Electronic Journal of Statistics
Vol. 17 (2023) 2539–2574
ISSN: 1935-7524
https://doi.org/10.1214/23-EJS2165

# Asymptotic normality of Gini correlation in high dimension with applications to the $K$-sample problem

**Yongli Sang**

*Department of Mathematics*
*University of Louisiana at Lafayette*
*Lafayette, LA 70504*
*USA*
*e-mail:* yongli.sang@louisiana.edu

**Xin Dang***

*Department of Mathematics*
*University of Mississippi*
*University, MS 38677*
*USA*
*e-mail:* xdang@olemiss.edu

**Abstract:** The categorical Gini correlation proposed by Dang et al. [7] is a dependence measure to characterize independence between categorical and numerical variables. The asymptotic distributions of the sample correlation under dependence and independence have been established when the dimension of the numerical variable is fixed. However, its asymptotic behavior for high dimensional data has not been explored. In this paper, we develop the central limit theorem for the Gini correlation in the more realistic setting where the dimensionality of the numerical variable is diverging. We then construct a powerful and consistent test for the $K$-sample problem based on the asymptotic normality. The proposed test not only avoids computation burden but also gains power over the permutation procedure. Simulation studies and real data illustrations show that the proposed test is more competitive to existing methods across a broad range of realistic situations, especially in unbalanced cases.

**MSC2020 subject classifications:** Primary 60H20; secondary 60H15.
**Keywords and phrases:** Asymptotic normality, categorical Gini correlation, distance correlation, high dimensional $K$-sample test.

Received February 2022.

## Contents



*Corresponding author.

## 1. Introduction

Recently, Dang et al. [7] proposed the categorical Gini covariance and correlation, $\mathrm{gCov}(\boldsymbol{X}, Y)$ and $\mathrm{gCor}(\boldsymbol{X}, Y)$, to measure dependence of a $p$-variate numerical variable $\boldsymbol{X}$ and a categorical variable $Y$. Suppose that the categorical variable $Y$ takes values $L_1, \ldots, L_K$ and its distribution $P_Y$ is $P(Y = L_k) = p_k > 0$ for $k = 1, 2, \ldots, K$. $\boldsymbol{X}$ is from $F$ and $\psi$ denotes its characteristic function. Assume that the conditional distribution of $\boldsymbol{X}$ given $Y = L_k$ is $F_k$ with the corresponding characteristic function $\psi_k$. The Gini covariance is defined as

$$\mathrm{gCov}(\boldsymbol{X}, Y) = c(p) \sum_{k=1}^{K} p_k \int_{\mathbb{R}^p} \frac{|\psi_k(\boldsymbol{t}) - \psi(\boldsymbol{t})|^2}{\|\boldsymbol{t}\|^{p+1}} d\boldsymbol{t}, \tag{1.1}$$

where $c(p)$ is a known constant. The Gini covariance measures dependence of $\boldsymbol{X}$ and $Y$ by quantifying the difference between the conditional and the unconditional characteristic functions. The corresponding Gini correlation standardizes the Gini covariance to have a range in [0,1]. Zero Gini covariance or correlation mutually implies independence.

Another dependence measure that could characterize independence between two random variables is the popular distance correlation proposed by Szekely, Rizzo and Bakirov [27]. It is flexible for $\boldsymbol{X}$ and $\boldsymbol{Y}$ in arbitrary dimensions and any types (numerical or categorical). It has attracted much attention since then, see e.g. [11, 19, 20, 28, 29, 30, 31, 35, 36] and references therein. In the case of $p$-variate $\boldsymbol{X}$ and categorical $Y$, the distance covariance becomes

$$\mathrm{dCov}(\boldsymbol{X}, Y) = c(p) \sum_{k=1}^{K} p_k^2 \int_{\mathbb{R}^p} \frac{|\psi_k(\boldsymbol{t}) - \psi(\boldsymbol{t})|^2}{\|\boldsymbol{t}\|^{p+1}} d\boldsymbol{t}. \tag{1.2}$$

Comparing (1.1) and (1.2), we see that the two covariances are closely related. When the categorical variable $Y$ takes two values ($K = 2$) or $P_Y$ is uniform,

they are only different with a scaling factor [7]. While for the general $K \geq 3$ and unbalanced $P_Y$, the Gini covariance is a better dependence measure than the distance covariance because the weight $p_k$ in (1.1) takes the nature of the categorical variable, while $\mathrm{dCov}(\boldsymbol{X}, Y)$, due to its squared weights, is dominated by the classes with large probabilities and the contribution from smaller classes is substantially reduced.

A fruitful research has been developed to study the asymptotic distributions of the sample distance statistics in different scenarios. Under independence of $\boldsymbol{X} \in \mathbb{R}^p$ and $\boldsymbol{Y} \in \mathbb{R}^q$, the standard sample distance covariance or correlation converges in distribution to a mixture of chi-squared distribution in the classical setting where the sample size $n \to \infty$ and $p, q$ are fixed [27, 15]; while in the high-dimension-low-sample size (HDLSS) setting when $p, q \to \infty$ and $n$ is fixed, Székely and Rizzo [30] derived a $t$-distribution limit of the unbiased sample covariance by assuming that the components of the high-dimensional vectors in $\boldsymbol{X}$ and in $\boldsymbol{Y}$ are exchangeable; Zhu et al. [35] relaxed that assumption and considered the high-dimensional medium-sample-size setting (HDMSS) where $n, p, q \to \infty$ but $p, q$ growing more rapidly than $n$; Gao et al. [11] have developed central limit theorems in a more realistic high dimensional high-sample-size setting (HDHSS) where $n$ and $p + q$ diverge in an arbitrary fashion. This result applies for the sample distance covariance of (1.2) in which $q = 1$ and $n, p \to \infty$. However, there is no literature to study limiting distributions of sample Gini covariance and correlation in high dimension.

In the classical setting, Dang et al. [7] studied the asymptotic distributions of the $V$-statistic sample Gini covariance and correlation with the dimension of $\boldsymbol{X}$ fixed. They admit normal limits when $\boldsymbol{X}$ and $Y$ are dependent and converge in distribution to a quadratic form of centered Gaussian random variables when $\boldsymbol{X}$ and $Y$ are independent. In this paper, we will work with unbiased $U$-statistic covariance estimator and associated sample Gini correlation in high dimension. The first objective of this paper is to establish their asymptotic distributions for sample Gini covariance and correlation under independence of $\boldsymbol{X}$ and $Y$ in the HDHSS setting.

The derived asymptotic distributions can be used for independence test. In other words, it is to test the equality of $K$ conditional distributions, which is the classical $K$-sample problem encountered in almost every research field. Due to its fundamental importance and wide applications, research for this $K$-sample problem has been kept active since 1940's. For example, the widely used and well-studied tests such as Cramér-von Mises test [17, 6], Anderson-Darling test [8, 25] and their variations utilize different norms on the difference of empirical distribution functions, while some [1, 21] are based on the comparison of density estimators if the underlying distributions are continuous. Other tests [26, 10] are based on the difference between characteristic functions. Indeed, the test in [26] is equivalent to ours, but their test only considers the case of $K = 2$. Another equivalent test is the DISCO [23] whose test statistic is the ratio of the between Gini variation and the within Gini variation, while our Gini correlation is the ratio of the between Gini variation and the total Gini variation. Heller, Heller and Gorfine [13] and Heller et al. [14] proposed a dependence test

based on rank distances. All those distance-based tests require a permutation procedure to determine the critical values. Sang, Dang and Zhao [24] developed a nonparametric test which applies the jackknife empirical likelihood and has a standard limiting chi-squared distribution. Other tests viewing the $K$-sample test as an independent test between a numerical and categorical variable can be found in [4, 16, 33]. However, most the afore-mentioned work focuses on the fixed dimensional case and perform poorly or may even fail in high dimension.

Recently, several distance-based tests for two-sample problem have been proposed in high dimension, see [3, 5, 19, 36]. Li [19] constructed a test based on interpoint distances under HDLSS. Zhu and Shao [36] studied the two-sample problem using energy distance (ED) and maximum mean discrepancy with Gaussian and Laplacian kernels under HDLSS and HDMSS, in which they have shown that all these tests are inconsistent under some scenarios. The general $K$-sample testing in high dimension is more challenging and results in literature are very scarce. Mukhopadhyay and Wang [22] constructed a graph-based nonparametric approach under HDLSS. However, the power for the test is extremely low under some settings. Gao et al. [11] tested the $K$-sample problem in high dimension based the distance correlation.

Our second objective of this paper is to use the asymptotic result of the Gini covariance or correlation to construct a new consistent $K$-sample test in high dimension. The advantages of the new test include

- Computational efficiency. It avoids a permutation procedure which is computationally expansive.
- Statistical efficiency. It gains power over the nonparametric tests.
- Robustness for class imbalance. It is more appropriate than the distance based tests in dealing with unbalanced data.

Throughout this paper, if not mentioned otherwise, the letter $C$, with or without a subscript, denotes a generic positive finite constant whose exact value is independent of sample sizes and may change from line to line. $\|\cdot\|$ represents the Euclidean norm, that is, $\|\boldsymbol{a}\| = \sqrt{a_1^2 + a_2^2 + \cdots + a_p^2}$ for a $p$-vector $\boldsymbol{a} = (a_1, a_2, \ldots, a_p)^T \in \mathbb{R}^p$. 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 Section 2, we first briefly review the other representation of Gini correlation and review the existing statistical inference, then we present a $U$-estimator for the Gini correlation and the central limit theorem for the $U$-estimator when both the sample sizes and dimensionality are diverging. The $K$-sample test is proposed and its consistency is established. In Section 3, we conduct simulation studies to evaluate the performance of the proposed test. A real data analysis is illustrated in Section 4 to compare the proposed test with current existing approaches. We conclude and discuss future works in Section 5. All technical proofs are provided in Appendix.

## 2. Inference of Gini covariance and correlation in high dimension

### *2.1. Categorical Gini correlation*

The Gini covariation between $\boldsymbol{X}$ and $Y$ defined in (1.1) can be represented in the multivariate Gini mean differences (GMD). Let $(\boldsymbol{X}_1, \boldsymbol{X}_2)$ and $(\boldsymbol{X}_1^{(k)}, \boldsymbol{X}_2^{(k)})$ be independent pair variables independently from $F$ and $F_k$ respectively. Define $\Delta = \mathbb{E}\|\boldsymbol{X}_1 - \boldsymbol{X}_2\|$ as the GMD of $F$ and $\Delta_k = \mathbb{E}\|\boldsymbol{X}_1^{(k)} - \boldsymbol{X}_2^{(k)}\|$ as the GMD of $F_k$. From [7], we have

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

Then the Gini correlation is

$$\mathrm{gCor}(\boldsymbol{X}, Y) = \frac{\Delta - \sum_{k=1}^{K} p_k \Delta_k}{\Delta}. \tag{2.1}$$

This representation allows a nice interpretation. The Gini covariance is the between Gini variation and the Gini correlation is the ratio of the between and the total variation. Also from this representation, it is straightforward to obtain sample estimators.

Dang et al. [7] used V-statistic estimators and derived limiting distributions of the estimators under the classical setting when the dimension of $\boldsymbol{X}$ is fixed. More specifically, suppose a sample $\mathcal{D} = \{(\boldsymbol{X}_1, Y_1), (\boldsymbol{X}_2, Y_2), \ldots, (\boldsymbol{X}_n, Y_n)\}$ is drawn from the joint distribution of $\boldsymbol{X}$ and $Y$. Write $\mathcal{D} = \mathcal{D}_1 \cup \mathcal{D}_2 \cdots \cup \mathcal{D}_K$, where $\mathcal{D}_k = \{\boldsymbol{X}_1^{(k)}, \boldsymbol{X}_2^{(k)}, \ldots, \boldsymbol{X}_{n_k}^{(k)}\}$ is the sample with $Y_i = L_k$ and $n_k$ is the number of sample observations in the $k^{th}$ class. Plugging in $\hat{p}_k = n_k/n$ and V-statistics

$$\tilde{\Delta}_k = n_k^{-2} \sum_{1 \le i,j \le n_k} \|\boldsymbol{X}_i^{(k)} - \boldsymbol{X}_j^{(k)}\|, \quad \tilde{\Delta} = n^{-2} \sum_{1 \le i,j \le n} \|\boldsymbol{X}_i - \boldsymbol{X}_j\|$$

to (2.1), the estimator $\hat{\rho}_g(\boldsymbol{X}, Y)$ is obtained. Under the assumption of $\mathbb{E}\|\boldsymbol{X}\|^2 < \infty$ with $p$ fixed and $n \to \infty$, $\hat{\rho}_g(\boldsymbol{X}, Y)$ has the limiting distributions as below.

1. If $\mathrm{gCor}(\boldsymbol{X}, Y) \neq 0$, then

$$\sqrt{n}\big(\hat{\rho}_g(\boldsymbol{X}, Y) - \mathrm{gCor}(\boldsymbol{X}, Y)\big) \xrightarrow{d} \mathcal{N}\big(0, \sigma_g^2\big),$$

   where $\sigma_g^2$ is the asymptotic variance.
2. If $\mathrm{gCor}(\boldsymbol{X}, Y) = 0$, then

$$n\hat{\rho}_g(\boldsymbol{X}, Y) \xrightarrow{d} \frac{4}{\Delta}\left[\sum_{s=1}^{\infty}\sum_{k=1}^{K}(1 - p_k)\lambda_s Z_{s,k}^2 + \sum_{s=1}^{\infty}\sum_{k<l}\sqrt{p_k p_l}\lambda_s Z_{s,k} Z_{s,l}\right], \tag{2.2}$$

   where $Z_{s,k}$ $(k = 1, \ldots, K, s = 1, 2, \ldots)$ are independent standard normal variates and $\lambda_s$ are nonnegative coefficients.

Under independence of $\boldsymbol{X}$ and $Y$, $\hat{\rho}_g(\boldsymbol{X}, Y)$ converges to a quadratic form of normal random variables. This result is difficult to be applied for the independence test, and hence one has to rely on a permutation procedure to determine a critical value of the test, which is computationally expansive.

This result is obtained under the classical setting. The inference for the Gini correlation in high dimension has not been explored and we will fill this gap by developing the asymptotic distributions when both the sample sizes and the dimensionality diverge to infinity.

