Statistica Sinica Preprint No: SS-2023-0299	
Title	Testing for the Equality of Distributions in High Dimension
Manuscript ID	SS-2023-0299
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202023.0299
Complete List of Authors	Xu Li, Gongming Shi and Baoxue Zhang
Corresponding Authors	Baoxue Zhang
E-mails	zhangbaoxue@cueb.edu.cn

Statistica Sinica

TESTING FOR THE EQUALITY OF DISTRIBUTIONS IN HIGH DIMENTION

Xu Li, Gongming Shi, Baoxue Zhang

Capital University of Economics and Business, Beijing 100070, China.

Shanxi Normal University, Taiyuan 030000, China.

Abstract: In this paper, we propose a new homogeneous test for two high-dimensional random vectors. Our test is built on a new measure, the so-called characteristic distance, which can completely characterize the homogeneity of two distributions. The newly proposed metric has some desirable properties, for example, it possesses a clear and intuitive probabilistic interpretation, and can be used to address the high-dimensional distance inference. Theoretically, the limiting behaviors under the conventional fixed dimension and high-dimensional distance inference are thoroughly investigated. Simulation studies and real data analysis are presented to illustrate the finite-sample performance of the proposed test statistic.

Key words and phrases: Characteristic distance, High Dimensionality, Test of Homogeneity, U-statistic, Permutation procedure.

1. Introduction

Over the past decades, the problem of assessing the homogeneity of two high-dimensional data has often appeared in various research areas. In some specific situations, the researchers want to measure whether two samples are generated from the same population. One example can be found in clustering analysis, where before constructing the groups, it is recommended to verify whether it is really necessary. For this, a formal test of the null hypothesis that two samples have been drawn from the single population is essential to prevent misjudgment.

The research on measuring and testing the homogeneity of two populations has a long history. For univariate data, the most traditional tools are the Smirnov maximum deviation test (Smirnoff, 1939) and Wald Wolfowitz runs (Wald, 1940), whose multivariate and multidimensional extensions have been widely discussed, examples include the Darling (1957), Bickel (1969), Friedman (1979), among others. Recent years see another attempt to address the homogeneity between two random vectors by using the empirical characteristic function. Fernández (2008) based on the empirical characteristic function proposed a class of tests for the two sample problems. Liu (2015, 2019) exploited jackknife empirical likelihood with empirical characteristic function to study the homogeneity test between two random vectors. Lee (2020) considered another scenario where the observations were subject to measurement error. Despite the aforementioned existing results, the methods still have some defects. For example, the test statistics are often implicit and therefore cannot be used as a general measure of homogeneity.

Another line of work is based on distance and kernel-based tests for equality of distributions. Székely (2004) and Baringhaus (2004) independently introduced energy distance to test whether two populations are identically distributed. Gretton (2012) proposed the maximum mean difference to measure the difference between two probability distributions. Sejdinovic (2013) through the corresponding relationship between positive definite kernels and semimetrics of negative type, established the equivalent relationship between generalized energy distance and maximum mean difference. For a further review of the kernel-based two-sample test, we refer the readers to Harchaoui (2013).

Homogeneity test of two populations based on the energy distance and maximum mean difference, due to its attractive characteristics, the zero metric completely describes the homogeneity between two random vectors, and nonparametric method has the characteristics of fast calculation speed, not only been widely used in multivariable cases, but also further developed in high-dimensional statistical inference. Biswas (2014) modified the energy distance and proposed a two sample test in the case of high dimension low sample size. Li (2018) proposed the location and scale difference test statistics, and studied its asymptotic distribution when the dimensionality diverges with sample size fixed. Further, Chakraborty (2021) pointed out that the energy distance based on ordinary euclidean distance could not detect the difference between distributions in high-dimensional cases, therefore, it improved the energy distance and proposed a new high-dimensional two sample test. Zhu (2021) based on the permutation test, investigated the properties of energy distance and maximum mean difference in the case of high dimensional and low/medium sample size setting. Gao (2023) proposed to investigate the properties of the sample maximum mean difference under the situation of high-dimensionality and developed a new studentized test statistic. In addition, Kim (2020) generalized the energy distance through projection-averaging to obtain a robust test for the high dimensional two-sample problem. Pan (2018) introduced a new metric, called the ball divergence, to detect the difference between two probability measures in separable Banach spaces. Sarkar (2020) introduced a new high-dimensional two sample homogeneity test by modifying some popular graph-based two-sample tests. Qiu (2021) proposed a robust high-

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0299

dimensional heterogeneity test for two populations based on the Cramér-von Mises test. Liu (2022) generalized the classic Wilcoxon–Mann–Whitney test through using pairwise distances of all observations. Other recent developments include Ramdas (2015), Zhao (2015), Zhou (2017), Sarkar (2018), Mukhopadhyay (2020), Yan (2023), among others.

The tests based on energy distance performs well in detecting the location shift, and the ball divergence enjoys a comparable power in detecting the scale difference. Thus, inspired by Székely's energy distance and Pan's ball divergence, we are very much interested in constructing a homogeneous test for two high-dimensional random vectors, which can not only maintain desirable performance in detecting the location shift, but also be comparable in detecting the scale difference.

To achieve this purpose, a new two-sample test is introduced in the present study. Specifically, we first construct a metric with integration and standardization skills in a high-dimensional vector space, called characteristic distance, to measure and test the homogeneity between two random vectors. This new metric eliminates the divergence of high-dimensional inner products and the periodicity of feature functions, and can be successfully used for high-dimensional distance inference. The characteristic distance is nonnegative and is equal to zero if and only if two populations are identically distributed. Then, an empirical estimator of the characteristic distance is defined and the asymptotic theory of our test is studied systematically, including the consistency and the asymptotic distributions under the null hypothesis and the alternative hypothesis. Simulation studies and real data applications show that the new test is comparable to existing methods and more powerful in many cases. Next, we will list the main contributions.

- **Characteristic distance**: A new metric is proposed to characterize the homogeneity between two vectors, with some attractive features. Firstly, zero distance completely characterizes the homogeneity of two distributions. Secondly, this approach possesses a clear and intuitive probabilistic interpretation. Thirdly, this metric can be used to address the high-dimensional distance inference.

- **Closed form expression**: By building on the relevant theory of moment estimation and U-statistic, our test statistic has a simple closed-form expression.

- **Multivariate limiting distribution**: Based on the asymptotic properties of the general U-statistic, the limiting distribution of the proposed test statistic is discussed when the sample size tends to infinity and the dimension fixed.

- **High-dimensional behavior**: A high-dimension regime is considered where the sample size and the dimension tend to infinity simultaneously. Under this regime, the consistency and the asymptotic distribution are investigated systematically.

The rest of the paper is organized as follows. Section 2 introduces a new metric and defines its corresponding empirical version. In Section 3 and 4, the limiting behavior under the conventional fixed dimension and high-dimensional distance inference are studied, including the strong consistency and the asymptotic distributions under the null hypothesis and the alternative hypothesis. In Section 5, simulation studies and real data analysis are carried out to illustrate the usefulness of our proposed method. A brief discussion is presented in Section 6.

