Statistics and Its Interface Volume 15 (2022) 335–348

Paired-sample tests for homogeneity with/without confounding variables

Minqiong Chen, Ting Tian, Jin Zhu,
Wenliang Pan, and Xueqin Wang*

In this article, we are concerned about testing the homogeneity on paired samples with or without confounding variables. These problems usually arise in clinical trials, psychological or sociological studies. We introduce new nonparametric tests for equality of two distributions or two conditional distributions of random vectors on paired samples. We show that their test statistics are consistent but have different asymptotic distributions under the null hypothesis, depending on whether confounding variables exist. The limit distribution of the test statistic is a mixed χ^2 distribution when testing the equality of two paired distributions, while it is a normal distribution when testing the equality of two conditional distributions of paired samples. We conduct several simulation studies to evaluate the finite-sample performance of our tests. Finally, we apply our tests on real data to illustrate their usefulness in the applications.

Keywords and phrases: Homogeneity, Conditional distribution, Paired samples, Energy distance.

1. INTRODUCTION

Detecting whether two paired samples differ in their distribution functions is a common problem in statistical inference. Such a problem often arises in clinical trials, psychological, or sociological studies. Typical paired data include: (1) repeated measurements of the same subjects in a longitudinal study, for example, determining whether a drug has any effect on blood pressure after measuring the blood pressure of the same set of patients before and after taking the drug; (2) binate observations on the same individuals, for example, examining whether there is a significant difference in the distribution of sight in the right and left eyes of adolescents, or investigating whether tooth size profile is the same for the left and right sides around central incisors [10]; and (3) observations of different individuals that have been matched based on some set of characteristics, for instance, testing whether fathers and their sons have the same height. Please see [16, 14] for more examples.

Let (X, Y) be a paired random vector in Euclidean space $\mathbb{R}^p \times \mathbb{R}^p$. A paired-sample testing problem intends to test whether the distributions of X and Y are identical even when X and Y are dependent. Indeed, the null hypothesis of this problem is given as follows,

$$H_0 : X \overset{d}{=} Y, \tag{1}$$

where the notation "$\overset{d}{=}$" means "identically distributed". Many classical methods have been used to test the identity of X and Y for $p = 1$. When (X, Y) follows a bivariate normal distribution, the paired t-test is used to determine whether the difference between the two paired population means is significant. Suppose the normality assumption of (X, Y) is not satisfied. In that case the well-known Wilcoxon signed-rank test is frequently used to test whether $X - Y$ is symmetric about zero, especially, the M-test and Q-test [23] for the distributions of X and Y being both Weibull distributions. Also, the McNemar test [15] for X and Y being binary variables, and the test of Stuart [20] for categorical variables. When X and Y are multivariate, few paired-sample tests have been proposed, for example, the paired Hotelling's T^2 test can be used when (X, Y) is normally distributed. It is still an interesting problem to develop flexible paired-sample tests for $p > 1$.

Furthermore, we consider the case that (X, Y) could be affected by a confounding factor Z in $\mathbb{R}^r$, for instance, X and Y are the blood pressure before and after the drug treatment, Z is the factor of the age of the patients, which may affect both X and Y. To assess the treatment, we want to eliminate the effect of Z when comparing the distributional difference between X and Y. Thus, the problem of interest can be formulated to test the identity of two conditional distributions. That is, we need to test the following null hypothesis

$$H_0 : X|Z = z \overset{d}{=} Y|Z = z, \text{for all } z \in S(Z), \tag{2}$$

where $S(Z)$ is the support of the density function of Z. Few works of literature discussed this problem while comparing the difference of two conditional distributions for independent or paired samples via regression models. See [4, 7, 19] for more details. Lee [9] presented a covariate-matched Mann–Whitney statistic for comparing two conditional distributions for two independent samples, which is convenient to be modified to paired samples. Koul and

*Corresponding author.

Schick [6] addressed this testing problem as a "common design" when testing the equality of two nonparametric regression curves. Among the four types of tests in [6], the first and third types provided two Z-controlled versions of the t test. The first class used a kernel estimator of the regression function, while the third class avoided this by matching the covariates. The asymptotic properties and asymptotically optimal tests for these two classes were also discussed. Recently, Guo et al. [2] proposed a test for comparing two conditional means in paired samples via empirical characteristic functions, avoiding using kernel smoothing and no assumption of a specified regression model. All the tests listed above focused on comparing conditional means or other location parameters. Here we are interested in the overall conditional distributions, rather than only in their means. Li et al. [11] considered the testing for the equality of two conditional density functions for two independent samples by a new kernel smoothing approach. However, their approach only considered the cases of discrete confounders.

In this paper, we propose nonparametric homogeneity tests for the null hypotheses (1) and (2), and investigate their statistical properties. We first apply the concept of energy distance [22] directly but use paired samples to propose a novel test statistic for testing the equality of distributions. Like the energy distance-test statistic for two independent samples, our test statistic is consistent and has a mixed χ^2 distribution in asymptotic under the null hypothesis but is normal under the alternative hypothesis. For testing the equality of conditional distributions, we further extend our test statistic to test the equality of two conditional distributions by first extending the concept of energy distance to conditional energy distance, and then provide its statistical properties. We find that these two test statistics have different asymptotic distributions under the null hypothesis. Numerical studies are conducted to illustrate their usefulness in the applications.

The rest of this paper is organized as follows. In Section 2, we briefly introduce the modified version of energy distance for two dependent random vectors and give the test statistic for paired samples. We obtain the asymptotic properties of the test statistic. We also provide a bootstrap method to approximate the p-value of the test and prove the validity of the bootstrap. In Section 3, we extend the homogeneity test to test the equality of conditional distributions. And then, we discuss the asymptotic properties of the new test statistic. Simulation experiments are conducted to examine the performance of our tests in finite samples in Section 4. In Section 5, we apply both the homogeneity and the conditional homogeneity test on Student Performance data. A summary and discussion are presented in Section 6. All proofs of the main results are put in Appendix A.

2. PAIRED SAMPLES TEST

We first derived a test statistic for (1) based on the concept of Energy distance. Suppose that (X', Y') is an i.i.d. copy of (X, Y), and both $E|X|$ and $E|Y|$ are finite, then we define

$$V(X,Y) := 2E|X-Y'| - E|X-X'| - E|Y-Y'|, \tag{3}$$

where $|u| = \sqrt{u^T u}$ is the Euclidean norm of u in $\mathbb{R}^p$.

By Lemma 1 in Appendix A, $V(X,Y) \geq 0$ for any two paired random vectors X and Y with finite moments, and $V(X,Y) = 0$ if and only if X and Y are identically distributed. Hence it's natural to utilize the sample estimator of $V(X,Y)$ to test the null hypothesis (1). Specifically, given an i.i.d. sample $\mathcal{S}_n = \{(X_1,Y_1),(X_2,Y_2),\ldots,(X_n,Y_n)\}$ from (X,Y), we define the moment estimator of $V(X,Y)$ as

$$\hat{V}_n(X,Y) = \frac{1}{n^2}\sum_{i,j=1}^{n}\big(2|X_i-Y_j| - |X_i-X_j| - |Y_i-Y_j|\big). \tag{4}$$

Proposition 2 in Appendix A validates that under the null hypothesis (1), $n\hat{V}_n(X,Y)$ converges in law to a quadratic form $\sum_{v=1}^{\infty}\lambda_v \mathcal{Z}_v^2$, while Proposition 1 in Appendix A implies that $n\hat{V}_n(X,Y)$ converges to $+\infty$ almost surely under the alternative hypothesis. So we reject (1) whenever $n\hat{V}_n(X,Y) > c_\gamma$ at the significance level of γ, where c_γ is the upper γ-quantile of $\sum_{v=1}^{\infty}\lambda_v\mathcal{Z}_v^2$. However, it is hard to get the exact c_γ in practice since λ_v's depend on the unknown joint distribution of (X,Y) and hence are difficult to compute. Alternatively, we consider a bootstrap method to approximate the p-value. Different from the procedure proposed in [17], which resamples from the pooled sample $\{X_1,\ldots,X_{n_1},Y_1,\ldots,Y_{n_2}\}$ to obtain the bootstrap samples, we have to keep the dependence between X and Y when resampling. Inspired by the bootstrap procedures proposed by Konietschke et al. [5] for paired t-test, we construct a similar bootstrap approach to approximate the p-value for $n\hat{V}_n(X,Y)$ as follows:

Step 1. Calculate the statistic $n\hat{V}_n(X,Y)$ based on the original sample $\mathcal{S}_n$, denoted as $n\hat{V}_n(\mathcal{S}_n)$.
Step 2. Sample with replacement from the origin sample $\mathcal{S}_n$ to obtain $\{(X_{\pi(i)},Y_{\pi(i)})\}_{i=1}^n$, where $\{\pi(1),\ldots,\pi(n)\}$ denotes a resample with replacement of $\{1,2,\ldots,n\}$. Then, for each $i=1,2,\ldots,n$, generate e_i from the uniform distribution on $\{0,1\}$, that is $P(e_i=0)=P(e_i=1)=1/2$. Finally set $X_i^\star = e_iX_{\pi(i)}+(1-e_i)Y_{\pi(i)}, Y_i^\star = (1-e_i)X_{\pi(i)}+e_iY_{\pi(i)}$ to get the bootstrap sample $\{(X_i^\star,Y_i^\star)\}_{i=1}^n$ and the corresponding bootstrap statistic of $n\hat{V}_n(X,Y)$ is $n\hat{V}_n^\star$.
Step 3. Repeat *Step 2* for B times (say $B=399$), to obtain $\{n\hat{V}_{nb}^\star, 1\leq b\leq B\}$, then the p-value of $n\hat{V}_n$ is given by

$$p \approx \frac{1+\sum_{b=1}^{B} I(n\hat{V}_{nb}^\star > n\hat{V}_n(\mathcal{S}_n))}{1+B},$$

where $I(\cdot)$ is the indicator function.

The next theorem shows the validity of the proposed bootstrap method in approximating the null distribution of the test statistic $n\hat{V}_n(X,Y)$.