### 2.2. *U*-estimators and projection representation

When the dimension $p$ is large, the V-statistic Gini covariance and correlation estimators may have issues about bias. Therefore, we will estimate the GMDs by unbiased $U$-statistics. That is,

$$\begin{aligned}\hat{\Delta} &= \binom{n}{2}^{-1} \sum_{1\le i<j\le n} \|\boldsymbol{X}_i - \boldsymbol{X}_j\| := U_n;\\ \hat{\Delta}_k &= \binom{n_k}{2}^{-1} \sum_{1\le i<j\le n_k} \|\boldsymbol{X}_i^{(k)} - \boldsymbol{X}_j^{(k)}\| := U_{n_k}.\end{aligned} \tag{2.3}$$

Thus Gini covariance and correlation can be estimated by

$$\mathrm{gCov}_n(\boldsymbol{X}, Y) = \hat{\Delta} - \sum_{k=1}^{K} \hat{p}_k \hat{\Delta}_k = U_n - \sum_{k=1}^{K} \hat{p}_k U_{n_k} \tag{2.4}$$

and

$$\mathrm{gCor}_n(\boldsymbol{X}, Y) = \frac{\mathrm{gCov}_n(\boldsymbol{X}, Y)}{\hat{\Delta}} = \frac{U_n - \sum_{k=1}^{K} \hat{p}_k U_{n_k}}{U_n},$$

respectively. Both of them are functions of $U$-statistics $U_n$ and $U_{n_k}$'s. We shall focus on the asymptotic distribution of $\mathrm{gCov}_n(\boldsymbol{X}, Y)$. The application of Slutsky's theorem allows us to obtain the result on $\mathrm{gCor}_n(\boldsymbol{X}, Y)$ immediately.

Under independence of $\boldsymbol{X}$ and $Y$, the sample Gini covariance $\mathrm{gCov}_n$ in (2.4) is a linear combination of $U$-statistics with first-order degeneracy. By classical theory about $U$ statistics in the fixed dimensional asymptotic (fixed dimension with sample sizes diverge to infinity), a non-normal limiting distribution holds, a similar result as (2.2). However, as both the the dimension and the sample size go large, the degenerate $U$-statistic will admit a normal limit. To establish this result, we first take decompositions of U-statistics in (2.3) and rewrite (2.4).

By the Hoeffding decomposition, we have

$$U_n = \Delta + \frac{2}{n}\sum_{i=1}^{n}\{\mathbb{E}(\|\boldsymbol{X} - \boldsymbol{X}_i\||\boldsymbol{X}_i) - \Delta\} + \binom{n}{2}^{-1} \sum_{1\le i<j\le n} d(\boldsymbol{X}_i, \boldsymbol{X}_j),$$

where

$$d(\boldsymbol{X}_1, \boldsymbol{X}_2) = \|\boldsymbol{X}_1 - \boldsymbol{X}_2\| - \mathbb{E}(\|\boldsymbol{X}_1 - \boldsymbol{X}_2\| \mid \boldsymbol{X}_1) - \mathbb{E}(\|\boldsymbol{X}_1 - X_2\| \mid \boldsymbol{X}_2) + \mathbb{E}\|\boldsymbol{X}_1 - \boldsymbol{X}_2\| \tag{2.5}$$

is called the double centered distance and it is actually the second order centered projection of the kernel function of $U_n$. Analogously,

$$U_{n_k} = \Delta_k + \frac{2}{n_k}\sum_{i=1}^{n_k}\{\mathbb{E}(\|\boldsymbol{X}^{(k)} - \boldsymbol{X}_i^{(k)}\||\boldsymbol{X}_i^{(k)}) - \Delta_k\} + \binom{n_k}{2}^{-1}\sum_{1\le i<j\le n_k} d(\boldsymbol{X}_i^{(k)}, \boldsymbol{X}_j^{(k)}).$$

Under independence of $\boldsymbol{X}$ and $Y$, we have $F_1 = F_2 = \cdots = F_K = F$. Hence $\Delta = \Delta_k,\ k = 1, 2, \ldots, K$ and

$$\sum_{k=1}^{K}\hat{p}_k\frac{2}{n_k}\sum_{i=1}^{n_k}\mathbb{E}(\|\boldsymbol{X}^{(k)} - \boldsymbol{X}_i^{(k)}\||\boldsymbol{X}_i^{(k)}) = \frac{2}{n}\sum_{i=1}^{n}\mathbb{E}(\|\boldsymbol{X} - \boldsymbol{X}_i\||\boldsymbol{X}_i).$$

Then we can represent (2.4) as

$$\mathrm{gCov}_n(\boldsymbol{X}, Y) = \binom{n}{2}^{-1}\sum_{1\le i<j\le n} d(\boldsymbol{X}_i, \boldsymbol{X}_j) - \sum_{k=1}^{K}\hat{p}_k\binom{n_k}{2}^{-1}\sum_{1\le i<j\le n_k} d(\boldsymbol{X}_i^{(k)}, \boldsymbol{X}_j^{(k)}), \tag{2.6}$$

under the null that $\boldsymbol{X}$ and $Y$ are independent.

The representation of (2.6) has advantages over (2.4) due to appealing orthogonal properties of $d(\boldsymbol{X}_1, \boldsymbol{X}_2)$ as stated in Lemmas A.1 and A.2 in Appendix. Those properties largely simplify the calculation of specific moments involved.

### *2.3. Asymptotic normality*

We study the asymptotic distributions of the $U$-estimators in this section. Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3$ and $\boldsymbol{X}_4$ be i.i.d copies of $\boldsymbol{X}$. The following conditions will be needed to facilitate the proofs.

**C**1. $\mathbb{E}\|\boldsymbol{X}\|^4 < \infty$;
**C**2. $\frac{\mathbb{E}d^4(\boldsymbol{X}_1,\boldsymbol{X}_2)}{n(\mathbb{E}d^2(\boldsymbol{X}_1,\boldsymbol{X}_2))^2} \to 0$;
**C**3. $\frac{\mathbb{E}d(\boldsymbol{X}_1,\boldsymbol{X}_3)d(\boldsymbol{X}_2,\boldsymbol{X}_3)d(\boldsymbol{X}_1,\boldsymbol{X}_4)d(\boldsymbol{X}_2,\boldsymbol{X}_4)}{(\mathbb{E}d^2(\boldsymbol{X}_1,\boldsymbol{X}_2))^2} \to 0$;
**C**4. $\sqrt{n}\mathrm{gCov}(\boldsymbol{X}, Y) \to \infty$.

*Remark* 2.1. Our conditions **C**2 and **C**3 are corresponding to conditions (18) and (19) in [11] when $\tau = 1$. In fact, the condition **C**2 can be weaken to be $\frac{\mathbb{E}(|d(\boldsymbol{X}_1,\boldsymbol{X}_2)|^{2+2\alpha})}{n^\alpha(\mathbb{E}d^2(\boldsymbol{X}_1,\boldsymbol{X}_2))^2} \to 0$ for some constant $0 < \alpha \le 1$. However, it is hard to check the condition when $0 < \alpha < 1$, so we take the stronger but simple condition.

Applying Martingale central limit theorem, we establish the limiting distribution of the sample Gini covariance in the following theorem.

**Theorem 2.1.** *Under independence of* $\boldsymbol{X}$ *and* $Y$*, and conditions* ***C1–C3****, as* $\min\{n_1, n_2, \ldots, n_k\} \to \infty$ *and* $p \to \infty$*, we have*

$$\frac{gCov_n(\boldsymbol{X}, Y)}{\sigma_0} \xrightarrow{d} \mathcal{N}(0,1),$$

*where* $\sigma_0^2 = (\sum_k \hat{p}_k^2 \binom{n_k}{2}^{-1} - \binom{n}{2}^{-1}) \mathbb{E}d^2(\boldsymbol{X}_1, \boldsymbol{X}_2)$ *is the variance of* $gCov_n(\boldsymbol{X}, Y)$.

Theorem 2.1 reveals that a degenerate $U$-statistic admits a normal limit due to the high dimensionality. This is surprisingly inspiring to deal with problems which can be estimated by $U$-statistics in high dimension.

To make inference feasible, we need to estimate $\sigma_0^2$. A consistent estimator $\hat{\sigma}_0^2$ is

$$\hat{\sigma}_0^2 = \left( \sum_{k=1}^{K} \hat{p}_k^2 \binom{n_k}{2}^{-1} - \binom{n}{2}^{-1} \right) V_n^2(\boldsymbol{X}), \tag{2.7}$$

where $V_n^2(\boldsymbol{X})$ is the bias-corrected estimator for the squared distance variance in [30]. That is,

$$V_n^2(\boldsymbol{X}) = \frac{1}{n(n-3)} \sum_{1 \le k \ne l \le n} A_{k,l}^2$$

with $A_{k,l}$ being the centered sample distance, which is

$$A_{k,l} = \|\boldsymbol{X}_k - \boldsymbol{X}_l\| - \frac{1}{n-2} \sum_{i=1}^{n} \|\boldsymbol{X}_i - \boldsymbol{X}_l\| - \frac{1}{n-2} \sum_{j=1}^{n} \|\boldsymbol{X}_k - \boldsymbol{X}_j\| + \frac{1}{(n-1)(n-2)} \sum_{1 \le i,j \le n} \|\boldsymbol{X}_i - \boldsymbol{X}_j\|.$$

**Theorem 2.2.** *Under independence of* $\boldsymbol{X}$ *and* $Y$*, and conditions* ***C1–C3****, as* $\min\{n_1, n_2, \ldots, n_k\} \to \infty$ *and* $p \to \infty$*, we have*

$$\frac{gCov_n(\boldsymbol{X}, Y)}{\hat{\sigma}_0} \xrightarrow{d} \mathcal{N}(0,1).$$

The estimators in (2.3) are $U$-statistics and hence the ratio is consistent with $\hat{\Delta}/\Delta \to 1$ in probability. By applying Slutsky's theorem, we have the CLT for the Gini correlation.

**Corollary 2.1.** *Under independence of* $\boldsymbol{X}$ *and* $Y$*, and conditions* ***C1–C3****, as* $\min\{n_1, n_2, \ldots, n_k\} \to \infty$ *and* $p \to \infty$*, we have*

$$\frac{\hat{\Delta}}{\hat{\sigma}_0} gCor_n(\boldsymbol{X}, Y) \xrightarrow{d} \mathcal{N}(0,1).$$

From the result of (2.2) and Corollary 2.1, we see that when $\boldsymbol{X}$ and $Y$ are independent, as the dimensionality of the numerical variable goes large and under some conditions on the fourth moment, the complicate quadratic form of normal distributions converges to a normal distribution.

### *2.4. High-dimensional K-sample test*

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

The independence test is stated as

$$\mathcal{H}_0 : \mathrm{gCov}(\boldsymbol{X}, Y) = 0, \quad \text{vs} \quad \mathcal{H}_1 : \mathrm{gCov}(\boldsymbol{X}, Y) > 0. \tag{2.8}$$

Note that the null hypothesis of the test in (2.8) is equivalent to the null of the $K$-sample test

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

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 $\mathcal{D} = \mathcal{D}_1 \cup \mathcal{D}_2 \cdots \cup \mathcal{D}_K$ has the distribution $F$, which is the average distribution of $F_k$'s.

By Theorem 2.1, we can reject $\mathcal{H}_0$ or $\mathcal{H}_0'$ if $\mathrm{gCov}_n(\boldsymbol{X}, Y) > Z_\alpha \hat{\sigma}_0$ at level $\alpha$, where $Z_\alpha$ is the $(1-\alpha)100\%$ percentile of the standard normal distribution.

For $K = 2$, the two sample problem, the proposed test is asymptotically equivalent to the test based on distance covariance because $gCov(\boldsymbol{X}, Y) = dCov(\boldsymbol{X}, Y)/\sqrt{dCov(Y, Y)}$. This is the result of Remark 9 in [7]. And hence two test statistics estimate a same population quantity. They are also asymptotically equivalent to Székely's energy test [26, 2] that is based on energy statistic between $F_1$ and $F_2$.

Theorem 2.1 allows us to avoid computation burden of the permutation tests. As demonstrated in the simulation, the test based on the limiting normality is more powerful than the permutation tests. The power function for the proposed test is

$$P_n(\alpha) = P\big(\mathrm{gCov}_n(\boldsymbol{X}, Y) > Z_\alpha \hat{\sigma}_0 \mid \mathcal{H}_1\big).$$

The test consistency is established in the below theorem.

**Theorem 2.3.** *For any alternative $\mathcal{H}_1$ satisfying conditions **C1** and **C4**, as $\min\{n_1, n_2, \ldots, n_k\} \to \infty$, $p_k > 0$ and $p \to \infty$, we have*

$$P_n(\alpha) = P\big(gCov_n(\boldsymbol{X}, Y) > Z_\alpha \hat{\sigma}_0 \mid \mathcal{H}_1\big) \to 1.$$

Condition **C**1 is the usual assumption on the finite fourth moment. Condition **C**4, $\sqrt{n} gCov(\boldsymbol{X}, Y) \to \infty$, requires dependence of $\boldsymbol{X}$ and $Y$ cannot be too weak. We might state a local alternative as

$$\mathcal{H}_1' : \mathrm{gCov}(\boldsymbol{X}, Y) \geq Cn^{-t}, \quad \text{for } t < 1/2.$$

The proposed test is able to detect the dependence under $\mathcal{H}_1'$ with power going to 1 as sample sizes increase.

## 3. Simulation study

In this section, we conduct three simulation studies to verify the theoretical properties of the standardized Gini covariance statistic and compare its performance in $K$-sample tests with others.

