Statistical Papers
https://doi.org/10.1007/s00362-024-01575-9

REGULAR ARTICLE

Asymptotic normality of a modified estimator of Gini distance correlation

Yongli Sang[1] · **Xin Dang**[2]

Received: 7 February 2024 / Revised: 30 April 2024

Abstract
Recently, the Gini distance correlation (GDC), ρ_g, was proposed to measure dependence between numerical and categorical variables (Dang et al. 2021). This new dependence measure can mutually characterize independence between the random variables. That is, $\rho_g = 0$ if and only only the categorical variable and the numerical variable are independent. Limiting distributions of the naive estimator of GDC have been established in Dang et al. (2021). It has been shown that under independence, the empirical GDC admits a degenerating limit which is an infinite weighted sum of Chi-squared distributions. In this paper, we propose a modified estimator of the GDC that is asymptotically normal under independence between the numerical and the categorical variables. We also extend this method to the generalized GDC Zhang et al. (2019) in reproducing kernel Hilbert space (RKHS). Both the modified GDC and generalized GDC can be applied to test the K-sample problem. Simulations studies are conducted to examine the finite sample performance of the new K-sample test based on the modified estimators.

Keywords Asymptotic normality · K-sample test · Modified Gini distance correlation · Reproducing kernel Hilbert space

Mathematics Subject Classification 62G35 · 62G20

1 Introduction

The Gini distance correlation in Dang et al. (2021) is proposed to measure dependence between a numerical random variable, $\boldsymbol{X}$ in $\mathbb{R}^q$ and a categorical variable Y in $\mathbb{R}$. Suppose that the categorical variable Y takes values $L_1, ..., L_K$ with its distribution P_Y is $P(Y = L_k) = p_k > 0$ for $k = 1, 2, ..., K$. $\boldsymbol{X}$ is from F and assume that the conditional distribution of $\boldsymbol{X}$ given $Y = L_k$ is F_k. Let $(\boldsymbol{X}, \boldsymbol{X}')$ and $(\boldsymbol{X}^{(k)}, \boldsymbol{X}^{(k)'})$ be

✉ Yongli Sang
yongli.sang@louisiana.edu

[1] Department of Mathematics, University of Louisiana at Lafayette, Lafayette, LA 70504, USA

[2] Department of Mathematics, University of Mississippi, University, MS 38677, USA

Published online: 06 June 2024

independent pair variables from F and F_k, respectively, then the Gini covariance is defined as

$$gCov(\boldsymbol{X}, Y) = \sum_{k=1}^{K} p_k T(\boldsymbol{X}^{(k)}, \boldsymbol{X}), \tag{1}$$

where $T(\boldsymbol{X}^{(k)}, \boldsymbol{X}) = 2\mathbb{E}\|\boldsymbol{X}^{(k)} - \boldsymbol{X}\| - \mathbb{E}\|\boldsymbol{X}^{(k)} - \boldsymbol{X}^{(k)'}\| - \mathbb{E}\|\boldsymbol{X} - \boldsymbol{X}'\|$ is the energy distance between F_k and F Székely and Rizzo (2013, 2017). The Gini distance covariance is the weighted average of energy distance between F_k and F, which implies that $gCov(\boldsymbol{X}, Y) = 0$ if and only if $F_1 = F_2 = \cdots = F_K = F$. That is, zero Gini distance covariance mutually implies independence between $\boldsymbol{X}$ and Y. The Gini distance correlation standardizes the Gini distance covariance by

$$\rho_g(\boldsymbol{X}, Y) = \frac{\sum_{k=1}^{K} p_k T(\boldsymbol{X}^{(k)}, \boldsymbol{X})}{\mathbb{E}\|\boldsymbol{X} - \boldsymbol{X}'\|}, \tag{2}$$

which takes values in [0, 1]. The naive estimator for the Gini distance covariance in (1) is a linear combination of U-statistics or V-statistics. Under independence between $\boldsymbol{X}$ and Y, the estimators are degenerate and hence converge to a infinite sum of quadratic form of centered Gaussian random variables (Dang et al. 2021). This cannot be easily applied to test the equality of K distributions because it is an infinite sum, and finding the weights in the degenerating limit is also a difficult problem. In high dimension, as q diverges, this degenerate estimator admit a normal limit (Sang and Dang 2023). In this paper, we aim to establish a normal limit under the regular setting where q is fixed.

Ahmad (1993) provided a method to testing goodness-of-fit by adding weights to the Cramér-von Mises statistic. Then the modified estimator is asymptotically normal under the null of the goodness-of-fit problem. The Cramér-von Mises statistic is an estimator of the L_2 distance between a completely specified distribution and the underlying distribution. The Gini distance covariance and the correlation are Gini distance based dependence measures. In order to achieve asymptotic normality under the null of independence between $\boldsymbol{X}$ and Y, we make an appropriate modification of the aforementioned V-estimator by adopting the approach proposed in Ahmad (1993).

Zhang et al. (2019) extended the Gini distance covariance and GDC to the RKHS by a Mercer kernel induced distance. The generalized covariance and correlation also characterize independence between $\boldsymbol{X}$ and Y. Same as GDC, the empirical parts of the generalized measures are degenerate under independence between $\boldsymbol{X}$ and Y. We provide modified estimators for the generalized Gini distance covariance and GDC in RKHS which admit normal limits under the null of independence. Makigusa and Naito (2020) constructed a consistent estimator of the maximum mean discrepancy in the Hilbert space to make it yield a normal limit when the the maximum mean discrepancy is zero. And their result has been generalized to solve K-sample problem in Balogoun et al. (2021). Manfoumbi Djonguet et al. (2024) has also adopted this method for independence testing between two functional variables.

Throughout this paper, $\|\cdot\|$ represents the Euclidean norm, that is, $\|\boldsymbol{a}\| = \sqrt{a_1^2 + a_2^2 + \cdots + a_q^2}$ for a q-vector, $\boldsymbol{a} = (a_1, a_2, \cdots, a_q)^T$, in $\mathbb{R}^q$. For two sequences,

a_n, b_n, of real numbers, $a_n = o(b_n)$ means $\lim_{n\to\infty} a_n/b_n = 0$, and $a_n = O(b_n)$ means $L \le a_n/b_n \le U$ for some finite constants L and U. For random variable sequences, similar notations $o_p(n)$ and $O_p(n)$ are used to stand for the relationships holding in probability.

The remainder of the paper is organized as follows. In Sect. 2, we provide the modified estimator for the Gini distance covariance and the asymptotic distribution. Section 3 is devoted to the modified estimator for the generalized Gini distance covariance in RKHS. In Sect. 4, we conduct simulation studies to evaluate the performance of the proposed modified test statistics. We conclude and discuss future works in Sect. 5. All technical proofs are provided in Appendix.

2 Modified Gini distance covariance estimator

There is an alternative representation for the Gini distance covariance and correlation using multivariate Gini mean differences (GMD) defined as below

$$\Delta = \mathbb{E}\|\boldsymbol{X} - \boldsymbol{X}'\|, \quad \Delta_k = \mathbb{E}\|\boldsymbol{X}^{(k)} - \boldsymbol{X}^{(k)'}\|, \ k = 1, 2, ..., K,$$
$$\Delta_{kl} = \mathbb{E}\|\boldsymbol{X}^{(k)} - \boldsymbol{X}^{(l)}\|, \ k \neq l, k, l = 1, 2, ..., K.$$

where Δ and Δ_k are GMDs for F and F_k, respectively. Gini mean difference was introduced as an alternative measure of variability to the standard deviation (Gini 1914; Yitzhaki and Schechtman 2013). The Gini covariation between $\boldsymbol{X}$ and Y defined in (1) can be represented in the GMD,

$$\mathrm{gCov}(\boldsymbol{X}, Y) = \Delta - \sum_{k=1}^{K} p_k \Delta_k, \tag{3}$$

and the Gini correlation is

$$\rho_g(\boldsymbol{X}, Y) = \frac{\Delta - \sum_{k=1}^{K} p_k \Delta_k}{\Delta}. \tag{4}$$

This representation not only shows a nice interpretation of the new dependence measurement (Dang et al. 2021) but also makes the analytical calculation feasible. In the proof of Theorem 1 in Dang et al. (2021), it has been shown that

$$\mathrm{gCov}(\boldsymbol{X}, Y) = 2 \sum_{1\le k<l\le K} p_k p_l \Delta_{kl} - \sum_{k=1}^{K} p_k(1 - p_k)\Delta_k. \tag{5}$$

All the three representations (1), (3) and (5) are equivalent (Dang et al. 2021). We will use the equation (5) to develop new estimators as it has the distance between different groups where we will add the weights.