Theorem 1. *Given the original sample $\mathcal{S}_n$, the bootstrap statistic $n\hat{V}_n^\star$ converges in law to $\sum_{v=1}^{\infty}\lambda_v\mathcal{Z}_v^2$ in Proposition 2, that is*

$$n\hat{V}_n^\star\Big|\mathcal{S}_n \rightsquigarrow \sum_{v=1}^{\infty}\lambda_v\mathcal{Z}_v^2.$$

The proof of Theorem 1 is given in Appendix A.

3. PAIRED SAMPLE TEST WITH CONFOUNDERS

If the confounder vector Z exists, we propose a conditional homogeneity test for the null hypothesis (2). "Conditionally identically distributed" implies "identically distributed", but not vice versa. In line with the spirit of energy distance, we introduce the conditional energy distance between X and Y given $Z=z$, and derive its estimator, which can be used to test the null hypothesis (2). More precisely, if we denote $\phi_{X|Z=z}(t), \phi_{Y|Z=z}(t)$ as the conditional characteristic functions of X and Y given $Z=z$ respectively, that is $\phi_{X|Z=z}(t) = E[e^{i\langle t,X\rangle}|Z=z]$ and $\phi_{Y|Z=z}(t) = E[e^{i\langle t,Y\rangle}|Z=z]$, where i is the imaginary unit, and $\langle\cdot,\cdot\rangle$ represents the inner product of the two corresponding vectors, then we have the following definition.

Definition 1. The conditional energy distance $\varepsilon(X,Y|Z=z)$ between X and Y with finite norm moments given $Z=z$ is defined as

$$\begin{aligned}\varepsilon(X,Y|z) &= \|\phi_{X|Z=z}(t)-\phi_{Y|Z=z}(t)\|^2 \\ &= \frac{1}{c(p)}\int_{\mathbb{R}^p}\frac{|\phi_{X|Z=z}(t)-\phi_{Y|Z=z}(t)|^2}{|t|^{p+1}}dt,\end{aligned} \tag{5}$$

where $c(p)=\frac{\pi^{(p+1)/2}}{\Gamma((p+1)/2)}$.

Typically, $\varepsilon(X,Y|z)\geq 0, \forall z\in S(Z)$ where the equality holds if and only if the null hypothesis (2) holds. If $\varepsilon(X,Y|z)$ is estimated directly from the definition, it seems to be complicated because it involves the estimations of the two conditional characteristic functions. Fortunately, we can derive its variant in the form of conditional expectation, which gives us a concise estimator. Next, let $W_i=(X_i,Y_i,Z_i), i=1,2,\ldots,n$ be an i.i.d. sample from the distribution of (X,Y,Z), and

$$\begin{aligned}&g((x_1,y_1),(x_2,y_2))\\ &=|x_1-y_2|+|x_2-y_1|-|x_1-x_2|-|y_1-y_2|,\end{aligned}$$

then $\varepsilon(X,Y|z)$ can be estimated by the kernel smoothing method according to Lemma 2 in Appendix A,

$$\begin{aligned}&\hat{\varepsilon}_n(X,Y|z)\\ &=\frac{\sum_{i<j}g((X_i,Y_i),(X_j,Y_j))\mathcal{K}_H(Z_i-z)\mathcal{K}_H(Z_j-z)}{\sum_{i<j}\mathcal{K}_H(Z_i-z)\mathcal{K}_H(Z_j-z)},\end{aligned}$$

where $\mathcal{K}_H(\cdot)$ is a kernel function in $\mathbb{R}^r$ with H being the bandwidth matrix. $\hat{\varepsilon}_n(X,Y|z)$ is a so-called conditional U-statistic, a concept proposed in [21] and can be used for testing the conditionally identically distributed of X and Y given $Z=z\in S(Z)$.

To test the null hypothesis (2), the conditionally identically distributed of X and Y given $Z=z$ for all $z\in S(Z)$, a natural way is to weight $\varepsilon(X,Y|z)$ by a nonnegative function, say the density function of Z, $f_Z(z)$, to get

$$\mathcal{U} := E[\varepsilon(X,Y|Z)f_Z(Z)].$$

Consequently, the null hypothesis (2) is true if and only if $\mathcal{U}=0$. Let

$$\hat{\mathcal{U}}_n := \frac{1}{C_n^2}\sum_{i<j}g((X_i,Y_i),(X_j,Y_j))\mathcal{K}_H(Z_i-Z_j), \tag{6}$$

then as will be shown below, $\hat{\mathcal{U}}_n$ is a consistent estimator of $\mathcal{U}$, and is therefore a candidate of the test statistic for the null hypothesis (2).

We next present the asymptotic properties of the test statistic $\hat{\mathcal{U}}_n$, which can be proved using the theory of U-statistics with random kernels.

For simplicity, we choose $\mathcal{K}_H$ to be the Gaussian kernel

$$\begin{aligned}\mathcal{K}_H(\mathbf{u}) &= |H|^{-1}\mathcal{K}(H^{-1}\cdot\mathbf{u})\\ &= (2\pi)^{-r/2}|H|^{-1}\exp\Big(-\frac{1}{2}\mathbf{u}^TH^{-2}\mathbf{u}\Big)\end{aligned}$$

in $\mathbb{R}^r$, where H is a diagonal matrix $diag\{h_1,h_2,\ldots,h_r\}$ determined by the bandwidths $h_1,h_2,\ldots,h_r$. Assume the regularity conditions (C1)–(C3) in Appendix A hold. The following theorems state the limiting distributions of $\hat{\mathcal{U}}_n$ under the null and alternative hypotheses respectively, as well as its consistency.

Theorem 2 (Consistency)**.** *Suppose $E|X|^2<\infty$ and $E|Y|^2<\infty$, and assume the conditions (C1)–(C3) hold, then we have*

$$\hat{\mathcal{U}}_n \xrightarrow{P} \mathcal{U}.$$

The proof of Theorem 2 is shown in Appendix A.

Moreover, using the theory of U-statistics discussed in Hall [3] and Lee [8], we obtain the following asymptotic distributions for $\hat{\mathcal{U}}_n$.

Theorem 3 (Weak convergence under null hypothesis)**.** *Suppose $E|X|^2<\infty$ and $E|Y|^2<\infty$, and assume that*

conditions (C1)–(C3) hold, then, under the null hypothesis (2), we have

$$\hat{T}_n := n|H|^{1/2}\hat{\mathcal{U}}_n/\hat{\sigma}_n \rightsquigarrow N(0,1), \tag{7}$$

where

$$\hat{\sigma}_n^2 = \frac{2|H|}{C_n^2}\sum_{i<j} g^2((X_i,Y_i),(X_j,Y_j))\mathcal{K}_H^2(Z_i - Z_j).$$

Theorem 4 (Weak convergence under alternative hypothesis)**.** *Suppose $E|X|^2 < \infty$ and $E|Y|^2 < \infty$, and assume the conditions (C1)–(C3) hold, in addition, $n|H|^4 \to 0$ as $n \to \infty$, then, under the alternative hypothesis, we have*

$$\sqrt{n}(\hat{\mathcal{U}}_n - \mathcal{U}) \rightsquigarrow N(0, 4\sigma^2), \tag{8}$$

where σ^2 will be given in (12) in Appendix A.

The proofs of Theorems 3 and 4 are given in Appendix A.

Theorem 3 implies that the asymptotic null distribution of $\hat{T}_n$ is normal, which is very different from $n\hat{V}_n(X,Y)$ in Proposition 2. This allows us to approximate p-values without using the bootstrap method.

According to Theorem 3, we reject the null hypothesis (2) whenever $|\hat{T}_n| > z_{\gamma/2}$ at the significance level of γ, where $z_{\gamma/2}$ is the upper $\gamma/2$ quantile of the standard normal distribution. Combining this and Theorem 4, we can obtain the asymptotic power of test $\hat{T}_n$ as

$$1 - \Phi\Big(\frac{\hat{\sigma}_n z_{\gamma/2}}{2\sqrt{n|H|}}\sigma^{-1} - \sqrt{n}\mathcal{U}\sigma^{-1}/2\Big) + \Phi\Big(-\frac{\hat{\sigma}_n z_{\gamma/2}}{2\sqrt{n|H|}}\sigma^{-1} - \sqrt{n}\mathcal{U}\sigma^{-1}/2\Big),$$

which converges to 1 as $n \to \infty$. Hence the test is consistent.

Remark 1. Theorems 2–4 also hold for X and Y being two discrete random vectors, if we replace $f(x,y|z)$ in (C3) by $p(x,y|z)$, the conditional joint probability of (X,Y) given $Z = z$.

Remark 2. Our test statistic relies on the density estimator of Z. Here the kernel density estimator we use may work for low-dimensional Z. For high-dimensional Z, we could use some alternative estimators such as the estimators in [12, 13].

4. SIMULATIONS

In this section, we conduct comprehensive simulation studies **S.1-S.3** to evaluate the finite-sample performance of the tests we proposed for testing the homogeneity with and without confounding variables for paired samples. We use the bootstrap procedure presented in Section 2 with the number of bootstrap samples $B = 399$ to obtain the p-value of $\hat{V}_n$, while we calculate the p-value of $\hat{T}_n$ directly by the normal approximation. Performance of the tests is based on the Type-I error and power. We use the significance level of 0.05 and each simulation is replicated 1000 times.

In **S.1**, we wish to show the difference between the two concepts of "identically distributed" and "conditionally identically distributed" combining $\hat{V}_n$ and $\hat{T}_n$. Consider the following five cases: (1) $X = Z + \epsilon_1, Y = Z + \epsilon_2$ with $Z \sim N(0, 0.5^2)$ and $\epsilon_1, \epsilon_2 \sim N(0,1)$; (2) The same setting as in (1) except that $Y = -Z + \epsilon_2$; (3) $X = Z + \epsilon_1, Y = 1 - Z + \epsilon_2$ with $Z \sim U(0,1)$ and $\epsilon_1, \epsilon_2 \sim U(-1,1)$; (4) $X = (1+\epsilon)Z, Y = (1+\epsilon)(1-Z)$ with $\epsilon \sim B(1, 0.3)$ and $Z \sim U(0,1)$; (5) $X = Z + \epsilon_1, Y = -Z + \epsilon_2$ with Z following the multivariate normal distribution $N_3(0,\Sigma)$ and $\epsilon_1, \epsilon_2 \sim N_3(0, I_3)$, where Σ is the matrix with elements equal to 1 on the diagonal and 0.5^2 everywhere else and I_3 is the identity matrix. $Z, \epsilon_1, \epsilon_2, \epsilon$ are independently generated in all the settings. Obviously, X and Y are both identically distributed and conditionally identically distributed given Z in case (1), whereas, they are identically distributed but not conditionally identically distributed given Z in cases (2)–(5). Table 1 presents the simulation results for each case with the sample sizes varying from 30 to 200. It can be seen from Table 1 that $\hat{V}_n$ controls the Type-I errors well around 0.05 in all the cases, while $\hat{T}_n$ accurately detects the difference between the conditional distributions of X and Y given Z with desirable powers in cases (2)–(5) as well as controls reasonably the sizes in case (1).