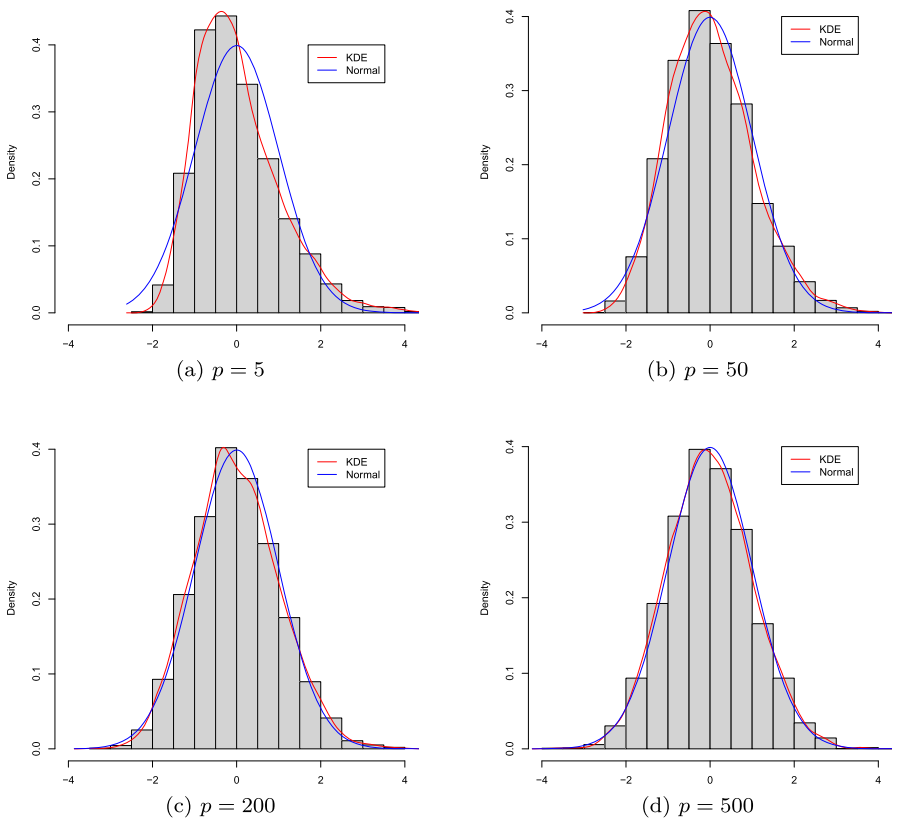



FIG 1. *Histograms of the standardized Gini covariance statistic in Example 1 with kernel density estimation curves in red and standard normal density curves in blue.*

### *3.1. Limiting normality*

We generate independent $K$ samples from the same multivariate normal distributions and compute the standardized Gini covariance statistic. The procedure is repeated 5000 times. The setup parameters are listed below.

*Example* 1. $K = 5$ samples of sizes $\boldsymbol{n} = (30, 40, 50, 60, 70)$ are generated from $\mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$, where $p = 5, 50, 200, 500$ and $\boldsymbol{\Sigma} = (\Sigma_{ij}) \in \mathbb{R}^{p\times p}$ with $\Sigma_{ij} = 0.7^{|i-j|}$.

For each dimension $p$, the histogram of 5000 standardized Gini covariance statistics is plotted in Figure 1. Also the kernel density estimation (KDE) curve

and the standard normal density curve are added to the histogram plot to visualize closeness between empirical density and asymptotical density functions. For $p = 5$ in Figure 1(a), the histogram is slightly right-skewed and there is some discrepancy between KDE and the normal curve. But when dimension increases, the discrepancy becomes less and diminishes as shown in Figure 1(b)–(d). We also calculate the maximum point distance between KDE and Normal density function as a measure of discrepancy in Table 1. It is clear that the difference decreases with dimensionality. Gao et al. [11] developed the limiting normal distribution for distance correlation, so we also involve the maximum point distance between KDE for the distance measure. Comparing with the scaled distance covariance statistic, the Gini one has a better normal approximation in each dimension.

TABLE 1

*The maximum point distances between the kernel density estimation function and standard normal density function. $KDE_g$ is for rescaled $gCov_n$ and $KDE_d$ for $dCov_n$.*

| Distance | $p = 5$ | $p = 50$ | $p = 200$ | $p = 500$ |
|---|---|---|---|---|
| Dist($KDE_g$, Normal) | 0.1176 | 0.0478 | 0.0294 | 0.0177 |
| Dist($KDE_d$, Normal) | 0.1290 | 0.0493 | 0.0338 | 0.0207 |

### *3.2. Size and power in K-sample tests*

In this simulation, we compare five methods for $K$ sample problem. Two of them are permutation tests. The one based on distance covariance in high dimension has been studied in [36] for $K = 2$. Here we examine both permutation tests for $K$ sample problem in high dimension. Five methods are

**gCov:** our proposed method using rescaled Gini covariance statistic and the normal percentile as the critical value.

**gCov-perm:** permutation test using Gini covariance statistic. This test is asymptotically equivalent to the one-way DISCO method [23].

**dCov:** the method using rescaled distance covariance statistic using the percentile of the standard normal as the critical value [11].

**dCov-perm:** permutation test using distance covariance statistic.

**GLP:** graphic LP polynomial basis function method proposed in [22].

We consider $K = 3$ case in dimensions $p = 200, 500$ with the equal size $\boldsymbol{n} = (40, 40, 40)$, slightly unbalanced size $\boldsymbol{n} = (50, 40, 30)$ and heavily unbalanced size $\boldsymbol{n} = (72, 36, 12)$. Let

$$\boldsymbol{\mu}_1 = \mathbf{0}_p;\ \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma} = (\Sigma_{ij}) \in \mathbb{R}^{p\times p},\ \text{where } \Sigma_{ij} = 0.7^{|i-j|};$$

$$\boldsymbol{\mu}_2 = \big(0.1 \times \mathbf{1}_{\beta p}^T, \mathbf{0}_{(1-\beta)p}^T\big)^T;\ \boldsymbol{\Sigma}_2 = \boldsymbol{D}_1\boldsymbol{\Sigma}\boldsymbol{D}_1 \text{ with } \boldsymbol{D}_1 = \operatorname{diag}\big(1.1 \times \mathbf{1}_{\beta p}^T, \mathbf{1}_{(1-\beta)p}^T\big);$$

$$\boldsymbol{\mu}_3 = \big(0.2 \times \mathbf{1}_{\beta p}^T, \mathbf{0}_{(1-\beta)p}^T\big)^T;\ \boldsymbol{\Sigma}_2 = \boldsymbol{D}_2\boldsymbol{\Sigma}\boldsymbol{D}_2 \text{ with } \boldsymbol{D}_2 = \operatorname{diag}\big(1.2 \times \mathbf{1}_{\beta p}^T, \mathbf{1}_{(1-\beta)p}^T\big).$$

Here $\beta \in [0, 1]$ is the proportion of the $p$ components for which 3 samples differ in mean and in variance.

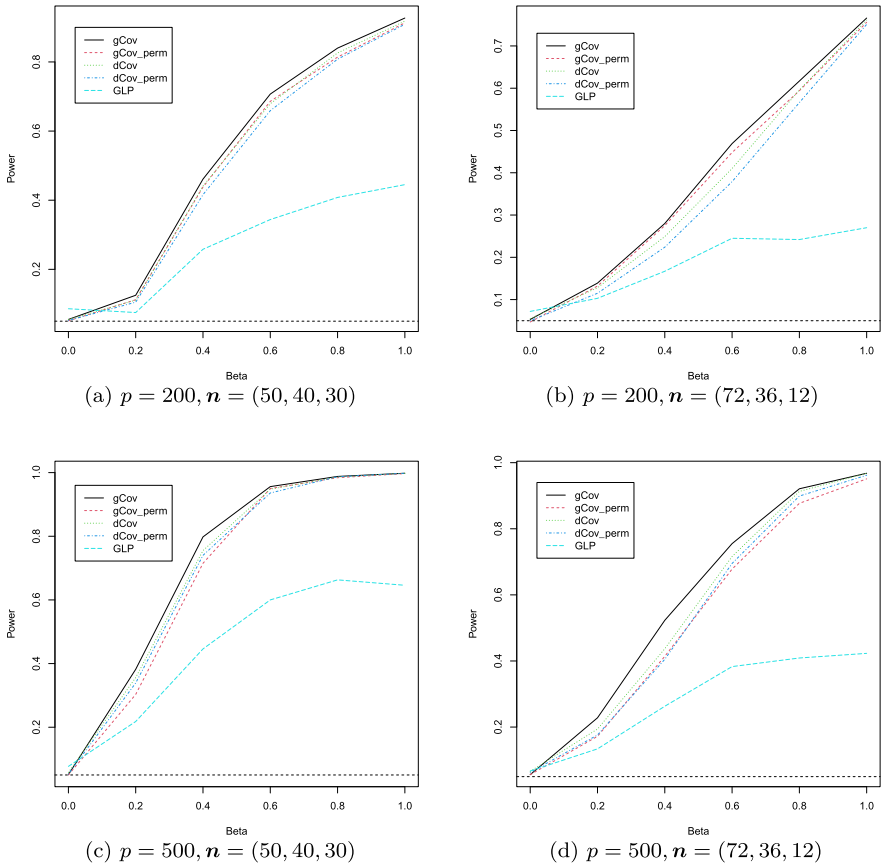
TABLE 2
*Size and Power of Tests for $K = 3$ samples in Example 2.*

| $p$ | $\boldsymbol{n}$ | method | $\beta = 0$ | $\beta = .2$ | $\beta = .4$ | $\beta = .6$ | $\beta = .8$ | $\beta = 1$ |
|---|---|---|---|---|---|---|---|---|
| 200 | (40,40,40) | gCov | .052 | .171 | .421 | .692 | .864 | .966 |
| | | gCov-perm | .050 | .123 | .327 | .609 | .815 | .942 |
| | | dCov | .053 | .172 | .423 | .691 | .867 | .966 |
| | | dCov-perm | .043 | .159 | .416 | .665 | .852 | .954 |
| | | GLP | .060 | .098 | .254 | .466 | .720 | .875 |
| | (50,40,30) | gCov | .065 | .183 | .484 | .718 | .882 | .949 |
| | | gCov-perm | .061 | .133 | .402 | .621 | .823 | .914 |
| | | dCov | .068 | .170 | .454 | .699 | .873 | .948 |
| | | dCov-perm | .062 | .160 | .417 | .664 | .858 | .948 |
| | | GLP | .069 | .096 | .241 | .455 | .687 | .845 |
| | (72,36,12) | gCov | .058 | .155 | .282 | .476 | .632 | .814 |
| | | gCov-perm | .049 | .110 | .212 | .391 | .555 | .749 |
| | | dCov | .063 | .112 | .233 | .444 | .606 | .802 |
| | | dCov-perm | .060 | .104 | .215 | .419 | .571 | .780 |
| | | GLP | .066 | .090 | .178 | .264 | .403 | .577 |
| 500 | (40,40,40) | gCov | .061 | .268 | .665 | .942 | .997 | 1.00 |
| | | gCov-perm | .063 | .207 | .587 | .904 | .993 | 1.00 |
| | | dCov | .063 | .274 | .667 | .943 | .997 | 1.00 |
| | | dCov-perm | .060 | .269 | .654 | .934 | .998 | 1.00 |
| | | GLP | .049 | .143 | .455 | .812 | .971 | .999 |
| | (50,40,30) | gCov | .052 | .340 | .801 | .972 | .997 | .999 |
| | | gCov-perm | .055 | .280 | .727 | .950 | .990 | .999 |
| | | dCov | .058 | .313 | .776 | .961 | .995 | .999 |
| | | dCov-perm | .051 | .308 | .762 | .956 | .994 | .998 |
| | | GLP | .059 | .156 | .428 | .800 | .956 | .993 |
| | (72,36,12) | gCov | .051 | .231 | .493 | .769 | .923 | .979 |
| | | gCov-perm | .055 | .154 | .399 | .671 | .901 | .968 |
| | | dCov | .054 | .175 | .426 | .721 | .916 | .978 |
| | | dCov-perm | .052 | .172 | .420 | .711 | .909 | .976 |
| | | GLP | .047 | .109 | .240 | .450 | .688 | .853 |

*Example* 2. Generate samples of $\boldsymbol{X}^{(1)} \sim \mathcal{N}_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$, $\boldsymbol{X}^{(2)} \sim \mathcal{N}_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ and $\boldsymbol{X}^{(3)} \sim \mathcal{N}_p(\boldsymbol{\mu}_3, \boldsymbol{\Sigma}_3)$.

We conduct 1000 simulations. The size and power of each test are computed and reported in Table 2. The column $\beta = 0.0$ corresponds to the size of tests. Several observations can be drawn. All tests maintain the nominal level 5% quite well. Permutation tests are slightly less powerful than their corresponding counterparts. GLP test is inferior to others in all cases. In the equal size case, Gini method $gCov$ produces almost the same size and power as $dCov$, which is an expected result since the Gini covariance and distance covariance are asymptotically equivalent. While in the unbalanced cases, our Gini method gains 1%–6% power advantage over the distance one. An intuitive interpretation of the advantage is that $gCov$ is a better measure than $dCov$ in unbalanced distributions as stated in the Introduction section.

*Example* 3. Let $\boldsymbol{Z}_k = (Z_{k1}, Z_{k2}, \ldots, Z_{kp})^T - \mathbf{1}_p$, where for $k = 1, 2, 3$ and




(a) $p = 200, \boldsymbol{n} = (50, 40, 30)$ (b) $p = 200, \boldsymbol{n} = (72, 36, 12)$

(c) $p = 500, \boldsymbol{n} = (50, 40, 30)$ (d) $p = 500, \boldsymbol{n} = (72, 36, 12)$

FIG 2. *Size and power of tests in Example 3. Dashed horizontal line is the nominal level 0.05.*

$j = 1, \ldots, p$, $Z_{kj}$'s are i.i.d. from Exp(1). Then generate $\boldsymbol{X}_1 \sim \boldsymbol{\Sigma}_1^{1/2}\boldsymbol{Z}_1$, $\boldsymbol{X}_2 \sim \boldsymbol{\Sigma}_2^{1/2}\boldsymbol{Z}_2$, $\boldsymbol{X}_3 \sim \boldsymbol{\Sigma}_3^{1/2}\boldsymbol{Z}_3$ samples.

Although the distributions are not elliptically symmetric, the patterns and observations from this simulation are very similar to those in Example 2 for all tests but GLP. We present the results in Figure 2. GLP seems sensitive to the asymmetry of distributions not only in terms of performance as well as in terms of computation. The GLP algorithm includes a middle step to perform $K$-mean clustering, and that step occasionally stops especially for unbalanced sample sizes. The GLP is slightly oversized and its power is extremely low.