Suppose a sample $\mathcal{D} = \{(\boldsymbol{X}_1, Y_1), (\boldsymbol{X}_2, Y_2), ..., (\boldsymbol{X}_n, Y_n)\}$ is drawn from the joint distribution of $\boldsymbol{X}$ and Y. We can write $\mathcal{D} = \mathcal{D}_1 \cup \mathcal{D}_2 ... \cup \mathcal{D}_K$, where

$\mathcal{D}_k = \left\{ \boldsymbol{X}_1^{(k)}, \boldsymbol{X}_2^{(k)}, ..., \boldsymbol{X}_{n_k}^{(k)} \right\}$ is the sample with $Y_i = L_k$ and n_k is the number of sample points in the k^{th} class. Then the Gini distance covariance in (5) can be estimated by

$$T_n = \sum_{1 \le k \ne l \le K} \hat{p}_k \hat{p}_l \frac{1}{n_k n_l} \sum_{i=1}^{n_k} \sum_{j=1}^{n_l} \|\boldsymbol{X}_i^{(k)} - \boldsymbol{X}_j^{(l)}\| - \sum_{k=1}^{K} \hat{p}_k (1 - \hat{p}_k) \binom{n_k}{2}^{-1} \sum_{1 \le i < j \le n_k} \|\boldsymbol{X}_i^{(k)} - \boldsymbol{X}_j^{(k)}\|, \tag{6}$$

where $\hat{p}_k = \dfrac{n_k}{n}$. Under independence of $\boldsymbol{X}$ and Y, T_n is a degenerate statistic and hence converges to a infinite sum of weighted Chi-squared random variables (Dang et al. 2021).

In order to overcome the degeneracy of the naive estimator, T_n, under independence, we propose a modified estimator as

$$T_{n,\gamma} = \sum_{1 \le k \ne l \le K} \hat{p}_k \hat{p}_l \frac{1}{n_k n_l} \sum_{i=1}^{n_k} \sum_{j=1}^{n_l} \omega_{i,n_k}(\gamma) \|\boldsymbol{X}_i^{(k)} - \boldsymbol{X}_j^{(l)}\| - \sum_{k=1}^{K} \hat{p}_k (1 - \hat{p}_k) \frac{1}{\binom{n_k}{2}} \sum_{1 \le i < j \le n_k} \|\boldsymbol{X}_i^{(k)} - \boldsymbol{X}_j^{(k)}\|, \tag{7}$$

where the weights $\{\omega_{i,s}(\gamma)\}_{i=1}^{s}$ are triangular array of positive real numbers depending on a parameter $\gamma (0 < \gamma \le 1)$ and satisfy the following conditions Makigusa and Naito (2020):

C1. There exists a real number $\kappa (> 0)$ and a positive integer s_0 such that

$$s \left| \frac{1}{s} \sum_{i=1}^{s} \omega_{i,s}(\gamma) - 1 \right| \le \kappa$$

for all $s > s_0$;

C2. There exists c_k such that $\max_{1 \le i \le s} \omega_{i,s}(\gamma) < c_k$ for all s and $0 < \gamma \le 1$;

C3. For all $0 < \gamma < 1$, $\lim_{s \to \infty} \dfrac{1}{s} \sum_{i=1}^{s} \left(\omega_{i,s}(\gamma) - 1\right)^2 = \eta(\gamma) > 0$.

Then the corresponding modified estimator for GDC is

$$\hat{\rho}_{g,\gamma} = \frac{T_{n,\gamma}}{\hat{\Delta}}, \tag{8}$$

where $\hat{\Delta} = \binom{n}{2}^{-1} \sum_{1 \le i < j \le n} \|\boldsymbol{X}_i - \boldsymbol{X}_j\|$.

A typical choice of the weights $\{\omega_{i,s}(\gamma)\}_{i=1}^{s}$ suggested by Ahmad (1993) is $\omega_{i,s}(\gamma) = 1 + (-1)^i\gamma$, and this has been adopted to develop the modified maximum mean discrepancy estimators in Balogoun et al. (2021) and Makigusa and Naito (2020). Manfoumbi Djonguet et al. (2024) also provided some other examples of $\omega_{i,s}(\gamma)$ satisfying the above conditions **C**1-**C**3: $\omega_{i,s}(\gamma) = 1 + \sin(i\pi\gamma)$ and $\omega_{i,s}(\gamma) = 1 + \cos(i\pi\gamma)$. The first choice of weights yields $\eta(\gamma) = \gamma^2$ and the latter two weights generate $\eta(\gamma) = 1/2$.

Applying the weights satisfying the conditions **C**1-**C**3 to $T_{n,\gamma}$ in (7), we provide the asymptotic normality of this modified estimator in the following theorem.

Define $h(\boldsymbol{x}, \boldsymbol{x}') = \|\boldsymbol{x} - \boldsymbol{x}'\|$, $h_1(\boldsymbol{x}) = \mathbb{E}\|\boldsymbol{x} - \boldsymbol{X}_1\|$ and $\sigma_g^2 = \mathrm{Var}(h_1(\boldsymbol{X})) > 0$.

Theorem 2.1 *Under independence of* $\boldsymbol{X}$ *and* Y*, and conditions* ***C****1-****C****3, if* $\mathbb{E}\|\boldsymbol{X}\|^2 < \infty$*, as* $\min\{n_1, n_2, ..., n_k\} \to \infty$*, we have*

$$\sqrt{n}T_{n,\gamma} \xrightarrow{d} \mathcal{N}(0, \sigma_\gamma^2),$$

with $\sigma_\gamma^2 = \sum_{k=1}^{K} p_k(1-p_k)^2\sigma_1^2(\gamma)$ *where* $\sigma_1^2(\gamma) = \eta(\gamma)\sigma_g^2$.

Theorem 2.1 shows that the modified estimator, $T_{n,\gamma}$, for the Gini distance covariance has a normal limit which can be applied to test independence between $\boldsymbol{X}$ and Y, and hence to test the equality of K-distributions.

Applying Slustky's theorem, we have central limit theorem (CLT) for the modified GDC estimator, $\hat{\rho}_{g,\gamma}$, defined in (8).

Corollary 2.1 *Under independence of* $\boldsymbol{X}$ *and* Y*, and conditions* ***C****1-****C****3, if* $\mathbb{E}\|\boldsymbol{X}\|^2 < \infty$*, as* $\min\{n_1, n_2, ..., n_k\} \to \infty$*, we have*

$$\sqrt{n}\hat{\rho}_{g,\gamma} \xrightarrow{d} \mathcal{N}(0, \sigma_{\rho_g,\gamma}^2),$$

where $\sigma_{\rho_g,\gamma}^2 = \sum_{k=1}^{K} p_k(1-p_k)^2\sigma_1^2(\gamma)/\Delta^2$.

