Under the null, it holds for any n, m, p and $0 < \tau \le 1$ that*

$$\sup_{x\in\mathbb{R}}\left|\mathbb{P}\left(T^k_{n,m,p}\le x\right)-\Phi(x)\right| \le C(\rho,\tau)\left\{\frac{\mathbb{E}\left[\left|\tilde{k}(Z_1,Z_2)\right|^{2+2\tau}\right]}{N^\tau(\mathcal{V}^2_k(Z))^{1+\tau}} + \frac{\left(\mathbb{E}\left[g^k(Z_1,Z_2,Z_3,Z_4)\right]\right)^{\frac{1+\tau}{2}}}{(\mathcal{V}^2_k(Z))^{1+\tau}}\right\}^{\frac{1}{3+2\tau}}.$$

Theorem 17 states a non-asymptotic Berry-Esseen bound of the proposed test statistic. The two terms in the bound, $\frac{\mathbb{E}\left[|\tilde{k}(Z_1,Z_2)|\right]^{2+2\tau}}{N^\tau(\mathcal{V}^2_k(Z))^{1+\tau}}$ and $\frac{\left(\mathbb{E}\left[g^k(Z_1,Z_2,Z_3,Z_4)\right]\right)^{(1+\tau)/2}}{(\mathcal{V}^2_k(Z))^{1+\tau}}$ jointly determine the accuracy of normal approximation. As we have mentioned after Theorem 16, the second term is solely determined by p, whereas the first term might depend on both N and p. Although the bound established in Theorem 17 is valid for any n, m, p, the accuracy of normal approximation is guaranteed only when both quantities are close to zero, which might put some restrictions on the way the dimension p diverges with respect to N. Such restriction is implicit for a general kernel k, but under some assumptions we can explicitly calculate the order of each term on the right-hand side, which enables us to derive the specific regime where the bound goes to zero. To this end, we first present a computational formula for $\mathbb{E}\left[g^k(Z_1,Z_2,Z_3,Z_4)\right]$ in the following proposition.

Proposition 18 *Assume that $\mathbb{E}\left[k^4(Z_1,Z_2)\right] < \infty$, it holds that*

$$\mathbb{E}\left[g^k(Z_1,Z_2,Z_3,Z_4)\right] = G_1+G_2+G_3+G_4,$$

where $G_1 = \mathbb{E}\left[k(Z_1,Z_2)k(Z_1,Z_3)k(Z_2,Z_4)k(Z_3,Z_4)\right] - 4\mathbb{E}\left[k(Z_1,Z_2)k(Z_1,Z_3)k(Z_2,Z_4)k(Z_4,Z_5)\right] + 2\mathbb{E}\left[k(Z_1,Z_2)k(Z_1,Z_3)\right]^2$, $G_2 = 4\mathbb{E}\left[k(Z_1,Z_2)\right]\mathbb{E}\left[k(Z_1,Z_2)k(Z_1,Z_3)k(Z_2,Z_4)\right]$, $G_3 = -4\mathbb{E}\left[k(Z_1,Z_2)\right]^2 \times \mathbb{E}\left[k(Z_1,Z_2)k(Z_1,Z_3)\right]$, and $G_4 = \mathbb{E}\left[k(Z_1,Z_2)\right]^4$.

If we restrict our attention to the kernels of the form $k(x,y) = f(|x-y|)$ for some smooth function f, we can derive the explicit rate of convergence. To this end, in the following we state the technical assumptions on f and the distributions of X and Y.

Assumption 1 *Assume that for each $k^{(p)} \in \mathcal{C}$, there exists some function $f^{(p)}$, such that $k^{(p)}(x,y) = f^{(p)}(|x-y|)$ for any $x, y \in \mathbb{R}^p$. Let D be the domain of $f^{(p)}$ and $D_0 \subseteq D$ be a set that contains $A_0 = \mathbb{E}\left[|Z_1 - Z_2|^2\right]^{1/2}$ and $A_0^{XY} = \mathbb{E}\left[|X_1 - Y_1|^2\right]^{1/2}$.*

Additionally, assume that

(i) for each $f^{(p)}$ and any $s \in D$ and $s_0 \in D_0$, it holds that

$$f^{(p)}(s) = \sum_{i=0}^{6} \frac{1}{i!} f_i^{(p)}(s_0)(s-s_0)^i + f_7^{(p)}(\xi(s,s_0))(s-s_0)^7,$$

where $f_i^{(p)}$ denotes the i-th order derivative of $f^{(p)}$, and $\xi(s,s_0)$ denotes some point between s and s_0.

(ii) these exists a positive constant $\tilde{M} < \infty$, such that for any $f^{(p)}$ and any $s_0 \in D_0$, it holds that

$$\max_{1 \le i \le 7} \sup_{s \in D} |f_i^{(p)}(s)| \cdot |s_0^i| \le \tilde{M} |f_0^{(p)}(s_0)|.$$

(iii) there exists a positive constant $\hat{M} < \infty$, such that for any $f^{(p)}$ and any $s_0 \in D_0$, it holds that

$$|f_0^{(p)}(s)| \le \hat{M} \min\{|\frac{1}{2} f_1^{(p)}(s)s|, |-\frac{1}{8} f_1^{(p)}(s)s + \frac{1}{8} f_2^{(p)}(s)s^2|\}.$$

(iv) there exists a positive constant $\hat{M} < \infty$, such that for any $f^{(p)}$ and any $s_0 \in D_0$, it holds that

$$\begin{aligned} |f_0^{(p)}(s)| \quad \le \quad & \hat{M} \min\{|\frac{1}{16} f_1^{(p)}(s)s - \frac{1}{16} f_2^{(p)}(s)s^2 + \frac{1}{48} f_3^{(p)}(s)s^3|, \\ & |-\frac{5}{128} f_1^{(p)}(s)s + \frac{5}{128} f_2^{(p)}(s)s^2 - \frac{1}{64} f_3^{(p)}(s)s^3 + \frac{1}{384} f_4^{(p)}(s)s^4|\}. \end{aligned}$$

Assumption 1(i) is mild and it only requires that the function f is smooth enough and has continuous derivatives up to the 7-th order. Assumption 1(ii)-(iv) further regulates the smoothness of the derivatives of f and will be used to determine the exact orders of $\mathcal{E}^k(X,Y)$ and $\mathcal{V}_k^2(Z)$. Later in Section 3.5, we will use the Gaussian kernel as a special example to demonstrate the verification of Assumption 1. Additional examples for L_2 norm and the Laplacian kernel can be found in the online supplement.

Before stating the next assumption, we introduce some useful notations. Define $A = \mathbb{E}\left[|Z_1 - Z_2|^2\right]$, $A^X = \mathbb{E}\left[|X_1 - X_2|^2\right]$, $A^{XY} = \mathbb{E}\left[|X_1 - Y_1|^2\right]$ and $A^Y = \mathbb{E}\left[|Y_1 - Y_2|^2\right]$. Let $A_0 = A^{1/2}$ and define A_0^X, A_0^{XY} and A_0^Y in the same way. For each X and Y, let $\mu_X = \mathbb{E}[X]$ and $\mu_Y = \mathbb{E}[Y]$ be the mean vectors, and use $\tilde{X} = X - \mu_X$ and $\tilde{Y} = Y - \mu_Y$ to denote the centered version of X, Y respectively. Additionally, we use $\Delta = \mu_X - \mu_Y$ to denote the mean difference between X and Y and we denote the covariance matrices of X, Y by $\Sigma_X = (\sigma^2_{X,j_1j_2})_{j_1,j_2}$ and $\Sigma_Y = (\sigma^2_{Y,j_1j_2})_{j_1,j_2}$.

Assumption 2 *For any fixed p and $X = (x_1, \ldots, x_p)^\top$ and $Y = (y_1, \ldots, y_p)^\top$, assume that*

(i) there exists an integer $1 \leq \alpha(p) \prec p$*, such that* X *and* Y *have* $\alpha(p)$*-dependent components, respectively. Specifically, for any* $1 \leq j \leq p - \alpha(p) - 1$ *and* $\ell > \alpha(p)$*,* $\{x_i\}_{i \leq j}$ *is independent of* $\{x_i\}_{i \geq j+\ell}$*, and* $\{y_i\}_{i \leq j}$ *is independent of* $\{y_i\}_{i \geq j+\ell}$*.*

(ii) there exists a constant $0 < U^* < \infty$*, such that*

$$\max_{1 \leq j \leq p} \max_{1 \leq r \leq 128} \{\mathbb{E}[|x_j|^r], \mathbb{E}[|y_j|^r]\} < U^*.$$

(iii) there exists some positive constants $0 < L_0, U_0 < \infty$*, such that*

$$L_0 p \leq \min\{A^X, A^{XY}, A^Y\} \leq \max\{A^X, A^{XY}, A^Y\} \leq U_0 p.$$

(iv) there exists some positive constants $0 < L_0^*, U_0^* < \infty$*, such that*

$$L_0^* \alpha(p) p \leq \min\{\|\Sigma_X\|_F^2, \|\Sigma_Y\|_F^2\} \leq \max\{\|\Sigma_X\|_F^2, \|\Sigma_Y\|_F^2\} \leq U_0^* \alpha(p) p.$$

Assumption 2(i) imposes some condition on the weak componentwise dependence within X and Y and it only needs to hold for some permutation of components of X and Y, as our test statistic is permutation-invariant when the kernel $k(x, y) = f(|x - y|)$. It is worth noting that $\alpha(p)$ may vary w.r.t. p and thus the range of dependence is allowed to grow when p increases. Assumption 2(ii) requires a uniform bound of the componentwise moments of both distributions, which can be relaxed at the expense of lengthy proofs. Assumption 2(iii) requires both $\mathbb{E}\big[|\tilde{X}|^2\big]$ and $\mathbb{E}\big[|\tilde{Y}|^2\big]$ are strictly of order p, which is a mild condition. Finally, Assumption 2(iv) specifies the order of $\|\Sigma_X\|_F^2$ and $\|\Sigma_Y\|_F^2$, which seems reasonable in views of the $\alpha(p)$-dependent assumption. With Assumption 2, we are able to calculate the orders of the quantities involved in our main theorems, which lead to a specific convergence rate of normal approximation and some explicit power results to be stated in the next section. Note that it is not our intention to showcase the convergence rate of normal approximation under the weakest possible assumption, as that is at the expense of very complicated arguments. Assumption 2 is quite reasonable to illustrate the convergence rate in a case of broad interest.