## 4. Real data analysis

Two data sets from UCI machine learning repository [9] are studied for $K$ sample tests.

### 4.1. LSVT voice rehabilitation data

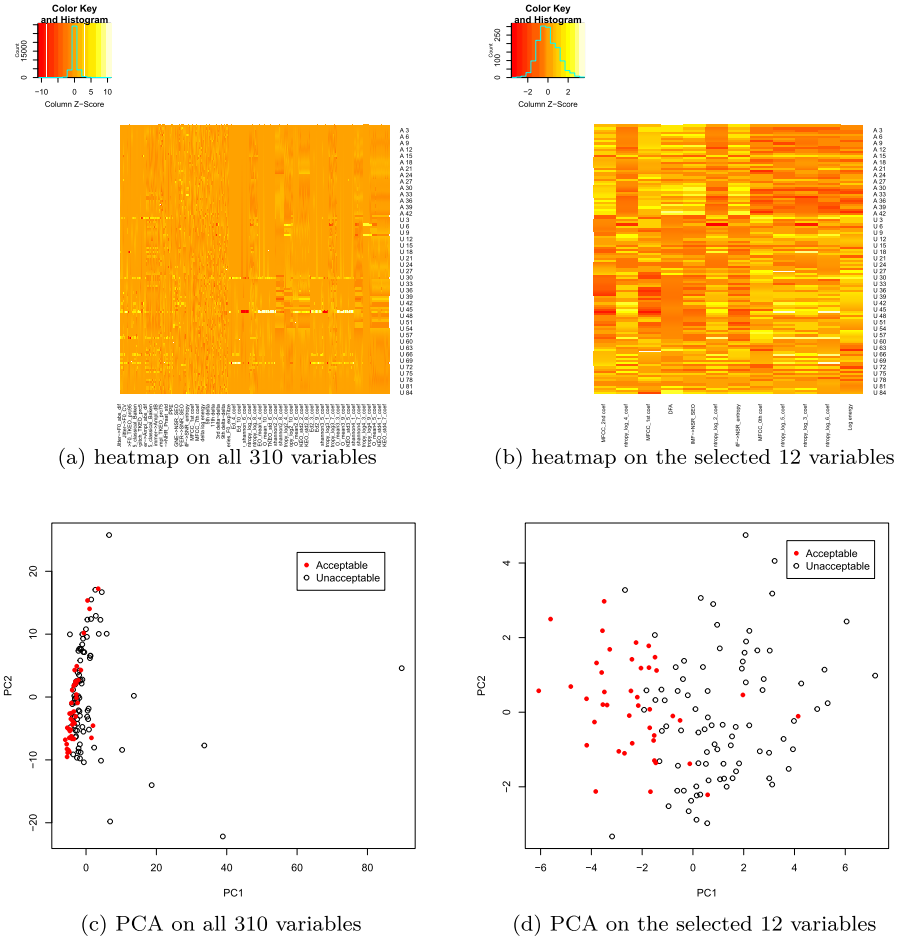
The first data is LSVT Voice Rehabilitation dataset. After speech rehabilitation treatments in Parkinson's disease, 126 patients were evaluated based on 310 attributes. Refer to [32] for details of the data set and dysphonia measure attributes. Phonations of 42 patients were evaluated as 'acceptable', while 84 patients had 'unacceptable' phonations. This data set has the dimension larger than the sample size. Our goal is to test whether or not phonation features have a same distribution in the 'acceptable' group and the 'unacceptable' group, which is a $K = 2$ sample problem. Before we preform the test, we do some exploratory data analysis to visualize the data in the original high dimensional space and the data projected in low dimensional space.

A heatmap on all 310 variables is plotted in Figure 3(a) in which the values are centered and scaled by each column variable. The top third rows are for the acceptable group, while the bottom two thirds for the unacceptable group. It is quite difficult to view differences between two groups. However, the difference shows in the heatmap on the selected 12 variables in Figure 3(b). The selected 12 variables are those with its categorical Gini correlation greater than 0.1.

We also conduct principal component analysis (PCA) on all variables. The proportions of variance of first two principal components (PC) are 32.29% and 19.87%, altogether accounting for 52.16% of the total variance. The data are projected on the plane of the first two PC's shown in he left panel of Figure 3(c) in which several patients with unacceptable evaluation are clearly outliers. We also plot data projection on the first two PC's when PCA is conducted on the selected 12 variables in Figure 3(d). From it, we can see that the unacceptable group tends to have larger values in the first PC. After a simple feature selection to reduce dimensionality, the separation of two groups is more evident. In the next, we perform formal tests on equality of distributions of two groups. The test of distributions on all 310 variables and the test of distributions on the 12 selected variables are conducted.

Besides the five methods considered in Section 3.2, five 2-sample test methods are added for comparison. Three methods are proposed in [19] and denoted as Li-loc, Li-scal and Li-both. Székely's energy test statistic in high dimension is also studied in [19]. It is asymptotically normally distributed, equivalent to gCov and dCov, but its variance estimation is different and quite complicate in [19] and we include it for comparing its efficiency on variance estimation. The last considered method denoted as BG is proposed by Biswas and Ghosh in [3]. The p-values of those ten methods are reported in Table 3.

With the feature selection to reduce dimension, all methods except for Li-scal strongly reject the equality of two distributions. While for the high dimension data, three methods GLP, Li-scal and BG fail to conclude different distributions in two groups. The gCov and dCov methods provide the most significant evidence on the differences of two groups.




(a) heatmap on all 310 variables

(b) heatmap on the selected 12 variables

(c) PCA on all 310 variables

(d) PCA on the selected 12 variables

FIG 3. *Heatmaps and 2 dimensional PCA projections of Voice rehabilitation data of all 310 variables and of the selected 12 variables.*

### *4.2. Arcene data*

The second data set we apply to is Arcene mass-spectrometric data for 900 patients from cancer group and healthy group. The data set was merged from three resources on ovarian cancer data and prostate cancer data. The preprocessing steps of limiting the mass range, averaging the technical repeats, removing the baseline, smoothing, rescaling and aligning the spectra were prepared to reduce disparity between data sources. Arcene data have 10000 features including 7000 real features and 3000 random probes. The dimension is much higher than the sample size. The data was formatted for benchmarking variable selection algorithms for the two class classification problem in 2003 NIPS, the top conference on machine mining and computational neuroscience. The data were partitioned

TABLE 3
*p-values of various 2 sample tests for all features and for the select 12 features in LSVT Voice rehabilitation data.*

| | gCov | gCov-perm | dCov | dCov-perm | GLP |
|---|---|---|---|---|---|
| all 310 variables | 0.0011 | 0.0211 | 0.0013 | 0.0193 | 0.5124 |
| 12 selected variables | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| | Li-loc | Li-scal | Li-both | Szekely | BG |
| all 310 variables | 0.0204 | 0.3190 | 0.0197 | 0.0166 | 0.3272 |
| 12 selected variables | 0.0000 | 0.0958 | 0.0000 | 0.0000 | 0.0009 |

to training, validation, and test sets. For the training and validation sets, each has 44 cancer positives and 56 negatives, while the test set has 310 positives and 390 negatives. Refer to [12] for details about the data preparation and NIPS challenge results.

TABLE 4
*p-values of testing whether training data, testing data and validation data in ARCENE have a same distribution.*

| | gCov | gCov-perm | dCov | dCov-perm | GLP |
|---|---|---|---|---|---|
| p-value | 0.5394 | 0.4530 | 0.4132 | 0.3160 | 0.0389 |

Rather than conducting two sample test, we perform 3 sample testing on distributional equality of the training data, validation data and test data. That is the assumption and the logic behind the procedure of using training data to build model, using validation data to select model and using test data to assess model. P-values of five methods are reported in Table 4. Only GLP rejects the equality with p-value 0.0389, while the other four methods with large p-values support the distribution equality assumption that makes the data mining challenge competition valid.

## 5. Conclusions and future work

The categorical Gini correlation is an alternative to the distance correlation to measure the correlation between a $p$-variate numeric variable $\boldsymbol{X}$ and a categorical variable $Y$. But the Gini one has more appealing properties such as nice presentation and better interpretation. When $p$ is fixed, Dang et al. [7] showed that the sample Gini correlation converges in distribution to a quadratic form of normal distributions under independence of $\boldsymbol{X}$ and $Y$. In this paper, we have studied the inference of the categorical Gini correlation in a more realistic setting where both the sample size and the dimensionality are diverging in an arbitrary fashion. One of our main results, Theorem 2.1, reveals that those complicated quadratic forms of normal random variables admit a normal limit as the dimensionality $p$ diverges to infinity, providing an intriguing example to understand the distinction between classical and high-dimensional theory.

Based on these asymptotic distributions, a new consistent $K$-sample test has been developed. Both simulation studies and real data illustrations have shown

the proposed test performs uniformly better than the distance correlation based test for unbalanced cases.

The Gini covariance has been generalized to a reproducing kernel Hilbert space (RKHS) in [34] as follows.

$$\mathrm{gCov}(\boldsymbol{X}, Y; d_\kappa) = \mathbb{E}\big\{d_\kappa(\boldsymbol{X}_1, \boldsymbol{X}_2)\big\} - \sum_{k=1}^{K} p_k \mathbb{E}\big\{d_\kappa\big(\boldsymbol{X}_1^{(k)}, \boldsymbol{X}_2^{(k)}\big)\big\}, \tag{5.1}$$

where $d_\kappa(\boldsymbol{x}_1, \boldsymbol{x}_2) = \sqrt{\kappa(\boldsymbol{x}_1, \boldsymbol{x}_1) + \kappa(\boldsymbol{x}_2, \boldsymbol{x}_2) - 2\kappa(\boldsymbol{x}_1, \boldsymbol{x}_2)}$, the distance in the feature space induced by positive definite kernel $\kappa$. More specifically, a positive definite kernel, $\kappa : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$, implicitly defines an embedding map:

$$\phi : \boldsymbol{x} \in \mathbb{R}^p \mapsto \phi(\boldsymbol{x}) \in \mathcal{F},$$

via an inner product in the feature space $\mathcal{F}$:

$$\kappa(\boldsymbol{x}_1, \boldsymbol{x}_2) = \big\langle \phi(\boldsymbol{x}_1), \phi(\boldsymbol{x}_2) \big\rangle, \quad \boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{R}^p.$$

Replacing the expectations in (5.1) by the corresponding $U$-statistics and replacing $p_k$ by $\hat{p}_k$, we obtain the sample kernelized Gini covariance $\mathrm{gCov}_n(\boldsymbol{X}, Y; d_\kappa)$. With a choice of bounded kernel such as the popular radial basis function kernel (RBF), the moment condition **C**1 can be dropped. It will be interesting to derive similar results for kernel covariance and correlation.

As long as pairwise (dis)similarities are available, kernel Gini covariance can be used for complex data type. It is interesting to adopt kernel Gini covariance and correlation based on neural tangent kernel (NTK) in the study of deep artificial neural networks (ANN). Continuations of this work could take those directions as well as the following.

- The permutation test based on Gini covariances in high dimension has demonstrated its size and power empirically. A theoretical and rigorous treatment is needed.
- When $\boldsymbol{X}$ and $Y$ are dependent, the CLT holds for the sample Gini covariance $\mathrm{gCov}_n$. Under the null that $\boldsymbol{X}$ and $Y$ are independent, $\mathrm{gCov}_n$ is a $U$-statistic representation with first order degeneracy but admits a normal limit in the high dimension. Therefore, we would expect a non-null CLT for $\mathrm{gCov}_n$ when $p \to \infty$.
- In this study, the number of levels of $Y$ is fixed and finite. However, some applications like Poisson process have infinity levels. In some applications like discretization procedure, the number of levels might increase as sample size increases. It is interesting to study estimation of Gini correlation in those cases and explore its asymptotical distribution when $n$, $p$ and $K$ diverge.

## Appendix A: Appendix

Let $\boldsymbol{X}, \boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3$ and $\boldsymbol{X}_4$ be independent random variables from $F$. We will adopt the following notations through this section.

$$\begin{aligned}
&\xi(\boldsymbol{X}_1) = \mathbb{E}\big(d^2(\boldsymbol{X}, \boldsymbol{X}_1)|\boldsymbol{X}_1\big),\\
&\sigma^2 = \mathbb{E}\xi(\boldsymbol{X}_1),\\
&\gamma^4 = \mathbb{E}\big(\xi^2(\boldsymbol{X}_1)\big),\\
&\eta(\boldsymbol{X}_1, \boldsymbol{X}_2) = \mathbb{E}\big(\big(d(\boldsymbol{X}, \boldsymbol{X}_1) \bigm| \boldsymbol{X}_1\big)\big(d(\boldsymbol{X}, \boldsymbol{X}_2) \bigm| \boldsymbol{X}_2\big)\big),\\
&\tau^4 = \mathbb{E}\big(\eta(\boldsymbol{X}_1, \boldsymbol{X}_2)\big)^2,\\
&\omega^4 = \mathbb{E}d^4(\boldsymbol{X}_1, \boldsymbol{X}_2).
\end{aligned}$$

It is easy to check that $\gamma^4 > \sigma^4 > \tau^4$ and $\omega^4 > \sigma^4$ by Jensen's inequality.

### *A.1. Lemmas*

Before we prove the major result in Theorem 2.1, let us provide several necessary lemmas and their proofs. The remaining lemmas shall be given in the proof of Theorem 2.1. The double centered distance $d(\cdot,\cdot)$ in (2.5) has appealing orthogonal properties in the following Lemmas A.1 and A.2.