2. Characteristic distance

Suppose $X = (X_1, \cdots, X_p)$, $Y = (Y_1, \cdots, Y_p)$, $p \geq 1$, throughout this paper, we focus on testing whether X and Y are generated from the single population, that is,

$$H_0 : f_X(\mathbf{t}) = f_Y(\mathbf{t}) \qquad versus \qquad H_1 : f_X(\mathbf{t}) \neq f_Y(\mathbf{t}),$$

where $f_X(\mathbf{t})$ and $f_Y(\mathbf{t})$ stand for the characteristic functions of X and Y respectively. Below for ease of argument, we will introduce some notations.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0299

Given a random vector X, the expectation and covariance matrix of X are denoted by μ_X and Σ_X respectively. Similarly, μ_Y and Σ_Y stand for the expectation and covariance matrix with respect to Y. Suppose $\{X_n\}_{n=1}^{\infty}$ is a sequence of random variables and $\{a_n\}_{n=1}^{\infty}$ is a real sequences, we use $X_n = O_P(a_n)$ if, for any $\varepsilon > 0$, there exists $M > 0$ such that $P(|X_n/a_n| > M) < \varepsilon$ for large n. We write $X_n = o_P(a_n)$, if $X_n/a_n \xrightarrow{P} 0$. The symbol $\sum_{i_1,\cdots,i_m}^{*}$ denotes summation over the $m!$ permutations $(i_1, \cdots, i_m)$ of $(1, \cdots, n)$. Let $X \perp\!\!\!\perp Y$ indicate that X is independent of Y, and $X_1 \sim X$ represents that X_1 and X are identically distributed. For any vectors $\mathbf{t} \in R^p$ and $\mathbf{s} \in R^p$, we denote $\langle \mathbf{t}, \mathbf{s} \rangle$ the corresponding inner product, and $\| \cdot \|^2$ stands for the square module of a complex number.

Note if

$$f_X(\mathbf{t}) = f_Y(\mathbf{t}),$$

then

$$\int \|E(e^{i\langle X,\mathbf{t}\rangle}) - E(e^{i\langle Y,\mathbf{t}\rangle})\|^2 w(\mathbf{t}) d\mathbf{t} = 0, \qquad w(\mathbf{t}) \geq 0.$$

So for

$$f_X(\mathbf{t}) = f_Y(\mathbf{t}),$$

we have

$$E\|E(e^{i\langle X,X'\rangle} \mid X') - E(e^{i\langle Y,X'\rangle} \mid X')\|^2 = 0.$$

Further note that the support for X may be asymmetric but $\mathbf{t} \in R^p$, hence, we choose $X' - X''$ to replace X'. In addition, due to the fact that the high-dimensional vector inner product may not converge, to address this issue, we need to further standardize it. Inspired by these views, a new metric — characteristic distance is introduced to measure the homogeneity of two distributions, defined as follows.

Definition 1. (Characteristic distance)

Suppose $X', X'' \overset{i.i.d.}{\sim} X$, $Y', Y'' \overset{i.i.d.}{\sim} Y$, and $X \perp\!\!\!\perp Y$, $\underset{1 \leqslant i \leqslant p}{sup}\ EX_i^2 < \infty$, $\underset{1 \leqslant i \leqslant p}{sup}\ EY_i^2 < \infty$, the characteristic distance of X and Y is defined as

$$
\begin{aligned}
& CD(X,Y) \qquad\qquad (2.1)\\
&= E\left\{\left\| E\left(e^{i\frac{\langle X'', X-X'\rangle}{\sqrt{Var\langle X'', X-X'\rangle}}} \middle| X-X' \right) - E\left(e^{i\frac{\langle Y, X-X'\rangle}{\sqrt{Var\langle X'', X-X'\rangle}}} \middle| X-X' \right) \right\|^2\right\} \\
&+ E\left\{\left\| E\left(e^{i\frac{\langle X, Y-Y'\rangle}{\sqrt{Var\langle Y'', Y-Y'\rangle}}} \middle| Y-Y' \right) - E\left(e^{i\frac{\langle Y'', Y-Y'\rangle}{\sqrt{Var\langle Y'', Y-Y'\rangle}}} \middle| Y-Y' \right) \right\|^2\right\} \\
&:= A + C.
\end{aligned}
$$

Remark 1. As can be seen from (2.1), $CD(X,Y)$ is composed of two parts, and each of which can be used to test equality of distributions, the reason we do this is for fairness and to ensure that when constructing (2.1), the information extracted from X is just as numerous as Y.

The following proposition establishes the characteristic property of $CD(X,Y)$, which is the keystone to our testing procedure.

Proposition 1. *$CD(X,Y)$ is nonnegative and has the characteristic property*

$$CD(X,Y) = 0 \quad if\ and\ only\ if \quad X \sim Y.$$

Next, based on the theory of moment estimates and U-statistic, we will give the empirical form of $CD(X,Y)$. Note that $CD(X,Y)$ contains nuisance parameters $Var\langle X'', X - X'\rangle$ and $Var\langle Y'', Y - Y'\rangle$, therefore, when giving the estimator of $CD(X,Y)$, we need to give the estimates of $Var\langle X'', X-X'\rangle$ and $Var\langle Y'', Y-Y'\rangle$ firstly. In the existing methods, when the sample size and data dimension tend to infinity simultaneously, the estimators of the above parameters can be obtained by estimating the trace of the covariance matrix, but this method is relatively cumbersome. Hence, this paper will use the definition of U-statistic to estimate the unknown parameters.

Definition 2. (Empirical characteristic distance)

Suppose $X_1, \ldots, X_n$ is a sample of size n from a population X, and $Y_1, \ldots, Y_m$ is a sample of size m from a population Y, $X \perp\!\!\!\perp Y$, then the sample statistic is defined by

$$\begin{aligned}
&U_{n,m} \qquad (2.2)\\
&= \frac{1}{\binom{n}{4}\binom{m}{2}} \sum_{j<q<k<k'}^{n} \sum_{l<l'}^{m} \psi_A^s(X_j, X_q, X_k, X_{k'}; Y_l, Y_{l'})\\
&+ \frac{1}{\binom{n}{2}\binom{m}{4}} \sum_{k<k'}^{n} \sum_{j<q<l<l'}^{m} \psi_C^s(X_k, X_{k'}; Y_j, Y_q, Y_l, Y_{l'})\\
&:= A_{n,m} + C_{n,m}.
\end{aligned}$$

That is

$$A_{n,m} = \frac{1}{\binom{n}{4}\binom{m}{2}} \sum_{j<q<k<k'}^{n} \sum_{l<l'}^{m} \psi_A^s(X_j, X_q, X_k, X_{k'}; Y_l, Y_{l'}),$$

and

$$C_{n,m} = \frac{1}{\binom{n}{2}\binom{m}{4}} \sum_{k<k'}^{n} \sum_{j<q<l<l'}^{m} \psi_C^s(X_k, X_{k'}; Y_j, Y_q, Y_l, Y_{l'}),$$

where

$$\begin{aligned}
&\psi_A^s(X_j, X_q, X_k, X_{k'}; Y_l, Y_{l'})\\
&= \frac{1}{4!2!} \sum_{\tau \in \pi(j,q,k,k')} \sum_{\gamma \in \pi(l,l')} \psi_A(X_{\tau(1)}, X_{\tau(2)}, X_{\tau(3)}, X_{\tau(4)}; Y_{\gamma(1)}, Y_{\gamma(2)}),
\end{aligned}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0299

and

$$
\begin{aligned}
&\psi_A(X_j, X_q, X_k, X_{k'}; Y_l, Y_{l'}) \\
&= cos\frac{\langle X_k - X_{k'}, X_j - X_q\rangle}{\sqrt{U_n}} + cos\frac{\langle Y_l - Y_{l'}, X_j - X_q\rangle}{\sqrt{U_n}} \\
&- cos\frac{\langle X_k - Y_{l'}, X_j - X_q\rangle}{\sqrt{U_n}} - cos\frac{\langle X_{k'} - Y_l, X_j - X_q\rangle}{\sqrt{U_n}},
\end{aligned}
$$

$$
U_n = \frac{1}{\binom{n}{3}} \sum_{u<v<s}^{n} \varphi(X_u, X_v, X_s),
$$

$$
\varphi(X_u, X_v, X_s) = \frac{1}{3}(\langle X_u, X_v - X_s\rangle^2 + \langle X_v, X_u - X_s\rangle^2 + \langle X_s, X_v - X_u\rangle^2).
$$