Proposition 19 *Assume that* $\mathbb{E}\left[k^4(Z_1, Z_2)\right] < \infty$ *and* $n/N \to \rho$ *as* $n, m \to \infty$ *for some* $0 < \rho < 1$*. Suppose that Assumptions 1(i)-(ii) and Assumptions 2(i)-(ii) hold, then there exists some* $p_0 = p_0(\tilde{M}, \hat{M}, U^*, L_0, U_0, L_0^*, U_0^*)$*, such that for any* $p \geq p_0$*, it holds under the null that for any* $k = k^{(p)} \in \mathcal{C}$*,*

$$\sup_{x \in \mathbb{R}} \left| \mathbb{P}\left(T_{n,m,p}^k \leq x\right) - \Phi(x) \right| \leq C(\tilde{M}, \hat{M}, U^*, L_0, U_0, L_0^*, U_0^*) \left(\frac{1}{N} + \frac{\alpha(p)}{p}\right)^{1/5}, \tag{19}$$

where $\tilde{M}, \hat{M}$ *are defined in Assumption 1 and* $U^*, L_0, U_0, L_0^*, U_0^*$ *are defined in Assumption 2.*

Proposition 19 provides a uniform explicit rate of convergence for a class of kernels and for X and Y with weakly dependent components. In fact, the rate of convergence is determined only by N, p, and the parameters $\tilde{M}, \hat{M}$ from Assumption 1, as well as

$\alpha(p), U^*, L_0, U_0, L_0^*, U_0^*$ from Assumption 2. One implication of Proposition 19 is that, the empirical distribution of the proposed test statistic can be accurately approximated by the standard Gaussian distribution only when both N and p diverge to infinity, though no constraint is required regarding the divergence rate between N and p. Another implication is that, the dependence within X and Y is allowed to grow as p increases, but at the sacrifice of the accuracy of normal approximation. When the dependence within X and Y gets stronger, accurate normal approximation can only be obtained with larger p. This theoretical phenomenon is consistent with our empirical finding in Section 4.

The main theoretical tool we use to obtain the rate of convergence is the Berry-Esseen bound for martingale [Haeusler (1988)], as also used in Gao et al. (2021). One important difference between Gao et al. (2021) and our work is that the denominator of our test statistic is estimated over the pooled sample, and its leading term is a combination of several two-sample U-statistics, and the tools provided in Gao et al. (2021) are not sufficient for our theory. To this end, we generalize the moment inequality for the one-sample U-statistic to the two-sample U-statistic. Furthermore, Berry-Esseen bound obtained here is valid for a general kernel, and the rate of convergence can be explicitly derived under some mild conditions as shown in Proposition 19.

3.4 Power Analysis

Next we look into the power behavior of the studentized test statistic. In the following theorem, we can show that the power of the proposed test is asymptotically one under some conditions.

Theorem 20 *Assume that* $\mathbb{E}\left[k^4(Z_1, Z_2)\right] < \infty$ *and* $n/N = \rho + O(1/N^s)$ *as* $n, m \to \infty$ *for some* $0 < \rho < 1$ *and* $s > 0$. *If for some constant* $0 < \tau \leq 1$, *it holds that*

$$\frac{\left|\mathcal{E}^k(X,Y)\right|^{2+2\tau}}{N^\tau \left(\mathcal{V}_k^2(Z)\right)^{1+\tau}} \to 0, \qquad \frac{\mathbb{E}\left[k^2(Z_1,Z_2)\right]}{N^s \mathcal{V}_k^2(Z)} \to 0, \tag{20}$$

$$\frac{\mathbb{E}\left[\left|\tilde{k}(X_1,X_2)\right|^{2+2\tau} + \left|\tilde{k}(X_1,Y_1)\right|^{2+2\tau} + \left|\tilde{k}(Y_1,Y_2)\right|^{2+2\tau}\right]}{N^\tau \left(\mathcal{V}_k^2(Z)\right)^2} \to 0, \tag{21}$$

$$\frac{N\left(\mathcal{E}^k(X,Y)\right)^2}{\mathbb{E}\left[\left(h^k(X_1,X_2,Y_1,Y_2)\right)^2\right]} \to \infty, \qquad \frac{N\mathcal{E}^k(X,Y)}{\sqrt{\mathcal{V}_k^2(Z)}} \to \infty, \tag{22}$$

where $\tilde{k}$ *denotes the centered version of* k. *Then for any* $C > 0$, *we have* $\mathbb{P}\left(T_{n,m,p}^k > C\right) \to 1$ *as* $n, m, p \to \infty$.

Theorem 20 gives the conditions under which the power of the test can be asymptotically one for a general kernel. Note that conditions (20) and (21) are introduced in Proposition 15 to ensure the ratio-consistency of the pooled-sample estimate $\mathcal{V}_{n,m}^{k*}(X,Y)$. In the proof, we show that when $\frac{N\left(\mathcal{E}^k(X,Y)\right)^2}{\mathbb{E}\left[\left(h^k(X_1,X_2,Y_1,Y_2)\right)^2\right]} \to \infty$, the sample estimate $\mathcal{E}_{n,m}^k(X,Y)$ closely approximates its population counterpart $\mathcal{E}^k(X,Y)$ and the asymptotic divergence of $T_{n,m,p}^k(X,Y)$

is equivalent to $\frac{\mathcal{E}^k(X,Y)}{\sqrt{c_{n,m}\mathcal{V}_k^2(Z)}}$ diverging to infinity as N, p increases. It is then not difficult to see that the asymptotic power one of the proposed test can be achieved under the condition $\frac{N\mathcal{E}^k(X,Y)}{\sqrt{\mathcal{V}_k^2(Z)}} \to \infty$.

The conditions presented in Theorem 20 are sufficient but may not be necessary due to the technical arguments we employed. Nevertheless, Theorem 20 can provide us some interesting insights of the regimes where our proposed test has nontrivial power. Below we shall discuss multiple scenarios based on the leading terms of $\mathcal{E}^k(X,Y)$ and $\mathcal{V}_k^2(Z)$. For the sake of readability, we only present the results when $\alpha(p) = O(1)$ (i.e. fixed) and $s = 1$ and leave the general results when $\alpha(p) = p^{\delta_0}$ with $0 \le \delta_0 < 1$ and $s > 0$ in online appendices.

Assumption 3 *For any fixed p and $X = (x_1, \ldots, x_p)^\top$ and $Y = (y_1, \ldots, y_p)^\top$, assume that there exists some positive constants $L_1, U_1 < \infty$, such that*

$$L_1 p \le \max\{|\Delta|^2, \left|\mathbb{E}\big[|\tilde{X}_1|^2\big] - \mathbb{E}\big[|\tilde{Y}_1|^2\big]\right|\} \le U_1 p.$$

Assumption 3 focuses on the scenario when at least one of $|\Delta|^2$ and $|\mathbb{E}\big[|\tilde{X}_1|^2\big] - \mathbb{E}\big[|\tilde{Y}_1|^2\big]|$ is strictly of order p. It holds under Assumption 2(ii) that $|\Delta|^2 = \sum_{j=1}^p (\mathbb{E}\big[x_j\big] - \mathbb{E}\big[y_j\big])^2 \le 2\sum_{j=1}^p (\mathbb{E}\big[x_j\big]^2 + \mathbb{E}\big[y_j\big]^2) \le 4U^* p$ and $\left|\mathbb{E}\big[|\tilde{X}_1|^2\big] - \mathbb{E}\big[|\tilde{Y}_1|^2\big]\right| = \left|\sum_{j=1}^p (\mathrm{Var}(x_j) - \mathrm{Var}(y_j))\right| \le \sum_{j=1}^p (\mathrm{Var}(x_j) + \mathrm{Var}(y_j)) \le 2U^* p$, then Assumption 3 implies that the differences in componentwise mean or variance attain the highest possible order.

Proposition 21 *Suppose that Assumption 1(i)-(ii), Assumption 2(i)-(iii) and Assumption 3 hold, and additionally, for any $k = k^{(p)} \in \mathcal{C}$, assume that $\mathbb{E}\left[k^4(Z_1, Z_2)\right] < \infty$ and $n/N = \rho + O(1/N)$ as $n, m \to \infty$ for some $0 < \rho < 1$. When there exists some positive constant $L^* < \infty$, such that*

$$\left|2f(A_0^{XY}) - f(A_0^X) - f(A_0^Y)\right| \ge L^* |f(A_0^{XY})|, \tag{23}$$

then it holds that $\mathbb{P}\left(T_{n,m,p}^k > C\right) \to 1$ as $n, m, p \to \infty$.

The condition (23) requires that the leading term of $\left|2f(A_0^{XY}) - f(A_0^X) - f(A_0^Y)\right|$ can be lower bounded by $|f(A_0^{XY})|$ up to a multiplicative constant, which is a mild condition and can be satisfied by many kernel functions; see Section 3.5 for its verification of the Gaussian kernel and the online supplement for the verifications of the L_2 norm and the Laplacian kernel. Under the assumptions in Proposition 21, $2f(A_0^{XY}) - f(A_0^X) - f(A_0^Y)$ is the leading term of $\mathcal{E}^k(X,Y)$ and $(2f(A_0^{XY}) - f(A_0^X) - f(A_0^Y))^2$ is that of $\mathcal{V}_k^2(Z)$. It follows that $(\mathcal{E}^k(X,Y))^2$ and $\mathcal{V}_k^2(Z)$ are of the same order, and both of them dominate $\mathbb{E}\big[k^2(Z_1, Z_2)\big]$ and $\mathbb{E}\big[(h^k(X_1, X_2, Y_1, Y_2))^2\big]$. Additionally, $(\mathcal{V}_k^2(Z))^2$ dominated the numerator of condition (21). Consequently, all the conditions in Theorem 20 are naturally satisfied and the nontrivial power is obtained with no constraints on the order of p relative to N, which seems reasonable in view of significant differences in either the means and/or the sum of marginal variances. In comparison, Zhu and Shao (2021) obtained the asymptotic power one result for the MMD permutation test under the HDLSS and HDMSS settings only. The asymptotic power function for the studentized test proposed in Chakraborty and Zhang

(2021) is also derived only under the HDLSS setting. Additional comparison with Zhu and Shao (2021) under the special case when X and Y have either identical means or identical covariance matrices are discussed for the Gaussian kernel in Section 3.5; see Remark 26.