In **S.2**, we aim to examine the empirical performance of $\hat{V}_n$ with different dimensions of the two paired vectors. For the dimension $p = 1$, we generate two dependent variables X and Y with identical or distinct distributions by the Gaussian copula. Both continuous and discrete distributions are considered for X and Y. Seven distributions including the three normal distributions $N(0.5,1), N(1,1), N(0.5,2)$, the exponential law with mean 2, denoted as $Exp(2)$, the gamma distribution with parameters equal to 2 and 4, denoted as $Ga(2,4)$, the beta distribution with parameters equal to 2, denoted as $Beta(2,2)$, and the uniform distribution $U(0,1)$ are considered for the continuous cases. The binomial distribution $B(5, 0.4)$, the negative binomial $NB(2, 0.5)$ and the Poisson distribution $P(2)$ are considered for the discrete cases. Table 2 displays the performance of $\hat{V}_n$ compared with the classical paired t-test and the Wilcoxon sign-ranked test, denoted as W_n, with a sample size $n = 200$. As expected, both t-test and Wilcoxon sign-ranked test lose powers when X and Y differ in distribution but have the same location. In contrast, $\hat{V}_n$ has remarkable performance in detecting these differences.

For the multivariate cases, we consider the dimensions of X and Y to be $p = 2$ and $p = 5$. We firstly draw (X, Y_0) from the multivariate normal distribution $N_{2d}(0,\Sigma)$, where Σ is a matrix with elements equal to 1 on the diagonal and 0.3 everywhere else, then define Y as the following nine cases (1) $Y = Y_0$; (2) $Y = Y_0 + 0.5$; (3) $Y = Y_0 + 1$; (4) $Y = Y_0 + 1.5$; (5) $Y = \sqrt{2}Y_0$; (6) $Y = \sqrt{3}Y_0$; (7) $Y = 2Y_0$; (8) $Y = (Y^{(1)}, Y^{(2)}, \ldots, Y^{(p)})'$ with $Y^{(j)} = 2\Phi(Y_0^{(j)}) - 1, j =$

Table 1. Empirical sizes and powers of $\hat{V}_n$ and $\hat{T}_n$ for the five cases in **S.1**

Case	H_0	Test	$n=30$	$n=50$	$n=100$	$n=150$	$n=200$
1	$X \overset{d}{=} Y$	$\hat{V}_n$	0.032	0.050	0.032	0.046	0.043
	$X\|Z \overset{d}{=} Y\|Z$	$\hat{T}_n$	0.045	0.067	0.048	0.044	0.042
2	$X \overset{d}{=} Y$	$\hat{V}_n$	0.035	0.053	0.043	0.043	0.044
	$X\|Z \overset{d}{=} Y\|Z$	$\hat{T}_n$	0.153	0.252	0.495	0.692	0.819
3	$X \overset{d}{=} Y$	$\hat{V}_n$	0.040	0.038	0.055	0.048	0.044
	$X\|Z \overset{d}{=} Y\|Z$	$\hat{T}_n$	0.734	0.925	0.999	1.000	1.000
4	$X \overset{d}{=} Y$	$\hat{V}_n$	0.041	0.056	0.049	0.052	0.046
	$X\|Z \overset{d}{=} Y\|Z$	$\hat{T}_n$	1.000	1.000	1.000	1.000	1.000
5	$X \overset{d}{=} Y$	$\hat{V}_n$	0.035	0.038	0.030	0.040	0.054
	$X\|Z \overset{d}{=} Y\|Z$	$\hat{T}_n$	0.996	1.000	1.000	1.000	1.000

Table 2. Empirical sizes and powers of $\hat{V}_n$, compared with the classical t-test and the Wilcoxon sign-ranked test in **S.2** *with a sample size $n = 200$*

X	Y	$\hat{V}_n$	t	W_n
$N(0.5,1)$	$N(0.5,1)$	0.054	0.053	0.054
	$N(1,1)$	1.000	1.000	1.000
	$N(0.5,2)$	1.000	0.050	0.041
	$Exp(2)$	1.000	0.054	0.054
	$Ga(2,4)$	1.000	0.058	0.051
	$Beta(2,2)$	1.000	0.053	0.053
	$U(0,1)$	1.000	0.057	0.054
$Exp(2)$	$Exp(2)$	0.048	0.056	0.052
	$Ga(2,4)$	0.995	0.058	0.344
	$Beta(2,2)$	1.000	0.068	0.646
	$U(0,1)$	1.000	0.056	0.445
$Ga(2,4)$	$Ga(2,4)$	0.054	0.062	0.052
	$Beta(2,2)$	1.000	0.058	0.245
	$U(0,1)$	0.992	0.062	0.095
$Beta(2,2)$	$Beta(2,2)$	0.050	0.051	0.052
	$U(0,1)$	0.994	0.055	0.046
$U(0,1)$	$U(0,1)$	0.051	0.048	0.042
$B(5,0.4)$	$B(5,0.4)$	0.051	0.056	0.056
	$NB(2,0.5)$	1.000	0.038	0.111
	$P(2)$	0.869	0.044	0.048
$NB(2,0.5)$	$NB(2,0.5)$	0.050	0.060	0.053
	$P(2)$	0.992	0.050	0.077
$P(2)$	$P(2)$	0.044	0.053	0.051

$1,2,\ldots,p$, where $\Phi(\cdot)$ is the standard normal distribution function; (9) The same setting as in (8) except that $Y^{(j)} = (Y_0^{(j)})^2, j = 1,2,\ldots,p$. Table 3 shows the performance of $\hat{V}_n$, compared with the classical Hotelling's T^2 test. As expected from the theory, Hotelling's T^2 performs better in detecting the location shift of normal distributions in cases (1)–(4), however it seems to lose powers in cases (5)–(9), where X and Y differ in distribution but have the same means.

In **S.3**, we investigate the performance of $\hat{T}_n$ when testing the equality of two conditional distributions for paired samples. To implement $\hat{T}_n$, in all of the following studies, we choose the kernel function $\mathcal{K}_H(.)$ to be the normal density, and set the bandwidths $h_k = \hat{\sigma}_{Z^{(k)}}(\frac{4}{(r+2)n})^{\frac{1}{r+4}}$, $k = 1,2,\ldots,r$ following the Silverman's rule of thumb, with $\hat{\sigma}_{Z^{(k)}}$ being the sample standard deviation of the k-th coordinative of Z.

Firstly, we consider the case that X, Y are both univariate, i.e. $p = 1$. For the sake of comparison, the test proposed in Guo et al. [2] with the form

$$D_{n,\beta} := \frac{1}{n^2}\sum_{i,j=1}^{n}(X_i - Y_i)(X_j - Y_j)\exp\left(-|Z_i - Z_j|^{\beta}\right), \quad 0 < \beta \le 2,$$

is considered. Suppose $Z = (Z_1,\ldots,Z_r) \sim N_r(0, I_r)$, $\varepsilon_1 \sim N(0,1), \varepsilon_2 \sim N(0,1), u \sim U(-1,1)$, and $Z, \varepsilon_1, \varepsilon_2, u$ are independent, let $X = \beta^T Z + \varepsilon_1$, where $\beta = (1,1,\ldots,1)/\sqrt{r}$. Then Y is defined as follows:

- **Case 1** $Y = \beta^T Z + \varepsilon_2$;
- **Case 2** $Y = \beta^T Z + \varepsilon_2 + 0.2$;
- **Case 3** $Y = \beta^T Z + \varepsilon_2 + 0.4$;
- **Case 4** $Y = \beta^T Z + 0.3\varepsilon_2$;
- **Case 5** $Y = \beta^T Z + 2\varepsilon_2 + 0.3$;
- **Case 6** $Y = \beta^T Z + u$.

$r = 1,2,3$ are considered for the dimensions of Z.

Table 4 presents the empirical sizes and powers of $\hat{T}_n$ and $D_{n,\beta}(\beta = 0.5, 1, 1.5)$ with samples $n = 100, 200$, where the

Table 3. Empirical sizes and powers of $\hat{V}_n$ compared with the Hotelling's T^2 test for multivariate X and Y in **S.2**