**Lemma A.1.** *If $\mathbb{E}\|\boldsymbol{X}\|^4 < \infty$, then $d(\cdot,\cdot)$ in (2.5) satisfies*

1. $\mathbb{E}d(\boldsymbol{X}_1, \boldsymbol{X}_2) = 0$;
2. $\mathbb{E}(d(\boldsymbol{X}_1, \boldsymbol{X}_2) \bigm| \boldsymbol{X}_1) = \mathbb{E}(d(\boldsymbol{X}_1, \boldsymbol{X}_2) \bigm| \boldsymbol{X}_2) = 0$;
3. $\mathbb{E}(d(\boldsymbol{X}_1, \boldsymbol{X}_2)d(\boldsymbol{X}_1, \boldsymbol{X}_3)) = 0$;
4. $\mathbb{E}(d(\boldsymbol{X}_1, \boldsymbol{X})d(\boldsymbol{X}_2, \boldsymbol{X})d(\boldsymbol{X}_3, \boldsymbol{X})d(\boldsymbol{X}_4, \boldsymbol{X})) = 0$;
5. $\mathbb{E}(d^2(\boldsymbol{X}_1, \boldsymbol{X})d(\boldsymbol{X}_2, \boldsymbol{X})d(\boldsymbol{X}_3, \boldsymbol{X})) = 0$;
6. $\mathbb{E}(d^3(\boldsymbol{X}_1, \boldsymbol{X})d(\boldsymbol{X}_2, \boldsymbol{X})) = 0$.
7. $\mathbb{E}(d^2(\boldsymbol{X}_1, \boldsymbol{X}_2)) = \sigma^2$;
8. $\mathbb{E}(d^2(\boldsymbol{X}_1, \boldsymbol{X}_2)d^2(\boldsymbol{X}_1, \boldsymbol{X}_3)) = \gamma^4$.

*Proof.* It is straightforward to obtain that $\mathbb{E}d(\boldsymbol{X}_1, \boldsymbol{X}_2) = 0$ and

$$\mathbb{E}\big(d(\boldsymbol{X}_1, \boldsymbol{X}_2) \bigm| \boldsymbol{X}_1\big) = \mathbb{E}\big(d(\boldsymbol{X}_1, \boldsymbol{X}_2) \bigm| \boldsymbol{X}_2\big) = 0.$$

By the double expectation argument, we have

$$\begin{aligned}
\mathbb{E}\big(d(\boldsymbol{X}_1, \boldsymbol{X}_2)d(\boldsymbol{X}_1, \boldsymbol{X}_3)\big) &= \mathbb{E}\big\{\mathbb{E}\big(d(\boldsymbol{X}_1, \boldsymbol{X}_2)d(\boldsymbol{X}_1, \boldsymbol{X}_3) \bigm| \boldsymbol{X}_1\big)\big\}\\
&= \mathbb{E}\big\{\mathbb{E}\big(d(\boldsymbol{X}_1, \boldsymbol{X}_2) \bigm| \boldsymbol{X}_1\big)\mathbb{E}\big(d(\boldsymbol{X}_1, \boldsymbol{X}_3) \bigm| \boldsymbol{X}_1\big)\big\}\\
&= 0.
\end{aligned}$$

The other properties can be proved similarly. □

**Lemma A.2.** *If $\mathbb{E}\|\boldsymbol{X}\|^4 < \infty$, we have*

1. $\mathbb{E}\eta(\boldsymbol{X}_1, \boldsymbol{X}_2) = 0$,

2. $\mathbb{E}(\eta(\boldsymbol{X}_1,\boldsymbol{X}_2)\eta(\boldsymbol{X}_1,\boldsymbol{X}_3))=0$,
3. $\mathbb{E}[d(\boldsymbol{X}_1,\boldsymbol{X}_3)d(\boldsymbol{X}_2,\boldsymbol{X}_3)d(\boldsymbol{X}_1,\boldsymbol{X}_4)d(\boldsymbol{X}_2,\boldsymbol{X}_4)]=\tau^4$,
4. $\mathbb{E}(\xi(\boldsymbol{X}_1)\eta(\boldsymbol{X}_1,\boldsymbol{X}_2))^2=0$.

*Proof.* $\mathbb{E}\eta(\boldsymbol{X}_1,\boldsymbol{X}_2)=0$ follows directly from the property 3 in Lemma A.1. Using the double expectation argument and properties in Lemma A.1, we have

$$
\begin{aligned}
&\mathbb{E}\big(\eta(\boldsymbol{X}_1,\boldsymbol{X}_2)\eta(\boldsymbol{X}_1,\boldsymbol{X}_3)\big)\\
&=\mathbb{E}\big\{\mathbb{E}\big(d(\boldsymbol{X},\boldsymbol{X}_1)d(\boldsymbol{X},\boldsymbol{X}_2)\mid\boldsymbol{X}_1,\boldsymbol{X}_2\big)\mathbb{E}\big(d(\boldsymbol{X}',\boldsymbol{X}_1)d(\boldsymbol{X}',\boldsymbol{X}_3)\mid\boldsymbol{X}_1,\boldsymbol{X}_3\big)\big\}\\
&=\mathbb{E}\big(d(\boldsymbol{X},\boldsymbol{X}_1)d(\boldsymbol{X},\boldsymbol{X}_2)d(\boldsymbol{X}',\boldsymbol{X}_1)d(\boldsymbol{X}',\boldsymbol{X}_3)\big)\\
&=\mathbb{E}\big\{\mathbb{E}\big(d(\boldsymbol{X},\boldsymbol{X}_1)d(\boldsymbol{X},\boldsymbol{X}_2)d(\boldsymbol{X}',\boldsymbol{X}_1)d(\boldsymbol{X}',\boldsymbol{X}_3)\mid\boldsymbol{X},\boldsymbol{X}'\big)\big\}\\
&=\mathbb{E}\big\{\mathbb{E}\big(d(\boldsymbol{X},\boldsymbol{X}_1)d(\boldsymbol{X}',\boldsymbol{X}_1)\mid\boldsymbol{X},\boldsymbol{X}'\big)\mathbb{E}\big(d(\boldsymbol{X},\boldsymbol{X}_2)\mid\boldsymbol{X}\big)\mathbb{E}\big(d(\boldsymbol{X}',\boldsymbol{X}_3)\mid\boldsymbol{X}'\big)\big\}\\
&=0,\\
&\mathbb{E}\big[\eta(\boldsymbol{X}_1,\boldsymbol{X}_2)\big]^2\\
&=\mathbb{E}\big\{\mathbb{E}\big(d(\boldsymbol{X},\boldsymbol{X}_1)d(\boldsymbol{X},\boldsymbol{X}_2)|\boldsymbol{X}_1,\boldsymbol{X}_2\big)\mathbb{E}\big(d(\boldsymbol{X}',\boldsymbol{X}_1)d(\boldsymbol{X}',\boldsymbol{X}_2)|\boldsymbol{X}_1,\boldsymbol{X}_2\big)\big\}\\
&=\mathbb{E}\big(d(\boldsymbol{X}_1,\boldsymbol{X}_3)d(\boldsymbol{X}_2,\boldsymbol{X}_3)d(\boldsymbol{X}_1,\boldsymbol{X}_4)d(\boldsymbol{X}_2,\boldsymbol{X}_4)\big)=\tau^4,\\
&\mathbb{E}\big(\xi(\boldsymbol{X}_1)\eta(\boldsymbol{X}_1,\boldsymbol{X}_2)\big)^2\\
&=\mathbb{E}\big\{\mathbb{E}\big(d^2(\boldsymbol{X},\boldsymbol{X}_1)\mid\boldsymbol{X}_1\big)\mathbb{E}\big(d(\boldsymbol{X}',\boldsymbol{X}_1)\mid\boldsymbol{X}_1\big)\mathbb{E}\big(d(\boldsymbol{X}'',\boldsymbol{X}_2)\mid\boldsymbol{X}_2\big)\big\}\\
&=\mathbb{E}\big(d^2(\boldsymbol{X},\boldsymbol{X}_1)d(\boldsymbol{X}',\boldsymbol{X}_1)d(\boldsymbol{X}'',\boldsymbol{X}_2)\big)\\
&=\mathbb{E}\big\{\mathbb{E}\big(d^2(\boldsymbol{X},\boldsymbol{X}_1)d(\boldsymbol{X}',\boldsymbol{X}_1)\mid\boldsymbol{X},\boldsymbol{X}'\big)\mathbb{E}\big(d(\boldsymbol{X}_2,\boldsymbol{X}'')\mid\boldsymbol{X}''\big)\big\}=0.
\end{aligned}
$$

This completes the proof of Lemma A.2. □

**Lemma A.3.** *Under conditions* ***C2****,*

$$\frac{\gamma^4}{n\sigma^4}\to 0.$$

*Proof.* By the Cauchy-Schwarz inequality, it is easy to obtain that

$$
\begin{aligned}
\gamma^4&=\mathbb{E}\big(d^2(\boldsymbol{X}_1,\boldsymbol{X})d^2(\boldsymbol{X}_2,\boldsymbol{X})\big)\\
&\le\big(\mathbb{E}d^4(\boldsymbol{X}_1,\boldsymbol{X})\big)^{1/2}\big(\mathbb{E}d^4(\boldsymbol{X}_2,\boldsymbol{X})\big)^{1/2}\\
&=\mathbb{E}d^4(\boldsymbol{X}_1,\boldsymbol{X}_2).
\end{aligned}
$$

By condition **C**2, we have $\frac{\gamma^4}{n\sigma^4}\to 0$. □

### *A.2. Proof of Theorem 2.1*

Under independence of $\boldsymbol{X}$ and $Y$, by Lemma A.1, we have

$$\sigma_0^2=Var\left(\binom{n}{2}^{-1}\sum_{1\le i<j\le n}d(\boldsymbol{X}_i,\boldsymbol{X}_j)\right)\tag{A.1}$$

$$
\begin{aligned}
&+\sum_{k=1}^{K}\hat{p}_k^2 Var\left(\binom{n_k}{2}^{-1}\sum_{1\le i<j\le n_k} d\big(\boldsymbol{X}_i^{(k)},\boldsymbol{X}_j^{(k)}\big)\right)\\
&-2\binom{n}{2}^{-1}\sum_{k=1}^{K}\hat{p}_k\binom{n_k}{2}^{-1} Cov\Big(\sum_{1\le i<j\le n} d(\boldsymbol{X}_i,\boldsymbol{X}_j),\sum_{1\le i<j\le n_k} d\big(\boldsymbol{X}_i^{(k)},\boldsymbol{X}_j^{(k)}\big)\Big)\\
&=\binom{n}{2}^{-1} Var\big(d(\boldsymbol{X}_1,\boldsymbol{X}_2)\big)+\sum_{k=1}^{K}\hat{p}_k^2\binom{n_k}{2}^{-1} Var\big(d\big(\boldsymbol{X}_1^{(k)},\boldsymbol{X}_2^{(k)}\big)\big)\\
&-2\binom{n}{2}^{-1}\sum_{k=1}^{K}\hat{p}_k Var\big(d\big(\boldsymbol{X}_1^{(k)},\boldsymbol{X}_2^{(k)}\big)\big)\\
&=\left(\sum_{k=1}^{K}\hat{p}_k^2\binom{n_k}{2}^{-1}-\binom{n}{2}^{-1}\right)\mathbb{E}d^2(\boldsymbol{X}_1,\boldsymbol{X}_2) \qquad \text{(A.2)}\\
&=\left(\frac{2K-2}{n^2}+o\big(n^{-2}\big)\right)\mathbb{E}d^2(\boldsymbol{X}_1,\boldsymbol{X}_2), \qquad \text{(A.3)}
\end{aligned}
$$

where $\mathbb{E}d^2(\boldsymbol{X}_1,\boldsymbol{X}_2)=V^2(\boldsymbol{X})$ is the squared distance variance of $\boldsymbol{X}$ in [27].

For a short presentation, we denote $gCov_n(\boldsymbol{X},Y)$ as $G_n$, which is

$$
G_n:=\binom{n}{2}^{-1}\sum_{1\le i<j\le n} d(\boldsymbol{X}_i,\boldsymbol{X}_j)-\sum_{k=1}^{K}\hat{p}_k\binom{n_k}{2}^{-1}\sum_{1\le i<j\le n_k} d\big(\boldsymbol{X}_i^{(k)},\boldsymbol{X}_j^{(k)}\big).
$$

In order to show the asymptotic normality of $G_n$, we construct a martingale sequence as follows. Assume that $\boldsymbol{X}_i$'s have been sorted by $Y_i$'s, that is, $\boldsymbol{X}_i=\boldsymbol{X}_i^{(1)}$, for $i=1,2,\ldots,n_1$; $\boldsymbol{X}_{n_1+i}=\boldsymbol{X}_i^{(2)}$, for $i=1,\ldots,n_2;\ldots;\boldsymbol{X}_{n_1+\cdots+n_{k-1}+i}=\boldsymbol{X}_i^{(k)}$, for $i=1,\ldots,n_k$. Let $\mathscr{F}_0=\{\emptyset,\Omega\}$, $\mathscr{F}_l=\sigma\{\boldsymbol{X}_1,\ldots,\boldsymbol{X}_l\}$ with $l=1,2,\ldots,n$. $\mathbb{E}_l$ denotes the conditional expectation given $\mathscr{F}_l$. Define

$$
M_{n,l}=(\mathbb{E}_l-\mathbb{E}_{l-1})G_n.
$$

$\{M_{n,l},1\le l\le n\}$ is a martingale difference sequence with respect to the nested $\sigma$-fields $\{\mathscr{F}_l,1\le l\le n\}$. Also under the independence,

$$
\sum_{l=1}^{n}M_{n,l}=(\mathbb{E}_n-\mathbb{E}_0)G_n=G_n-\mathbb{E}G_n=G_n.
$$

We need to establish the asymptotic normality of $\sum_{l=1}^{n}M_{n,l}$. Without loss of generality, we will prove the case for $K=3$.

We first work out the representations of $M_{n,l}$ by using the properties in Lemmas A.1 and A.2. Depending on $l$, $M_{n,l}$ have three forms.