Similarly

$$
\begin{aligned}
&\psi_C^s(X_k, X_{k'}; Y_j, Y_q, Y_l, Y_{l'}) \\
&= \frac{1}{2!4!} \sum_{\tau\in\pi(k,k')} \sum_{\gamma\in\pi(j,q,l,l')} \psi_C(X_{\tau(1)}, X_{\tau(2)}; Y_{\gamma(1)}, Y_{\gamma(2)}, Y_{\gamma(3)}, Y_{\gamma(4)}),
\end{aligned}
$$

and

$$
\begin{aligned}
&\psi_C(X_k, X_{k'}; Y_j, Y_q, Y_l, Y_{l'}) \\
&= cos\frac{\langle X_k - X_{k'}, Y_j - Y_q\rangle}{\sqrt{U_m}} + cos\frac{\langle Y_l - Y_{l'}, Y_j - Y_q\rangle}{\sqrt{U_m}} \\
&- cos\frac{\langle X_k - Y_{l'}, Y_j - Y_q\rangle}{\sqrt{U_m}} - cos\frac{\langle X_{k'} - Y_l, Y_j - Y_q\rangle}{\sqrt{U_m}},
\end{aligned}
$$

$$
U_m = \frac{1}{\binom{m}{3}} \sum_{u<v<s}^{m} \frac{1}{3}(\langle Y_u, Y_v - Y_s\rangle^2 + \langle Y_v, Y_u - Y_s\rangle^2 + \langle Y_s, Y_v - Y_u\rangle^2).
$$

By (2.2), it is easy to establish the asymptotic behavior, however, when we do simulation studies, such a formula is computationally expensive for large-scale datasets, hence we give an equivalent form of it.

Definition 3. (Empirical characteristic distance)

Suppose $X_1, \dots, X_n$ is a sample of size n from a population X, and $Y_1, \dots, Y_m$ is a sample of size m from a population Y, $X \perp\!\!\!\perp Y$, then the sample statistic is defined by

$$T_{n,m} = \widetilde{A}_{n,m} + \widetilde{C}_{n,m}, \tag{2.3}$$

where

$$\begin{aligned} \widetilde{A}_{n,m} = & \frac{1}{n(n-1)(n-2)(n-3)} \sum_{j,q,k,k'}^{*} cos \frac{\langle X_j - X_q, X_k - X_{k'} \rangle}{\sqrt{U_n}} \\ & + \frac{1}{n(n-1)m(m-1)} \sum_{j,q}^{*} \sum_{l,l'}^{*} cos \frac{\langle X_j - X_q, Y_l - Y_{l'} \rangle}{\sqrt{U_n}} \\ & - \frac{2}{n(n-1)(n-2)m} \sum_{j,q,k}^{*} \sum_{l=1}^{m} cos \frac{\langle X_j - X_q, X_k - Y_l \rangle}{\sqrt{U_n}}, \end{aligned}$$

and

$$\begin{aligned} \widetilde{C}_{n,m} = & \frac{1}{m(m-1)(m-2)(m-3)} \sum_{j,q,l,l'}^{*} cos \frac{\langle Y_j - Y_q, Y_k - Y_{k'} \rangle}{\sqrt{U_m}} \\ & + \frac{1}{n(n-1)m(m-1)} \sum_{j,q}^{*} \sum_{k,k'}^{*} cos \frac{\langle Y_j - Y_q, X_k - X_{k'} \rangle}{\sqrt{U_m}} \\ & - \frac{2}{nm(m-1)(m-2)} \sum_{k=1}^{n} \sum_{j,q,l}^{*} cos \frac{\langle Y_j - Y_q, X_k - Y_l \rangle}{\sqrt{U_m}}. \end{aligned}$$

Below with some minor algebraic rearrangements, we will establish the equivalence between the sample statistics (2.2) and (2.3).

Proposition 2. *For any sample size n, m, we have $U_{n,m} = T_{n,m}$.*

3. Main Theorems in multivariate case

In this section, we will investigate the large sample properties of the proposed test statistic under the asymptotic regime where the dimension is fixed while the sample sizes n, m tend to infinity. More specifically, our first step is to establish the consistency of our approach, and then discuss the asymptotic distributions under nul hypothesis and alternative hypothesis. To proceed, we will establish the following theorem firstly.

Theorem 1. *Suppose $n, m \to \infty$, $\underset{1 \leqslant i \leqslant p}{sup} EX_i^2 < \infty$, $\underset{1 \leqslant i \leqslant p}{sup} EY_i^2 < \infty$, then we have*

$$\frac{\langle X_1 - X_2, Z_1^* - Z_2^* \rangle}{\sqrt{U_n}} = \frac{\langle X_1 - X_2, Z_1^* - Z_2^* \rangle}{\sqrt{Var\langle X'', X - X' \rangle}} + O_P\left(\frac{1}{\sqrt{n}}\right),$$

and

$$\frac{\langle Y_1 - Y_2, Z_3^* - Z_4^* \rangle}{\sqrt{U_m}} = \frac{\langle Y_1 - Y_2, Z_3^* - Z_4^* \rangle}{\sqrt{Var\langle Y'', Y - Y' \rangle}} + O_P\left(\frac{1}{\sqrt{m}}\right),$$

where $X_1, X_2, Z_1^*, Z_2^*, Z_3^*, Z_4^*$ are mutually independent, and $X_1, X_2 \sim X$, $Y_1, Y_2 \sim Y$, $Z_1^*, Z_2^*, Z_3^*, Z_4^*$ follow either F_X or F_Y.

Theorem 2. *Suppose* $n, m \to \infty$, $\underset{1\leqslant i\leqslant p}{sup}\ EX_i^2 < \infty$, $\underset{1\leqslant i\leqslant p}{sup}\ EY_i^2 < \infty$,*then*

$$U_{n,m} \xrightarrow{a.s.} CD(X, Y).$$

Theorem 2 illustrates that when the sample size tends to infinity, $U_{n,m}$ is the strongly consistent estimation of $CD(X, Y)$. In addition, $CD(X, Y) = 0$ if and only if X and Y are drawn from the single population, so the statistic we propose can be applied to test the homogeneity of two random vectors.

The next theorem discusses the asymptotic distribution of $U_{n,m}$. Note under the null hypothesis, the statistics $A_{n,m}$ and $C_{n,m}$ are two degenerate U-statistics. Therefore, based on Theorem 1 and the asymptotic theory of nonparametric statistics, we can obtain that the test statistic converges in distribution to a mixture of χ^2 distribution, stated as below.