In order to apply Theorem 2.1 to make inference, we provide a consistent estimator for σ_γ^2. σ_g^2 can be estimated by the empirical version,

$$\begin{aligned}\hat{\nu} &= \frac{1}{n-1}\sum_{i=1}^{n}\left\{\frac{1}{n-1}\sum_{j=1}^{n}\|\boldsymbol{X}_j - \boldsymbol{X}_i\| - \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{n-1}\sum_{j=1}^{n}\|\boldsymbol{X}_j - \boldsymbol{X}_i\|\right)\right\}^2 \\ &= \frac{1}{n-1}\sum_{i=1}^{n}\left\{\frac{1}{n-1}\sum_{j=1}^{n}\|\boldsymbol{X}_j - \boldsymbol{X}_i\| - \frac{1}{n(n-1)}\sum_{i,j=1}^{n}\|\boldsymbol{X}_j - \boldsymbol{X}_i\|\right\}^2.\end{aligned}$$

Then a consistent estimator for σ_γ^2 can be obtained by $\hat{\sigma}_0^2 = \eta(\gamma)\hat{\nu}\sum_{k=1}^{K}\hat{p}_k^2(1-\hat{p}_k)$.

Corollary 2.2 *Under independence of* $\boldsymbol{X}$ *and* Y*, and conditions* $\boldsymbol{C1}$*-*$\boldsymbol{C3}$*, if* $\mathbb{E}\|\boldsymbol{X}\| < \infty$*, as* $\min\{n_1, n_2, ..., n_k\} \to \infty$*, we have*

$$\frac{\sqrt{n}T_{n,\gamma}}{\hat{\sigma}_0} \xrightarrow{d} \mathcal{N}(0, 1).$$

These established CLTs can be applied to test the independence of $\boldsymbol{X}$ and Y. We will use the CLT for the Gini distance covariance to do the test. The one based on the Gini correlation is asymptotically equivalent. The independence test is stated as

$$\mathcal{H}_0 : \text{gCov}(\boldsymbol{X}, Y) = 0, \quad \text{vs} \quad \mathcal{H}_1 : \text{gCov}(\boldsymbol{X}, Y) > 0. \tag{9}$$

Note that the null hypothesis of the test in (9) is equivalent to the null of the K-sample test

$$\mathcal{H}_0' : F_1 = F_2 = ... = F_K = F.$$

In the K sample test, we can view sample point $(\boldsymbol{X}_i, Y_i)$ in such way. Y_i is the class label of $\boldsymbol{X}_i$. $Y_i = L_k$ indicates that $\boldsymbol{X}_i$ is drawn from F_k. The pooled sample $\mathcal{D} = \mathcal{D}_1 \cup \mathcal{D}_2 ... \cup \mathcal{D}_K$ has the distribution F, which is the average distribution of F_k's.

By Corollary 2.2, we can reject $\mathcal{H}_0$ or $\mathcal{H}_0'$ if $\sqrt{n}T_{n,\gamma} > Z_\alpha \hat{\sigma}_0$ at level α, where Z_α is the $(1-\alpha)100\%$ percentile of the standard normal distribution.

3 Modified Gini distance covariance estimator in RKHS

Distance based statistics can be generalized from a euclidean space to metric spaces. With a Mercer (1909), distributions can be mapped into a RKHS with a kernel induced distance. The Gini distance covariance has been generalized to a RKHS, $\mathcal{H}_M$, as Zhang et al. (2019)

$$\begin{aligned}\text{gCov}_{\mathcal{H}(M)}(\boldsymbol{X}, Y) =& 2 \sum_{1 \le k < l \le K} p_k p_l \mathbb{E} d_M(\boldsymbol{X}^{(k)}, \boldsymbol{X}^{(l)}) \\ & - \sum_{k=1}^{K} p_k (1 - p_k) \mathbb{E} d_M(\boldsymbol{X}_1^{(k)}, \boldsymbol{X}_2^{(k)}),\end{aligned} \tag{10}$$

where $M : \mathbb{R}^q \times \mathbb{R}^q \to \mathbb{R}$ is a Mercer kernel with the distance function $d : \mathbb{R}^q \times \mathbb{R}^q \to \mathbb{R}$. d defines a distance in $\mathcal{H}_M$ as

$$d_M(\boldsymbol{x}, \boldsymbol{x}') = \sqrt{M(\boldsymbol{x}, \boldsymbol{x}) + M(\boldsymbol{x}', \boldsymbol{x}') - 2M(\boldsymbol{x}, \boldsymbol{x}')}.$$

As the regular Gini distance covariance in $\mathbb{R}^p$, the generalized Gini distance covariance can also characterize independence in RKHS, $\text{gCov}_{\mathcal{H}(M)}(\boldsymbol{X}, Y) = 0$ if and only if $\boldsymbol{X}$ and Y are independent (Zhang et al. 2019).

The generalized Gini distance covariance can be estimated by

$$G_n = \sum_{1\le k\ne l\le K} \hat{p}_k\hat{p}_l \frac{1}{n_k n_l}\sum_{i=1}^{n_k}\sum_{j=1}^{n_l} d_M(\boldsymbol{X}_i^{(k)}, \boldsymbol{X}_j^{(l)}) - \sum_{k=1}^{K}\hat{p}_k(1-\hat{p}_k)\binom{n_k}{2}^{-1}\sum_{1\le i<j\le n_k} d_M(\boldsymbol{X}_i^{(k)}, \boldsymbol{X}_j^{(k)}), \tag{11}$$

which has been shown to be degenerate and converges to a mixture of infinite chi-square distributions under independence of $\boldsymbol{X}$ and Y Zhang et al. (2019).

We give a modified estimator as

$$G_{n,\gamma} = \sum_{1\le k\ne l\le K} \hat{p}_k\hat{p}_l \frac{1}{n_k n_l}\sum_{i=1}^{n_k}\sum_{j=1}^{n_l} \omega_{i,n_k}(\gamma) d_M(\boldsymbol{X}_i^{(k)}, \boldsymbol{X}_j^{(l)}) - \sum_{k=1}^{K}\hat{p}_k(1-\hat{p}_k)\binom{n_k}{2}^{-1}\sum_{1\le i<j\le n_k} d_M(\boldsymbol{X}_i^{(k)}, \boldsymbol{X}_j^{(k)}), \tag{12}$$

where the weights $\{\omega_{i,s}(\gamma)\}_{i=1}^{s}$ are chosen as the same in Sect. 2.

Theorem 3.1 *Assume M is a Mercer kernel over $\mathbb{R}^q\times\mathbb{R}^q\to\mathbb{R}$ that induces a distance function $d_M(\cdot,\cdot)$ with bounded range [0, 1). Under independence of $\boldsymbol{X}$ and Y, assume conditions **C1**-**C3**, as* $\min\{n_1, n_2, ..., n_k\}\to\infty$, *we have*

$$\sqrt{n}G_{n,\gamma}\xrightarrow{d}\mathcal{N}(0,\sigma_{M,\gamma}^2),$$

with $\sigma_{M,\gamma}^2=\sum_{k=1}^{K}p_k(1-p_k)^2\sigma_{2,M}^2(\gamma)$ where $\sigma_{2,M}^2(\gamma)$ is given in the proof.