Case	Test	$p=2$					$p=5$				
		$n=30$	$n=50$	$n=100$	$n=150$	$n=200$	$n=30$	$n=50$	$n=100$	$n=150$	$n=200$
1	$\hat{V}_n$	0.027	0.031	0.042	0.043	0.043	0.021	0.024	0.037	0.049	0.042
	Hotelling's T^2	0.051	0.040	0.049	0.052	0.049	0.063	0.050	0.044	0.056	0.054
2	$\hat{V}_n$	0.692	0.930	0.999	1.000	1.000	0.935	1.000	1.000	1.000	1.000
	Hotelling's T^2	0.803	0.966	1.000	1.000	1.000	0.969	1.000	1.000	1.000	1.000
3	$\hat{V}_n$	0.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Hotelling's T^2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
4	$\hat{V}_n$	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Hotelling's T^2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
5	$\hat{V}_n$	0.120	0.341	0.793	0.968	0.997	0.189	0.597	0.993	1.000	1.000
	Hotelling's T^2	0.048	0.044	0.048	0.049	0.045	0.057	0.049	0.048	0.059	0.052
6	$\hat{V}_n$	0.483	0.868	1.000	1.000	1.000	0.821	0.998	1.000	1.000	1.000
	Hotelling's T^2	0.051	0.047	0.045	0.050	0.045	0.056	0.055	0.047	0.056	0.046
7	$\hat{V}_n$	0.781	0.991	1.000	1.000	1.000	0.988	1.000	1.000	1.000	1.000
	Hotelling's T^2	0.044	0.048	0.047	0.049	0.046	0.058	0.054	0.046	0.057	0.047
8	$\hat{V}_n$	0.384	0.809	0.998	1.000	1.000	0.716	0.996	1.000	1.000	1.000
	Hotelling's T^2	0.042	0.039	0.059	0.045	0.047	0.048	0.053	0.047	0.050	0.045
9	$\hat{V}_n$	0.243	0.533	0.963	1.000	1.000	0.135	0.414	0.968	1.000	1.000
	Hotelling's T^2	0.074	0.078	0.051	0.052	0.054	0.071	0.079	0.075	0.058	0.062

p-values of $D_{n,\beta}$ are based on 399 times of Wild bootstrap used in Guo et al. [2]. From Table 4, we see that all of the tests control the Type-I errors well around 0.05 in **Case 1**, where X and Y are conditional identically distributed given Z. But the Type-I errors of $D_{n,\beta}$ are sometimes little lower than the nominal level when $r=3$. As expected, $D_{n,\beta}(\beta=0.5,1,1.5)$ perform better than $\hat{T}_n$ in **Case 2** and **Case 3**, where the conditional distributions of X and Y given Z only differ in location. It's not surprising since $D_{n,\beta}$ is a consistent test for detecting the difference between the two conditional means. In **Case 4** and **Case 6**, where the conditional distributions of X and Y are distinct but the mean is the same, $D_{n,\beta}(\beta=0.5,1,1.5)$ lose powers completely, whereas, our test $\hat{T}_n$ detects the distributional difference with desirable powers. In **Case 5**, where X and Y differ in mean and variance given Z, $\hat{T}_n$ and $D_{n,\beta}(\beta=0.5,1,1.5)$ can identify the differences well, but $\hat{T}_n$ is far superior to $D_{n,\beta}(\beta=0.5,1,1.5)$.

We also consider the cases that X and Y are both discrete variables. Two types of distributions are considered for $Z=(Z_1,Z_2,\ldots,Z_r)$. One is the normal distribution, with mean 0 and variance of 3 for $r=1$, and the multivariate normal distribution with mean vector of 0 and covariance matrix $\Sigma=3I_r$ for $r=2,3$. Another one is the r-variate distribution with i.i.d. uniform distribution on $[-2,2]$ marginals, denoted by $U[-2,2]^r$. Conditioning on Z, we generate X and Y from the Bernoulli distributions with $P(X=1|Z)=\Phi((Z_1+Z_2+\ldots+Z_r)/r)$ and $P(Y=1|Z)=\Phi(a(Z_1+Z_2+\ldots+Z_r)/r)$, respectively. We set $a=1,0.1,0.2,0.3,0.4$ to vary the difference between $P(X=1|Z)$ and $P(Y=1|Z)$. Results are presented in Table 5.

Table 5 reveals that $\hat{T}_n$ can control the Type-I errors well around 0.05 when $a=1$. The powers of $\hat{T}_n$ share a decreasing trend with the increase of a in the sense that the difference between X and Y reduces as a gets close to 1. Meanwhile, the powers of $\hat{T}_n$ decrease sharply as the dimension of Z increases from 1 to 3, which is a common phenomenon of the kernel-based test.

Next, we consider the cases that X, Y are multivariate. Suppose $Z\in\mathbb{R}^2, X_0\in\mathbb{R}^p, Y_0\in\mathbb{R}^p$. We generate (Z,X_0,Y_0) from the multivariate normal distribution $N_{2+2p}(0,\Sigma)$ with $\Sigma=(\sigma_{ij})_{(2+2p)\times(2+2p)}$, $\sigma_{ij}=1, i=j$ else $\sigma_{ij}=0.3^2, i\neq j, i,j=1,2,\ldots,2+2p$. Then we consider the following two settings:

- **Setting 1** $X=X_0,\ Y=a+\sqrt{1+a}Y_0$.
- **Setting 2** $X=\exp(X_0),\ Y=\exp(a+\sqrt{1+a}Y_0)$.

Let $a=0,0.15,0.30,0.45,0.60$ in both settings, and $p=2,4,8$ are considered for the dimensions of X and Y. Table 6 shows the performance of $\hat{T}_n$ with samples $n=100,200$. From Table 6, we find that $\hat{T}_n$ controls the Type-I errors reasonably around 0.05 when $a=0$, and it becomes more and more powerful as a increases from 0.15 to 0.6.

Finally, we are interested in the performance of $\hat{T}_n$ with a large sample size, say $n=10000$. Let $X=Z_1+Z_2+\varepsilon_1, Y=Z_1+Z_2+aZ_1Z_2+(1+aZ_1)\varepsilon_2$, where $Z=(Z_1,Z_2)\sim$

Table 4. Empirical sizes and powers of $\hat{T}_n$ compared with the test $D_{n,\beta}((0<\beta\leq 2)$ for univariate X and Y in **S.3**

Cases	Test	$r=1$		$r=2$		$r=3$	
		$n=100$	$n=200$	$n=100$	$n=200$	$n=100$	$n=200$
Case 1	$\hat{T}_n$	0.045	0.049	0.043	0.040	0.059	0.046
	$D_{n,0.5}$	0.046	0.050	0.040	0.047	0.038	0.044
	$D_{n,1}$	0.048	0.051	0.040	0.043	0.029	0.034
	$D_{n,1.5}$	0.051	0.052	0.038	0.043	0.018	0.033
Case 2	$\hat{T}_n$	0.177	0.331	0.126	0.200	0.100	0.151
	$D_{n,0.5}$	0.250	0.463	0.275	0.468	0.258	0.453
	$D_{n,1}$	0.233	0.444	0.229	0.416	0.186	0.390
	$D_{n,1.5}$	0.221	0.427	0.194	0.370	0.131	0.296
Case 3	$\hat{T}_n$	0.610	0.883	0.468	0.771	0.313	0.577
	$D_{n,0.5}$	0.777	0.970	0.773	0.973	0.766	0.972
	$D_{n,1}$	0.750	0.961	0.722	0.957	0.681	0.951
	$D_{n,1.5}$	0.732	0.954	0.663	0.933	0.540	0.901
Case 4	$\hat{T}_n$	1.000	1.000	0.996	1.000	0.918	1.000
	$D_{n,0.5}$	0.037	0.044	0.040	0.052	0.030	0.046
	$D_{n,1}$	0.037	0.046	0.032	0.045	0.025	0.045
	$D_{n,1.5}$	0.041	0.050	0.029	0.040	0.020	0.050
Case 5	$\hat{T}_n$	0.942	1.000	0.799	0.995	0.557	0.944
	$D_{n,0.5}$	0.221	0.440	0.235	0.432	0.231	0.422
	$D_{n,1}$	0.211	0.412	0.205	0.380	0.173	0.356
	$D_{n,1.5}$	0.209	0.393	0.182	0.335	0.136	0.268
Case 6	$\hat{T}_n$	0.465	0.919	0.257	0.617	0.168	0.380
	$D_{n,0.5}$	0.038	0.050	0.039	0.041	0.039	0.040
	$D_{n,1}$	0.047	0.050	0.041	0.041	0.031	0.038
	$D_{n,1.5}$	0.045	0.046	0.037	0.038	0.031	0.041

Table 5. Empirical sizes and powers of $\hat{T}_n$ with sample sizes $n=100,200$ for discrete X and Y in **S.3**

Z	a	$r=1$		$r=2$		$r=3$	
		$n=100$	$n=200$	$n=100$	$n=200$	$n=100$	$n=200$
$N_r(0,3I_r)$	1	0.057	0.042	0.052	0.050	0.041	0.047
	0.1	0.967	1.000	0.694	0.966	0.361	0.719
	0.2	0.863	0.998	0.546	0.873	0.268	0.521
	0.3	0.681	0.966	0.380	0.698	0.189	0.352
	0.4	0.449	0.813	0.244	0.457	0.127	0.220
$U[-2,2]^r$	1	0.049	0.053	0.049	0.048	0.053	0.057
	0.1	0.974	1.000	0.635	0.919	0.285	0.584
	0.2	0.906	1.000	0.471	0.813	0.217	0.441
	0.3	0.767	0.989	0.352	0.653	0.155	0.315
	0.4	0.577	0.906	0.235	0.455	0.121	0.227

$N_2(0,\Sigma)$ with $\Sigma=(\sigma_{ij})_{r\times r}$, $\sigma_{ij}=1, i=j$ else $\sigma_{ij}=0.6, i\neq j, i,j=1,2$; $\varepsilon_1,\varepsilon_2 \overset{i.i.d}{\sim} N(0,1)$, independent with Z, and a is a constant. Set $a=0,0.03,0.06,0.09,0.12,0.15$. Table 7 reports the empirical sizes and powers of $\hat{T}_n$ with a sample size $n=10000$. In this table, we see that $\hat{T}_n$ can detect accurately the subtle difference between the two conditional distributions and enjoys growing powers as a increases, as well as controls the Type-I error reasonably around 0.05 when $a=0$.

5. REAL DATA ANALYSIS

In this section, we use the Student Performance dataset to illustrate our proposed tests. Student achievement in secondary education of two Portuguese schools during the year 2005–2006 were approached in the dataset, which was collected using school reports and questionnaires. The original data is available from UCI Machine Learning Repository http://archive.ics.uci.edu/ml/datasets/Student+Performance. Attributes of the data in-

Table 6. Empirical sizes and powers of $\hat{T}_n$ with sample sizes $n = 100, 200$ for multivariate X and Y in **Setting 1** *and* **Setting 2**, **S.3**

	a	**Setting 1**		**Setting 2**	
		$n = 100$	$n = 200$	$n = 100$	$n = 200$
$p = 2$	0	0.050	0.046	0.040	0.038
	0.15	0.095	0.178	0.144	0.286
	0.30	0.399	0.747	0.567	0.886
	0.45	0.807	0.987	0.904	0.999
	0.60	0.969	1.000	0.995	1.000
$p = 4$	0	0.048	0.044	0.048	0.042
	0.15	0.155	0.271	0.272	0.491
	0.30	0.640	0.941	0.821	0.992
	0.45	0.959	1.000	0.994	1.000
	0.60	0.999	1.000	1.000	1.000
$p = 8$	0	0.046	0.041	0.049	0.050
	0.15	0.213	0.450	0.415	0.753
	0.30	0.863	0.998	0.983	1.000
	0.45	0.999	1.000	1.000	1.000
	0.60	1.000	1.000	1.000	1.000

Table 7. Empirical sizes and powers of $\hat{T}_n$ with a sample size $n = 10000$ in **S.3**

a	0	0.03	0.06	0.09	0.12	0.15
	0.044	0.085	0.370	0.910	1.000	1.000

clude student grades, demographic, social, and school-related features. Two datasets are provided regarding the performance in two distinct subjects: Mathematics and Portuguese language. In each subject, students were evaluated in three periods during the school year and got three corresponding grades denoted by G_1, G_2, and G_3 (the final grade). Cortez and Silva [1] used the data to predict student performance (G_3, the final grade) and analysed the factors that affect student achievement via data mining approaches. Different models and different data mining approaches showed the importance of previous grades G_1 and G_2 on predicting the final grade G_3.