**Case 1**, for $1\le l\le n_1$, $\mathscr{F}_l=\sigma\{\boldsymbol{X}_1^{(1)},\ldots,\boldsymbol{X}_l^{(1)}\}$. We have

$$
\mathbb{E}(G_n|\mathscr{F}_l)=\binom{n}{2}^{-1}\mathbb{E}\Big(\sum_{1\le i<j\le n} d(\boldsymbol{X}_i,\boldsymbol{X}_j)|\mathscr{F}_l\Big)
$$

$$
\begin{aligned}
&- \binom{n_1}{2}^{-1} \hat{p}_1 \mathbb{E}\Big( \sum_{1\le i<j\le n_1} d(\boldsymbol{X}_i^{(1)}, \boldsymbol{X}_j^{(1)}) | \mathscr{F}_l \Big) \\
&= \binom{n}{2}^{-1} \sum_{1\le i<j\le l} d(\boldsymbol{X}_i, \boldsymbol{X}_j) - \hat{p}_1 \binom{n_1}{2}^{-1} \sum_{1\le i<j\le l} d\big(\boldsymbol{X}_i^{(1)}, \boldsymbol{X}_j^{(1)}\big) \\
&= -\frac{2(n-n_1)}{n(n-1)(n_1-1)} \sum_{1\le i<j\le l} d\big(\boldsymbol{X}_i^{(1)}, \boldsymbol{X}_j^{(1)}\big)
\end{aligned}
$$

and

$$
\mathbb{E}(G_n|\mathscr{F}_{l-1}) = -\frac{2(n-n_1)}{n(n-1)(n_1-1)} \sum_{1\le i<j\le l-1} d\big(\boldsymbol{X}_i^{(1)}, \boldsymbol{X}_j^{(1)}\big).
$$

Thus,

$$
M_{n,l} = (\mathbb{E}_l - \mathbb{E}_{l-1}) G_n = -\frac{2(n-n_1)}{n(n-1)(n_1-1)} \sum_{j=1}^{l-1} d\big(\boldsymbol{X}_l, \boldsymbol{X}_j^{(1)}\big).
$$

**Case 2**, for $n_1 < l \le n_1 + n_2$, $\mathscr{F}_l = \sigma\{\boldsymbol{X}_1^{(1)}, \ldots, \boldsymbol{X}_{n_1}^{(1)}, \boldsymbol{X}_1^{(2)} \ldots, \boldsymbol{X}_{l-n_1}^{(2)}\}$. We have

$$
\begin{aligned}
\mathbb{E}(G_n|\mathscr{F}_l) &= \binom{n}{2}^{-1} \mathbb{E}\Big( \sum_{1\le i<j\le n} d(\boldsymbol{X}_i, \boldsymbol{X}_j) | \mathscr{F}_l \Big) \\
&\quad - \binom{n_1}{2}^{-1} \hat{p}_1 \mathbb{E}\Big( \sum_{1\le i<j\le n_1} d\big(\boldsymbol{X}_i^{(1)}, \boldsymbol{X}_j^{(1)}\big) | \mathscr{F}_l \Big) \\
&\quad - \binom{n_2}{2}^{-1} \hat{p}_2 \mathbb{E}\Big( \sum_{1\le i<j\le n_2} d\big(\boldsymbol{X}_i^{(2)}, \boldsymbol{X}_j^{(2)}\big) | \mathscr{F}_l \Big) \\
&= \binom{n}{2}^{-1} \sum_{1\le i<j\le n_1} d\big(\boldsymbol{X}_i^{(1)}, \boldsymbol{X}_j^{(1)}\big) + \sum_{1\le i<j\le l-n_1} d\big(\boldsymbol{X}_i^{(2)}, \boldsymbol{X}_j^{(2)}\big) \\
&\quad + \binom{n}{2}^{-1} \Big( \sum_{j=1}^{l-n_1} \sum_{i=1}^{n_1} d\big(\boldsymbol{X}_i^{(1)}, \boldsymbol{X}_j^{(2)}\big) \Big) \\
&\quad - \hat{p}_1 \binom{n_1}{2}^{-1} \sum_{1\le i<j\le n_1} d\big(\boldsymbol{X}_i^{(1)}, \boldsymbol{X}_j^{(1)}\big) \\
&\quad - \hat{p}_2 \binom{n_2}{2}^{-1} \sum_{1\le i<j\le l-n_1} d\big(\boldsymbol{X}_i^{(2)}, \boldsymbol{X}_j^{(2)}\big) \\
&= \left( \binom{n}{2}^{-1} - \hat{p}_1 \binom{n_1}{2}^{-1} \right) \sum_{1\le i<j\le n_1} d\big(\boldsymbol{X}_i^{(1)}, \boldsymbol{X}_j^{(1)}\big) \\
&\quad + \left( \binom{n}{2}^{-1} - \hat{p}_2 \binom{n_2}{2}^{-1} \right) \sum_{1\le i<j\le l-n_1} d\big(\boldsymbol{X}_i^{(2)}, \boldsymbol{X}_j^{(2)}\big)
\end{aligned}
$$

$$
\begin{aligned}
&+\binom{n}{2}^{-1}\sum_{j=1}^{l-n_1}\sum_{i=1}^{n_1} d\big(\boldsymbol{X}_i^{(1)},\boldsymbol{X}_j^{(2)}\big)\\
=&-\frac{2(n-n_1)}{n(n-1)(n_1-1)}\sum_{1\le i<j\le n_1} d\big(\boldsymbol{X}_i^{(1)},\boldsymbol{X}_j^{(1)}\big)\\
&-\frac{2(n-n_2)}{n(n-1)(n_2-1)}\sum_{1\le i<j\le l-n_1} d\big(\boldsymbol{X}_i^{(2)},\boldsymbol{X}_j^{(2)}\big)\\
&+\binom{n}{2}^{-1}\sum_{j=1}^{l-n_1}\sum_{i=1}^{n_1} d\big(\boldsymbol{X}_i^{(1)},\boldsymbol{X}_j^{(2)}\big).
\end{aligned}
$$

Therefore,

$$
M_{n,l}=-\frac{2(n-n_2)}{n(n-1)(n_2-1)}\sum_{j=1}^{l-n_1-1} d\big(\boldsymbol{X}_l,\boldsymbol{X}_j^{(2)}\big)+\binom{n}{2}^{-1}\sum_{i=1}^{n_1} d\big(\boldsymbol{X}_l,\boldsymbol{X}_i^{(1)}\big).
$$

**Case 3**, for $n_1+n_2<l\le n$, $\mathscr{F}_l=\sigma\{\boldsymbol{X}_1^{(1)},\dots,\boldsymbol{X}_{n_2}^{(2)},\boldsymbol{X}_1^{(3)},\dots,\boldsymbol{X}_{l-n_1-n_2}^{(3)}\}$.
We have

$$
\begin{aligned}
\mathbb{E}(G_n|\mathscr{F}_l)=&\binom{n}{2}^{-1}\mathbb{E}\Big(\sum_{1\le i<j\le n} d(\boldsymbol{X}_i,\boldsymbol{X}_j)|\mathscr{F}_l\Big)\\
&-\binom{n_1}{2}^{-1}\hat{p}_1\mathbb{E}\Big(\sum_{1\le i<j\le n_1} d\big(\boldsymbol{X}_i^{(1)},\boldsymbol{X}_j^{(1)}\big)|\mathscr{F}_l\Big)\\
&-\binom{n_2}{2}^{-1}\hat{p}_2\mathbb{E}\Big(\sum_{1\le i<j\le n_2} d\big(\boldsymbol{X}_i^{(2)},\boldsymbol{X}_j^{(2)}\big)|\mathscr{F}_l\Big)\\
&-\binom{n_3}{2}^{-1}\hat{p}_3\mathbb{E}\Big(\sum_{1\le i<j\le n_3} d\big(\boldsymbol{X}_i^{(3)},\boldsymbol{X}_j^{(3)}\big)|\mathscr{F}_l\Big)\\
=&\binom{n}{2}^{-1}\Big(\sum_{1\le i<j\le n_1} d\big(\boldsymbol{X}_i^{(1)},\boldsymbol{X}_j^{(1)}\big)+\sum_{1\le i<j\le n_2} d\big(\boldsymbol{X}_i^{(2)},\boldsymbol{X}_j^{(2)}\big)\Big)\\
&+\binom{n}{2}^{-1}\sum_{1\le i<j\le l-n_1-n_2} d\big(\boldsymbol{X}_i^{(3)},\boldsymbol{X}_j^{(3)}\big)+\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} d\big(\boldsymbol{X}_i^{(1)},\boldsymbol{X}_j^{(2)}\big)\\
&+\sum_{i=1}^{n_1}\sum_{j=1}^{l-n_1-n_2} d\big(\boldsymbol{X}_i^{(1)},\boldsymbol{X}_j^{(3)}\big)+\sum_{i=1}^{n_2}\sum_{j=1}^{l-n_1-n_2} d\big(\boldsymbol{X}_i^{(2)},\boldsymbol{X}_j^{(3)}\big)\\
&-\hat{p}_1\binom{n_1}{2}^{-1}\sum_{1\le i<j\le n_1} d\big(\boldsymbol{X}_i^{(1)},\boldsymbol{X}_j^{(1)}\big)-\hat{p}_2\binom{n_2}{2}^{-1}\sum_{1\le i<j\le n_2} d\big(\boldsymbol{X}_i^{(2)},\boldsymbol{X}_j^{(2)}\big)\\
&-\hat{p}_3\binom{n_3}{2}^{-1}\sum_{1\le i<j\le l-n_1-n_2} d\big(\boldsymbol{X}_i^{(3)},\boldsymbol{X}_j^{(3)}\big)
\end{aligned}
$$

$$
\begin{aligned}
&= \left(\binom{n}{2}^{-1} - \hat{p}_1\binom{n_1}{2}^{-1}\right) \sum_{1\le i<j\le n_1} d\big(\boldsymbol{X}_i^{(1)}, \boldsymbol{X}_j^{(1)}\big) \\
&+ \left(\binom{n}{2}^{-1} - \hat{p}_2\binom{n_2}{2}^{-1}\right) \sum_{1\le i<j\le n_2} d\big(\boldsymbol{X}_i^{(2)}, \boldsymbol{X}_j^{(2)}\big) \\
&+ \left(\binom{n}{2}^{-1} - \hat{p}_3\binom{n_3}{2}^{-1}\right) \sum_{1\le i<j\le l-n_1-n_2} d\big(\boldsymbol{X}_i^{(3)}, \boldsymbol{X}_j^{(3)}\big) \\
&+ \binom{n}{2}^{-1} \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} d\big(\boldsymbol{X}_i^{(1)}, \boldsymbol{X}_j^{(2)}\big) + \binom{n}{2}^{-1} \sum_{i=1}^{n_1}\sum_{j=1}^{l-n_1-n_2} d\big(\boldsymbol{X}_i^{(1)}, \boldsymbol{X}_j^{(3)}\big) \\
&+ \binom{n}{2}^{-1} \sum_{i=1}^{n_2}\sum_{j=1}^{l-n_1-n_2} d\big(\boldsymbol{X}_i^{(2)}, \boldsymbol{X}_j^{(3)}\big) \\
&= -\frac{2(n-n_1)}{n(n-1)(n_1-1)} \sum_{1\le i<j\le n_1} d\big(\boldsymbol{X}_i^{(1)}, \boldsymbol{X}_j^{(1)}\big) \\
&- \frac{2(n-n_2)}{n(n-1)(n_2-1)} \sum_{1\le i<j\le n_2} d\big(\boldsymbol{X}_i^{(2)}, \boldsymbol{X}_j^{(2)}\big) \\
&+ \frac{2(n-n_3)}{n(n-1)(n_3-1)} \sum_{1\le i<j\le l-n_1-n_2} d\big(\boldsymbol{X}_i^{(3)}, \boldsymbol{X}_j^{(3)}\big) \\
&+ \binom{n}{2}^{-1} \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} d\big(\boldsymbol{X}_i^{(1)}, \boldsymbol{X}_j^{(2)}\big) \\
&+ \binom{n}{2}^{-1} \sum_{i=1}^{n_1}\sum_{j=1}^{l-n_1-n_2} d\big(\boldsymbol{X}_i^{(1)}, \boldsymbol{X}_j^{(3)}\big) + \binom{n}{2}^{-1} \sum_{i=1}^{n_2}\sum_{j=1}^{l-n_1-n_2} d\big(\boldsymbol{X}_i^{(2)}, \boldsymbol{X}_j^{(3)}\big).
\end{aligned}
$$

Thus,

$$
\begin{aligned}
M_{n,l} = &-\frac{2(n-n_3)}{n(n-1)(n_3-1)} \sum_{j=1}^{l-n_1-n_2-1} d\big(\boldsymbol{X}_l, \boldsymbol{X}_j^{(3)}\big) + \binom{n}{2}^{-1} \sum_{i=1}^{n_1} d\big(\boldsymbol{X}_l, \boldsymbol{X}_i^{(1)}\big) \\
&+ \binom{n}{2}^{-1} \sum_{i=1}^{n_2} d\big(\boldsymbol{X}_l, \boldsymbol{X}_i^{(2)}\big).
\end{aligned}
$$

In order to apply martingale central limit theorem to the constructed martingale sequence, $M_{n,l}$, $l = 1, \ldots, n$, we need the following Lemma A.4 and Lemma A.5.

**Lemma A.4.** *Under conditions* ***C1***–***C3*** *and independence of* $\boldsymbol{X}$ *and* $Y$*, as* $\min\{n_1, n_2, \ldots, n_K\} \to \infty$*, we have*

$$
\frac{\sum_{l=1}^{n} \sigma_{n,l}^2}{\sigma_0^2} \to 1 \quad \textit{in probability},
$$

*where* $\sigma_{n,l}^2 = \mathbb{E}_{l-1}(M_{n,l}^2)$.

*Proof.* We first obtain three formulas of $\sigma_{n,l}^2$ according to $l$.