A consistent estimator for $\sigma_{M,\gamma}^2$ is $\hat{\sigma}_{M,0}^2=\eta(\gamma)\hat{\nu}_M\sum_{k=1}^{K}\hat{p}_k^2(1-\hat{p}_k)$, where

$$\hat{\nu}_M = \frac{1}{n-1}\sum_{i=1}^{n}\left\{\frac{1}{n-1}\sum_{j=1}^{n}d_M(\boldsymbol{X}_i,\boldsymbol{X}_j) - \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{n-1}\sum_{j=1}^{n}d_M(\boldsymbol{X}_i,\boldsymbol{X}_j)\right)\right\}^2$$
$$= \frac{1}{n-1}\sum_{i=1}^{n}\left\{\frac{1}{n-1}\sum_{j=1}^{n}d_M(\boldsymbol{X}_i,\boldsymbol{X}_j) - \frac{1}{n(n-1)}\sum_{i,j=1}^{n}d_M(\boldsymbol{X}_i,\boldsymbol{X}_j)\right\}^2.$$

Corollary 3.1 *Assume M is a Mercer kernel over $\mathbb{R}^q\times\mathbb{R}^q\to\mathbb{R}$ that induces a distance function $d_M(\cdot,\cdot)$ with bounded range [0, 1). Under independence of $\boldsymbol{X}$ and Y, assume conditions **C1**-**C3**, as* $\min\{n_1, n_2, ..., n_k\}\to\infty$, *we have*

$$\frac{\sqrt{n}G_{n,\gamma}}{\hat{\sigma}_{M,0}}\xrightarrow{d}\mathcal{N}(0,1).$$

The modified estimator of the generalized Gini distance covariance can also be used to test the equality of K populations. By Corollary 3.1, we can reject $\mathcal{H}_0$ or $\mathcal{H}_0'$ if $\sqrt{n}G_{n,\gamma} > Z_\alpha \hat{\sigma}_{M,0}$ at level α.

4 Simulation

In this section, we conduct simulation studies to verify the theoretical properties of the modified Gini covariance statistic and compare its performance in K-sample tests with others. Also based on empirical results, we discuss how to select the weight function.

4.1 Limiting normality

We generate independent K samples from the same multivariate normal distributions and compute the weighted Gini covariance statistic with weights $\omega_{i,n}(\gamma) = 1 + (-1)^i\gamma$, $i = 1, ..., n$. The procedure is repeated 10000 times.

Example 1 $K = 2$ samples of size $(n_1, n_2) = (200, 200)$ are generated from $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$, where $\mathbf{\Sigma} = (\Sigma_{ij}) \in \mathbb{R}^{q\times q}$ with $\Sigma_{ij} = 0.7^{|i-j|}$. We consider $q = 3, 5$ and $\gamma = 0.1, 0.2, 0.5, 0.8$, respectively.

For each dimension and each value of γ, the histogram of 10000 standardized weighted Gini covariance statistics is plotted in Fig. 1. The kernel density estimation (KDE) for the weighted Gini covariance and the standard Gini covariance are added to the plots. We also add the standard normal density curve to visualize the closeness between empirical density and asymptotic density functions. Firstly, we can see that the KDEs for the regular Gini covariance are always skewed to the right, which agrees well with its limiting distribution of the mixture of χ^2 distributions due to the degeneracy of the regular Gini covariance statistics. Then we notice that the histograms at $\gamma = 0.1$ for both dimensions are skewed to the right, and there is some discrepancy between KDE of the weighted Gini covariance and the normal curve. However, as γ increases, the discrepancy becomes less and diminishes. This suggests that larger γ values are preferred for this weight function. We use $\gamma = 0.8$ in the next subsection for performance comparison in K-sample tests. The impacts of the choice of γ in $\omega_{i,n}(\gamma) = 1+(-1)^i\gamma$ as well as in $\omega_{i,s}(\gamma) = 1 + \sin(i\pi\gamma)$ are explored in Subsection 4.3.

4.2 Size and power in *K*-sample tests

In this simulation, we compare three methods for K sample problem by computing the type I errors and the powers.

mmd: generalized maximum mean discrepancy method developed in Balogoun et al. (2021).

wrg: our proposed method using weighted Gini covariance statistic.

wkrg: our proposed method using weighted Gini covariance statistic in a RKHS where the distance function $d_M(\boldsymbol{x}, \boldsymbol{x}') = \sqrt{1 - e^{-\|\boldsymbol{x}-\boldsymbol{x}'\|^2/\sigma^2}}$ is induced by a weighted Gaussian kernel $M(\boldsymbol{x}, \boldsymbol{x}') = 0.5e^{-\|\boldsymbol{x}-\boldsymbol{x}'\|^2/\sigma^2}$ Zhang et al. (2019).

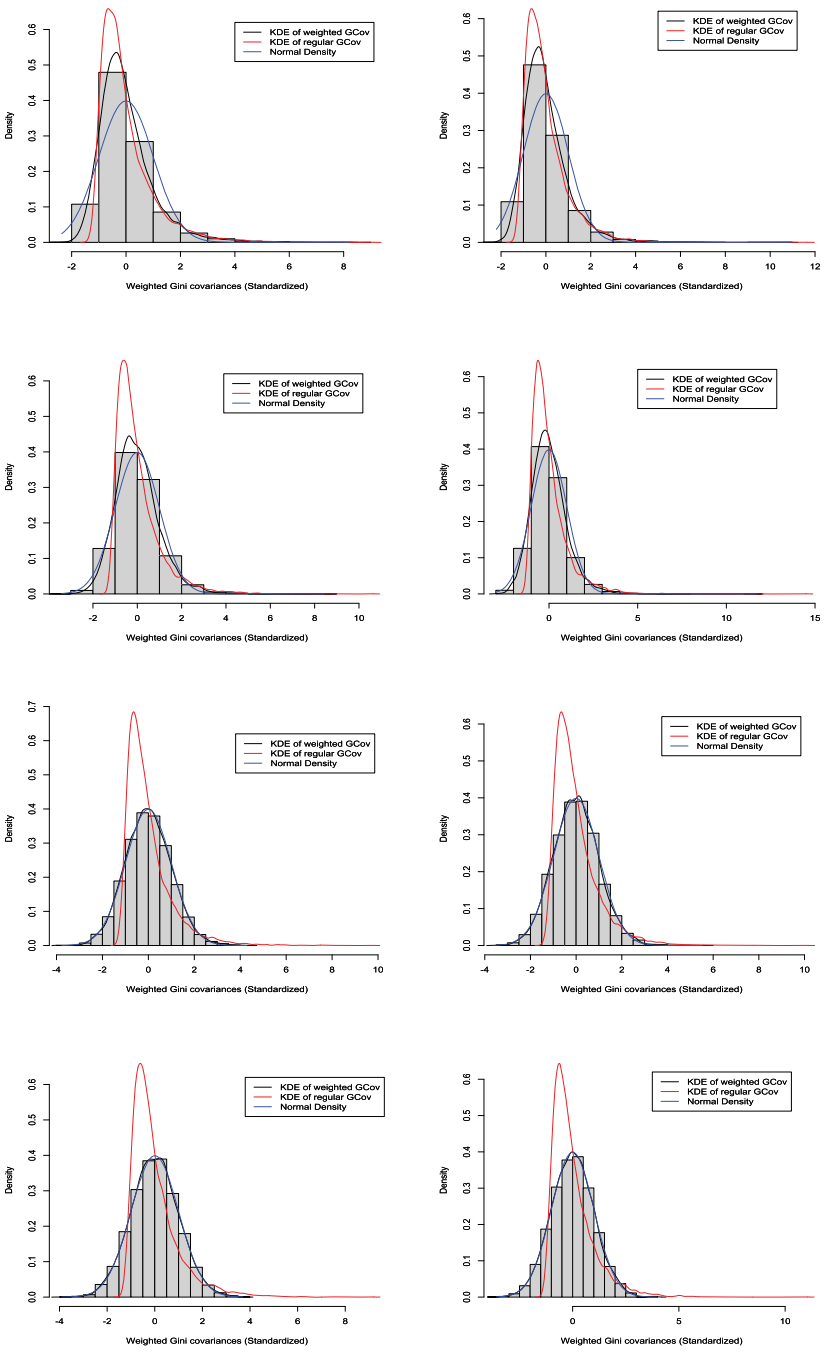

Fig. 1 Histograms of the proposed weighted Gini covariance under weight functions $\omega_{i,n}(\gamma) = 1+(-1)^i\gamma$ with different γ values of 0.1, 0.2, 0.5 and 0.8, respectively. The left plots are for dimension $q = 3$, and the right ones are for dimension $q = 5$