In this work, we aim to test the difference among the three evaluations. We use the merged data consisting of the three grades for both Mathematics and Portuguese language of 382 students. Each grade is standardized to eliminate the influence of different teachers who gave the scores. We use $zG_1, zG2, zG3$ to denote the standardized joint grades of the two subjects at baseline, the second, and the last evaluations, respectively. Figure 1 shows the scatter plots of Mathematics and Portuguese language grades for the three times evaluations, indicating the evident positive correlation of these two subjects. Moreover, in the baseline evaluation, the students' scores in both subjects were uniformly distributed, while in the second evaluation, students with unusually low scores in Mathematics were clearly isolated. In the last evaluation, students with unusually low scores

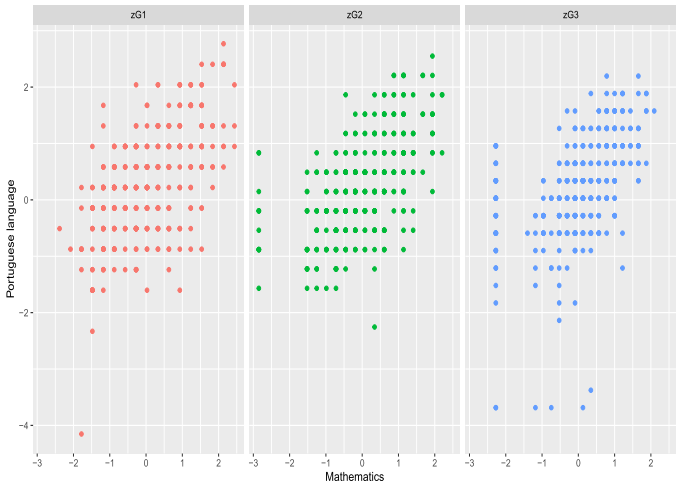

Figure 1. Scatter plots of Mathematics and Portuguese language for the three times of evaluations.

in Portuguese language were also clearly separated, with the grades of the remaining students being distributed more compact than those of $zG1$ and $zG2$. In summary, there exist apparent differences among the three evaluations. If we mistakenly ignore the dependence among the three grades, and use the *eqdist.etest()* in the R package *energy* [18] to test the pairwise differences among the three evaluations, we will get the resulting p-values of 0.316 for $zG1$ and $zG2$, 0.132 for $zG2$ and $zG3$, and 0.018 for $zG1$ and $zG3$, respectively. Consequently, we will conclude incorrectly no significant difference between $zG1$ and $zG2$, nor was there a significant difference between $zG2$ and $zG3$, while the difference only exists between $zG1$ and $zG3$.

Table 8 lists the bootstrap p-values of $n\hat{V}_n$ for testing the pairwise differences among the three evaluations on Mathematics and Portuguese language. The small p-values show that $n\hat{V}_n$ correctly detects the pairwise differences among the three evaluations, which coincides with what Figure 1 shows.

Finally, we test the difference between the last two evaluations of performances conditioning on the first performance to examine whether the students' basic level affects the second and the last performances in different ways. In doing so, we set the bandwidths by the same approach as in the previous simulation studies. The resulting test statistic $\hat{T}_n = 14.5483$ and the corresponding p-value is 5.98721×10^{-48}. Consequently, we reject the null hypothesis that the second and last performances are identically distributed given the first performance. This means that there is a significant difference between the second and last performances at the same basic level. This is not difficult to understand, because the dependency of the last performance on the basic level (that is the first performance) is not as strong as that of the second performance on the basic level.

Table 8. p-values for testing the differences among the three evaluations on Mathematics and Portuguese language. Each p-value is approximated based on $B = 399$ times of bootstrap resamples

Compared grades	$n\hat{V}_n$	p-values
zG_1 v.s. zG_2	3.715160	0.0025
zG_2 v.s. zG_3	5.591203	0.0025
zG_1 v.s. zG_3	10.61590	0.0025

6. SUMMARY AND DISCUSSION

In this paper, we discuss the problems of testing the homogeneity or conditional homogeneity of paired samples, formulated as $H_0 : X \overset{d}{=} Y$ and $H_0 : X|Z = z \overset{d}{=} Y|Z = z$, for all $z \in S(Z)$ respectively. We first introduce the modified version of energy distance for two paired random vectors and provide the test statistic $\hat{V}_n$ for the paired-sample test. We also present a bootstrap method to approximate the p-value of $\hat{V}_n$ and prove the validity of the bootstrap. Compared with the energy distance for two independent vectors and its application, $\hat{V}_n$ only differs in the bootstrap procedure. We further extend the concept of energy distance to conditional energy distance, and then derive the test statistic $\hat{T}_n$ to deal with the conditional homogeneity test. We have proved that $\hat{T}_n$ is asymptotically normal under both the null and alternative hypotheses as well as its consistency.

We adopt Euclidean distance for simplicity here, but the distance $|\cdot|$ involved in $\hat{V}_n$ and $\hat{T}_n$ can be extended for $|\cdot|^\alpha$ for $\alpha \in (0, 2)$. Proper α may potentially improve the power of the two tests with the Type-I error under control. And how to choose a proper α in practice is worth further investigation.

ACKNOWLEDGEMENT

Dr. Tian's research is partially supported by the National Natural Science Foundation of China (71991474, 12001554), the Fundamental Research Funds for the Central Universities, Sun Yat-sen University (2021qntd21, 19lgpy236), and the Natural Science Foundation of Guangdong Province, China (2021A1515010205, 2020A1515010617).

Dr. Pan's research is partially supported by the National Natural Science Foundation of China (12071494), and the Science and Technology Program of Guangzhou, China (202102080481).

Dr. Wang's research is partially supported by the National Natural Science Foundation of China (11771462), the National Key Research and Development Program of China (2018YFC1315401), the Science and Technology Program of Guangzhou, China (202002030129) and the Key Research and Development Program of Guangdong, China (2019B020228001).

APPENDIX A. TECHNICAL DETAILS

The following conditions (C1)–(C3) are assumed for the kernel function $\mathcal{K}(\mathbf{u})$ in Section 3.

(C1) $\int_{\mathbb{R}^r} \mathbf{u}\mathcal{K}(\mathbf{u})d\mathbf{u} = 0, \int_{\mathbb{R}^r} \mathcal{K}(\mathbf{u})d\mathbf{u} = 1, \int_{\mathbb{R}^r} |\mathcal{K}(\mathbf{u})|d\mathbf{u} < \infty, \int_{\mathbb{R}^r} \mathcal{K}^2(\mathbf{u})d\mathbf{u} < \infty, \int_{\mathbb{R}^r} |\mathbf{u}|^2\mathcal{K}(\mathbf{u})d\mathbf{u} < \infty$.

(C2) $|H| \to 0$ and $n|H| \to \infty$, as $n \to \infty$. This requires $h_1, h_2, \ldots, h_r$ to be chosen appropriately according to n.

(C3) The density function of Z and the conditional joint density of (X, Y) given $Z = z$, denoted by $f(x, y|z)$, are twice differentiable and all of the derivatives are bounded.

The following lemma gives an equivalent moment condition for X and Y to be identically distributed.

Lemma 1. *For a paired random vector (X, Y) in $\mathbb{R}^p \times \mathbb{R}^p$, $V(X, Y)$ is defined in (3) if their norm moments $E|X| < \infty$ and $E|Y| < \infty$, then $V(X, Y) \geq 0$ and the equality holds if and only if X and Y are identically distributed.*

Lemma 1 is straightforward since $V(X, Y)$ is indeed the energy distance $\varepsilon(X, Y')$ of two independent random vectors X and Y' proposed in [22]. Hence, $V(X, Y) \geq 0$ and it equals to zero if and only if X and Y' are identically distributed, i.e. X and Y are identically distributed. Lemma 1 states a discrepancy between two random variables can be expressed as a moment form.

Next, we investigate the statistical properties of $\hat{V}_n(X, Y)$ in (4). It is clear that $\hat{V}_n(X, Y)$ is an one-sample V-statistic, so we could derive its asymptotic behaviour expediently.

Proposition 1. *Suppose $E|X| < \infty$ and $E|Y| < \infty$, then $\hat{V}_n(X, Y)$ converges to $V(X, Y)$ almost surely, that is,*

$$\hat{V}_n(X, Y) \xrightarrow{a.s.} V(X, Y).$$

The proof of Proposition 1 is straightforward according to Theorem 3 in Chapter 3 of Lee [8] and thus omitted.

Proposition 2. *Suppose $E|X|^2 < \infty, E|Y|^2 < \infty$, then under the null hypothesis that X, Y are identically distributed, $n\hat{V}_n(X, Y)$ converges to a limit distribution in law, that is,*

$$n\hat{V}_n(X, Y) \rightsquigarrow \sum_{v=1}^{\infty} \lambda_v \mathcal{Z}_v^2, \tag{9}$$

where $\mathcal{Z}_v \sim N(0, 1), i.i.d.$, and λ_v's are non-negative constants that depend on the joint distribution of (X, Y).