**Case 1**, for $l \le n_1$, we have

$$\begin{aligned}
\sigma_{n,l}^2 = \mathbb{E}_{l-1}\big(M_{n,l}^2\big) &= \mathbb{E}\bigg\{\bigg(-\frac{2(n-n_1)}{n(n-1)(n_1-1)}\sum_{j=1}^{k-1} d\big(\boldsymbol{X}_l, \boldsymbol{X}_j^{(1)}\big)\bigg)^2 |\mathscr{F}_{l-1}\bigg\} \\
&= \frac{4(n-n_1)^2}{n^2(n-1)^2(n_1-1)^2}\mathbb{E}\bigg\{\sum_{i=1}^{l-1}\sum_{j=1}^{l-1} d\big(\boldsymbol{X}_l, \boldsymbol{X}_i^{(1)}\big)d\big(\boldsymbol{X}_l, \boldsymbol{X}_j^{(1)}\big)|\mathscr{F}_{l-1}\bigg\} \\
&= \frac{4(n-n_1)^2}{n^2(n-1)^2(n_1-1)^2}\sum_{i=1}^{l-1}\sum_{j=1}^{l-1}\mathbb{E}\big\{d\big(\boldsymbol{X}_l, \boldsymbol{X}_i^{(1)}\big)d\big(\boldsymbol{X}_l, \boldsymbol{X}_j^{(1)}\big)|\mathscr{F}_{l-1}\big\} \\
&= \frac{4(n-n_1)^2}{n^2(n-1)^2(n_1-1)^2}\sum_{i=1}^{l-1}\sum_{j=1}^{l-1}\mathbb{E}\big\{d\big(\boldsymbol{X}_l, \boldsymbol{X}_i^{(1)}\big)d\big(\boldsymbol{X}_l, \boldsymbol{X}_j^{(1)}\big)|\boldsymbol{X}_i^{(1)}, \boldsymbol{X}_j^{(1)}\big\} \\
&= \frac{4(n-n_1)^2}{n^2(n-1)^2(n_1-1)^2}\bigg\{\sum_{i=1}^{l-1}\xi\big(\boldsymbol{X}_i^{(1)}\big) + \sum_{1\le i\ne j\le l-1}\eta\big(\boldsymbol{X}_i^{(1)}, \boldsymbol{X}_j^{(1)}\big)\bigg\}.
\end{aligned}$$

**Case 2**, for $n_1 < l \le n_1 + n_2$, we have

$$\begin{aligned}
&\sigma_{n,l}^2 = \\
&\mathbb{E}\bigg\{\bigg(\frac{-2(n-n_2)}{n(n-1)(n_2-1)}\sum_{j=1}^{l-n_1-1} d\big(\boldsymbol{X}_l, \boldsymbol{X}_j^{(2)}\big) + \binom{n}{2}^{-1}\sum_{i=1}^{n_1} d\big(\boldsymbol{X}_l, \boldsymbol{X}_i^{(1)}\big)\bigg)^2 |\mathscr{F}_{l-1}\bigg\} \\
&= \mathbb{E}\bigg\{\bigg(-\frac{2(n-n_2)}{n(n-1)(n_2-1)}\bigg)\bigg(\sum_{i=1}^{l-n_1-1} d\big(\boldsymbol{X}_l, \boldsymbol{X}_i^{(2)}\big) + \binom{n}{2}^{-1}\sum_{i=1}^{n_1} d\big(\boldsymbol{X}_l, \boldsymbol{X}_i^{(1)}\big)\bigg) \\
&\bigg(-\frac{2(n-n_2)}{n(n-1)(n_2-1)}\bigg)\bigg(\sum_{j=1}^{l-n_1-1} d\big(\boldsymbol{X}_l, \boldsymbol{X}_j^{(2)}\big) + \binom{n}{2}^{-1}\sum_{j=1}^{n_1} d\big(\boldsymbol{X}_l, \boldsymbol{X}_j^{(1)}\big)\bigg)|\mathscr{F}_{l-1}\bigg\} \\
&= \binom{n}{2}^{-2}\sum_{1\le i\ne j\le n_1}\eta\big(\boldsymbol{X}_i^{(1)}, \boldsymbol{X}_j^{(1)}\big) + \frac{4(n-n_2)^2}{n^2(n-1)^2(n_2-1)^2}\sum_{i=1}^{l-n_1-1}\xi\big(\boldsymbol{X}_i^{(2)}\big) \\
&+ \frac{4(n-n_2)^2}{n^2(n-1)^2(n_2-1)^2}\sum_{1\le i\ne j\le l-n_1-1}\eta\big(\boldsymbol{X}_i^{(2)}, \boldsymbol{X}_j^{(2)}\big) \\
&- \frac{8(n-n_2)}{n^2(n-1)^2(n_2-1)}\sum_{i=1}^{n_1}\sum_{j=1}^{l-n_1-1}\eta\big(\boldsymbol{X}_i^{(1)}, \boldsymbol{X}_j^{(2)}\big) + \binom{n}{2}^{-2}\sum_{i=1}^{n_1}\xi\big(\boldsymbol{X}_i^{(1)}\big).
\end{aligned}$$

**Case 3**, for $n_1 + n_2 < l \le n$, we have

$$\sigma_{n,l}^2 =$$

$$
\begin{aligned}
&\mathbb{E}\left\{\left(-\frac{2(n-n_3)}{n(n-1)(n_3-1)}\sum_{j=1}^{l-n_1-n_2-1} d\big(\boldsymbol{X}_l,\boldsymbol{X}_j^{(3)}\big)+\binom{n}{2}^{-1}\sum_{i=1}^{n_1} d\big(\boldsymbol{X}_l,\boldsymbol{X}_i^{(1)}\big)\right.\right.\\
&\left.\left.+\binom{n}{2}^{-1}\sum_{i=1}^{n_2} d\big(\boldsymbol{X}_l,\boldsymbol{X}_i^{(2)}\big)\right)^2 |\mathscr{F}_{l-1}\right\}\\
&=\binom{n}{2}^{-2}\sum_{i=1}^{n_1}\xi\big(\boldsymbol{X}_i^{(1)}\big)+\binom{n}{2}^{-2}\sum_{1\le i\ne j\le n_1}\eta\big(\boldsymbol{X}_i^{(1)},\boldsymbol{X}_j^{(1)}\big)\\
&+2\binom{n}{2}^{-2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\eta\big(\boldsymbol{X}_i^{(1)},\boldsymbol{X}_j^{(2)}\big)+\binom{n}{2}^{-2}\sum_{i=1}^{n_2}\xi\big(\boldsymbol{X}_i^{(2)}\big)\\
&-2\binom{n}{2}^{-1}\frac{2(n-n_3)}{n(n-1)(n_3-1)}\sum_{i=1}^{n_1}\sum_{j=1}^{l-n_1-n_2-1}\eta\big(\boldsymbol{X}_i^{(1)},\boldsymbol{X}_j^{(3)}\big)\\
&+\binom{n}{2}^{-2}\sum_{1\le i\ne j\le n_2}\eta\big(\boldsymbol{X}_i^{(2)},\boldsymbol{X}_j^{(2)}\big)+\frac{4(n-n_3)^2}{n^2(n-1)^2(n_3-1)^2}\sum_{i=1}^{l-n_1-n_2-1}\xi\big(\boldsymbol{X}_i^{(3)}\big)\\
&-2\binom{n}{2}^{-1}\frac{2(n-n_3)}{n(n-1)(n_3-1)}\sum_{i=1}^{n_2}\sum_{j=1}^{l-n_1-n_2-1}\eta\big(\boldsymbol{X}_i^{(2)},\boldsymbol{X}_j^{(3)}\big)\\
&+\frac{4(n-n_3)^2}{n^2(n-1)^2(n_3-1)^2}\sum_{1\le i\ne j\le l-n_1-n_2-1}\eta\big(\boldsymbol{X}_i^{(3)},\boldsymbol{X}_j^{(3)}\big).
\end{aligned}
$$

Therefore, under independence of $\boldsymbol{X}$ and $Y$, we have

$$
\begin{aligned}
&\mathbb{E}\left(\sum_{l=1}^{n}\sigma_{n,l}^2\right)\\
&=\frac{4(n-n_1)^2}{n^2(n-1)^2(n_1-1)^2}\sum_{l=1}^{n_1}\sum_{i=1}^{l-1}\mathbb{E}\big(d\big(\boldsymbol{X}_l,\boldsymbol{X}_i^{(1)}\big)\big)^2\\
&\quad+\frac{4(n-n_2)^2}{n^2(n-1)^2(n_2-1)^2}\sum_{l=n_1+1}^{n_1+n_2}\sum_{i=1}^{l-n_1-1}\mathbb{E}\big(d\big(\boldsymbol{X}_l,\boldsymbol{X}_i^{(2)}\big)\big)^2\\
&\quad+\binom{n}{2}^{-2}\sum_{l=n_1+1}^{n_1+n_2}\sum_{i=1}^{n_1}\mathbb{E}\big(d\big(\boldsymbol{X}_l,\boldsymbol{X}_i^{(1)}\big)\big)^2\\
&\quad+\binom{n}{2}^{-2}\sum_{l=n_1+n_2+1}^{n}\sum_{i=1}^{n_1}\mathbb{E}\big(d\big(\boldsymbol{X}_l,\boldsymbol{X}_i^{(1)}\big)\big)^2\\
&\quad+\binom{n}{2}^{-2}\sum_{l=n_1+n_2+1}^{n}\sum_{i=1}^{n_2}\mathbb{E}\big(d\big(\boldsymbol{X}_l,\boldsymbol{X}_i^{(2)}\big)\big)^2\\
&\quad+\frac{4(n-n_3)^2}{n^2(n-1)^2(n_3-1)^2}\sum_{l=n_1+n_2+1}^{n}\sum_{i=1}^{l-n_1-n_2-1}\mathbb{E}\big(d\big(\boldsymbol{X}_l,\boldsymbol{X}_i^{(3)}\big)\big)^2
\end{aligned}
$$

$$
\begin{aligned}
&= \left( \frac{2n_1(n-n_1)^2}{n^2(n-1)^2(n_1-1)} + \frac{2n_2(n-n_2)^2}{n^2(n-1)^2(n_2-1)} + \frac{2n_3(n-n_3)^2}{n^2(n-1)^2(n_3-1)} \right. \\
&\quad \left. + \frac{4n_1n_2 + 4n_1n_3 + 4n_2n_3}{n^2(n-1)^2} \right) \mathbb{E}d^2(\boldsymbol{X}_1, \boldsymbol{X}_2) \\
&= \frac{2}{n^2(n-1)^2} \left\{ \frac{n_1(n-n_1)^2}{(n_1-1)} + \frac{n_2(n-n_2)^2}{(n_2-1)} + \frac{n_3(n-n_3)^2}{(n_3-1)} \right. \\
&\quad \left. + 2n_1n_2 + 2n_1n_3 + 2n_2n_3 \right\} \mathbb{E}d^2(\boldsymbol{X}_1, \boldsymbol{X}_2).
\end{aligned}
$$

It is not difficult to show that

$$
\sigma_0^2 = \operatorname{var}(G_n) = \mathbb{E}\left( \sum_{l=1}^{n} \sigma_{n,l}^2 \right). \tag{A.4}
$$

To complete the proof of Lemma A.4, it suffices to show that

$$
\frac{var(\sum_{l=1}^{n} \sigma_{n,l}^2)}{var^2(G_n)} \to 0. \tag{A.5}
$$

We partition $\sum_{l=1}^{n} \sigma_{n,l}^2$ into two parts, that is,

$$
\sum_{k=1}^{n} \sigma_{n,k}^2 := R_n^{(1)} + R_n^{(2)},
$$

where

$$
\begin{aligned}
R_n^{(1)} =& \frac{4(n-n_1)^2}{n^2(n-1)^2(n_1-1)^2} \sum_{k=1}^{n_1} \sum_{i=1}^{k-1} \xi\big(\boldsymbol{X}_i^{(1)}\big) \\
&+ \frac{4(n-n_2)^2}{n^2(n-1)^2(n_2-1)^2} \sum_{k=n_1+1}^{n_1+n_2} \sum_{i=1}^{k-n_1-1} \xi\big(\boldsymbol{X}_i^{(2)}\big) \\
&+ \binom{n}{2}^{-2} \sum_{k=n_1+1}^{n} \sum_{i=1}^{n_1} \xi\big(\boldsymbol{X}_i^{(1)}\big) + \binom{n}{2}^{-2} \sum_{k=n_1+n_2+1}^{n} \sum_{i=1}^{n_2} \xi\big(\boldsymbol{X}_i^{(2)}\big) \\
&+ \frac{4(n-n_3)^2}{n^2(n-1)^2(n_3-1)^2} \sum_{k=n_1+n_2+1}^{n} \sum_{i=1}^{k-n_1-n_2-1} \xi\big(\boldsymbol{X}_i^{(3)}\big), \\
R_n^{(2)} =& \frac{4(n-n_1)^2}{n^2(n-1)^2(n_1-1)^2} \sum_{k=1}^{n_1} \sum_{1 \le i \ne j \le k-1} \eta\big(\boldsymbol{X}_i^{(1)}, \boldsymbol{X}_j^{(1)}\big) \\
&+ \frac{4(n-n_2)^2}{n^2(n-1)^2(n_2-1)^2} \sum_{k=n_1+1}^{n_1+n_2} \sum_{1 \le i \ne j \le k-n_1-1} \eta\big(\boldsymbol{X}_i^{(2)}, \boldsymbol{X}_j^{(2)}\big) \\
&- \frac{8(n-n_2)}{n^2(n-1)^2(n_2-1)} \sum_{k=n_1+1}^{n_1+n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{k-n_1-1} \eta\big(\boldsymbol{X}_i^{(1)}, \boldsymbol{X}_j^{(2)}\big)
\end{aligned}
$$

$$
\begin{aligned}
&+\binom{n}{2}^{-2}\sum_{k=n_1+1}^{n}\sum_{1\le i\ne j\le n_1}\eta\big(\boldsymbol{X}_i^{(1)},\boldsymbol{X}_j^{(1)}\big)\\
&+2\binom{n}{2}^{-2}\sum_{k=n_1+n_2+1}^{n}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\eta\big(\boldsymbol{X}_i^{(1)},\boldsymbol{X}_j^{(2)}\big)\\
&-2\binom{n}{2}^{-1}\frac{2(n-n_3)}{n(n-1)(n_3-1)}\sum_{k=n_1+n_2+1}^{n}\sum_{i=1}^{n_1}\sum_{j=1}^{k-n_1-n_2-1}\eta\big(\boldsymbol{X}_i^{(1)},\boldsymbol{X}_j^{(3)}\big)\\
&+\binom{n}{2}^{-2}\sum_{k=n_1+n_2+1}^{n}\sum_{1\le i\ne j\le n_2}\eta\big(\boldsymbol{X}_i^{(2)},\boldsymbol{X}_j^{(2)}\big)\\
&-2\binom{n}{2}^{-1}\frac{2(n-n_3)}{n(n-1)(n_3-1)}\sum_{k=n_1+n_2+1}^{n}\sum_{i=1}^{n_2}\sum_{j=1}^{k-n_1-n_2-1}\eta\big(\boldsymbol{X}_i^{(2)},\boldsymbol{X}_j^{(3)}\big)\\
&+\frac{4(n-n_3)^2}{n^2(n-1)^2(n_3-1)^2}\sum_{k=n_1+n_2+1}^{n}\sum_{1\le i\ne j\le k-n_1-n_2-1}\eta\big(\boldsymbol{X}_i^{(3)},\boldsymbol{X}_j^{(3)}\big).
\end{aligned}
$$