Table 1 Size and Power of Tests in Example 2

$\boldsymbol{n}$	method	$\boldsymbol{\delta} = (\delta_1, \delta_2)$					
		(0.0, 0.0)	(0.1, 0.2)	(0.2, 0.4)	(0.3, 0.6)	(0.4, 0.8)	(0.5, 1.0)
(40,40,40)	mmd	.0705	.1650	.4759	.8829	.9947	.9999
	wrg	.0701	.1697	.4946	.8948	.9959	1.000
	wkrg	.0739	.1715	.4870	.8884	.9948	.9999
(30,40,50)	mmd	.0723	.1416	.4592	.8577	.9903	.9998
	wrg	.0725	.1537	.4917	.8810	.9936	.9999
	wkrg	.0792	.1593	.4857	.8741	.9925	.9998
(12,36,72)	mmd	.0611	.1131	.3068	.6432	.9014	.9911
	wrg	.0714	.1545	.4075	.7782	.9674	.9985
	wkrg	.0758	.1584	.4058	.7693	.9633	.9982
(80,80,80)	mmd	.0220	.0983	.5453	.9734	1.000	1.000
	wrg	.0655	.1960	.6956	.9906	1.000	1.000
	wkrg	.0644	.1960	.6842	.9889	1.000	1.000
(60,80,100)	mmd	.0221	.0858	.5157	.9620	1.000	1.000
	wrg	.0646	.1779	.6872	.9874	1.000	1.000
	wkrg	.0660	.1813	.6787	.9852	1.000	1.000
(24,72,144)	mmd	.0201	.0568	.3046	.7709	.9837	.9999
	wrg	.0664	.1598	.5700	.9431	.9991	1.000
	wkrg	.0690	.1622	.5368	.9396	.9989	1.000

Both **mmd** and **wkrg** are kernel methods. The bandwidth of the Gaussian kernel in both methods is chosen to be the median of pairwise distances, as used and suggested in Chen et al. (2009). All three methods use the weight function $\omega_{i,n}(\gamma) = 1+(-1)^i\gamma$ with $\gamma = 0.8$.

We consider cases for $K = 3$ and $q = 5$ with $\boldsymbol{p} = (p_1, p_2, ...p_K)$ where $p_k = P(Y_i = L_k)$: (I) balanced, $\boldsymbol{p} = (1/3, 1/3, 1/3)$; (II) slightly unbalanced, $\boldsymbol{p} = (3/12, 4/12, 5/12)$; (III) heavily unbalanced, $\boldsymbol{p} = (0.1, 0.3, 0.6)$. We conduct 10000 simulations for different sample sizes of $n = 120$ and $n = 240$, respectively. The type I error and the power of each test are computed at significance level $\alpha = 0.05$ for Example 2 and Example 3.

Example 2 Generate samples of $\boldsymbol{X}^{(k)} = \boldsymbol{\mu}_k + \boldsymbol{\epsilon}, k = 1, 2, 3$, where the mean vector $\boldsymbol{\mu}_1 = (0, 0, ..., 0)$, $\boldsymbol{\mu}_2 = (\delta_1, \delta_1, ..., \delta_1)$, $\boldsymbol{\mu}_3 = (\delta_2, \delta_2, ..., \delta_2)$, and $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, ...\epsilon_q)$ is a q-dimensional error term with ϵ_j's are iid from $N(0, 1)$. Here $\boldsymbol{\delta} = (\delta_1, \delta_2)$ measures differences in means.

Results of Example 2 are reported in Table 1. The column $\boldsymbol{\delta} = (0, 0)$ corresponds to the size of tests. At $n = 120$, all tests have slight over-size problems with size is 1–2% higher than the nominal level. And they all have higher powers for equal size case than unbalanced cases. For unbalanced cases, our method **wrg** gains 1%-4% power advantage over **mmd** at small values of $\boldsymbol{\delta}$, especially for heavily unbalanced cases. As sample size increases, the type I errors of our **wrg** and **wkrg** are getting closer

Table 2 Size and Power of Tests in Example 3

$\boldsymbol{n}$	method	$\boldsymbol{\delta} = (\delta_1, \delta_2)$					
		(1.0, 1.0)	(1.1, 1.5)	(1.2, 2.0)	(1.3, 2.5)	(1.4, 3.0)	(1.5, 3.5)
(40,40,40)	mmd	.0269	.0924	.2813	.4957	.6595	.7970
	wrg	.0645	.1822	.4131	.6118	.7403	.8264
	wkrg	.0868	.1983	.4334	.6472	.7836	.8836
(30,40,50)	mmd	.0225	.0719	.2246	.4617	.6537	.7845
	wrg	.0674	.1756	.4009	.6194	.7564	.8414
	wkrg	.0859	.2041	.4413	.6675	.8205	.9004
(12,36,72)	mmd	.0224	.0248	.0883	.2164	.3728	.5220
	wrg	.0665	.1662	.3784	.5887	.7362	.8246
	wkrg	.0887	.2209	.4551	.6718	.8134	.8975
(80,80,80)	mmd	.0089	.0768	.3461	.6572	.8661	.9443
	wrg	.0552	.2236	.5664	.7975	.9200	.9634
	wkrg	.0706	.2357	.5873	.8348	.9507	.9830
(60,80,100)	mmd	.0104	.0578	.3229	.6468	.8654	.9537
	wrg	.0556	.2289	.5734	.8144	.9294	.9687
	wkrg	.0686	.2466	.6176	.8591	.9611	.9887
(24,72,144)	mmd	.0096	.0166	.1111	.3559	.6404	.8206
	wrg	.0610	.1996	.5439	.7980	.9220	.9655
	wkrg	.0770	.2549	.6194	.8656	.9625	.9908

to the nominal level and powers get improved. However, **mmd** suffers from severe under-size problems with very low powers for unbalanced case when the difference in means is small.

Example 3 We generate samples from $\boldsymbol{X}^{(k)} = (Z_{k1}, Z_{k2}, ...Z_{kq})^T, k = 1, 2, 3$. For $k = 1$, $j = 1, ..., q$, Z_{kj}'s are i.i.d. from Exp(1); $k = 2$, $j = 1, ..., q$, Z_{kj}'s are i.i.d. from Exp(δ_1); $k = 3$, $j = 1, ..., q$, Z_{kj}'s are i.i.d. from Exp(δ_2).

We present results for this example in Table 2. The **mmd** seems sensitive to the asymmetry of distributions. It is undersized and its power is much lower than the weight Gini covariance based ones. Our **wrg** performs best with well-controlled size and higher powers.

4.3 Discussion on weights

Manfoumbi Djonguet et al. (2024) provided two weight schemes based on sine and cosine functions, but they didn't study the performance of those weights. In this simulation, we would like to compare this weight $\omega_{i,s}(\gamma) = 1 + \sin(i\pi\gamma)$ (**weight2**) with the previously used one $\omega_{i,s}(\gamma) = 1 + (-1)^i\gamma$ (**weight1**). We compare the effects of different γ values of 0.2, 0.4,0.6, 0.8 and 0.9 of each weight function on our **wrg** method. With empirical results, we provide some suggestions on the choice of weights.