Proof of Proposition 2. Denote $F(x, y)$ as the joint distribution function of (X, Y), and

$$g((x, y), (x', y')) = |x - y'| + |x' - y| - |x - x'| - |y - y'|,$$

then

$$\hat{V}_n(X, Y) = \frac{1}{n^2} \sum_{i,j=1}^{n} g((X_i, Y_i), (X_j, Y_j)).$$

Since $E[g((X,Y),(X',Y'))\big|(X,Y)] = 0$ under the null hypothesis that X and Y are identically distributed. $\hat{V}_n(X,Y)$ is a degenerate V-statistic of order 1. According to [8], we have

$$n\hat{V}_n(X,Y) \rightsquigarrow \sum_v \lambda_v \mathcal{Z}_v^2,$$

where λ_v's are the eigenvalues of the integral equation

$$\int g((x,y),(x',y'))\rho(x,y)dF(x,y) = \lambda\rho(x',y'),$$

and $\mathcal{Z}_v$'s are independent standard normal random variables. Moreover, since X and Y are identically distributed, which implies $F(x,y) = F(y,x)$ for any $(x,y) \in \mathbb{R}^p \times \mathbb{R}^p$, therefore, λ_v's are also the eigenvalues of the integral equation

$$\begin{aligned}(10)\quad &\int g((x,y),(x',y'))\rho(x,y)d\big(1/2F(x,y) + 1/2F(y,x)\big)\\ &= \lambda\rho(x',y').\end{aligned}$$

Proof of Theorem 1. Conditional on the original sample $\mathcal{S}_n$, $(X_i^\star, Y_i^\star), i = 1,2,\ldots,n$ are independent and identically distributed on the points

$$\{(X_1,Y_1),(Y_1,X_1),\ldots,(X_n,Y_n),(Y_n,X_n)\}$$

with equal probability of $\frac{1}{2n}$. Let $E^\star(\cdot)$ be $E(\cdot|\mathcal{S}_n)$, then for the V-statistic

$$\hat{V}_n^\star = \frac{1}{n^2}\sum_{i,j=1}^n g((X_i^\star,Y_i^\star),(X_j^\star,Y_j^\star)),$$

we have

$$\begin{aligned}&E^\star[g((X_1^\star,Y_1^\star),(X_2^\star,Y_2^\star))\big|(X_1^\star,Y_1^\star)]\\ &= \frac{1}{2n}\sum_{j=1}^n g((X_1^\star,Y_1^\star),(X_j,Y_j))\\ &+ \frac{1}{2n}\sum_{j=1}^n g((X_1^\star,Y_1^\star),(Y_j,X_j))\\ &= 0,\end{aligned}$$

due to $g((x,y),(x',y'))+g((x,y),(y',x')) \equiv 0$, and therefore

$$\begin{aligned}&E^\star[g((X_1^\star,Y_1^\star),(X_2^\star,Y_2^\star))]\\ &= E^\star\Big\{E^\star[g((X_1^\star,Y_1^\star),(X_2^\star,Y_2^\star))\big|(X_1^\star,Y_1^\star)]\Big\}\\ &= 0.\end{aligned}$$

Thus, $\hat{V}_n^\star$ is a degenerate V-statistic of order 1. Again, according to [8], we have

$$n\hat{V}_n^\star \rightsquigarrow \sum_v \lambda_v \mathcal{Z}_v^2,$$

where $\mathcal{Z}_v$'s are independent standard normal random variables, and λ_v's are the eigenvalues of the integral equation

$$\int g((x,y),(x',y'))\rho(x,y)dF^\star(x,y) = \lambda\rho(x',y'),$$

with $F^\star(x,y)$ being the limit joint distribution function of $(X_1^\star, Y_1^\star)$. Obviously

$$\begin{aligned}F^\star(x,y) &= \lim_{n\to\infty} 1/2(F_n(x,y) + F_n(y,x))\\ &= 1/2F(x,y) + 1/2F(y,x).\end{aligned}$$

Hence λ_v's are exact the eigenvalues of the integral equation (10), which implies that $n\hat{V}_n^\star$ converges in distribution to the null distribution of the test $n\hat{V}_n(X,Y)$ in Proposition 2.

The following lemma states that $\varepsilon(X,Y|z)$ in (5) has an equivalent expression.

Lemma 2. *$\varepsilon(X,Y|z)$ can be rewritten in the form of*

$$\varepsilon(X,Y|z) = E[g((X_1,Y_1),(X_2,Y_2))|Z_1 = Z_2 = z].$$

Proof of Lemma 2. Given the event $Z = z$, we have

$$\begin{aligned}&|\phi_{X|Z=z}(t) - \phi_{Y|Z=z}(t)|^2\\ &= \phi_{X|Z=z}(t)\overline{\phi_{X|Z=z}(t)} + \phi_{Y|Z=z}(t)\overline{\phi_{Y|Z=z}(t)}\\ &- \phi_{X|Z=z}(t)\overline{\phi_{Y|Z=z}(t)} - \phi_{Y|Z=z}(t)\overline{\phi_{X|Z=z}(t)}\\ &= E[\exp(i\langle t, X_1 - X_2\rangle)\big|Z_1 = z, Z_2 = z]\\ &+ E[\exp(i\langle t, Y_1 - Y_2\rangle)\big|Z_1 = z, Z_2 = z]\\ &- E[\exp(i\langle t, X_1 - Y_2\rangle)\big|Z_1 = z, Z_2 = z]\\ &- E[\exp(i\langle t, Y_1 - X_2\rangle)\big|Z_1 = z, Z_2 = z]\\ &= 1 - E[\exp(i\langle t, X_1 - Y_2\rangle)\big|Z_1 = z, Z_2 = z]\\ &+ 1 - E[\exp(i\langle t, Y_1 - X_2\rangle)\big|Z_1 = z, Z_2 = z]\\ &- (1 - E[\exp(i\langle t, X_1 - X_2\rangle)\big|Z_1 = z, Z_2 = z])\\ &- (1 - E[\exp(i\langle t, Y_1 - Y_2\rangle)\big|Z_1 = z, Z_2 = z]).\end{aligned}$$

According to the equation

$$\int_{\mathbb{R}^p} \frac{1 - \exp(i\langle t, X\rangle)}{|t|^{p+1}}dt = c(p)|X|$$

in [22], we obtain

$$\begin{aligned}&\varepsilon(X,Y|Z = z)\\ &= \int_{\mathbb{R}^p} \frac{|\phi_{X|Z=z}(t) - \phi_{Y|Z=z}(t)|^2}{c(p)|t|^{p+1}}dt\\ &= E\Big[\int_{\mathbb{R}^p}\Big(\frac{1 - \exp(i\langle t, X_1 - Y_2\rangle)}{c(p)|t|^{p+1}}\\ &+ \frac{1 - \exp(i\langle t, Y_1 - X_2\rangle)}{c(p)|t|^{p+1}}\end{aligned}$$

$$
\begin{aligned}
&- \frac{1-\exp(i\langle t, X_1 - X_2\rangle)}{c(p)|t|^{p+1}} \\
&- \frac{1-\exp(i\langle t, Y_1 - Y_2\rangle)}{c(p)|t|^{p+1}}\Big) dt \Big| Z_1 = z, Z_2 = z\Big] \\
&= E\big[|X_1 - Y_2| + |X_2 - Y_1| - |X_1 - X_2| \\
&- |Y_1 - Y_2| \Big| Z_1 = z, Z_2 = z\big] \\
&= E[g((X_1, Y_1), (X_2, Y_2)) | Z_1 = Z_2 = z].
\end{aligned}
$$

Proof of Theorem 2. Let

$$
\begin{aligned}
&P_n(W_1, W_2) = g((X_1, Y_1), (X_2, Y_2))\mathcal{K}_H(Z_1 - Z_2), \\
&P_{n1}(W_i) = E(P_n(W_1, W_2)|W_i), i = 1, 2,
\end{aligned}
$$

and

$$
\sigma_{n1}^2 = Var(P_{n1}(W_1)), \quad \sigma_{n2}^2 = Var(P_n(W_1, W_2)).
$$

We use two steps to prove the consistency of $\hat{\mathcal{U}}_n$.

step 1: $\hat{\mathcal{U}}_n = E[\hat{\mathcal{U}}_n] + o_p(1)$.

By Lee [8], we have

$$
\begin{aligned}
Var(\hat{\mathcal{U}}_n) &= \frac{C_2^1 C_{n-2}^1}{C_n^2}\sigma_{n1}^2 + \frac{C_2^2 C_{n-2}^0}{C_n^2}\sigma_{n2}^2 \\
&= \frac{4(n-2)}{n(n-1)}\sigma_{n1}^2 + \frac{2}{n(n-1)}\sigma_{n2}^2,
\end{aligned}
$$

where

$$
\begin{aligned}
\sigma_{n1}^2 &= Var(P_{n1}(W_1)) \\
&= Var(E(P_n(W_1, W_2)|W_1)] \\
&\le Var(P_n(W_1, W_2)) \\
&\le EP_n^2(W_1, W_2), \\
\sigma_{n2}^2 &= Var(P_n(W_1, W_2)) \le EP_n^2(W_1, W_2),
\end{aligned}
$$

and

$$
\begin{aligned}
&EP_n^2(W_1, W_2) \\
&= \int \Big[g((x_1, y_1), (x_2, y_2))\mathcal{K}_H(z_2 - z_1)\Big]^2 \\
&\quad \cdot f(x_1, y_1, z_1) f(x_2, y_2, z_2) dx_1 dy_1 dz_1 dx_2 dy_2 dz_2 \\
&= |H|^{-1} \int \Big[g((x_1, y_1), (x_2, y_2))\mathcal{K}(z_{12})\Big]^2 \\
&\quad \cdot f(x_1, y_1, z_1) f(x_2, y_2, z_1 + Hz_{12}) dx_1 dy_1 dz_1 dx_2 dy_2 dz_{12} \\
&= O(\frac{1}{|H|}).
\end{aligned}
$$

Therefore, we get

$$
\begin{aligned}
Var(\hat{\mathcal{U}}_n) &= \frac{4(n-2)}{n(n-1)}\sigma_{n1}^2 + \frac{2}{n(n-1)}\sigma_{n2}^2 \\
&\le O(\frac{1}{n|H|}) + O(\frac{1}{n})O(\frac{1}{n|H|}) \\
&= o(1).
\end{aligned}
$$

So

$$
\hat{\mathcal{U}}_n = E[\hat{\mathcal{U}}_n] + o_p(1),
$$

by the Chebyshev's inequality.

step 2: $E\hat{\mathcal{U}}_n = E[\varepsilon(X, Y|Z)f_Z(Z)] + O(|H|^2)$.