Next, we further investigate the scenarios where the differences in marginal mean or variance are weaker.

Assumption 4 *For any fixed p and $X = (x_1, \ldots, x_p)^\top$ and $Y = (y_1, \ldots, y_p)^\top$, assume that there exists some positive constants $L_2, U_2 < \infty$ and $0 \le \delta_1 < 1$, $-1 \le \delta_2 < 1$, such that $L_2\alpha(p)p \le \|\rho\Sigma_X + (1-\rho)\Sigma_Y\|_F^2 \le U_2\alpha(p)p$, and*

$$L_2 p^{\delta_1} \le |\Delta|^2 \le U_2 p^{\delta_1} \quad \text{and} \qquad L_2 p^{(1+\delta_2)/2} \le \left|\mathbb{E}\big[|\tilde{X}_1|^2\big] - \mathbb{E}\big[|\tilde{Y}_1|^2\big]\right| \le U_2 p^{(1+\delta_2)/2}.$$

Assumption 4 considers the case where the orders of both $|\Delta|^2$ and $\left|\mathbb{E}\big[|\tilde{X}_1|^2\big] - \mathbb{E}\big[|\tilde{Y}_1|^2\big]\right|$ are strictly smaller than p but are no smaller than a constant. In this case, the differences in the marginal mean and variance still dominate those in higher moments as long as $\max\{\delta_1, \delta_2\} > 0$, resulting in high power under certain rate constraints on p. The condition $L_2\alpha(p)p \le \|\rho\Sigma_X + (1-\rho)\Sigma_Y\|_F^2 \le U_2\alpha(p)p$ is mild under Assumption 2(iv).

Proposition 22 *Suppose that Assumption 1(i)-(iii) and Assumption 2(i)-(iv) hold, Assumption 4 holds with $\delta_1 \ne \delta_2$ and $\max\{\delta_1, \delta_2\} > 0$, and additionally, for any $k = k^{(p)} \in \mathcal{C}$, assume that $\mathbb{E}\left[k^4(Z_1, Z_2)\right] < \infty$ and $n/N = \rho + O(1/N)$ as $n, m \to \infty$ for some $0 < \rho < 1$, then it holds that $\mathbb{P}\left(T^k_{n,m,p} > C\right) \to 1$ when $n, m, p \to \infty$ as long as $p = o\left(N^{1/(2-2\max\{\delta_1,\delta_2,1/2\})}\right)$.*

When the leading difference between the distributions of X and Y lies in marginal mean or variance, the $\mathcal{E}^k(X, Y)$ is of order $|f_0(A_0^{XY})|p^{\max\{\delta_1,\delta_2\}-1}$, while $\mathcal{V}_k^2(Z)$ is of order $f_0^2(Z_0^{XY})p^{2\max\{\delta_1,\delta_2,1/2\}-2}$. When $\max\{\delta_1, \delta_2\} > 1/2$, $(\mathcal{E}^k(X, Y))^2$ has the same order as of $\mathcal{V}_k^2(Z)$, thus condition (21)-(22) are naturally satisfied. To have $\frac{\mathbb{E}\left[k^2(Z_1,Z_2)\right]}{N\mathcal{V}_k^2(Z)} \to 0$, we need additional constraint between N and p since $\mathcal{V}_k^2(Z)$ is dominated by $\mathbb{E}\big[k^2(Z_1, Z_2)\big]$. For the case that $\max\{\delta_1, \delta_2\} \le 1/2$, the order of $\mathcal{V}_k^2(Z)$ becomes $f_0^2(A_0^{XY})p^{-1}$ but that of $\mathcal{E}^k(Z_1, Z_2)$ remains unchanged. Hence $(\mathcal{E}_k^2(X, Y))^2$ is no longer capable of dominating $\mathcal{V}_k^2(Z)$ and $\mathbb{E}\big[(h^k(X_1, X_2, Y_1, Y_2))^2\big]$, which leads to the constraint $p = o(N)$ to make condition (22) hold.

Intuitively, as the disparities in marginal mean and variance between X and Y weakens to the point $\max(\delta_1, \delta_2) \le 1/2$, our proposed test has nontrivial power only when the growth rate of p is strictly smaller than that of N. When $\delta_1 = \delta_2$, similar power results can be attained given a specific kernel function following some lengthy analysis, but we exclude this case for simplicity.

Next we investigate the scenario when the differences in the marginal mean and variance between X and Y further diminish.

Assumption 5 *For any fixed p and $X = (x_1, \ldots, x_p)^\top$ and $Y = (y_1, \ldots, y_p)^\top$, assume that $|\Delta| = 0$ and there exists some positive constants $L_3, U_3 < \infty$ and $0 \le \delta_3, \delta_4 < 1$, such that*

$$L_3 p^{\delta_3/2} \le \left|\mathbb{E}\big[|\tilde{X}_1|^2\big] - \mathbb{E}\big[|\tilde{Y}_1|^2\big]\right| \le U_3 p^{\delta_3/2} \quad \text{and} \quad L_3 p^{\delta_4/2} \le \|\Sigma_X - \Sigma_Y\|_F \le U_3 p^{\delta_4/2}.$$

Assumption 5 targets at the case when X and Y have the identical mean, and their leading disparities fall within the covariances. Note that with $\alpha(p) = O(1)$ and Assumption 2(iv), the order of $\|\Sigma_X - \Sigma_Y\|_F^2$ won't exceed p, and under Assumption 5 we set it to be p^{δ_4}, where $\delta_4 \in (0, 1)$.

Proposition 23 *Suppose that Assumption 1(i)-(iii) and Assumption 2(i)-(iv) hold, Assumption 5 holds with $\delta_3 \neq \delta_4$ and $\max\{\delta_3, \delta_4\} > 0$, and additionally, for any $k = k^{(p)} \in \mathcal{C}$, assume that $\mathbb{E}\left[k^4(Z_1, Z_2)\right] < \infty$ and $n/N = \rho + O(1/N)$ as $n, m \to \infty$ for some $0 < \rho < 1$, then it holds that $\mathbb{P}\left(T^k_{n,m,p} > C\right) \to 1$ when $n, m, p \to \infty$ as long as $p = o\left(N^{1/(3-2\max\{\delta_3,\delta_4\})}\right)$.*

Under Assumption 5, the order of $\mathcal{E}^k(X, Y)$ decreases to $|f_0(A_0^{XY})|p^{\max\{\delta_3,\delta_4\}-2}$, as compared to the second scenario (under Assumption 4). Recall that $\mathcal{E}^k(X, Y)$ characterizes the disparity between the distributions X and Y, then it is not surprising that its order decreases as the leading disparities between X and Y move to some higher moment quantities. In this case, $\mathcal{E}^k(X, Y)$ may not dominate $\mathbb{E}\big[(h^k(X_1, X_2, Y_1, Y_2))^2\big]$, and furthermore, $\mathcal{E}^k(X, Y)$ is dominated by $\mathcal{V}_k^2(Z)$, whose order stays at $f_0^2(A_0^{XY})p^{-1}$. Therefore, additional constraints on p are required to satisfy condition (22).

Zhu and Shao (2021) showed that in the HDMSS setting, when $|\Delta|^2 = o(\sqrt{p}/N)$, $\left|\mathbb{E}\big[|\tilde{X}_1|^2\big] - \mathbb{E}\big[|\tilde{Y}_1|^2\big]\right| = o(\sqrt{p}/N)$ and $\|\Sigma_X - \Sigma_Y\|_F = o(\sqrt{p})$, MMD permutation test has trivial power. In the special case $\delta_3 = 0, \delta_4 \in (0, 1)$, it is easy to see that both the condition in Zhu and Shao (2021) and our Assumption 5 can be satisfied for different sets of (p, N). The resulting power phenomenon is strikingly different with the MMD permutation test being powerless and our studentized test being power one asymptotically. This difference is not a contradiction but is mainly attributed to the different regimes, since HDMSS setting implies $p \gg N$ whereas our Proposition 23 requires $p \ll N$. This is an example that shows that even for the same alternative, the order of p relative to N can play an important role in determining the power behavior.

Finally, we look into the scenario when X and Y have identical means and covariance matrices to complete the discussions in this section. With $\mathrm{cum}(\cdot)$ denoting the cumulant, we propose the following assumption.