Theorem 3. *Suppose* $n, m \to \infty$, *and* $\frac{n}{n+m} \to \theta \in [0, 1]$, $\underset{1\leqslant i\leqslant p}{sup}\ EX_i^2 < \infty$, $\underset{1\leqslant i\leqslant p}{sup}\ EY_i^2 < \infty$, *then under the null hypothesis, we have*

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

where

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

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0299

and

$$\begin{aligned}
&\phi_A^{(2,0)}(\boldsymbol{x}, \boldsymbol{x}') \\
&= E\left(cos\frac{\langle X_1 - X_2, \boldsymbol{x} - \boldsymbol{x}'\rangle}{\sqrt{Var\langle X'', X - X'\rangle}}\right) + E\left(cos\frac{\langle X_1 - X_2, Y_1 - Y_2\rangle}{\sqrt{Var\langle X'', X - X'\rangle}}\right) \\
&- E\left(cos\frac{\langle X_1 - X_2, \boldsymbol{x} - Y_2\rangle}{\sqrt{Var\langle X'', X - X'\rangle}}\right) - E\left(cos\frac{\langle X_1 - X_2, \boldsymbol{x}' - Y_1\rangle}{\sqrt{Var\langle X'', X - X'\rangle}}\right),
\end{aligned}$$

$$\begin{aligned}
&\phi_A^{(1,1)}(\boldsymbol{x}, \boldsymbol{y}) \\
&= E\left(cos\frac{\langle X_1 - X_2, \boldsymbol{x} - X\rangle}{\sqrt{Var\langle X'', X - X'\rangle}}\right) + E\left(cos\frac{\langle X_1 - X_2, Y - \boldsymbol{y}\rangle}{\sqrt{Var\langle X'', X - X'\rangle}}\right) \\
&- E\left(cos\frac{\langle X_1 - X_2, \boldsymbol{x} - \boldsymbol{y}\rangle}{\sqrt{Var\langle X'', X - X'\rangle}}\right) - E\left(cos\frac{\langle X_1 - X_2, X - Y\rangle}{\sqrt{Var\langle X'', X - X'\rangle}}\right),
\end{aligned}$$

$$\begin{aligned}
&\phi_A^{(0,2)}(\boldsymbol{y}, \boldsymbol{y}') \\
&= E\left(\frac{cos\langle X_1 - X_2, X_3 - X_4\rangle}{\sqrt{Var\langle X'', X - X'\rangle}}\right) + E\left(\frac{cos\langle X_1 - X_2, \boldsymbol{y} - \boldsymbol{y}'\rangle}{\sqrt{Var\langle X'', X - X'\rangle}}\right) \\
&- E\left(\frac{cos\langle X_1 - X_2, X_3 - \boldsymbol{y}'\rangle}{\sqrt{Var\langle X'', X - X'\rangle}}\right) - E\left(\frac{cos\langle X_1 - X_2, X_4 - \boldsymbol{y}\rangle}{\sqrt{Var\langle X'', X - X'\rangle}}\right).
\end{aligned}$$

Furthermore, the function $Q(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{x}', \boldsymbol{y}')$ *mentioned above has the following spectral decomposition*

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

where λ_k *and* f_k *are the eigenvalues and eigenfunctions of* $Q(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{x}', \boldsymbol{y}')$ *respectively, and*

$$a_k^2(\theta) = (1-\theta)E_X[E_Y f_k(X,Y)]^2, \quad b_k^2(\theta) = \theta E_Y[E_X f_k(X,Y)]^2,$$

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0299

$$Z_{1k}, Z_{2k} \overset{i.i.d.}{\sim} N(0,1), k = 1, 2, \cdots .$$

To exploit characteristic distance for homogeneity testing between two probability distribution, it is very important to determine the significance threshold. However, it can be seen from Theorem 3 that there are some unknown parameters in the asymptotic distribution, thus it is difficult to obtain the significance threshold. Therefore, in our simulation studies, we will use the permutation test procedure to get the empirical significance threshold.

The next theorem establishes the asymptotic normality of the test statistic under the alternative hypothesis.

Theorem 4. *Assume* $n, m \to \infty$*, and* $\frac{n}{n+m} \to \theta \in [0,1]$*,* $\underset{1 \leqslant i \leqslant p}{sup} EX_i^2 < \infty$*,* $\underset{1 \leqslant i \leqslant p}{sup} EY_i^2 < \infty$*, then under the alternative hypothesis, we can obtain*

$$\sqrt{\frac{nm}{n+m}}(U_{n,m} - CD(X,Y)) \xrightarrow{D} N(0, (1-\theta)\delta_{1,0}^2 + \theta\delta_{0,1}^2) + \tilde{\zeta},$$

where

$$g^{(1,0)}(X_k) = \phi_{A,1}^{(1,0)}(X_k) + \phi_{A,2}^{(1,0)}(X_k) + \phi_{A,3}^{(1,0)}(X_k) + \phi_{A,4}^{(1,0)}(X_k) + \phi_{C,1}^{(1,0)}(X_k) + \phi_{C,2}^{(1,0)}(X_k),$$

$$g^{(0,1)}(Y_l) = \phi_{A,1}^{(0,1)}(Y_l) + \phi_{A,2}^{(0,1)}(Y_l) + \phi_{C,1}^{(0,1)}(Y_l) + \phi_{C,2}^{(0,1)}(Y_l) + \phi_{C,3}^{(0,1)}(Y_l) + \phi_{C,4}^{(0,1)}(Y_l),$$

and

$$Var(g^{(1,0)}(X_k)) = \delta_{1,0}^2, \qquad Var(g^{(0,1)}(Y_l)) = \delta_{0,1}^2,$$

$$\tilde{\zeta} = -\sqrt{1-\theta} E H_A(X_1, X_2, X_3, X_4; Y_3, Y_4) - \sqrt{\theta} E H_C(X_3, X_4; Y_1, Y_2, Y_3, Y_4).$$

The limiting distribution of the test under null hypothesis is a mixture of χ^2 distribution, however, it is very different under H_1. That's because in this case, $A_{n,m}$ and $C_{n,m}$ are nondegenerate. Therefore, under suitable regular conditions, our proposed statistic converges in distribution to a normal distribution.

4. Main Theorems in high dimensional case

In this section, we will provide the asymptotic behavior of the test statistic in the high-dimensional case. In order to draw these conclusions, some technical assumptions are listed below.

(A1) $(\mu_X - \mu_Y)^\tau (\Sigma_X + \Sigma_Y)(\mu_X - \mu_Y) = O(p)$, $tr(\Sigma_X + \Sigma_Y)^2 = O(p)$.

(A2) $E\langle X_1, X_2 - X_3 \rangle^4 + E\langle Y_1, Y_2 - Y_3 \rangle^4 = O(p^2)$, $X_1, X_2, X_3 \overset{i.i.d.}{\sim} X$, $Y_1, Y_2, Y_3 \overset{i.i.d.}{\sim} Y$.

Remark 2. Assumption **(A1)** made in Kim (2020) is satisfied when $\mu = (u, \cdots, u)^\tau$ and the covariance matrix is unit matrix, it is very important in proving the asymptotic convergence. For assumption **(A2)**, when $\mu = (0, \cdots, 0)^\tau$ and covariance matrix is Σ, by proposition A.1 (see Chen, 2010), it can be proved that **(A2)** is true. Details are as follows.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0299

Note that

$$E\left\langle X_1, X_2 - X_3\right\rangle^4 = 4\{3tr^2\Sigma^2 + 6tr(\Sigma^4) + 6\Delta tr(\Sigma^2 \circ \Sigma^2) + \Delta^2 \sum_{i,j=1}^{n} (\Sigma_{ij})^4\},$$

and

$$tr(\Sigma^2 \circ \Sigma^2) \le tr(\Sigma^4) = o(tr^2(\Sigma^2)), \qquad \sum_{i,j=1}^{n} (\Sigma_{ij})^4 \le tr^2(\Sigma^2),$$

thus

$$E\left\langle X_1, X_2 - X_3\right\rangle^4 = O(tr^2(\Sigma_X^2)) = O(p^2).$$

In a similar way, we have $E\left\langle Y_1, Y_2 - Y_3\right\rangle^4 = O(p^2)$. This implies the assumption **(A2)** holds.