It is easy to verify that

$$
\begin{aligned}
&E\hat{\mathcal{U}}_n = EP_n(W_1, W_2) \\
&= E[g((X_1, Y_1), (X_2, Y_2))\mathcal{K}_H(Z_1 - Z_2)] \\
&= |H|^{-1} \int g((x_1, y_1), (x_2, y_2))\mathcal{K}(H^{-1}(z_2 - z_1)) \\
&\quad \cdot f(x_1, y_1, z_1) f(x_2, y_2, z_2) dx_1 dx_2 dy_1 dy_2 dz_1 dz_2 \\
&= \int g((x_1, y_1), (x_2, y_2))\mathcal{K}(z_{12}) f(x_1, y_1|z_1) f_Z(z_1) \\
&\quad \cdot f_Z(z_1 + Hz_{12}) f(x_2, y_2|z_1 + Hz_{12}) dx_1 dx_2 dy_1 dy_2 dz_1 dz_{12} \\
&= \int g((x_1, y_1), (x_2, y_2))\mathcal{K}(z_{12}) f(x_1, y_1|z_1) f_Z^2(z_1) \\
&\quad \cdot f(x_2, y_2|z_1) dx_1 dx_2 dy_1 dy_2 dz_1 dz_{12} + O(|H|^2) \\
&= \int \Big[g((x_1, y_1), (x_2, y_2)) f(x_1, y_1|z_1) \\
&\quad \cdot f(x_2, y_2|z_1) dx_1 dx_2 dy_1 dy_2\Big] f_Z^2(z_1) dz_1 + O(|H|^2) \\
&= \int E[g((X_1, Y_1), (X_2, Y_2))|Z_1 = z, Z_2 = z] f_Z^2(z) dz \\
&\quad + O(|H|^2) \\
&= E[\varepsilon(X, Y|Z)f_Z(Z)] + O(|H|^2).
\end{aligned}
$$

Thus, combining the results in step 1 and step 2, we finally obtain

$$
\hat{\mathcal{U}}_n \xrightarrow{P} \mathcal{U} = E[\varepsilon(X, Y|Z)f_Z(Z)].
$$

Proof of Theorem 3. We can rewrite $\hat{\mathcal{U}}_n$ as

$$
\hat{\mathcal{U}}_n = \frac{1}{C_n^2 |H|^{1/2}} \sum_{i<j} \varphi_n(W_i, W_j),
$$

where

$$
\begin{aligned}
\varphi_n(W_i, W_j) &= |H|^{1/2} P_n(W_i, W_j) \\
&= |H|^{1/2} g((X_1, Y_1), (X_2, Y_2))\mathcal{K}_H(Z_i - Z_j).
\end{aligned}
$$

We use Theorem 1 in Hall [3] to derive the asymptotic distribution of $\hat{\mathcal{U}}_n$ under H_0 in the following steps.

Step 1: Under H_0, $E(\varphi_n(W_1, W_2)|W_1) = 0$.

Note that

$$
\begin{aligned}
&E(\varphi_n(W_1, W_2)|W_1) \\
&= |H|^{1/2} E(P_n(W_1, W_2)|W_1) \\
&= |H|^{-1/2} \int g((x_1, y_1), (x_2, y_2))\mathcal{K}(H^{-1}(z_2 - z_1)) \\
&\quad \cdot f(x_2, y_2, z_2) dx_2 dy_2 dz_2
\end{aligned}
$$

$$
\begin{aligned}
&= |H|^{1/2} \int g((x_1,y_1),(x_2,y_2))\mathcal{K}(z_{12}) \\
&\quad \cdot f_Z(z_1+Hz_{12}) f(x_2,y_2|z_1+Hz_{12}) dx_2 dy_2 dz_{12} \\
&= |H|^{1/2} \int \Big\{ \int g((x_1,y_1),(x_2,y_2)) \\
&\quad \cdot f(x_2,y_2|z_1+Hz_{12}) dx_2 dy_2 \Big\} \mathcal{K}(z_{12}) f_Z(z_1+Hz_{12}) dz_{12} \\
&= 0.
\end{aligned}
$$

Therefore, we also have

$$
E\varphi_n(W_1,W_2) = E[E(\varphi_n(W_1,W_2)|W_1)] = 0.
$$

Step 2: As $n \longrightarrow \infty$,

$$
\frac{E[G_n^2(W_1,W_2)] + n^{-1}E[\varphi_n^4(W_1,W_2)]}{(E[\varphi_n^2(W_1,W_2)])^2} \longrightarrow 0,
$$

where

$$
\begin{aligned}
G_n(w_1,w_2) &= E[\varphi_n(w_1,W)\varphi_n(w_2,W)] \\
&= |H| E[P_n(w_1,W)P_n(w_2,W)].
\end{aligned}
$$

Obviously,

$$
\begin{aligned}
&E[\varphi_n^2(W_1,W_2)] \\
&= |H| \int \Big[g((x_1,y_1),(x_2,y_2)) \mathcal{K}_H(z_2-z_1) \Big]^2 \\
&\quad \cdot f(x_1,y_1,z_1) f(x_2,y_2,z_2) dx_1 dy_1 dz_1 dx_2 dy_2 dz_2 \\
&= \int \Big[g((x_1,y_1),(x_2,y_2)) \mathcal{K}(z_{12}) \Big]^2 \\
&\quad \cdot f(x_1,y_1,z_1) f(x_2,y_2,z_1+Hz_{12}) dx_1 dy_1 dz_1 dx_2 dy_2 dz_{12} \\
&= \int \mathcal{K}^2(u) du \int \big[g((x_1,y_1),(x_2,y_2)) \big]^2 \\
&\quad \cdot f(x_1,y_1,z) f(x_2,y_2,z) dx_1 dx_2 dy_1 dy_2 dz + O(|H|).
\end{aligned}
$$

Therefore $(E[\varphi_n^2(W_1,W_2)])^2 = O(1)$.

Analogously to $E[\varphi_n^2(W_1,W_2)]$, we can obtain that $E[\varphi_n^4(W_1,W_2)] = O(\frac{1}{|H|})$. Furthermore,

$$
\begin{aligned}
&E[G_n^2(W_1,W_2)] \\
&\quad = |H|^2 E\Big[\int (g((x_1,y_1),(x,y)) g((x_2,y_2),(x,y)) \\
&\quad \mathcal{K}_H(Z_1-z)\mathcal{K}_H(Z_2-z) f(x,y,z) dx dy dz \Big]^2 \\
&\quad = \frac{1}{|H|^2} \int \Big[\int g((x_1,y_1),(x,y)) g((x_2,y_2),(x,y)) \\
&\quad \mathcal{K}(H^{-1}(z-z_1)) \mathcal{K}(H^{-1}(z-z_2)) f(x,y,z) dx dy dz \Big]^2 \\
&\quad \cdot f(x_1,y_1,z_1) f(x_2,y_2,z_2) dx_1 dx_2 dy_1 dy_2 dz_1 dz_2 \\
&\quad = \int \Big[\int g((x_1,y_1),(x,y)) g((x_2,y_2),(x,y)) \\
&\quad \mathcal{K}(z_3)\mathcal{K}(H^{-1}(z_1-z_2)+z_3) f(x,y,z_1+Hz_3) dx dy dz_3 \Big]^2 \\
&\quad \cdot f(x_1,y_1,z_1) f(x_2,y_2,z_2) dx_1 dx_2 dy_1 dy_2 dz_1 dz_2 \\
&\quad = |H| \int \Big[\int g((x_1,y_1),(x,y)) g((x_2,y_2),(x,y)) \\
&\quad K(z_3)\mathcal{K}(z_4+z_3) f(x,z_2+H(z_3+z_4)) dx dy dz_3 \Big]^2 \\
&\quad \cdot f(x_1,y_1,z_2+Hz_4) f(x_2,y_2,z_2) dx_1 dx_2 dy_1 dy_2 dz_2 dz_4 \\
&\quad = O(|H|).
\end{aligned}
$$

Therefore, under the conditions $n|H| \longrightarrow \infty$ and $|H| \longrightarrow 0$, we obtain that

$$
\begin{aligned}
\frac{EG_n^2(W_1,W_2)) + n^{-1}E\varphi_n^4(W_1,W_2)}{(E\varphi_n^2(W_1,W_2))^2} &= \frac{O(|H|) + O(\frac{1}{n|H|})}{O(1)} \\
&\longrightarrow 0.
\end{aligned}
$$

According to Theorem in Hall [3], it follows that

$$
\frac{n|H|^{1/2}\hat{\mathcal{U}}_n}{\sqrt{2E\varphi_n^2(W_1,W_2)}} \rightsquigarrow N(0,1).
$$

Finally, we show that $\hat{\sigma}_n^2$ is a consistent estimator of $2E\varphi_n^2(W_1,W_2)$. Note that

$$
\begin{aligned}
\hat{\sigma}_n^2 &= \frac{2|H|}{C_n^2} \sum_{i<j} \big[g((X_i,Y_i),(X_j,Y_j)) \big]^2 \mathcal{K}_H^2(Z_i-Z_j) \\
&= \frac{1}{C_n^2} \sum_{i<j} 2\varphi_n^2(W_i,W_j),
\end{aligned}
$$

which implies that $E\hat{\sigma}_n^2 = 2E\varphi_n^2(W_1,W_2)$.

Moreover,

$$
\begin{aligned}
&Var(\hat{\sigma}_n^2) \\
&= \frac{C_2^1 C_{n-2}^1}{C_n^2} Var(\varphi_{n1}(W_1)) + \frac{C_2^2 C_{n-2}^0}{C_n^2} Var(2\varphi_n^2(W_1,W_2)) \\
&= \frac{4(n-2)}{n(n-1)} Var(\varphi_{n1}(W_1)) + \frac{2}{n(n-1)} Var(2\varphi_n^2(W_1,W_2)),
\end{aligned}
$$

where $\varphi_{n1}(W_1) = E(2\varphi_n^2(W_1,W_2)|W_1)$.