Assumption 6 *For any fixed p and $X = (x_1, \ldots, x_p)^\top$ and $Y = (y_1, \ldots, y_p)^\top$, assume that $|\Delta| = \|\Sigma_X - \Sigma_Y\|_F = 0$ and there exists some positive constants $L_4, U_4 < \infty$ and $0 \leq \delta_5, \delta_6, \delta_7 < 1$, such that*

$$
\begin{aligned}
&L_4 p^{\delta_5} \leq \sum_{j_1,j_2,j_3=1}^{p} \left(cum(\tilde{x}_{1j_1}, \tilde{x}_{1j_2}, \tilde{x}_{1j_3}) - cum(\tilde{y}_{1j_1}, \tilde{y}_{1j_2}, \tilde{y}_{1j_3})\right)^2 \leq U_4 p^{\delta_5},\\
&L_4 p^{\delta_6} \leq \sum_{j_2=1}^{p} \left(\sum_{j_1=1}^{p} \{Cov(\tilde{x}^2_{1j_1} - \sigma_{X,j_1}, \tilde{x}_{1j_2}) - Cov(\tilde{y}^2_{1j_1} - \sigma_{Y,j_1}, \tilde{y}_{1j_2})\} \right)^2 \leq U_4 p^{\delta_6},\\
&L_4 p^{2\delta_7} \leq \left(\sum_{j_1,j_2=1}^{p} \{Cov(\tilde{x}^2_{1j_1} - \sigma_{X,j_1}, \tilde{x}^2_{1j_2} - \sigma_{X,j_2}) - Cov(\tilde{y}^2_{1j_1} - \sigma_{Y,j_1}, \tilde{y}^2_{1j_2} - \sigma_{Y,j_2})\} \right)^2 \leq U_4 p^{2\delta_7}.
\end{aligned}
$$

Proposition 24 *Suppose that Assumption 1(i)-(iv) and Assumption 2(i)-(iv) hold, Assumption 6 holds with $1 + \max\{\delta_5, \delta_6\} \neq 2\delta_7$, $\max\{\delta_5, \delta_6, \delta_7\} > 0$, and additionally, for*

any $k = k^{(p)} \in \mathcal{C}$*, assume that* $\mathbb{E}\left[k^4(Z_1, Z_2)\right] < \infty$ *and* $n/N = \rho + O(1/N)$ *as* $n, m \to \infty$ *for some* $0 < \rho < 1$*, then it holds that* $\mathbb{P}\left(T^k_{n,m,p} > C\right) \to 1$ *when* $n, m, p \to \infty$ *as long as* $p = o\left(N^{1/(7-2\max\{1+\max\{\delta_5,\delta_6\},2\delta_7\})}\right)$.

Proposition 24 implies that, when $\delta_7 < (1 + \max\{\delta_5, \delta_6\})/2$, nontrivial power against the alternative is obtained when $p = o\left(N^{1/(5-2(\delta_5 \vee \delta_6))}\right)$ and otherwise the corresponding regime is $p = o\left(N^{1/(7-4\delta_7)}\right)$. In fact, the order of $\mathcal{V}_k^2(Z)$ remains $f_0^2(A_0^{XY})p^{-1}$ while that of $\mathcal{E}^k(X, Y)$ drops to $|f_0(A_0^{XY})|p^{\max\{\delta_5,\delta_6,2\delta_7-1\}-3}$, and following some similar arguments as in the previous scenario, we obtain the constraint between N and p for this case.

To summarize, Proposition 21-Proposition 23 jointly investigate the cases when the discrepancy between two distributions is dominated by their differences in the mean and/or covariance matrices, which correspond to S1 in Section 4 of Yan and Zhang (2023). Proposition 24 corresponds to the scenario where two distributions have identical first and second moments, and the difference lies in the third and/or fourth moments/cumulants. This scenario corresponds to S2 with $\ell = 3$ in Section 4 of Yan and Zhang (2023). The latter authors provided a comprehensive description of when their test has trivial power, nontrivial power and asymptotic power one based on non-null CLT obtained. In general, we feel it is difficult to directly compare the power results in Yan and Zhang (2023) with ours due to the different settings and regimes we explored. In particular, we mainly focus on the regime $p \ll N^{\omega_1}$ for some $\omega_1 > 0$ as stated in Proposition 22- Proposition 24, whereas Yan and Zhang (2023) focus on the regime where $N \ll p^{\omega_2}$ for some $\omega_2 \geq 1/2$. The two regimes may have overlap (i.e., the intersection is nonzero), their power one results and ours complement each other, and both contribute to the understanding of the space of alternatives for which the MMD-based test has high power.

As revealed by the four propositions above, our test is powerful against a wide range of alternatives, including the differences in means, variances, covariances and high-order features associated with the distributions.

The intuition behind all these propositions is that, the disparities that fall within lower moments between X and Y are easier to be detected by our proposed test. When the leading differences move to higher moment quantities, stricter constraints between N and p are required to make the test powerful. This phenomenon is consistent with that found by Yan and Zhang (2023), who provided an asymptotic exact power analysis and revealed a delicate interplay between the detectable moment discrepancy and the dimension-and-sample orders (see Table 1 therein).

3.5 An Illustrative Example with the Gaussian Kernel

As shown in Table 1, a special case covered by our setup is the Gaussian kernel multiplied by -1, that is, $k(x, y) = -\exp\left(-|x - y|^2/(2\gamma^2)\right)$, where γ is a pre-specified tuning parameter. We note that many technical assumptions and theoretical results are presented in the previous sections, which may be difficult to digest. In this section, we use the Gaussian kernel as a special example to demonstrate the verification of Assumption 1 and condition (23) in the previous section.

We define $D = [0, \infty)$ and $D_0 = \left[\mathbb{E}[|Z_1 - Z_2|^2]^{1/2}, \mathbb{E}[|X_1 - Y_1|^2]^{1/2}\right]$. For each fixed p and the tuning parameter γ that depends on p, we consider the Gaussian kernel $k^{(p)}(x, y) =$

$-\exp\left(-\frac{|x-y|^2}{2\gamma^2}\right)$. Here, different choices of γ^2 lead to different Gaussian kernels, and we restrict our interest to $k^{(p)}$ with specific γ, that is,

$$\mathcal{C} = \mathcal{C}(\ell, u) := \left\{ k^{(p)} : \text{for each } p, \frac{\mathbb{E}\left[|X_1 - Y_1|^2\right]^{1/2}}{u} \leq \gamma \leq \frac{\mathbb{E}\left[|Z_1 - Z_2|^2\right]^{1/2}}{\ell} \right\}, \tag{24}$$

where $0 < \ell, u < \infty$ are some specified constants such that $\mathcal{C}(\ell, u)$ is well defined. Note that $\mathcal{C}$ is a set of Gaussian kernel sequences with growing p, for each $k^{(p)} \in \mathcal{C}$, we define $f^{(p)}(s) = -\exp\left(-\frac{s^2}{2\gamma^2}\right)$ to be the unique smooth function associated with $k^{(p)}$ and for simplicity, we drop the superscript hereafter. With the explicit expression of f, we obtain the derivatives of f up to the 7th order, that is

$$f_0(s) = -\exp\left(-\frac{s^2}{2\gamma^2}\right), \qquad f_1(s) = \frac{s}{\gamma^2}\exp\left(-\frac{s^2}{2\gamma^2}\right),$$
$$f_2(s) = \left(\frac{1}{\gamma^2} - \frac{s^2}{\gamma^4}\right)\exp\left(-\frac{s^2}{2\gamma^2}\right), \qquad f_3(s) = \left(-\frac{3s}{\gamma^4} + \frac{s^3}{\gamma^6}\right)\exp\left(-\frac{s^2}{2\gamma^2}\right),$$
$$f_4(s) = \left(-\frac{3}{\gamma^4} + \frac{6s^2}{\gamma^6} - \frac{s^4}{\gamma^8}\right)\exp\left(-\frac{s^2}{2\gamma^2}\right), \quad f_5(s) = \left(\frac{15s}{\gamma^6} - \frac{10s^3}{\gamma^8} + \frac{s^5}{\gamma^{10}}\right)\exp\left(-\frac{s^2}{2\gamma^2}\right),$$

$$f_6(s) = \left(\frac{15}{\gamma^6} - \frac{45s^2}{\gamma^8} + \frac{15s^4}{\gamma^{10}} - \frac{s^6}{\gamma^{12}}\right)\exp\left(-\frac{s^2}{2\gamma^2}\right),$$
$$f_7(s) = \left(-\frac{105s}{\gamma^8} + \frac{105s^3}{\gamma^{10}} - \frac{21s^5}{\gamma^{12}} + \frac{s^7}{\gamma^{14}}\right)\exp\left(-\frac{s^2}{2\gamma^2}\right).$$

It follows from the Taylor theorem with the Lagrange form of remainder that $\mathcal{C}$ satisfies Assumption 1(i)

To verify Assumption 1(ii), we note that

$$\sup_{s\in D}|f_1(s)| = \sup_{s\geq 0}\left|\frac{s}{\gamma^2}\exp\left(-\frac{s^2}{\gamma^2}\right)\right| = \frac{1}{\gamma}\sup_{t\geq 0}|t\exp(-t^2/2)| = \frac{0.607}{\gamma}.$$

It follows from similar steps that

$$\sup_{s\in D}|f_2(s)| = \frac{1}{\gamma^2}, \quad \sup_{s\in D}|f_3(s)| = \frac{1.38}{\gamma^3}, \quad \sup_{s\in D}|f_4(s)| = \frac{3}{\gamma^4},$$
$$\sup_{s\in D}|f_5(s)| = \frac{5.783}{\gamma^5}, \quad \sup_{s\in D}|f_6(s)| = \frac{15}{\gamma^6}, \quad \sup_{s\in D}|f_7(s)| = \frac{35.539}{\gamma^7}.$$

Recall that $\mathcal{C}$ has restrictions on the tuning parameter γ associated with $k^{(p)}$ such that $\mathbb{E}\left[|X_1 - Y_1|^2\right]^{1/2}/u \leq \gamma \leq \mathbb{E}\left[|Z_1 - Z_2|^2\right]^{1/2}/\ell$, then it holds that $\sup_{s_0\in D_0}\frac{s_0}{\gamma} \leq u$, thus $\max_{1\leq i\leq 7}\sup_{s\in D}|f_i(s)|\cdot|s_0^i| \leq \tilde{M}|f_0(s_0)|$ holds for any $s_0 \in D_0$ when

$$\tilde{M} = \max\{0.607u\exp(u^2/2), u^2\exp(u^2/2), 1.38u^3\exp(u^2/2), 3u^4\exp(u^2/2),$$

$$5.783u^5\exp(u^2/2), 15u^6\exp(u^2/2), 35.539u^7\exp(u^2/2)\},$$

which completes the verification of Assumption 1(ii).