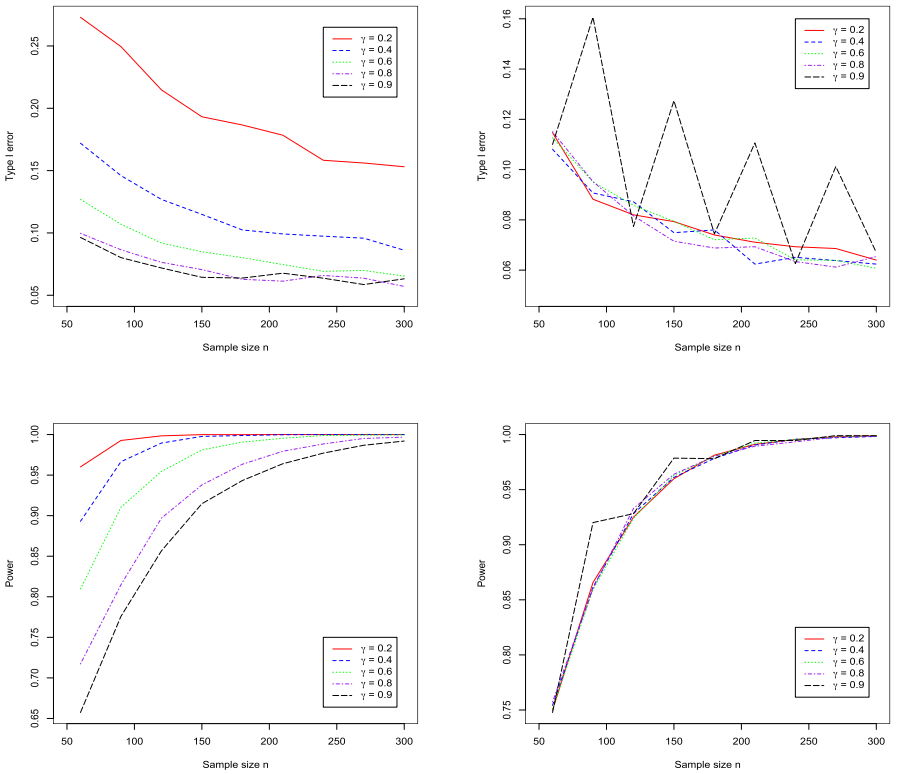

Fig. 2 Empirical size and power versus total sample size at different γ values in Example 2 with $\boldsymbol{p} = (1/3, 1/3, 1/3)$. The left plots are for $\omega_(i, n)(\gamma) = 1 + (-1)^i\gamma$, and the right ones are for $\omega_(i, n)(\gamma) = 1 + \sin(i\pi\gamma)$

The balanced $\boldsymbol{p} = (1/3, 1/3, 1/3)$ and unbalanced $\boldsymbol{p} = (0.1, 0.3, 0.6)$ scenarios of Example 2 are considered. At different sample sizes of $n = 60, 90, 120, 150, 180, 210, 240, 270, 300$, the type I errors and/or powers of the tests under different weighting schemes are calculated based on $M = 10000$ repetitions and reported in plots.

Figure 2 is for the balanced case with the top two plots on type I error and the bottom two plots on power at $(\delta_1, \delta_2) = (0.3, 0.6)$. The left panels are for **weight1**, while the right ones are for **weight2**.

The nominal size is 0.05. From Fig. 2, we observe that the tests based on two weight functions have the over-size problem but this issue becomes less serious as sample size increases. For **weight1** with γ values 0.8 and 0.9, the type I errors decrease from 0.10 to 0.06 when the sample size increases from 60 to 150. But small γ values of 0.2 and 0.4, even at sample size 300, produce unacceptable type I errors of 0.16 and 0.10 respectively. For **weight2** with a large range of gamma values from 0.2 to 0.8, the type I errors decrease from 0.12 to 0.08 when the sample sizes increase from 60 to150. Further reduction type I error to 0.06 requires a sample size to be as large as 300. As sample size increases, γ value 0.9 yields a relatively large zig-zag oscillation in type I errors, which is undesired.

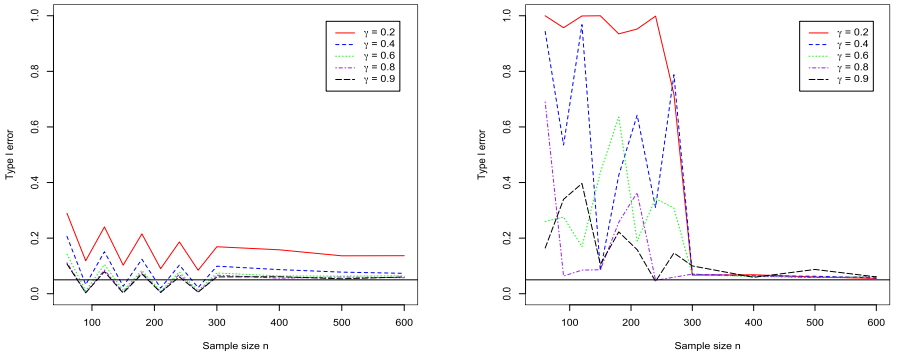

Fig. 3 Empirical size versus the sample size at different γ values when $\boldsymbol{p} = (0.1, 0.3, 0.6)$. The left plot is for $\omega_(i, n)(\gamma) = 1 + (-1)^i \gamma$, and the right one is for $\omega_(i, n)(\gamma) = 1 + \sin(i\pi\gamma)$

The tests based on both weight schemes produce relatively high power. The power of all tests is over 0.90 when the sample size is 150. For **weight1**, the power decreases in γ. With consideration of controllable type I error and high power, $\gamma = 0.8$ is recommended. This suggestion is also supported by the previous empirical results in Subsections 4.1 and 4.2. **weight2** performs very well in terms of power. Except for 0.9, all other γ values produce an almost same power at each sample size. It is quite robust with the choice of γ in the balanced case. However, in the unbalanced case, **weight2** tests fail badly not only in terms of huge type I errors but also in terms of sensitivity of γ choice.

The type I errors of tests under the unbalanced scenarios are reported in Fig. 3. The left plot is for **weight1**, while the right one is for **weight2**. Both plots have a same scale in order to provide a fair visual comparison. From Fig. 3, we see that for **weight2**, all gamma values cause unacceptable type I errors when sample size is less than 300. With $\gamma = 0.2$ and sample size less than 240, the test is meaningless with type I error higher than 0.9. For $\gamma = 0.4$, the type I errors jump up and down in a wide range, reaching 0.94 followed by 0.53, then up to 0.97 followed by 0.08, and then 0.43 followed by 0.64 when sample size changes from 60 to 210. Oscillation patterns in large ranges of type I errors also present for larger γ values. However, when sample size is as large as 300, all tests have a good size. We increase sample size up to 600 and find out **weight2** maintains the nominal size well. For **weight1** with sample size less than 300, the type I errors also show zig-zag oscillations, but in much smaller ranges. Except for $\gamma = 0.2$, all tests yield a reasonably good empirical size when the sample size is 300 or larger.

Overall, for balanced case, both weights are acceptable depending on the choice of γ. **weight2** performs better than **weight1** with a wide range of choices for γ. For unbalanced cases, **weight2** is not applicable unless the sample size is sufficiently large (at least 300). **weight1** with $\gamma = 0.8$ is recommended due to its controllable type I errors and high powers for both balanced and unbalanced cases.

5 Conclusions and future work

We have proposed a modified estimator of the Gini distance correlation. By adding weights to the distances between different groups, the modified estimator admits a normal limit under independence of numerical and categorical variables. We have also generalized the results into RKHS, where normal limits also hold. All of the asymptotic results have been applied to test the equality of K distributions. With a proper choice of weight function, the modified Gini correlation estimator performs well. We have studied two types of weights $\omega_{i,n}(\gamma) = 1 + (-1)^i \gamma$ and $\omega_{i,s}(\gamma) = 1 + \sin(i\pi\gamma)$. The second weight works well with a wide range of choices for γ for balanced K sample problem. However, it is not applicable unless the sample size is sufficiently large for unbalanced cases. The first weight with a large γ value like 0.8 is recommended. It could control type I error close to nominal level for all cases including balanced and unbalanced, and keep reasonable high power.