Theorem 5. *Assume $n, m, p \to \infty$, if condition **(A1)**, **(A2)** hold, then*

$$\sqrt{\frac{Var(\langle X - X', X''\rangle)}{U_n}} - 1 = O_P\left(\frac{1}{\sqrt{n}}\right),$$

and

$$\sqrt{\frac{Var(\langle Y - Y', Y''\rangle)}{U_m}} - 1 = O_P\left(\frac{1}{\sqrt{m}}\right).$$

Furthermore

$$\frac{\langle X_1 - X_2, Z_1^* - Z_2^*\rangle}{\sqrt{U_n}} = \frac{\langle X_1 - X_2, Z_1^* - Z_2^*\rangle}{\sqrt{Var\langle X'', X - X'\rangle}} + O_P\left(\frac{1}{\sqrt{n}}\right),$$

and

$$\frac{\langle Y_1 - Y_2, Z_3^* - Z_4^*\rangle}{\sqrt{U_m}} = \frac{\langle Y_1 - Y_2, Z_3^* - Z_4^*\rangle}{\sqrt{Var\langle X'', X - X'\rangle}} + O_P\left(\frac{1}{\sqrt{m}}\right),$$

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0299

where $X_1, X_2, Z_1^*, Z_2^*, Y_1, Y_2, Z_3^*, Z_4^*$ *are mutually independent,* $X_1, X_2 \sim X$, $Y_1, Y_2 \sim Y$, $Z_1^*, Z_2^*, Z_3^*, Z_4^*$ *follow either* F_X *or* F_Y.

Theorem 6. *Assume* $n, m, p \to \infty$, *if condition* ***(A1)***, ***(A2)*** *hold, then*

$$U_{n,m} \xrightarrow{a.s.} CD(X,Y).$$

Theorem 7. *Suppose* $n, m, p \to \infty$ $\frac{n}{n+m} \to \theta \in [0,1]$, *if condition* ***(A1)***, ***(A2)*** *hold, then under the null hypothesis, we can obtain*

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

where $a_k(\theta), b_k(\theta), \lambda_k,\ Z_{1k}, Z_{2k}, \cdots$ *are the same as those provided in Theorem* 3, $k = 1, 2, \cdots$.

It is easy to see from Theorem 7 that in this case, the test statistic proposed in this paper has the same asymptotic behavior as Theorem 3. In fact, under the null hypothesis, by using Taylor expansion, we can find that the dominating part of $U_{n,m}$ in high-dimensional distance inference is the same as the conventional fixed dimension, and this part is asymptotically chi-square via H-decomposition (see Koroljuk, 1994), while the other parts convergence to zero in probability. Therefore, based on Slutsky's theorem, we can come to this conclusion.

The next theorem discusses the asymptotic theory of the proposed method under the alternative hypothesis.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0299

Theorem 8. *Assume $n, m, p \to \infty$ and $\frac{n}{n+m} \to \theta \in [0,1]$, if condition **(A1)**, **(A2)** hold, then under the alternative hypothesis, we can obtain*

$$\sqrt{\frac{nm}{n+m}}(U_{n,m} - CD(X,Y)) \xrightarrow{D} N(0, (1-\theta)\delta_{1,0}^2 + \theta\delta_{0,1}^2) + \bar{\zeta},$$

where $\delta_{1,0}^2, \delta_{0,1}^2$ are the same as those provided in Theorem 4, and

$$\bar{\zeta} = -\sqrt{1-\theta}E\widetilde{H}_A(X_1, X_2, X_3, X_4; Y_3, Y_4) - \sqrt{\theta}E\widetilde{H}_C(X_3, X_4; Y_1, Y_2, Y_3, Y_4).$$

Theorem 8 illustrates that the limiting alternative distribution under the situation of high-dimensionality is determined by the random vector with the smaller sample size.

Remark 3. Theorems 5 -8 listed in this section only cover a small portion of the regime in high dimensions. The reason is that regime of high dimensions implies the situation of "big p, small n, m" or "big p, big n, m". The former refers to the fact that when the dimension grows to infinity while the sample sizes n and m are held fixed, i.e., high dimension low sample size (HDLSS). While the latter assumes that p, n and m tend to infinity simultaneously. Furthermore, it should be noted that Theorems 5-8 established in this study are valid when $n, m, p \to \infty$ and assumptions **(A1)** and **(A2)** satisfy. Therefore, they cannot cover the full regime of high dimensions.

5. Numerical Study

In this section, we will conduct some simulation studies and real data analysis to compare the empirical performance of the proposed test statistic with other competitive nonparametric two sample tests, such as the energy distance in Székely (2004), the ball divergence in Pan (2018) and the maximum mean discrepancy in Gretton (2012). We call them the ED test, the BD test and the MMD test, respectively. Throughout the simulation studies, we will use the Gaussian kernel for the MMD test, and the bandwidth is chosen to be the median distance between pairs of points in the aggregate samples. In addition, under the 0.05 significance level, we will use the permutation procedure to obtain the P-value of different tests with 200 permutations. The simulations will be repeated 400 times to approximate the empirical size and power of each test.

5.1 Simulation studies

In the simulation study, several numerical examples are considered. Specifically, when the sample size tend to infinity but the dimension is fixed, we will use multivariate normal distributions to evaluate the finite sample performance of different test procedures with means

$$\mu^{(0)} = (0, \cdots, 0)^{\tau}, \qquad \mu^{(1)} = (0.5, \cdots, 0.5)^{\tau}, \qquad and$$

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0299

when p is odd number

$$\mu^{(2)} = (\underbrace{0.5, \cdots, 0.5}_{\lfloor p/2 \rfloor}, 0, \underbrace{-0.5, \cdots, -0.5}_{\lfloor p/2 \rfloor})^{\tau},$$

when p is even number

$$\mu^{(2)} = (\underbrace{0.5, \cdots, 0.5}_{\lfloor p/2 \rfloor}, \underbrace{-0.5, \cdots, -0.5}_{\lfloor p/2 \rfloor})^{\tau},$$

and covariance matrices:

1. Identity matrix (denoted by I_p), where $\sigma_{i,i} = 1$ and $\sigma_{i,j} = 0$ for $i \neq j$.

2. Banded matrix (denoted by Σ_{Band}), where $\sigma_{i,i} = 1$, $\sigma_{i,j} = 0.6$ for $|i-j| = 1$, $\sigma_{i,j} = 0.2$ for $|i-j| = 2$ and $\sigma_{i,j} = 0$ otherwise.

3. Autocorrelation matrix (denoted by Σ_{Auto}), where $\sigma_{i,i} = 1$ and $\sigma_{i,j} = 0.2^{|i-j|}$ when $i \neq j$.

For Type-I error rates evaluation, the following three models are considered.

Example **5.1**. Suppose $X_k = (X_{k1}, \cdots, X_{kp})$ and $Y_l = (Y_{l1}, \cdots, Y_{lp})$ with $k = 1, \cdots, n$ and $l = 1, \cdots, m$. We generate independent identically distributed samples from the models:

1. $X_k \sim N(\mu^{(0)}, I_p)$ and $Y_l \sim N(\mu^{(0)}, I_p)$.

2. $X_k \sim N(\mu^{(0)}, \Sigma_{Band})$ and $Y_l \sim N(\mu^{(0)}, \Sigma_{Band})$.

3. $X_k \sim N(\mu^{(0)}, \Sigma_{Auto})$ and $Y_l \sim N(\mu^{(0)}, \Sigma_{Auto})$.

For the power evaluation, we consider several models as below.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0299

Example **5.2**. Suppose $X_k = (X_{k1}, \cdots, X_{kp})$ and $Y_l = (Y_{l1}, \cdots, Y_{lp})$ with $k = 1, \cdots, n$ and $l = 1, \cdots, m$. We generate independent identically distributed samples from the models:

1. $X_k \sim N(\mu^{(0)}, I_p)$ and $Y_l \sim N(\mu^{(1)}, I_p)$.
2. $X_k \sim N(\mu^{(0)}, \Sigma_{Band})$ and $Y_l \sim N(\mu^{(1)}, \Sigma_{Band})$.
3. $X_k \sim N(\mu^{(0)}, \Sigma_{Auto})$ and $Y_l \sim N(\mu^{(1)}, \Sigma_{Auto})$.
4. $X_k \sim N(\mu^{(0)}, I_p)$ and $Y_l \sim N(\mu^{(2)}, I_p)$.
5. $X_k \sim N(\mu^{(0)}, \Sigma_{Band})$ and $Y_l \sim N(\mu^{(2)}, \Sigma_{Band})$.
6. $X_k \sim N(\mu^{(0)}, \Sigma_{Auto})$ and $Y_l \sim N(\mu^{(2)}, \Sigma_{Auto})$.