As for Assumption 1(iii), it follows from direct computation that $|f_0(s)| = \exp\left(-\frac{1}{2}\left(\frac{s}{\gamma}\right)^2\right)$, $\frac{1}{2}|f_1(s)s| = \frac{1}{2}\left(\frac{s}{\gamma}\right)^2\exp\left(-\frac{1}{2}\left(\frac{s}{\gamma}\right)^2\right)$, and

$$|-\frac{1}{8}f_1(s)s + \frac{1}{8}f_2(s)s^2| = \frac{1}{8}\left(\frac{s}{\gamma}\right)^4\exp\left(-\frac{1}{2}\left(\frac{s}{\gamma}\right)^2\right).$$

Again, it follows from the definition of $\mathcal{C}$ that $\inf_{s_0\in D_0}(s_0/\gamma) \geq \ell$. Therefore, Assumption 1(iii) holds with $\hat{M} = \max\{2/\ell^2, 8/\ell^4\}$.

Lastly, we look into Assumption 1(iv). Note that $|f_0(s)| = \exp\left(-\frac{1}{2}\left(\frac{s}{\gamma}\right)^2\right)$, and

$$|\frac{1}{16}f_1(s)s - \frac{1}{16}f_2(s)s^2 + \frac{1}{48}f_3(s)s^3| = \frac{1}{48}\left(\frac{s}{\gamma}\right)^6\exp\left(-\frac{1}{2}\left(\frac{s}{\gamma}\right)^2\right),$$
$$|-\frac{5}{128}f_1(s)s + \frac{5}{128}f_2(s)s^2 - \frac{1}{64}f_3(s)s^3 + \frac{1}{384}f_4(s)s^4| = \frac{1}{384}\left(\frac{s}{\gamma}\right)^8\exp\left(-\frac{1}{2}\left(\frac{s}{\gamma}\right)^2\right),$$

then it is trivial Assumption 1(iv) is satisfied with $\hat{M} = \max\{48/\ell^6, 384/\ell^8\}$.

To conclude, we present the results in the following proposition.

Proposition 25 *Let $\mathcal{C}$ denote the set of Gaussian kernel sequences as defined in Equation (24), it holds that $\mathcal{C}$ satisfies Assumption 1.*

Next, we verify condition 23 in Proposition 21, that is

$$\left|2f(A_0^{XY}) - f(A_0^X) - f(A_0^Y)\right| \geq L^*|f(A_0^{XY})|,$$

Again, we restrict the analysis to the set $\mathcal{C}$. For each $k^{(p)} \in \mathcal{C}$, we define $f(s) = -\exp\left(-\frac{s^2}{2\gamma^2}\right)$, and it follows from the definition of $\mathcal{C}$ that $\mathbb{E}\left[|X_1 - Y_1|^2\right]^{1/2}/u \leq \gamma \leq \mathbb{E}\left[|Z_1 - Z_2|^2\right]^{1/2}/\ell$. Note that

$$\left(2f(A_0^{XY}) - f(A_0^X) - f(A_0^Y)\right)/f(A_0^{XY}) = 2 - \exp\left(-\frac{A^X - A^{XY}}{2\gamma^2}\right) - \exp\left(-\frac{A^Y - A^{XY}}{2\gamma^2}\right),$$

where

$$\exp\left(-\frac{A^X - A^{XY}}{2\gamma^2}\right) = \exp\left(\frac{|\Delta|^2}{2\gamma^2}\right)\exp\left(-\frac{\mathbb{E}\left[|\tilde{X}_1|^2\right] - \mathbb{E}\left[|\tilde{Y}_1|^2\right]}{2\gamma^2}\right),$$
$$\exp\left(-\frac{A^Y - A^{XY}}{2\gamma^2}\right) = \exp\left(\frac{|\Delta|^2}{2\gamma^2}\right)\exp\left(\frac{\mathbb{E}\left[|\tilde{X}_1|^2\right] - \mathbb{E}\left[|\tilde{Y}_1|^2\right]}{2\gamma^2}\right),$$

then it follows from the fact $\exp(s) > 1$ and $\exp(s) + \exp(-s) > 2$ for any $s > 0$ that

$$\begin{aligned} & \exp\left(-\frac{A^X - A^{XY}}{2\gamma^2}\right) + \exp\left(-\frac{A^Y - A^{XY}}{2\gamma^2}\right) \\ = \; & \exp\left(\frac{|\Delta|^2}{2\gamma^2}\right)\left(\exp\left(-\frac{\mathbb{E}\big[|\tilde{X}_1|^2\big] - \mathbb{E}\big[|\tilde{Y}_1|^2\big]}{2\gamma^2}\right) + \exp\left(\frac{\mathbb{E}\big[|\tilde{X}_1|^2\big] - \mathbb{E}\big[|\tilde{Y}_1|^2\big]}{2\gamma^2}\right)\right) \\ > \; & 2. \end{aligned}$$

Consequently, the condition is naturally satisfied with

$$\begin{aligned} & L^* \\ = \; & \inf_{A_0^{XY}/u \le \gamma \le A_0/\ell} \left\{ \exp\left(\frac{|\Delta|^2}{2\gamma^2}\right)\left(\exp\left(-\frac{\mathbb{E}\big[|\tilde{X}_1|^2\big] - \mathbb{E}\big[|\tilde{Y}_1|^2\big]}{2\gamma^2}\right) + \exp\left(\frac{\mathbb{E}\big[|\tilde{X}_1|^2\big] - \mathbb{E}\big[|\tilde{Y}_1|^2\big]}{2\gamma^2}\right)\right) - 2 \right\}. \end{aligned}$$

Suppose that Assumption 2 holds, and we have $\mathbb{E}[k^4(Z_1, Z_2)] < \infty$ for each $k = k^{(p)} \in \mathcal{C}$, we summarize a few regimes where the asymptotic power of the Gaussian kernels in $\mathcal{C}$ is one. We want to emphasize that Table 2 only includes a few special cases, whereas our proposed test is guaranteed to obtain full power asymptotically across a wider range of regimes.

$\|\Delta\|^2$	$O_s(p)$		$O_s(p^{1/2})$	0
$\left\|\mathbb{E}\big[\|\tilde{X}_1\|^2\big] - \mathbb{E}\big[\|\tilde{Y}_1\|^2\big]\right\|$		$O_s(p)$	$O_s(p^{1/2})$	$O_s(p^{1/4})$
$\|\rho\Sigma_X + (1-\rho)\Sigma_Y\|_F^2$			$O_s(\alpha(p)p)$	
$\|\Sigma_X - \Sigma_Y\|_F$				$O_s(p^{1/4})$
Regime	$N, p \to \infty$	$N, p \to \infty$	$p = o(N)$	$p = o(N^{1/2})$

Table 2: Selected regimes where the power of the Gaussian kernels is asymptotically one.

In the following remark, we compare the sufficient conditions for asymptotically power one derived in Zhu and Shao (2021) and in this article under the special case when X and Y have either identical means or identical covariance matrices.

Remark 26 *Both Zhu and Shao (2021) and our work aim to test for the distributional discrepancy, that is, to test for $H_0 : X =^d Y$ versus $H_1 : X \neq^d Y$. The discussion in Zhu and Shao (2021) is limited to MMD with a user-specified kernel $\hat{k}$ of the following expression*

$$\hat{k}(X, Y) = \varphi\left(\frac{1}{p}\sum_{j=1}^{p} \psi(x_j, y_j)\right),$$

where $\psi \geq 0$ and φ has continuous second order derivative on $(0, \infty)$. It is trivial that the Gaussian kernel is covered by the set of $\hat{k}$.

Both the high dimensional low sample size setting (HDLSS) when n, m are fixed but $p \to \infty$ and the high dimensional medium sample size (HDMSS) setting when $p \to \infty$ and $n := n(p) \to \infty$ are investigated in Zhu and Shao (2021), but here we only focus on HDMSS

setting. It is shown that the permutation test in Zhu and Shao (2021) is consistent within the following consistency space $\mathcal{H}$:

$$\mathcal{H}_c = \{(X, Y) : \ 2\varphi(e_{XY}) \neq \varphi(e_X) + \varphi(e_Y)\},$$

where

$$e_X = \lim_{p\to\infty} \mathbb{E}[\bar{\psi}(X, X')], \qquad e_Y = \lim_{p\to\infty} \mathbb{E}[\bar{\psi}(Y, Y')], \qquad e_{XY} = \lim_{p\to\infty} \mathbb{E}[\bar{\psi}(X, Y)],$$

are all assumed to exist, with X', Y' being independent copies of X, Y and $\bar{\psi}$ denoting the average distance over components

$$\bar{\psi}(Z_i, Z_j) = \frac{1}{p}\sum_{s=1}^{p} \psi(z_{is}, z_{js}).$$

For the Gaussian kernels, $\mathcal{H}_c$ can be further characterized as

$$\mathcal{H}_c = \left\{ (X, Y) : \ \sum_{j=1}^{p} (\mathbb{E}[x_j] - \mathbb{E}[y_j])^2 = o(p), \ |\sum_{j=1}^{p} (\, Var(x_j) - \, Var(y_j))\,| = o(p) \right\}^c.$$

As shown in Theorem 3.5 of Zhu and Shao (2021), it holds under the HDMSS that

$$\lim_{p\to\infty} \mathbb{P}_{\mathcal{H}_c}\left(\mathcal{E}^k_{n,m}(X, Y) > c\right) = 1$$

for any $c \in \{Q_{\hat{R},1-\alpha}, Q_{\tilde{R},1-\alpha}\}$, where $Q_{\hat{R},1-\alpha}$ is the critical value obtained from $(n+m)!$ permutations with $\hat{R}$ being the randomization distribution of $\mathcal{E}^k_{n,m}(X, Y)$, and $Q_{\hat{R},1-\alpha}$ is the critical value obtained from a fixed number S of permutations with $\tilde{R}$ being the counterpart of $\hat{R}$ with only S permutations. In other words, it is shown in Zhu and Shao (2021) that the asymptotic power of their permutation-based test is one within $\mathcal{H}_c$.