Notice that

$$
\begin{aligned}
Var(\varphi_{n1}(W_1)) &= Var(E(2\varphi_n^2(W_1,W_2)|W_1)) \\
&\le Var(2\varphi^2(W_1,W_2)) \\
&\le 4E(\varphi_n^4(W_1,W_2)) \\
&= O(\frac{1}{|H|}).
\end{aligned}
$$

We therefore obtain that $Var(\hat{\sigma}_n^2) = o(1)$, which implies that $\hat{\sigma}_n^2$ is a consistent estimator of $E\hat{\sigma}_n^2 = 2E\varphi_n^2(W_1,W_2)$. Thus (7) holds.

Proof of Theorem 4. According to the H-decomposition [8], $\hat{\mathcal{U}}_n$ can be rewritten as

$$\begin{aligned}\hat{\mathcal{U}}_n &= E\hat{\mathcal{U}}_n + C_2^1 H_n^{(1)} + C_2^2 H_n^{(2)} \\ &= E\hat{\mathcal{U}}_n + 2\cdot\frac{1}{n}\sum_{i=1}\psi_n(W_i) + \frac{1}{C_n^2}\sum_{i<j}\psi_n(W_i,W_j),\end{aligned} \tag{11}$$

where

$$\begin{aligned}\psi_n(W_i) &= E(P_n(W_i,W_j)|W_i) - E\hat{\mathcal{U}}_n, \\ \psi_n(W_i,W_j) &= P_n(W_i,W_j) - \psi_n(W_i) - \psi_n(W_j) - E\hat{\mathcal{U}}_n.\end{aligned}$$

Since $E\hat{\mathcal{U}}_n = \mathcal{U} + O(|H|^2)$ under the alternative hypothesis, so

$$\begin{aligned}\sqrt{n}(\hat{\mathcal{U}}_n - \mathcal{U}) &= O(\sqrt{n}|H|^2) + \sqrt{n}\cdot\frac{2}{n}\sum_{i=1}\psi_n(W_i) \\ &\quad + \frac{\sqrt{n}}{C_n^2}\sum_{i<j}\psi_n(W_i,W_j),\end{aligned}$$

where $O(\sqrt{n}|H|^2) = o(1)$ due to the assumption that $n|H|^4 \longrightarrow 0$ as $n \longrightarrow \infty$.

Notice that

$$\begin{aligned}Var\Big(\frac{\sqrt{n}}{C_n^2}\sum_{i<j}\psi_n(W_i,W_j)\Big) &= \frac{2}{n-1}Var(\psi_n(W_1,W_2)) \\ &= \frac{2}{n-1}(\sigma_{2n}^2 - 2\sigma_{1n}^2) \\ &\le \frac{2}{n-1}(\sigma_{2n}^2 + 2\sigma_{1n}^2) \\ &\le \frac{2}{n-1}O(\frac{1}{|H|}) \\ &= o(1),\end{aligned}$$

which implies that

$$\frac{\sqrt{n}}{C_n^2}\sum_{i<j}\psi_n(W_i,W_j) \xrightarrow[n\to\infty]{P} 0.$$

Moreover,

$$\begin{aligned}&Var(\psi_n(W_1)) = Var(E(P_n(W_1,W_2)|W_1)) \\ &= E[E(P_n(W_1,W_2)|W_1)]^2 - (E(P_n(W_1,W_2)))^2 \\ &= E[E(P_n(W_1,W_2)|W_1)]^2 - (E\hat{\mathcal{U}}_n)^2 \\ &= \int\Big[\int g((x_1,y_1),(x_2,y_2))\mathcal{K}_H(z_1-z_2) \\ &\cdot f(x_2,y_2,z_2)dx_2dy_2dz_2\Big]^2 f(x_1,y_1,z_1)dx_1dy_1dz_1 \\ &- \mathcal{U}^2 + O(|H|^2) \\ &= \int\Big[\int g((x_1,y_1),(x_2,y_2))\mathcal{K}(z_{12}) \\ &\cdot f(x_2,y_2,z_1 - Hz_{12})dx_2dy_2dz_{12}\Big]^2 f(x_1,y_1,z_1)dx_1dy_1dz_1 \\ &- \mathcal{U}^2 + O(|H|^2) \\ &= \int\Big[\int g((x_1,y_1),(x_2,y_2))f(x_2,y_2|z_1))dx_2dy_2\Big]^2 \\ &\cdot f_Z^2(z_1)f(x_1,y_1,z_1)dx_1dy_1dz_1 - \mathcal{U}^2 + O(|H|^2) \\ &= \sigma^2 + O(|H|^2)\end{aligned}$$

with

$$\begin{aligned}\sigma^2 = &\int\Big[\int g((x_1,y_1),(x_2,y_2))f(x_2,y_2|z_1))dx_2dy_2\Big]^2 \\ &\cdot f_Z^2(z_1)f(x_1,y_1,z_1)dx_1dy_1dz_1 - \mathcal{U}^2.\end{aligned} \tag{12}$$

Therefore

$$\begin{aligned}\sqrt{n}(\hat{\mathcal{U}}_n - \mathcal{U}) &\overset{d}{=} \sqrt{n}\cdot\frac{2}{n}\sum_{i=1}\psi_n(W_i) \\ &\rightsquigarrow N(0,\ 4\sigma^2)\end{aligned}$$

by the central limit theorem and Slutsky's Theorem.

Received 13 October 2020

REFERENCES

[1] Cortez, P. and Silva, A. M. G. (2008). Using data mining to predict secondary school student performance. In *Proceedings of 5th Annual Future Business Technology Conference* (A. Brito and J. Teixeira, eds.) 5–12.

[2] Guo, X., Zhang, J. and Fang, Y. (2020). Regression function comparison for paired data. *Journal of Systems Science and Complexity* **33** 1558–1570. MR4169077

[3] Hall, P. (1984). Central limit theorem for integrated square error of multivariate nonparametric density estimators. *Journal of Multivariate Analysis* **14** 1–16. MR0734096

[4] King, E., Hart, J. D. and Wehrly, T. E. (1991). Testing for the equality of two nonparametric regression curves using linear smoothers. *Statistics and Probability Letters* **12** 239–247. MR1130364

[5] Konietschke, F. and Pauly, M. (2014). Bootstrapping and permuting paired t-test type statistics. *Statistics and Computing* **24** 283–296. MR3192257

[6] Koul, H. L. and Schick, A. (1997). Testing for the equality of two nonparametric regression curves. *Journal of statistical planning and inference* **65** 293–314. MR1622790

[7] Koul, H. L. and Schick, A. (2003). Testing for superiority among two regression curves. *Journal of Statistical Planning and Inference* **117** 15–33. MR2001140

[8] Lee, A. J. (1990). *U-Statistics: Theory and Practice. Statistics: Textbooks and Monographs*. M. Dekker. MR1075417

[9] Lee, M. (2009). Non-parametric tests for distributional treatment effect for randomly censored responses. *Journal of The Royal Statistical Society Series B-statistical Methodology* **71** 243–264. MR2655532

[10] Li, E., Lim, J., Kim, K. and Lee, S. J. (2012). Distribution-free tests of mean vectors and covariance matrices for multivariate paired data. *Metrika* **75** 833–854. MR2956279

[11] Li, Q., Maasoumi, E. and Racine, J. S. (2009). A nonparametric test for equality of distributions with mixed categorical and continuous data. *Journal of Econometrics* **148** 186–200. MR2500656

[12] Liu, L. X. and Wong, W. H. (2021). Multivariate Density Estimation via Adaptive Partitioning (I): Sieve MLE. *arXiv preprint* arXiv:1401.2597.

[13] Liu, Q., Xu, J., Jiang, R. and Wong, W. H. (2021). Density estimation using deep generative neural networks. *Proceedings of the National Academy of Sciences* **118**. DOI: https://doi.org/10.1073/pnas.2101344118. MR4294064
[14] Martínez-Camblor, P., Corral, N. and De La Hera, J. M. (2013). Hypothesis test for paired samples in the presence of missing data. *Journal of Applied Statistics* **40** 76–87. MR3042232
[15] McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* **12** 153–157.
[16] Rietveld, T. and van Hout, R. (2017). The paired t test and beyond: recommendations for testing the central tendencies of two paired samples in research on speech, language and hearing pathology. *Journal of Communication Disorders* **66** 44–57.
[17] Rizzo, M. L. (2002). A test of homogeneity for two multivariate populations. In *2002 Proceedings of the American Statistical Association, Physical and Engineering Sciences Section. American Statistical Association, Alexandria, VA*.
[18] Rizzo, M. L. and Székely, G. J. (2019). energy: E-Statistics: Multivariate Inference via the Energy of Data. R package version 1.7-7.
[19] Srihera, R. and Stute, W. (2010). Nonparametric comparison of regression functions. *Journal of Multivariate Analysis* **101** 2039–2059. MR2671200
[20] Stuart, A. (1955). A test for homogeneity for marginal distribution in a two-way classification. *Biometrika* **42** 412–416. MR0072413
[21] Stute, W. (1991). Conditional U-Statistics. *The annals of probability* **19** 812–825. MR1106287
[22] Székely, G. J. (2003). E-Statistics: the energy of statistical samples. In *Bowling Green State University, Department of Mathematics and Statistics Technical Report*.
[23] Wang, Z. and Ng, H. K. T. (2006). A comparative study of tests for paired lifetime data. *Lifetime Data Analysis* **12** 505–522. MR2338964

Minqiong Chen
Southern China Research Center of Statistical Science
School of Mathematics
Sun Yat-Sen University
Guangzhou, 510275
China
E-mail address: mcp04chm@mail3.sysu.edu.cn

Ting Tian
Southern China Research Center of Statistical Science
School of Mathematics
Sun Yat-Sen University
Guangzhou, GD 510275
China
E-mail address: tiant55@mail.sysu.edu.cn

Jin Zhu
Southern China Research Center of Statistical Science
School of Mathematics
Sun Yat-Sen University
Guangzhou, GD 510275
China
E-mail address: zhuj37@mail2.sysu.edu.cn

Wenliang Pan
Southern China Research Center of Statistical Science
School of Mathematics
Sun Yat-Sen University
Guangzhou, GD 510275
China
E-mail address: panwliang@mail.sysu.edu.cn

Xueqin Wang
Department of Statistics and Finance
School of Management
University of Science and Technology of China
Hefei, AH 230026
China
E-mail address: wangxq20@ustc.edu.cn