Example **5.3**. Suppose $X_k = (X_{k1}, \cdots, X_{kp})$ and $Y_l = (Y_{l1}, \cdots, Y_{lp})$ with $k = 1, \cdots, n$ and $l = 1, \cdots, m$. We generate independent identically distributed samples from the models:

1. $X_k \sim N(\mu^{(0)}, I_p)$ and $Y_l \sim N(\mu^{(0)}, 0.5 \cdot I_p)$.
2. $X_k \sim N(\mu^{(0)}, \Sigma_{Band})$ and $Y_l \sim N(\mu^{(0)}, 0.5 \cdot \Sigma_{Band})$.
3. $X_k \sim N(\mu^{(0)}, \Sigma_{Auto})$ and $Y_l \sim N(\mu^{(0)}, 0.5 \cdot \Sigma_{Auto})$.

Under the multivariate case, the empirical size and power of each test are reported in Table 1. As expected, when two populations are identically distributed, the Type-I error rates could be well-controlled for each test. In Examples 5.2 and 5.3, it is evident that variables X and Y follow different distributions. The results reported in Table 1 show that, among all multivariate normal location alternatives, whether the signal is in the same direction or not, the ED and our proposed method are more powerful than those based on ball divergence and maximum mean difference. For example, when $n = m = 50$, and $p = 5$, for data samples from Example 5.2, in (3), the empirical power of ED, BD, MMD and our approach are 0.9975, 0.9575, 0.9525 and 0.9925 respectively. However, when detecting the scale shift, the results are very different. The BD test is powerful in all cases, the performance of the MMD test is equivalent to or better than the test based on the proposed homogeneity metric, while the ED test performs worst.

For high-dimensional distance inference, the data-generating scheme is similar to the conventional fixed dimension but with

$$\mu^{(1)} = (0.2, \cdots, 0.2)^\tau, \mu^{(2)} = (\underbrace{0.2, \cdots, 0.2}_{\lfloor p/2 \rfloor}, \underbrace{-0.2, \cdots, -0.2}_{\lfloor p/2 \rfloor})^\tau,$$

and in example 5.3, the models are changed to

$$Y_l \sim N(\mu^{(0)}, 0.75 \cdot I_p), Y_l \sim N(\mu^{(0)}, 0.75 \cdot \Sigma_{Band}), Y_l \sim N(\mu^{(0)}, 0.75 \cdot \Sigma_{Auto}),$$

respectively.

In the high-dimensional case, the empirical size and power for different test procedures are reported in Table 2, Table 3 and Table 4. As shown in Table 2, at the significance level of 0.05, the empirical sizes of the aforementioned test procedures are all close to 0.05, indicating that these methods

Table 1: Empirical size and power for different test procedures.

		$n = m$	p	CD	ED	BD	MMD
5.1	(1)	50	4	0.0575	0.0650	0.0500	0.0650
	(1)	50	5	0.0500	0.0525	0.0600	0.0450
	(2)	50	4	0.0600	0.0675	0.0450	0.0500
	(2)	50	5	0.0600	0.0600	0.0675	0.0525
	(3)	50	4	0.0550	0.0475	0.0400	0.0550
	(3)	50	5	0.0375	0.0450	0.0500	0.0400
5.2	(1)	50	4	0.9900	0.9950	0.9575	0.9675
	(1)	50	5	0.9925	0.9975	0.9775	0.9950
	(2)	50	4	0.8525	0.9000	0.7975	0.8025
	(2)	50	5	0.8825	0.9400	0.8350	0.8250
	(3)	50	4	0.9650	0.9700	0.9125	0.9125
	(3)	50	5	0.9925	0.9975	0.9575	0.9525
	(4)	50	4	0.9900	0.9950	0.9675	0.9725
	(4)	50	5	0.9750	0.9850	0.9275	0.9575
	(5)	50	4	0.8850	0.9800	0.8700	0.9650
	(5)	50	5	0.8550	0.9175	0.7625	0.8400
	(6)	50	4	0.9825	0.9950	0.9375	0.9700
	(6)	50	5	0.9725	0.9700	0.9200	0.9475
5.3	(1)	50	4	0.8375	0.5025	0.9900	0.8675
	(1)	50	5	0.8750	0.5575	1.0000	0.9275
	(2)	50	4	0.5200	0.3575	0.9200	0.7050
	(2)	50	5	0.5400	0.3900	0.9675	0.8400
	(3)	50	4	0.7700	0.4700	0.9950	0.8475
	(3)	50	5	0.8425	0.5750	0.9975	0.9175

Table 2: Empirical size for different test procedures.

Model	Methods	$n=m=50$		$n=m=70$		$n=m=90$	
		$p=50$	$p=100$	$p=50$	$p=100$	$p=50$	$p=100$
(1)	CD	0.0600	0.0575	0.0500	0.0375	0.0650	0.0550
	ED	0.0650	0.0700	0.0500	0.0650	0.0600	0.0675
	BD	0.0700	0.0500	0.0400	0.0475	0.0425	0.0450
	MMD	0.0575	0.0500	0.0500	0.0575	0.0675	0.0550
(2)	CD	0.0525	0.0425	0.0625	0.0550	0.0650	0.0650
	ED	0.0750	0.0675	0.0375	0.0550	0.0500	0.0450
	BD	0.0625	0.0400	0.0550	0.0650	0.0600	0.0625
	MMD	0.0475	0.0300	0.0700	0.0625	0.0600	0.0450
(3)	CD	0.0600	0.0575	0.0500	0.0500	0.0625	0.0425
	ED	0.0725	0.0475	0.0550	0.0650	0.0550	0.0550
	BD	0.0650	0.0600	0.0650	0.0500	0.0575	0.0600
	MMD	0.0525	0.0650	0.0650	0.0625	0.0525	0.0525

can control the type I error well.

For power evaluation, when detecting the location shift in multivariate normal distribution (Example 5.2), it can be seen from Table 3: (i) When p is fixed, the power for all test procedures increases with the increase of n; (ii) In most settings, regardless of the signal is in the same direction or in the opposite direction, the empirical power of the ED test and our approach is similar and within the highest-power group, although they are drawn from a very different perspective; (iii) The power of BD test procedure is poor, and the MMD test is somewhat in between.

When a scale difference is detected (Example 5.3), the performance completely changes. Now BD testing has maintained desirable performance, and always belongs to the group with the highest power. The ED test performs the worst in all cases. As for our approach, when n, m and p are smaller, it is less competitive than the BD and MMD test in terms of power. However, as the sample size n, m and dimension p increase, this difference gradually decreases.

5.2 Real Data Example

In this section, we will apply the proposed test procedure to two scenarios: gene-set testing and benchmark dataset testing.

Table 3: Empirical power for different test procedures.