Now we are able to compare the sufficient conditions for consistent power derived in Zhu and Shao (2021) and ours when X and Y have either identical means or identical covariance matrices. When $\mu_X = \mu_Y$, the condition in Zhu and Shao (2021) reduces to $|\sum_{j=1}^{p} (\, Var(x_j) - \, Var(y_j))\,| \gtrsim p$, which is equivalent to $\left|\mathbb{E}[|\tilde{X}_1|^2] - \mathbb{E}[|\tilde{Y}_1|^2]\right| = O_s(p)$ in our article. When $\Sigma_X = \Sigma_Y$, the condition in Zhu and Shao (2021) is simplified as $\sum_{j=1}^{p} (\mathbb{E}[x_j] - \mathbb{E}[y_j])^2 \gtrsim p$, which exactly matches the condition $|\mu_X - \mu_Y|^2 = O_s(p)$ in our work. In summary, the sufficient conditions derived in both articles are equivalent under the special case that X and Y have either the same means or the same covariance matrices. However, it is worth mentioning that both works require some additional regularity conditions, which are not enumerated here.

4. Numerical Experiments

In this section, we carry out several simulation studies to examine the finite-sample performance of the proposed test statistics and compare with permutation-based counterparts.

4.1 Normal Approximation Accuracy

We generate two independent random samples $\mathbf{X} = \{X_1, \ldots, X_n\}$ and $\mathbf{Y} = \{Y_1, \ldots, Y_m\}$ as follows.

Example 1 *Generate independent samples:* $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(0, \Sigma)$, $Y_1, \ldots, Y_m \overset{iid}{\sim} \mathcal{N}(0, \Sigma)$, *where* $\Sigma = (\sigma_{ij}) \in \mathbb{R}^{p \times p}$ *with* $\sigma_{ij} = \rho^{|i-j|}$ *and* $\rho = 0.5$. *We set the sample size ratio* $m/n = 1$, *and consider the setting that* $n \in \{25, 50, 100, 200, 400\}$ *and the data dimensionality* $p \in \{25, 50, 100, 200\}$. *As for the kernel* k, *we consider the* L_2*-norm* $k_{L_2}(x, y) = |x - y|$, *the Gaussian kernel multiplied by -1, that is,,* $k_G(x, y) = -\exp\left(-|x - y|^2/(2\gamma^2)\right)$ *with* $\gamma^2 = Median\{|X_{i_1} - X_{i_2}|^2, |X_i - Y_j|^2, |Y_{j_1} - Y_{j_2}|^2\}$, *and the Laplacian kernel multiplied by -1, that is,,* $k_L(x, y) = -\exp\left(-|x - y|/\gamma\right)$ *with* $\gamma = Median\{|X_{i_1} - X_{i_2}|, |X_i - Y_j|, |Y_{j_1} - Y_{j_2}|\}$. *The median heuristic is a popular way of choosing* γ*; see Gretton et al. (2012).*

Throughout the simulations, our proposed methods are averaged over 5000 Monte Carlo replications, whereas those of the permutation tests are averaged over 1000 Monte Carlo replications owing to the high computational cost; see subsequent section. 300 permutations are conducted for each replication. Given the 5000 replicates of the studentized test statistic $T^k_{n,m,p}$, we plot the kernel density estimates (KDE) for the three kernels and the standard normal density function for each combination of sample size and dimension, see Figure 1. Each row of Figure 1 corresponds to a fixed pair of (n, m) whereas each column represents a fixed choice of p.

As shown in Figure 1, when p is fixed, the improvement of normal approximation accuracy is minimal as N increases. However, we do observe significant improvement in the accuracy as p grows for fixed $n = m$. The three kernels correspond to very similar empirical distributions suggesting the insensitivity to the kernel choice in terms of size. It is worth noticing that normal approximation is already quite accurate when the sample size and the dimensionality are relatively small, say $N = 50$ and $p = 100$, and higher accuracy is achieved with larger N and p. Such requirements of N and p are usually not demanding in real-world applications, which shows the applicability of the proposed test.

Additional simulation results regarding normal approximation accuracy can be found in online appendices, including the results when the sample sizes n, m are unequal with the difference beyond a constant, and the Kolmogorov-Smirnov distance as well the Wasserstein distance between the standard normal distribution and the empirical distribution of our proposed test statistic under the null. The overall finding from the unbalanced setting is qualitatively similar to what we observe here.

4.2 Empirical Size

Under the significance level $\alpha = 0.05$, we reject the null hypothesis if $T^k_{n,m,p} > \Phi(1-\alpha)$. As a comparison, we also consider the permutation test based on the sample MMD studied in Zhu and Shao (2021). For our test statistic, we consider the three kernels introduced in Example 1, while for the permutation test, we additionally consider the L_1-norm $k_{L_1}(x, y) = |x-y|_1$, which is advocated in Zhu and Shao (2021).

In this section, we consider a simulated example that mimics Example 4.1 in Zhu and Shao (2021).

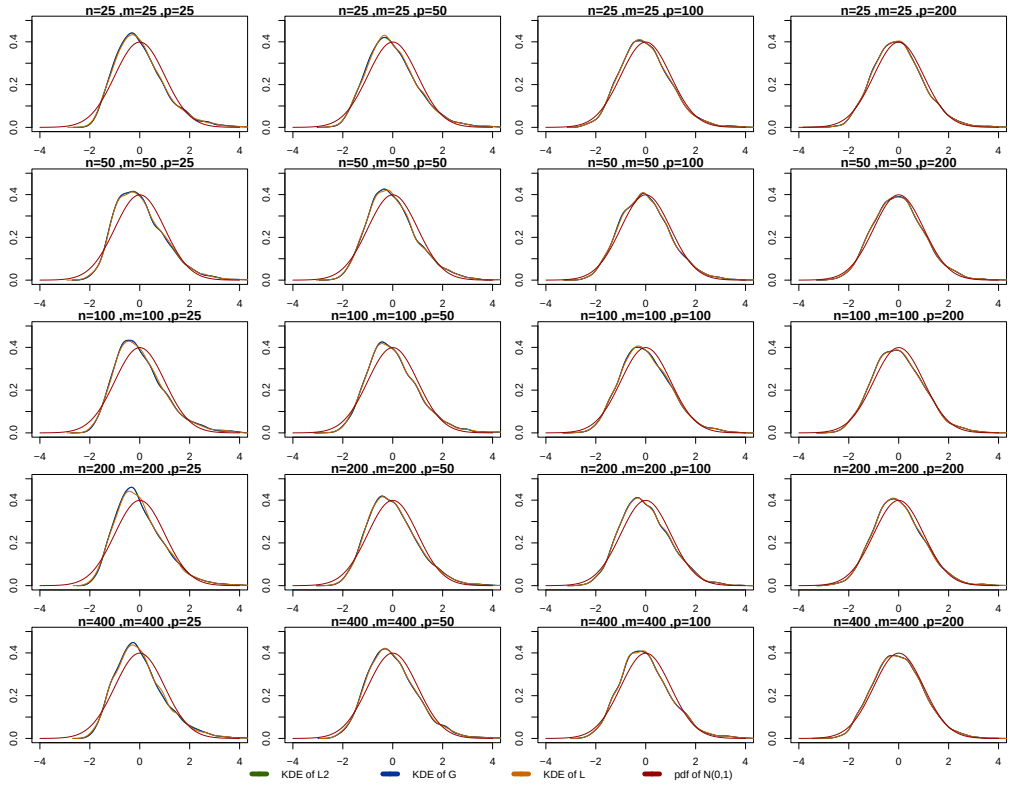

Figure 1: Kernel density estimates of the studentized test statistic $T^k_{n,m,p}$ with different kernels for Example 1 when $m/n = 1$. The four columns correspond to different p's and the five rows correspond to different pairs of (n, m).

Example 2 *Generate independent samples:* $X_1, \ldots, X_n \overset{iid}{\sim} (V^{1/2}\Sigma V^{1/2})^{1/2} Z_X$ *and* $Y_1, \ldots, Y_m \overset{iid}{\sim} (V^{1/2}\Sigma V^{1/2})^{1/2} Z_Y$, *where* $\Sigma = (\sigma_{ij}) \in \mathbb{R}^{p\times p}$ *with* $\sigma_{ij} = \rho^{|i-j|}$ *and* $\rho \in \{0.4, 0.7\}$. *We consider the setting that* $(n, m) \in \{(25, 25), (50, 50), (50, 100), (100, 100), (200, 200)\}$ *and* $p \in \{50, 100\}$. *Here,* V *is a diagonal matrix with* $V_{ii}^{1/2} = 1$ *or uniformly drawn from the interval* $(1, 5)$. Z_X, Z_Y *are iid copies of* Z *drawn from the following two distributions:*

(i) $Z = (z_1, \ldots, z_p)$ *with* $z_1, \ldots, z_p \overset{iid}{\sim} \mathcal{N}(0, 1)$.

(ii) $Z = (z_1 - 1, \ldots, z_p - 1)$ *with* $z_1, \ldots, z_p \overset{iid}{\sim}$ *Exponential*(1).

As reported in Table 3, our test exhibit some mild size distortion due to the inaccuracy of normal approximation in finite sample.

However, even when $n = m = 25$ and $p = 50$, the rejection rate is only slightly higher than the nominal level 5%, suggesting that our test is practically useful for small sample and moderate dimensional setting. The size distortion tends to increase when the componentwise

dependence gets stronger, which matches the theory developed in previous sections; see Proposition 19. As we expect, the permutation tests exhibit accurate size, but at the cost of computation. Comparing all three kernels, it seems that no kernel dominates the other in terms of size accuracy.