In the real application of K sample problem, most of the existing omnibus tests are permutation procedures based. The permutation tests have some optimum properties (Gebhard and Schmitz 1998a). However, they are computation-intensive (Gebhard and Schmitz 1998b). By considering all permutations, exact tests such as in Neuhäuser (2005) are only feasible for very small sample sizes. For larger sample sizes, a large number of random permutation procedures are repeated in order to approximate the p-value or to determine the critical value. The test based on the regular Gini distance covariance (Dang et al. 2021) relies on such a computationally expensive procedure. Our proposed weighted Gini covariance statistic admits a normal limit and can be directly applied to real-world analysis for K-sample problem by avoiding permutation procedures.

In this paper, we used the median of pairwise distances as a bandwidth in the Gaussian kernel for the **wkrg** and **mmd**. Such a choice makes a half of pairwise distances in the induced feature space greater than 0.8871 and 0.6065 for the **wkrg** method and **mmd** method, respectively. This choice is simple and seems to be effective, but is by no means "optimal". How to select an optimal bandwidth (in terms of some criteria) is always a challenge for any kernel method. For **wkrg** and also **mmd** methods, the task is particularly difficult because the bandwidth shall be selected optimally with consideration of weight function. How to jointly select the optimal kernel parameter and weight scheme in **wkrg** and **mmd** is worthy of further investigation.

6 Appendix

Proof of Theorem 2.1

$$\begin{aligned} T_{n,\gamma} &= T_{n,\gamma} - T_n + T_n \\ &= \sum_{1\le k\ne l\le K} \hat{p}_k \hat{p}_l \frac{1}{n_k n_l} \sum_{i=1}^{n_k}\sum_{j=1}^{n_l} \Big(\omega_{i,n_k}(\gamma) - 1\Big) \|\boldsymbol{X}_i^{(k)} - \boldsymbol{X}_j^{(l)}\| + T_n \\ &:= D_n + T_n, \end{aligned}$$

where

$$D_n = \sum_{1\le k\ne l\le K} \hat{p}_k \hat{p}_l \frac{1}{n_k n_l} \sum_{i=1}^{n_k}\sum_{j=1}^{n_l} \Big(\omega_{i,n_k}(\gamma) - 1\Big) \|\boldsymbol{X}_i^{(k)} - \boldsymbol{X}_j^{(l)}\|.$$

From Dang et al. (2021), T_n is degenerate under independence of $\boldsymbol{X}$ and Y with $\mathrm{Var}(T_n) = O_p(n^{-2})$. We will show that D_n dominates with a normal limit.

Define $\hat{D}_{k,l}(\gamma) = \frac{1}{n_k n_l} \sum_{i=1}^{n_k}\sum_{j=1}^{n_l} \Big(\omega_{i,n_k}(\gamma) - 1\Big) \|\boldsymbol{X}_i^{(k)} - \boldsymbol{X}_j^{(l)}\|$, then $D_n = \sum_{1\le k\ne l\le K} \hat{p}_k \hat{p}_l \hat{D}_{k,l}(\gamma)$. Then we can decompose $\hat{D}_{k,l}(\gamma)$ as

$$\begin{aligned}
\hat{D}_{k,l}(\gamma) =& \frac{1}{n_k n_l} \sum_{i=1}^{n_k}\sum_{j=1}^{n_l} \Big(\omega_{i,n_k}(\gamma) - 1\Big)\Delta_{kl} \\
&+ \frac{1}{n_k} \sum_{i=1}^{n_k} \left[\Big(h_1\big(\boldsymbol{X}_i^{(k)}\big) - \Delta_{kl}\Big)\big(\omega_{i,n_k}(\gamma) - 1\big)\right] \\
&+ \frac{1}{n_l} \sum_{j=1}^{n_l} \Big(h_1\big(\boldsymbol{X}_j^{(l)}\big) - \Delta_{kl}\Big) \frac{1}{n_k} \sum_{i=1}^{n_k} \big(\omega_{i,n_k}(\gamma) - 1\big) \\
&+ \frac{1}{n_k n_l} \sum_{i=1}^{n_k}\sum_{j=1}^{n_l} \left[\big(\omega_{i,n_k}(\gamma) - 1\big)\phi_{k,l}(\boldsymbol{X}_i^{(k)}, \boldsymbol{X}_j^{(l)})\right], \qquad (13)
\end{aligned}$$

where $\phi_{k,l}(\boldsymbol{X}^{(k)}, \boldsymbol{X}^{(l)}) = h(\boldsymbol{X}^{(k)}, \boldsymbol{X}^{(l)}) - h_1(\boldsymbol{X}^{(k)}) - h_1(\boldsymbol{X}^{(l)}) + \Delta_{kl}$ is the degenerate part with $\mathrm{Var}\big(\phi_{k,l}(\boldsymbol{X}^{(k)}, \boldsymbol{X}^{(l)})\big) = O(\frac{1}{n_k n_l})$.

We will show that $\hat{D}_{k,l}(\gamma)$ is dominated by $\frac{1}{n_k} \sum_{i=1}^{n_k} \big(h_1\big(\boldsymbol{X}_i^{(k)}\big) - \Delta_{kl}\big)\big(\omega_{i,n_k}(\gamma) - 1\big)$.

First of all, the first sum on the right of (13) is not random and is bounded. That is,

$$\begin{aligned}
\left|\frac{1}{n_k n_l} \sum_{i=1}^{n_k}\sum_{j=1}^{n_l} \Big(\omega_{i,n_k}(\gamma) - 1\Big)\Delta_{kl}\right| &= \frac{\Delta_{kl}}{n_k} \left|\sum_{i=1}^{n_k} \Big(\omega_{i,n_k}(\gamma) - 1\Big)\right| \\
&\le \frac{\kappa}{n_k}\Delta_{kl} \\
&\to 0
\end{aligned}$$

by condition **C**1.

Then the second and the third sums on the right side of (13) are the first-projection parts. By Theorem 3 in O'Neil, K.A. and Redner, R.A. (1993), we have

$$\mathrm{Var}\left(\frac{1}{n_k} \sum_{i=1}^{n_k} \left[\Big(h_1\big(\boldsymbol{X}_i^{(k)}\big) - \Delta_{kl}\Big)\big(\omega_{i,n_k}(\gamma) - 1\big)\right]\right) = O(\frac{1}{n_k^2}) n_k \sum_{i=1}^{n_k} \big(\omega_{i,n_k}(\gamma) - 1\big)^2,$$

$$\mathrm{Var}\Big(\frac{1}{n_l}\sum_{j=1}^{n_l}\Big(h_1\big(\boldsymbol{X}_j^{(l)}\big)-\Delta_{kl}\Big)\frac{1}{n_k}\sum_{i=1}^{n_k}\big(\omega_{i,n_k}(\gamma)-1\big)\Big)$$
$$=O\big(\frac{1}{n_l^2n_k^2}\big)n_l\Big[\sum_{i=1}^{n_k}\big(\omega_{i,n_k}(\gamma)-1\big)\Big]^2.$$

By conditions **C**1–**C**3, we have

$$\sum_{i=1}^{n_k}\big(\omega_{i,n_k}(\gamma)-1\big)^2\to\infty,$$

and

$$\Big[\sum_{j=1}^{n_k}\Big(\omega_{j,n_k}(\gamma)-1\Big)\Big]^2\le\kappa^2.$$

Therefore, $\frac{1}{n_k}\sum_{i=1}^{n_k}h_1(\boldsymbol{X}_i^{(k)})\big(\omega_{i,n_k}(\gamma)-1\big)$ dominates $\hat{D}_{k,l}(\gamma)$ and admits a normal limit, and hence D_n has a normal limit.