Model	Methods	$n = m = 50$		$n = m = 70$		$n = m = 90$	
		$p = 50$	$p = 100$	$p = 50$	$p = 100$	$p = 50$	$p = 100$
(1)	CD	0.8950	0.9575	0.9625	0.9950	0.9900	0.9975
	ED	0.9775	1.0000	0.9925	1.0000	1.0000	1.0000
	BD	0.5925	0.8475	0.7350	0.9050	0.8100	0.9575
	MMD	0.9075	0.9825	0.9475	1.0000	0.9800	1.0000
(2)	CD	0.8200	0.9500	0.8925	0.9800	0.9550	1.0000
	ED	0.8125	0.9800	0.9125	0.9800	0.9575	1.0000
	BD	0.5400	0.7225	0.6075	0.8025	0.7475	0.9225
	MMD	0.6100	0.8325	0.7025	0.8900	0.7950	0.9500
(3)	CD	0.9500	0.9950	0.9700	0.9950	0.9900	1.0000
	ED	0.9800	1.0000	0.9750	1.0000	0.9975	1.0000
	BD	0.6600	0.8050	0.7400	0.9250	0.8200	0.9625
	MMD	0.8725	0.9725	0.9075	0.9925	0.9425	0.9975
(4)	CD	0.8700	0.9625	0.9575	0.9925	0.9925	1.0000
	ED	0.9600	1.0000	0.9925	1.0000	0.9950	1.0000
	BD	0.5975	0.8450	0.6925	0.9375	0.7775	0.9600
	MMD	0.8925	0.9925	0.9625	1.0000	0.9900	1.0000
(5)	CD	0.8000	0.9300	0.8975	0.9800	0.9275	0.9975
	ED	0.8425	0.9800	0.9000	0.9900	0.9550	1.0000
	BD	0.5125	0.6975	0.6375	0.8250	0.7300	0.8600
	MMD	0.6250	0.8200	0.7375	0.8975	0.8125	0.9500
(6)	CD	0.9325	0.9800	0.9775	0.9975	0.9900	1.0000
	ED	0.9600	1.0000	0.9925	1.0000	0.9950	1.0000
	BD	0.6825	0.8150	0.7650	0.9050	0.8475	0.9575
	MMD	0.8250	0.9750	0.9200	0.9950	0.9550	0.9975

Table 4: Empirical power for different test procedures.

Model	Methods	$n = m = 50$		$n = m = 70$		$n = m = 90$	
		$p = 50$	$p = 100$	$p = 50$	$p = 100$	$p = 50$	$p = 100$
(1)	CD	0.5850	0.7075	0.7700	0.8850	0.8400	0.9575
	ED	0.3550	0.4725	0.4225	0.6700	0.5150	0.7900
	BD	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	MMD	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
(2)	CD	0.3400	0.4975	0.4450	0.6050	0.5000	0.7550
	ED	0.2225	0.3375	0.2600	0.4600	0.3250	0.5500
	BD	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	MMD	0.9525	1.0000	0.9725	1.0000	0.9875	1.0000
(3)	CD	0.5125	0.6375	0.6225	0.8325	0.7550	0.9175
	ED	0.2825	0.4725	0.3975	0.6675	0.5575	0.7600
	BD	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	MMD	0.9950	1.0000	0.9975	1.0000	1.0000	1.0000

5.2.1 Gene-set testing

Identifying differentially expressed genes is of great significance for some treatments, and it is the latest development in genetics research (see Barry, 2005; Nettleton, 2008). In this section, we will apply the proposed homogeneity test to reanalyze the acute lymphoblastic leukemia (ALL) dataset, which can be downloaded via the R package (see Li, 2009). For this data set, it consists of 128 patients with T-cell or B-cell leukemia, and each with 12625 genes expression. Since these two types of cells are different tissues, we treat them separately, usually focusing on B-cell tumors. Note that in B-cell type leukemia, there are two different molecular classes: B-cell ALL with the BCR/ABL fusion and cytogenetically normal NEG B-cell ALL. Therefore, in this article, we are interested in identifying differentially expressed genes between BCR/ABL (sample size $n = 37$) and NEG (sample size $m = 42$). Next, for convenience's sake, we carry out a data preprocessing for gene-filtering based on the strategy of Gentleman (2005), leaving 2391 genes for analysis. For each of these genes, at the significance level of 0.05, we will use the proposed two-sample test with false discovery rate (FDR) control (Benjamini, 1995), as well as ED, BD, and MMD with FDR, to detect the differences between BCR/ABL and NEG, respectively.

For each of the 2391 genes, at the significance level of 0.05, the number of differentially expressed genes detected by ED, BD, MMD, and CD tests is 131, 101, 121, and 130, respectively. This illustrates that the ED test and our approach are more powerful than the other two tests.

5.2.2 Benchmark dataset testing

We also consider exploring the potential difference between two high-dimensional distributions on the following two benchmark data sets: Strawberry data and SmallKitchenAppliances data. These two datasets can be downloaded from UCR Time Series Classification Archive (`https://www.cs.ucr.edu/~eamonn/time_series_data_2018/`). For Strawberry data, there are 351 strawberry and 632 non-strawberry purees, with a length of 235 for each data point. For SmallKitchenAppliances data, it contains three classes. Here for simplicity, we only use the data: Kettle and Microwave, with observation values of 250 for both types, and a length of 720 for each data point. In addition, to facilitate analysis of the problems, following the procedures of Biswas (2014) and Sarkar (2020), for each $m = n \in \{20, 30, 40, 50, 60, 70\}$, randomly sample n points from each class, and use the tests mentioned above, to analyze whether these two classes come from the single population.

Power comparison for the two datasets is shown in Fig. 1. The left panel

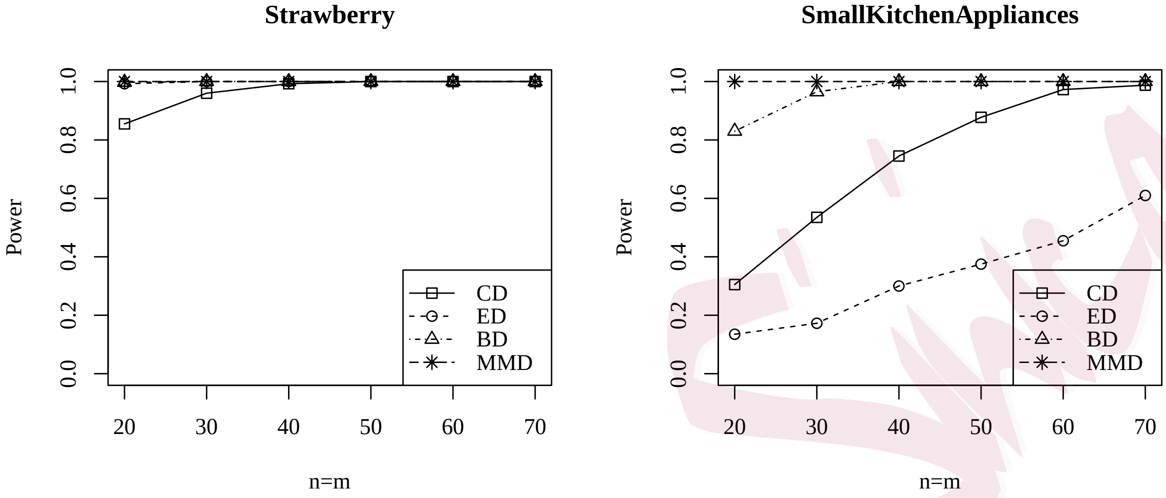

Figure 1: benchmark dataset testing

of Fig. 1 demonstrates that the ED, BD, and MMD have very high power for Strawberry data with relatively low sample size. For our method, with the increase of sample size, it quickly catches up with the above three test procedures. In addition, from the right panel of Fig. 1, we can clearly see that BD and MMD tests have a remarkable performance in high-dimensional and low sample size, while the ED test is less sensitive. As for our approach, as the sample size increases, its performance is greatly improved.

5.3 Comparison of computational complexity

In this section, we will provide a comparison between the new approach and the existing methods concerning computation burden.

The computational complexity of CD is $O((n+m)^4p)$ if we compute it exactly from (2.3), where n, m and p are the size and dimension of the sample set, respectively. As compared with the complexities of the previous test procedures, such as $O((n+m)^2p)$ for exact ED, $O((n+m)^2p)$ for MMD, and $O(n^2logn+m^2logm)$ for the BD test, the proposed test procedure evidently loses a speed gain. Therefore, in the simulation studies and real data analysis, our approach is the most time-consuming.