(n,m)	p	ρ	V	Example 2(i)							Example 2(ii)						
				Proposed			Permutation				Proposed			Permutation			
				L_2	G	L	L_2	G	L	L_1	L_2	G	L	L_2	G	L	L_1
(25,25)	50	0.4	Id	6.98	6.92	6.80	4.60	4.60	4.20	3.80	5.66	5.80	5.80	4.50	4.60	4.50	4.90
			Unif	6.54	6.54	6.50	4.60	4.60	4.10	4.50	6.72	6.72	6.60	5.40	5.40	5.50	5.50
		0.7	Id	5.86	5.80	6.00	5.20	5.00	5.00	5.20	6.38	6.26	6.00	6.00	5.80	4.90	4.80
			Unif	6.50	6.46	6.10	5.00	4.90	4.90	4.80	6.98	6.78	6.66	4.90	5.00	5.20	5.10
	100	0.4	Id	5.52	5.54	5.52	4.90	4.90	4.60	4.20	6.12	6.04	6.08	5.00	5.00	5.10	5.70
			Unif	6.42	6.36	6.34	5.50	5.40	5.70	5.10	6.06	6.04	6.08	4.90	4.80	5.40	4.60
		0.7	Id	6.90	6.86	7.10	5.80	5.70	5.10	4.90	6.52	6.56	6.46	5.80	5.70	5.40	5.50
			Unif	6.60	6.60	6.74	5.30	5.40	5.50	6.30	6.62	6.62	6.70	5.50	5.40	5.50	5.40
(50,50)	50	0.4	Id	6.06	6.06	6.00	5.80	5.60	5.50	5.20	6.36	6.44	6.64	4.00	4.10	4.20	4.10
			Unif	5.86	5.94	5.82	4.00	3.90	3.80	4.00	6.46	6.30	6.26	4.00	3.90	3.80	3.80
		0.7	Id	7.36	7.28	7.00	5.50	5.30	5.50	5.20	7.00	6.88	6.96	5.10	5.30	5.10	5.00
			Unif	6.56	6.66	6.68	5.90	6.30	6.30	6.50	6.46	6.36	6.22	5.10	5.30	5.40	5.20
	100	0.4	Id	5.98	6.00	6.18	5.10	5.00	5.30	5.30	5.82	5.72	5.74	4.30	4.30	4.10	4.00
			Unif	6.42	6.40	6.42	4.90	4.90	4.70	4.60	5.62	5.50	5.50	5.40	5.50	5.10	5.30
		0.7	Id	6.90	6.86	6.74	5.30	5.30	5.60	4.80	6.38	6.46	6.68	5.60	5.80	6.00	6.10
			Unif	6.68	6.68	6.50	6.30	6.00	5.50	5.70	6.10	6.12	6.10	4.20	4.20	4.60	4.60
(50,100)	50	0.4	Id	6.34	6.26	6.16	5.10	5.10	4.80	4.90	6.38	6.38	6.04	5.40	5.20	5.00	5.00
			Unif	6.66	6.76	6.80	6.00	6.00	5.50	6.00	6.08	6.20	5.92	4.90	5.10	5.40	4.10
		0.7	Id	6.44	6.40	6.38	5.20	5.10	4.70	5.10	6.50	6.36	6.30	4.30	4.10	4.10	4.00
			Unif	6.74	6.62	6.30	5.20	5.20	4.90	5.10	5.98	5.94	6.02	3.30	3.30	3.50	3.80
	100	0.4	Id	5.84	5.86	5.92	5.30	5.30	5.00	4.60	5.86	5.84	5.70	4.90	4.80	5.00	4.30
			Unif	6.14	6.22	6.24	5.50	5.60	6.20	6.80	6.04	6.06	6.14	5.70	5.70	5.50	5.10
		0.7	Id	6.50	6.54	6.30	5.40	5.30	4.20	5.30	6.16	6.22	6.08	5.30	5.30	5.30	5.20
			Unif	6.48	6.38	6.42	4.70	4.40	4.70	4.10	6.64	6.72	6.72	5.20	5.40	5.40	5.30
(100,100)	50	0.4	Id	6.24	6.14	6.16	5.30	5.30	5.80	5.10	5.54	5.56	5.68	4.80	4.50	4.90	4.40
			Unif	6.12	6.08	6.06	5.50	5.50	5.20	4.90	7.04	7.04	6.90	6.40	6.40	6.40	5.60
		0.7	Id	6.56	6.66	6.44	4.00	4.10	4.00	4.40	6.00	6.02	6.14	5.10	5.10	5.00	5.00
			Unif	6.40	6.36	6.42	5.10	5.00	5.30	4.90	6.66	6.60	6.70	4.50	4.30	4.70	4.40
	100	0.4	Id	5.98	5.94	6.08	5.50	5.30	5.60	5.00	5.92	5.86	5.76	4.40	4.60	4.20	3.90
			Unif	6.24	6.22	6.10	5.00	5.10	5.50	5.70	6.04	6.00	5.94	4.10	4.10	3.50	3.40
		0.7	Id	6.14	6.08	6.06	4.90	4.90	5.20	5.00	6.46	6.36	6.44	4.50	4.40	4.40	4.40
			Unif	6.60	6.66	6.54	4.50	4.30	4.30	4.70	6.76	6.88	6.82	5.30	5.40	5.50	4.60
(200,200)	50	0.4	Id	5.60	5.60	5.42	4.30	4.20	4.30	4.30	6.12	6.22	6.28	6.20	6.60	5.70	5.60
			Unif	6.38	6.52	6.32	5.20	5.20	4.80	5.10	6.14	6.24	6.26	5.00	5.20	5.00	5.20
		0.7	Id	6.78	6.84	6.52	4.50	4.40	4.10	4.10	6.66	6.60	6.56	5.60	5.50	4.80	5.40
			Unif	6.56	6.68	6.64	4.10	4.00	4.40	4.20	6.94	6.88	6.74	6.10	5.90	6.10	5.80
	100	0.4	Id	6.18	6.18	6.06	5.30	5.30	4.80	5.00	5.60	5.60	5.66	5.40	5.40	5.50	5.80
			Unif	6.74	6.74	6.66	5.10	5.10	5.00	4.70	5.50	5.52	5.56	4.30	4.50	4.10	3.90
		0.7	Id	7.22	7.24	7.10	5.70	5.70	5.90	5.80	6.42	6.50	6.16	4.40	4.30	3.60	3.40
			Unif	6.56	6.50	6.32	5.70	5.40	5.40	4.40	5.56	5.54	5.76	3.90	4.10	4.30	4.60

Table 3: Size comparison for Example 2. All the empirical sizes are reported in percentage.

4.3 Power Behavior

Next we investigate the power behavior. The simulated example is adopted from the setting of Example 4.2 in Zhu and Shao (2021). We present the simulation results for the alternative

of mean difference in this section, and defer to online appendices the results when two distributions differ in the covariance matrices.

Example 3 *Generate independent samples:* $X_1, \ldots, X_n \overset{iid}{\sim} (V^{1/2}\Sigma V^{1/2})^{1/2} Z_X$ *and* $Y_1, \ldots, Y_m \overset{iid}{\sim} (V^{1/2}\Sigma V^{1/2})^{1/2} Z_Y + \big(0.15 \times \mathbf{1}_{\beta p}, \mathbf{0}_{(1-\beta)p}\big)$, *where* Σ, Z_X, Z_Y *are defined the same as in Example 2 and* V *is chosen as the identity matrix. Here, we fix* $\rho = 0.5$, *consider* $(n, m) \in \{(25, 25), (50, 50), (100, 100), (200, 200)\}$, $p \in \{50, 100\}$ *and* $\beta \in \{0, 0.1, \ldots, 1\}$.

We plot the size-adjusted power curves against β in Figure 2. Note that we only made critical value adjustment in calculating the size-adjusted power for our method, as there is little distortion for permutation-based test. As can be seen from Figure 2, when there is a mean shift, our test statistic and permutation-based counterpart have almost identical power for all kernels. The use of L_1 norm brings some power gain in some cases. As N increases, we do observe a significant improvement in power, regardless of the choice of p, which is consistent with our intuition. When p increases from 50 to 100, the power increases noticeably for fixed (n, m), as the alternative gets farther away from the null.

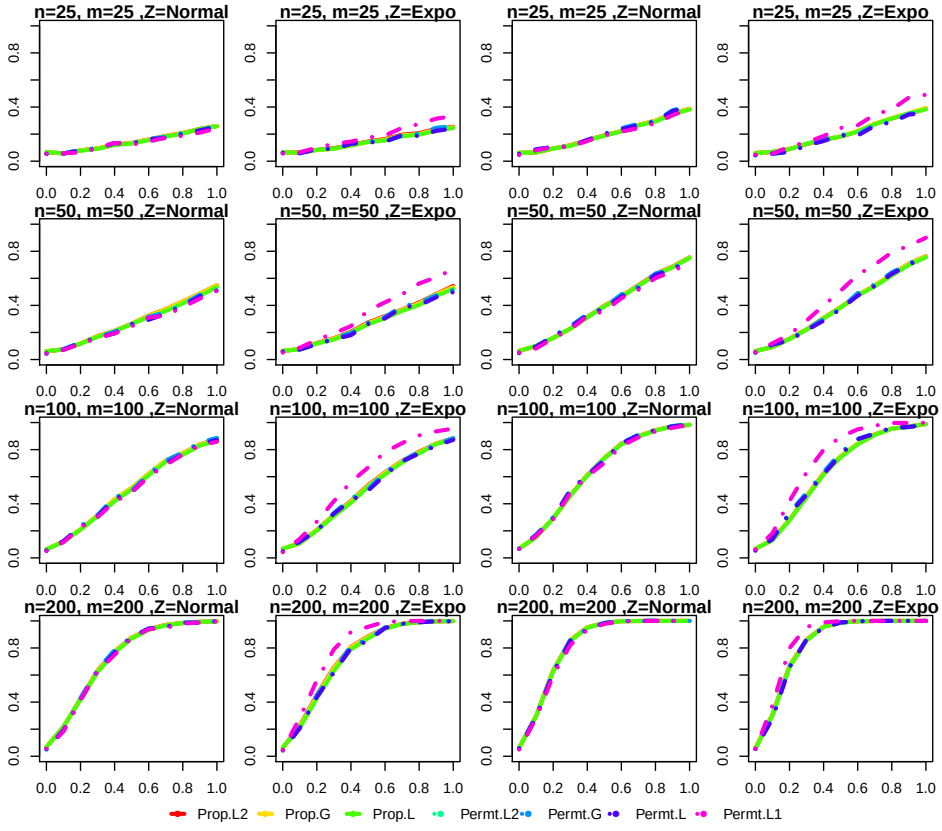

Figure 2: Size-adjusted Power Curves for Example 3. The first two columns correspond to $p = 50$ while the last two columns correspond to $p = 100$.