Next we find the variance of $T_{n,\gamma}$. We have

$$T_{n,\gamma}=\sum_{1\le k\ne l\le K}\hat{p}_k\hat{p}_l\left\{\frac{1}{n_k}\sum_{i=1}^{n_k}\Big(h_1\big(\boldsymbol{X}_i^{(k)}\big)-\Delta_{kl}\Big)\big(\omega_{i,n_k}(\gamma)-1\big)\right\}+O_P(n^{-1})$$
$$=\sum_{k=1}^{K}\hat{p}_k(1-\hat{p}_k)\left\{\frac{1}{n_k}\sum_{i=1}^{n_k}\Big(h_1\big(\boldsymbol{X}_i^{(k)}\big)-\Delta_{kl}\Big)\big(\omega_{i,n_k}(\gamma)-1\big)\right\}+O_P(n^{-1}).$$

By Theorem A of Section 6.4, Serfling (1980), we have

$$\sqrt{n}T_{n,\gamma}\to N\Big(0,\sum_{k=1}^{K}p_k(1-p_k)^2\sigma_1^2(\gamma)\Big),$$

where $\sigma_1^2(\gamma)=\mathrm{Var}\big(h_1(\boldsymbol{X})\big)\eta(\gamma)$. □

Proof of Theorem 3.1 Define

$$\Delta^M=\mathbb{E}d_M(\boldsymbol{X},\boldsymbol{X}'),$$
$$\Delta_k^M=\mathbb{E}d_M(\boldsymbol{X}^{(k)},\boldsymbol{X}^{(k)'}),\ k=1,2,...,K,$$
$$\Delta_{kl}^M=\mathbb{E}d(\boldsymbol{X}^{(k)},\boldsymbol{X}^{(l)}),\ k\ne l,k,l=1,2,...,K.$$

The proof of Theorem 3.1 is similar to the proof of Theorem 2.1 by replacing the the euclidean distance $\|\cdot\|$ by the induced distance $d_M(\cdot,\cdot)$. Here we provide a sketchy proof.

$$
\begin{aligned}
G_{n,\gamma} &= G_{n,\gamma} - G_n + G_n \\
&= \sum_{1\le k\ne l\le K} \hat{p}_k \hat{p}_l \frac{1}{n_k n_l} \sum_{i=1}^{n_k}\sum_{j=1}^{n_l} \Big(\omega_{i,n_k}(\gamma) - 1\Big) d_M(\boldsymbol{X}_i^{(k)}, \boldsymbol{X}_j^{(l)}) + G_n \\
&:= \Gamma_n + G_n.
\end{aligned}
$$

By Theorem 11 of Zhang et al. (2019), G_n is degenerate under independence of $\boldsymbol{X}$ and Y with $\mathrm{Var}(G_n) = O_p(n^{-2})$. We will show that Γ_n dominates with a normal limit.

Define $\hat{\Gamma}_{k,l}(\gamma) = \dfrac{1}{n_k n_l} \sum_{i=1}^{n_k} \sum_{j=1}^{n_l} \Big(\omega_{i,n_k}(\gamma) - 1\Big) d_M(\boldsymbol{X}_i^{(k)}, \boldsymbol{X}_j^{(l)})$, then $\Gamma_n = \sum_{1\le k\ne l\le K} \hat{p}_k \hat{p}_l \hat{\Gamma}_{k,l}(\gamma)$. Let $g(\boldsymbol{x}, \boldsymbol{x}') = d_M(\boldsymbol{x}, \boldsymbol{x}')$, and $g_1(\boldsymbol{x}) = \mathbb{E} d_M(\boldsymbol{x}, \boldsymbol{X}_1)$.

Under conditions **C**1–**C**3, $\hat{\Gamma}_{k,l}(\gamma)$ is dominated by $\dfrac{1}{n_k} \sum_{i=1}^{n_k} \Big(g_1\big(\boldsymbol{X}_i^{(k)}\big) - \Delta_{kl}^M\Big) \big(\omega_{i,n_k}(\gamma) - 1\big)$.

Therefore,

$$
\begin{aligned}
G_{n,\gamma} &= \sum_{1\le k\ne l\le K} \hat{p}_k \hat{p}_l \left\{ \frac{1}{n_k} \sum_{i=1}^{n_k} \Big(g_1\big(\boldsymbol{X}_i^{(k)}\big) - \Delta_{kl}^M\Big) \big(\omega_{i,n_k}(\gamma) - 1\big) \right\} + O_P(n^{-1}) \\
&= \sum_{k=1}^{K} \hat{p}_k (1 - \hat{p}_k) \left\{ \frac{1}{n_k} \sum_{i=1}^{n_k} \Big(g_1\big(\boldsymbol{X}_i^{(k)}\big) - \Delta_{kl}^M\Big) \big(\omega_{i,n_k}(\gamma) - 1\big) \right\} + O_P(n^{-1}).
\end{aligned}
$$

Applying Theorem A of Section 6.4 of Serfling (1980) again, we have

$$
\sqrt{n} G_{n,\gamma} \to N\Big(0, \sum_{k=1}^{K} p_k (1 - p_k)^2 \sigma_{2,M}^2(\gamma)\Big),
$$

where $\sigma_{2,M}^2(\gamma) = \mathrm{var}\big(g_1(\boldsymbol{X})\big)\eta(\gamma)$.

References

Ahmad IA (1993) Modification of some goodness-of-fit statistics to yield asymptotic normal null distributions. Biometrika 80:466–472

Balogoun AKS, Nkiet GM, Ogouyandjou C (2021) Asymptotic normality of a generalized maximum mean discrepancy estimator. Statist Probab Lett 169:108961

Chen Y, Dang X, Peng H, Bart H (2009) Outlier detection with the kernelized spatial depth function. IEEE Trans Pattern Anal Mach Intell 31(2):288–305

Dang X, Nguyen D, Chen X, Zhang J (2021) A new Gini correlation between quantitative and qualitative variables. Scand J Stat 48(4):1314–1343

Gebhard J, Schmitz N (1998) Permutation tests-a revival?! I Optimum properties. Stat Pap 39:75–85

Gebhard J, Schmitz N (1998) Permutation tests-a revival?! II. An efficient algorithm for computing the critical region. Stat Pap 39:87–96

Gini C (1914) On the measurement of concentration and variability of characters. Metron LXII I(1):3–8

Makigusa N, Naito K (2020) Asymptotic normality of a consistent estimator of maximum mean discrepancy in Hilbert space. Statist Probab Lett 156:108596

Manfoumbi Djonguet TK, Mbina Mbina A, Nkiet GM (2024) Testing independence of functional variables by an Hilbert–Schmidt independence criterion estimator. Statist Probab Lett 207:110016
Mercer J (1909) Functions of positive and negative type, and their connection the theory of integral equations. Philos Trans Roy Soc A 209:415–446
Neuhäuser M (2005) Exact tests based on the Baumgartner–Weiβ-Schindler statistic-a survey. Stat Pap 46:1–30
(1993) Asymptotic distributions of weighted U-statistics of degree 2. Ann Probab 21(2):1159–1169
Ryman N, Jorde PE (2001) Statistical power when testing for genetic differentiation. Mol Ecol 10:2361–2373
Sang Y, Dang X (2023) Asymptotic normality of Gini correlation in high dimension with applications to the K-sample problem. Electron J Stat 17(2):2539–2574
Serfling R (1980) Approximation theorems of mathematical statistics. Wiley, New York
Székely GJ, Rizzo ML (2013) Energy statistics: a class of statistics based on distances. J Stat Plan Infer 143:1249–1272
Székely GJ, Rizzo ML (2017) The energy of data. Ann Rev Stat Appl 4(1):447–479
Yitzhaki S, Schechtman E (2013) The Gini Methodology. Springer, New York
Zhang S, Dang X, Nguyen D, Wilkins D, Chen Y (2019) Estimating feature—label dependence using Gini distance statistics. IEEE Trans Pattern Anal Mach Intell 43(6):1947–1963