Therefore, driven by the potential computational burden associated with characteristic distance, in high-dimensional settings, we only discuss the cases of $n = m = 50, 70, 90$ and $p = 50, 100$. Also, considering the trends in empirical size and power of different tests, which are already visible in the above simulation settings, we did not consider scenarios where the sample size and the dimension are much lager.

Although computational complexity does play an important role in practical implementation, we opt not to pursue this further in this article. In future work, the authors will strive to improve the shortcomings in computational complexity.

6. Discussion

In this paper, we introduce a new two sample test, which is very useful in high-dimensional distance inference. The new metric proposed in this paper has several appealing features, including a zero distance that characterizes the homogeneity between two random vectors, as well as consistency against any alternative hypothesis. The simulation results show that the proposed test performs better in most cases. The study of acute lymphoblastic leukemia data and benchmark data sets further illustrates the feasibility and practicability of the proposed test procedure.

Besides the homogeneous test problems, our proposed method can also be applied to other issues. For example, the goodness-of-fit test, clustering, change-point detection, and so on. We expect the characteristic distance to be more effective for these statistical problems.

Supplementary Materials

Additional supporting materials can be found in the Supplementary Materials, including proof of the theoretical results presented in Sections 2-4, as well as some additional simulation results.

Acknowledgements

Zhang's work is supported by the National Natural Science Foundation of China (12271370). Li's work is supported by the National Natural Science Foundation of China (12401356), and the Natural Science Foundation of Shanxi Province, China (202203021222223, 20210302124262, 20210302124531).

References

Baringhaus, L. and Franz, C. (2004). On a new multivariate two-sample test. *Journal of Multivariate Analysis* **88**, 190-206.

Barry, W., Nobel, A. and Wright, F. (2005). Significance analysis of functional categories in gene expression studies: a structured permutation approach. *Bioinformatics* **21**, 1943-1949.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of Royal Statistical Society: Series B (Statistical Methodology)* **57**, 289-300.

Bickel, P. J. (1969). A distribution free version of the Smirnov two sample test in the p-variate case. *Annals of Mathematical Statistics* **40**, 1-23.

Biswas, M. and Ghosh, A. K. (2014). A nonparametric two-sample test applicable to high dimensional data. *Journal of Multivariate Analysis* **123**, 160-171.

Chakraborty, S. and Zhang, X. (2021). A new framework for distance and kernel-based metrics in high dimensions. *Electronic Journal of Statistics* **15**, 5455-5522.

Chen, S. X., Zhang, L. X. and Zhong, P. S. (2010). Tests for high-dimensional covariance matrices. *Journal of the American Statistical Association* **105**, 810-819.

Darling, D. A. (1957). The Kolmogorov-Smirnov, Cramer-von Mises Tests. *Annals of Mathematical Statistics* **29**, 842-851.

Fernández, V. A., Gamero, M. D. J. and García, J. M. (2008). A test for the two-sample problem based on empirical characteristic functions. *Computational Statistics and Data Analysis* **52**, 3730-3748.

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

Gao, H. J. and Shao, X. F. (2023). Two sample testing in high dimension via maximum mean discrepancy. *Journal of Machine Learning Research* **24**, 14406-14438.

Gentleman, R., Irizarry, R. A., Carey, V. J., Dudoit, S. and Huber, W. (2005). Bioinformatics and Computational Biology Solutions Using R and Bioconductor. *Springer, New York. MR2201836.*

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B. and Smola, A. J. (2012). A kernel two-sample test. *Journal of Machine Learning Research* **13**, 723-773.

Harchaoui, Z., Bach, F., Cappe, O. and Moulines, E. (2013). Kernel-based methods for hypothesis testing: A unified view. *IEEE Signal Processing Magazine* **30**, 87-97.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0299

Kim, I., Balakrishnan, S. and Wasserman, L. (2020). Robust multivariate nonparametric tests via projection averaging. *Annals of Statistics* **48**, 3417-3441.

Koroljuk, V. S. and Borovskich, Y. V. (1994). Theory of U- Statistics. *Kluwer Academic Publisher, Amsterdam.*

Lee, D. H., Lahiri, S. N. and Sinha, S. (2020). A test of homogeneity of distributions when observations are subject to measurement errors. *Biometrics* **76**, 821-833.

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

Li, X. C. (2009). ALL: A data package. *R package version 1.22.0.*

Liu, Y. M., Liu, Z. and Zhou, Wang. (2019). A test for equality of two distributions via integrating characteristic functions. *Statistica Sinica* **29**, 1779-1801.

Liu, J. M., Ma, S. G., Xu, W. L. and Zhu, L. P. (2022). A generalized Wilcoxon–Mann–Whitney type test for multivariate data through pairwise distance. *Journal of Multivariate Analysis* **190**, 104946.

Liu, Z., Xia, X. and Zhou, W. (2015). A test for equality of two distributions via jackknife empirical likelihood and characteristic functions. *Computational Statistics and Data Analysis* **92**, 97-114.

Mukhopadhyay, S. and Wang, K. J. (2020). A nonparametric approach to high-dimensional k-sample comparison problems. *Biometrika* **107**, 555-572.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0299

Nettleton, D., Recknor, J. and Reecy, J. M. (2008). Identification of differentially expressed gene categories in microarray studies using nonparametric multivariate analysis. *Bioinformatics* **24**, 192-201.

Pan, W., Tian, Y., Wang, X. and Zhang, H. (2018). Ball Divergence: Nonparametric two sample test. *Annals of Statistics* **46**, 1109-1137.

Qiu, T., Xu, W. L. and Zhu, L. P. (2021). A robust and nonparametric two-sample test in high dimensions. *Statistica Sinica* **31**, 1853-1869.

Ramdas, A., Reddi, S. J., Poczos, B., Singh, A. and Wasserman, L. (2015). On the decreasing power of kernel and distance based nonparametric hypothesis tests in high dimensions. *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 3571-3577.

Sarkar, S. and Ghosh, A. K. (2018). On some high-dimensional two-sample tests based on averages of inter-point distances. *Stat* **7**, 1-16.

Sarkar, S., Biswas, R. and Ghosh, A. K. (2020). On some graph-based two-sample tests for high dimension, low sample size data. *Machine Learning* **109**, 279-306.

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

Smirnoff, N. (1939). On the estimation of the discrepancy between empirical curves of distribution for two independent samples. *Bulletin de lUniversite de Moscow, Serie internationale (Mathematiques)* **2**, 3-14.

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

Wald, A. and Wolfowitz, J. (1940). On a test whether two samples are from the same population. *Annals of Mathematical Statistics* **11**, 147-162.

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

Zhao, J. and Meng, D.Y. (2015). FastMMD: Ensemble of circular discrepancy for efficient two-sample test. *Neural Computation* **27**, 1345-1372.

Zhou, W. X., Zheng, C. and Zhang, Z. (2017). Two-sample smooth tests for the equality of distributions. *Bernoulli* **23**, 951-989.

Zhu, C. and Shao, X. (2021). Interpoint distance based two sample tests in high dimension. *Bernoulli* **27**, 1189-1211.

School of Statistics, Capital University of Economics and Business, Beijing 100070, China.

School of Mathematics and Computer Science, Shanxi Normal University, Taiyuan 030000, China.

E-mail: (sxsdlixu2004@sina.com)

School of Statistics, Capital University of Economics and Business, Beijing 100070, China.

E-mail: (gmshi@cueb.edu.cn)

School of Statistics, Capital University of Economics and Business, Beijing 100070, China.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0299

E-mail: (zhangbaoxue@cueb.edu.cn)