n	m	p	Proposed			Permutation			
			L_2	G	L	L_2	G	L	L_1
25	25	50	0.12	0.21	0.19	20.87	22.55	22.64	22.04
		100	0.15	0.25	0.25	33.32	35.20	35.12	36.03
50	50	50	0.36	0.69	0.68	82.95	94.39	95.05	86.64
		100	0.58	1.12	1.12	148.75	158.73	159.32	152.56
50	100	50	0.59	1.11	1.12	155.40	171.97	173.00	168.64
		100	0.89	1.74	1.74	263.34	278.31	276.89	283.76
100	100	50	1.35	2.57	2.59	328.13	369.62	367.83	332.62
		100	2.31	4.30	4.28	576.19	625.81	626.61	601.24
200	200	50	5.04	9.98	10.05	1229.59	1380.77	1387.88	1241.33
		100	6.69	13.43	13.24	1879.47	1989.71	1991.56	2001.98

Table 4: Computational cost under multiple settings. All the numerical results are counted in seconds.

4.4 Computational Cost

One of the major advantage of our proposed method over the permutation test in Zhu and Shao (2021) is the computational efficiency. In this section we compare the computational cost per 100 replications of our method with that of the permutation test under multiple settings; see Table 4. The number of permutations per replication is set to be 300. As shown in Table 4, it is obvious that our method is much more computationally efficient compared to permutation-based counterpart, which makes up for the slight size distortion of our test under the null.

5. Discussion

In this paper, we have obtained the central limit theorems for studentized sample MMD and derived the explicit rates of convergence under the null hypothesis of equal distributions when both sample size and dimensionality are diverging. Furthermore, we have also developed a general power theory for the studentized sample MMD and demonstrated its ability of detecting the difference in distributions. Our proof is built on the argument in Gao et al. (2021) but we need to develop some new theoretical tools owing to the fact that we are dealing with a general class of kernels and a two sample U-statistic with high-dimensional observations. In particular, the pooled sample estimate $\mathcal{V}_{n,m}^{k*}(X,Y)$ is proposed and its ratio-consistency as an estimator of $\mathcal{V}_k^2(Z)$ is shown using a newly developed moment inequality for the multi-sample U-statistics. To deal with a general class of kernels, we also develop new bounds for the moments of the in-sample and between-sample distances. From a practical viewpoint, our proposed test is simple and easy to implement with much less computational cost compared to permutation test in Zhu and Shao (2021). Finite sample simulations suggest that the size is quite accurate and there is no power loss compared to the permutation-based counterpart.

As a part of future work, we expect our theory to be useful to the study of asymptotic behavior of sample HSIC in high-dimension, and to test for distributional change in a sequence of high-dimensional data. We leave these topics for future investigation.

Acknowledgments

We would like to thank Changbo Zhu for some early discussion on this topic. We would also like to thank the action editor and two reviewers for their constructive comments, which have substantially improved the presentation of the paper. The research is partially supported by NSF-DMS 16-07489. The online supplement is available at `https://arxiv.org/abs/2109.14913`.

References

Theodore W Anderson and Donald A Darling. Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes. *The Annals of Mathematical Statistics*, 23(2): 193–212, 1952.

Michael Arbel, Danica J Sutherland, Mikołaj Bińkowski, and Arthur Gretton. On gradient regularizers for mmd gans. *arXiv preprint arXiv:1805.11565*, 2018.

Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018.

Karsten M Borgwardt, Arthur Gretton, Malte J Rasch, Hans-Peter Kriegel, Bernhard Schölkopf, and Alex J Smola. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics*, 22(14):e49–e57, 2006.

Shubhadeep Chakraborty and Xianyang Zhang. A new framework for distance and kernel-based metrics in high dimensions. *Electronic Journal of Statistics, to appear, arXiv preprint arXiv:1909.13469*, 2021.

Song Xi Chen and Ying-Li Qin. A two-sample test for high-dimensional data with applications to gene-set testing. *The Annals of Statistics*, 38(2):808–835, 2010.

Harald Cramér. On the composition of elementary errors: First paper: Mathematical deductions. *Scandinavian Actuarial Journal*, 1928(1):13–74, 1928.

Gintare Karolina Dziugaite, Daniel M Roy, and Zoubin Ghahramani. Training generative neural networks via maximum mean discrepancy optimization. *arXiv preprint arXiv:1505.03906*, 2015.

Jerome H Friedman and Lawrence C Rafsky. Multivariate generalizations of the wald-wolfowitz and smirnov two-sample tests. *The Annals of Statistics*, 7(4):697–717, 1979.

Kenji Fukumizu, Arthur Gretton, Xiaohai Sun, and Bernhard Schölkopf. Kernel measures of conditional dependence. In *Proceedings of the 20th International Conference on Neural Information Processing Systems*, volume 20, pages 489–496. Curran Associates Inc., 2007. ISBN 9781605603520.

Lan Gao, Yingying Fan, Jinchi Lv, and Qiman Shao. Asymptotic distributions of high-dimensional distance correlation inference. *The Annals of Statistics*, 49(4):1999–2020, 2021.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *Proceedings of the 16th International Conference on Algorithmic Learning Theory*, pages 63–77. Springer, 2005.

Arthur Gretton, Kenji Fukumizu, Choon Hui Teo, Le Song, Bernhard Schölkopf, Alexander J Smola, et al. A kernel statistical test of independence. In *Advances in Neural Information Processing Systems*, volume 20, pages 585–592. Curran Associates, Inc., 2007.

Arthur Gretton, Kenji Fukumizu, Zaid Harchaoui, and Bharath K Sriperumbudur. A fast, consistent kernel two-sample test. In *Advances in Neural Information Processing Systems*, volume 23, pages 673–681, 2009.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(1):723–773, 2012.

Erich Haeusler. On the rate of convergence in the central limit theorem for martingales with discrete and continuous time. *The Annals of Probability*, 16(1):275–299, 1988.

Qiyang Han and Yandi Shen. Generalized kernel distance covariance in high dimensions: non-null clts and power universality. *arXiv preprint arXiv:2106.07725*, 2021.

Cheng Huang and Xiaoming Huo. An efficient and distribution-free two-sample test based on energy statistics and random projections. *arXiv preprint arXiv:1707.04602*, 2017.

Jared D Huling and Simon Mak. Energy balancing of covariate distributions. *arXiv preprint arXiv:2004.13962*, 2020.

Andrey Kolmogorov. Sulla determinazione empirica di una lgge di distribuzione. *Inst. Ital. Attuari, Giorn.*, 4:83–91, 1933.

Henry B Mann and Donald R Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 18(1):50–60, 1947.

David S Matteson and Nicholas A James. A nonparametric approach for multiple change point analysis of multivariate data. *Journal of the American Statistical Association*, 109 (505):334–345, 2014.

Dino Sejdinovic, Bharath Sriperumbudur, Arthur Gretton, and Kenji Fukumizu. Equivalence of distance-based and rkhs-based statistics in hypothesis testing. *The Annals of Statistics*, 41(5):2263–2291, 2013.

Nikolai V Smirnov. On the estimation of the discrepancy between empirical curves of distribution for two independent samples. *Bull. Math. Univ. Moscou*, 2(2):3–14, 1939.

Gábor J Székely and Maria L Rizzo. The distance correlation t-test of independence in high dimension. *Journal of Multivariate Analysis*, 117:193–213, 2013a.

Gábor J Székely and Maria L Rizzo. Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference*, 143(8):1249–1272, 2013b.

Gábor J Székely and Maria L Rizzo. Partial distance correlation with methods for dissimilarities. *The Annals of Statistics*, 42(6):2382–2412, 2014.

Gábor J Székely, Maria L Rizzo, et al. Testing for equal distributions in high dimension. *InterStat*, 5(16.10):1249–1272, 2004.

Gabor J Szekely, Maria L Rizzo, et al. Hierarchical clustering via joint between-within distances: Extending ward's minimum variance method. *Journal of Classification*, 22(2): 151–184, 2005.

Gábor J Székely, Maria L Rizzo, and Nail K Bakirov. Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35(6):2769–2794, 2007.

Abraham Wald and Jacob Wolfowitz. On a test whether two samples are from the same population. *The Annals of Mathematical Statistics*, 11(2):147–162, 1940.

Jian Yan and Xianyang Zhang. Kernel two-sample tests in high dimensions: interplay between moment discrepancy and dimension-and-sample orders. *Biometrika*, 110(2):411–430, 2023.

Zhixun Zhao, Hui Peng, Xiaocai Zhang, Yi Zheng, Fang Chen, Liang Fang, and Jinyan Li. Identification of lung cancer gene markers through kernel maximum mean discrepancy and information entropy. *BMC Medical Genomics*, 12(8):1–10, 2019.

Changbo Zhu and Xiaofeng Shao. Interpoint distance based two sample tests in high dimension. *Bernoulli*, 27(2):1189–1211, 2021.

Changbo Zhu, Xianyang Zhang, Shun Yao, and Xiaofeng Shao. Distance-based and rkhs-based dependence metrics in high dimension. *The Annals of Statistics*, 48(6):3366–3394, 2020.

Xiaofeng Zhu, Kim-Han Thung, Ehsan Adeli, Yu Zhang, and Dinggang Shen. Maximum mean discrepancy based multiple kernel learning for incomplete multimodality neuroimaging data. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 72–80. Springer, 2017.