Under independence of $\boldsymbol{X}$ and $Y$ and by the properties in Lemmas A.1 and A.2, $R^{(1)}$ and $R^{(2)}$ are orthogonal, that is,

$$\mathbb{E}\big(R_n^{(1)}R_n^{(2)}\big)=0.$$

Also,

$$
\begin{aligned}
\mathbb{E}\big(R_n^{(1)}\big)^2=&\mathbb{E}\Bigg\{\frac{4(n-n_1)^2}{n^2(n-1)^2(n_1-1)^2}\sum_{k=1}^{n_1}\sum_{i=1}^{k-1}\xi\big(\boldsymbol{X}_i^{(1)}\big)\\
&+\frac{4(n-n_2)^2}{n^2(n-1)^2(n_2-1)^2}\sum_{k=n_1+1}^{n_1+n_2}\sum_{i=1}^{k-n_1-1}\xi\big(\boldsymbol{X}_i^{(2)}\big)\\
&+\binom{n}{2}^{-2}\sum_{k=n_1+1}^{n}\sum_{i=1}^{n_1}\xi\big(\boldsymbol{X}_i^{(1)}\big)+\binom{n}{2}^{-2}\sum_{k=n_1+n_2+1}^{n}\sum_{i=1}^{n_2}\xi\big(\boldsymbol{X}_i^{(2)}\big)\\
&+\frac{4(n-n_3)^2}{n^2(n-1)^2(n_3-1)^2}\sum_{k=n_1+n_2+1}^{n}\sum_{i=1}^{k-n_1-n_2-1}\xi\big(\boldsymbol{X}_i^{(3)}\big)\Bigg\}^2\\
:=&\mathbb{E}\big\{A^2+B^2+C^2+D^2+E^2+2AB+2AC+2AD+2AE\\
&+2BC+2BD+2BE+2CD+2CE+2DE\big\},
\end{aligned}
$$

where

$$
\begin{aligned}
\mathbb{E}A^2&=\mathbb{E}\Bigg\{\frac{4(n-n_1)^2}{n^2(n-1)^2(n_1-1)^2}\sum_{k=1}^{n_1}\sum_{i=1}^{k-1}\xi\big(\boldsymbol{X}_i^{(1)}\big)\Bigg\}^2\\
&=\frac{16(n-n_1)^4}{n^4(n-1)^4(n_1-1)^4}\Bigg\{\frac{(n_1-1)(2n_1-1)n_1}{6}\gamma^4
\end{aligned}
$$

$$+\left(\frac{n_1(n_1-1)^2(n_1-2)}{4}+\frac{n_1(n_1-1)(n_1-2)}{6}\right)\sigma^4\Big\},$$

$$\mathbb{E}B^2=\mathbb{E}\Bigg\{\frac{4(n-n_2)^2}{n^2(n-1)^2(n_2-1)^2}\sum_{k=n_1+1}^{n_1+n_2}\sum_{i=1}^{k-n_1-1}\xi(\boldsymbol{X}_i^{(2)})\Bigg\}^2$$

$$=\frac{16(n-n_2)^4}{n^4(n-1)^4(n_2-1)^4}\Big\{\frac{(n_2-1)(2n_2-1)n_2}{6}\gamma^4$$

$$+\left(\frac{n_2(n_2-1)^2(n_2-2)}{4}+\frac{n_2(n_2-1)(n_2-2)}{6}\right)\sigma^4\Big\},$$

$$\mathbb{E}C^2=\mathbb{E}\left(\binom{n}{2}^{-2}\sum_{k=n_1+1}^{n_1+n_2}\sum_{i=1}^{n_1}\xi(\boldsymbol{X}_i^{(1)})\right)^2$$

$$=\binom{n}{2}^{-4}(n_2+n_3)^2\{n_1\gamma^4+n_1(n_1-1)\sigma^4\},$$

$$\mathbb{E}D^2=\mathbb{E}\left(\binom{n}{2}^{-2}\sum_{k=n_1+n_2+1}^{n}\sum_{i=1}^{n_2}\xi(\boldsymbol{X}_i^{(2)})\right)^2=\binom{n}{2}^{-4}n_3^2\{n_2\gamma^4+n_2(n_2-1)\sigma^4\},$$

$$\mathbb{E}E^2=\mathbb{E}\Bigg\{\frac{4(n-n_3)^2}{n^2(n-1)^2(n_3-1)^2}\sum_{k=n_1+n_2+1}^{n}\sum_{i=1}^{k-n_1-n_2-1}\xi(\boldsymbol{X}_i^{(3)})\Bigg\}^2$$

$$=\frac{16(n-n_3)^4}{n^4(n-1)^4(n_3-1)^4}\Big\{\frac{(n_3-1)(2n_3-1)n_3}{6}\gamma^4$$

$$+\left(\frac{n_3(n_3-1)^2(n_3-2)}{4}+\frac{n_3(n_3-1)(n_3-2)}{6}\right)\sigma^4\Big\},$$

$$\mathbb{E}AB=\frac{16(n-n_1)^2(n-n_2)^2}{n^4(n-1)^4(n_1-1)^2(n_2-1)^2}\frac{n_1(n_1-1)}{2}\frac{n_2(n_2-1)}{2}\sigma^4,$$

$$\mathbb{E}AC=\binom{n}{2}^{-2}\frac{4(n-n_1)^3}{n^2(n-1)^2(n_1-1)^2}\Bigg(\frac{n_1(n_1-1)}{2}\gamma^4$$

$$+\left(\frac{n_1(n_1-1)(n_1-2)}{2}+\frac{n_1(n_1-1)}{2}\right)\sigma^4\Bigg),$$

$$\mathbb{E}AD=\frac{4(n-n_1)^2}{n^2(n-1)^2(n_1-1)^2}\binom{n}{2}^{-2}n_3n_2\frac{n_1(n_1-1)}{2}\sigma^4,$$

$$\mathbb{E}AE=\frac{4(n-n_1)^2}{n^2(n-1)^2(n_1-1)^2}\frac{4(n-n_3)^2}{n^2(n-1)^2(n_3-1)^2}\frac{n_1(n_1-1)}{2}\frac{n_3(n_3-1)}{2}\sigma^4,$$

$$\mathbb{E}BC=\frac{4(n-n_2)^2}{n^2(n-1)^2(n_2-1)^2}\binom{n}{2}^{-2}n_1(n-n_1)\frac{n_2(n_2-1)}{2}\sigma^4,$$

$$\mathbb{E}BD=\binom{n}{2}^{-2}\frac{4n_3(n-n_2)^2}{n^2(n-1)^2(n_2-1)^2}\Bigg(\frac{n_2(n_2-1)}{2}\gamma^4$$

$$+\left(\frac{n_2(n_2-1)(n_2-2)}{2}+\frac{n_2(n_2-1)}{2}\right)\sigma^4\Bigg),$$

$$\mathbb{E}BE = \frac{4(n-n_2)^2}{n^2(n-1)^2(n_2-1)^2}\frac{4(n-n_3)^2}{n^2(n-1)^2(n_3-1)^2}\frac{n_2(n_2-1)}{2}\frac{n_3(n_3-1)}{2}\sigma^4,$$

$$\mathbb{E}CD = \binom{n}{2}^{-4} n_1n_2n_3(n_2+n_3)\sigma^4,$$

$$\mathbb{E}CF = \frac{4(n-n_3)^2}{n^2(n-1)^2(n_3-1)^2}\binom{n}{2}^{-2} n_1(n_2+n_3)\frac{n_3(n_3-1)}{2}\sigma^4,$$

$$\mathbb{E}DE = \frac{4(n-n_3)^2}{n^2(n-1)^2(n_3-1)^2}\binom{n}{2}^{-2} n_2n_3\frac{n_3(n_3-1)}{2}\sigma^4.$$

Therefore,

$$\begin{aligned}
\mathbb{E}\big(R_n^{(1)}\big)^2 &= \frac{4}{n^4(n-1)^4}\bigg\{\frac{n_1^2(n-n_1)^4}{(n_1-1)^2} + \frac{n_2^2(n-n_2)^4}{(n_2-1)^2} + \frac{n_3^2(n-n_3)^4}{(n_3-1)^2} \\
&\quad + 4n_1^2n_2^2 + 4n_1^2n_3^2 + 4n_2^2n_3^2 + 8n_1^2n_2n_3 + 8n_1n_2^2n_3 + 8n_1n_2n_3^2 \\
&\quad + \frac{2n_1n_2(n-n_1)^2(n-n_2)^2}{(n_1-1)(n_2-1)} + \frac{2n_1n_3(n-n_1)^2(n-n_3)^2}{(n_1-1)(n_3-1)} \\
&\quad + \frac{2n_2n_3(n-n_2)^2(n-n_3)^2}{(n_2-1)(n_3-1)} \\
&\quad + \frac{4n_1^2n_2 + 4n_1^2n_3 + 4n_1n_2n_3}{n_1-1}(n-n_1)^2 \\
&\quad + \frac{4n_1n_2^2 + 4n_2^2n_3 + 4n_1n_2n_3}{n_2-1}(n-n_2)^2 \\
&\quad + \frac{4n_1n_3^2 + 4n_2n_3^2 + 4n_1n_2n_3}{n_3-1}(n-n_3)^2 + o\big(n^{-4}\big)\bigg\}\sigma^4 + O\big(n^{-5}\big)\gamma^4 \\
&= \big(16n^{-4} + o\big(n^{-4}\big)\big)\sigma^4 + O\big(n^{-5}\big)\gamma^4.
\end{aligned}$$

Similarly, after a tedious evaluation, we have

$$\begin{aligned}
\mathbb{E}\big(R_n^{(2)}\big)^2 &= \tau^4\binom{n}{2}^{-4}\bigg\{n_1^1(n_2+n_3)^2 + n_3^2n_2^2 + 4n_1n_2n_3^2 + \frac{n_1(n_2+n_3)^3}{3} \\
&\quad + \frac{n_2(n_1+n_3)^3}{3} - 4n_1n_2n_3(n_1+n_3)\bigg\} + o\big(n^{-4}\big) \\
&= O\big(n^{-4}\big)\tau^4 + o\big(n^{-4}\big).
\end{aligned}$$

Now we have

$$\begin{aligned}
\operatorname{var}\left(\sum_{k=1}^n \sigma_{n,k}^2\right) &= \mathbb{E}\left(\sum_{k=1}^n \sigma_{n,k}^2\right)^2 - \left\{\mathbb{E}\left(\sum_{k=1}^n \sigma_{n,k}^2\right)\right\}^2 \\
&= \mathbb{E}\big(R_n^{(1)}\big)^2 + \mathbb{E}\big(R_n^{(2)}\big)^2 - \operatorname{var}^2(G_n).
\end{aligned}$$

To prove (A.5), we only need to show that

$$\frac{\mathbb{E}(\sum_{l=1}^n \sigma_{n,l}^2)^2}{var^2(G_n)} \to 1.$$

This is true, because

$$\begin{aligned}\mathbb{E}\left(\sum_{k=1}^{n}\sigma_{n,k}^2\right)^2 &= \mathbb{E}\big(R_n^{(1)}\big)^2 + \mathbb{E}\big(R_n^{(2)}\big)^2 \\ &= \frac{4}{n^4(n-1)^4}\bigg\{\frac{n_1^2(n-n_1)^4}{(n_1-1)^2} + \frac{n_2^2(n-n_2)^4}{(n_2-1)^2} + \frac{n_3^2(n-n_3)^4}{(n_3-1)^2} \\ &+ 4n_1^2n_2^2 + 4n_1^2n_3^2 + 4n_2^2n_3^2 + 8n_1^2n_2n_3 + 8n_1n_2^2n_3 + 8n_1n_2n_3^2 \\ &+ \frac{2n_1n_2(n-n_1)^2(n-n_2)^2}{(n_1-1)(n_2-1)} + \frac{2n_1n_3(n-n_1)^2(n-n_3)^2}{(n_1-1)(n_3-1)} \\ &+ \frac{2n_2n_3(n-n_2)^2(n-n_3)^2}{(n_2-1)(n_3-1)} + \frac{4n_1^2n_2+4n_1^2n_3+4n_1n_2n_3}{n_1-1}(n-n_1)^2 \\ &+ \frac{4n_1n_2^2+4n_2^2n_3+4n_1n_2n_3}{n_2-1}(n-n_2)^2 \\ &+ \frac{4n_1n_3^2+4n_2n_3^2+4n_1n_2n_3}{n_3-1}(n-n_3)^2\bigg\}\sigma^4 \\ &+ O\big(n^{-5}\big)\gamma^4 + o\big(n^{-4}\big) + O\big(n^{-4}\big)\tau^4 \\ &= \frac{16\sigma^4}{n^4} + o(1),\end{aligned}$$

where the last equality is obtained under conditions **C**2 and **C**3 and Lemma (A.3). From (A.2), we have

$$\begin{aligned}\operatorname{var}^2(G_n) &= \frac{4}{n^4}\left(\frac{n_1}{n_1-1} + \frac{n_2}{n_2-1} + \frac{n_3}{n_3-1} - \frac{n}{n-1}\right)^2\sigma^4 \\ &= \frac{16\sigma^4}{n^4} + o(1).\end{aligned}$$

Therefore, as $\min\{n_1, n_2, n_3\} \to \infty$,

$$\frac{\mathbb{E}(\sum_{l=1}^{n}\sigma_{n,l}^2)^2}{\operatorname{var}^2(G_n)} \to 1 \quad \text{and} \quad \frac{\operatorname{var}(\sum_{l=1}^{n}\sigma_{n,l}^2)}{\operatorname{var}^2(G_n)} \to 0.$$

The last step of the proof is to apply Chebyshev's inequality together with (A.4) and (A.5). More specifically, for any $\varepsilon > 0$,

$$\begin{aligned}&P\left(\left|\frac{\sum_{l=1}^{n}\sigma_{n,l}^2}{\sigma_0^2} - 1\right| > \varepsilon\right) = P\left(\left|\sum_{l=1}^{n}\sigma_{n,l}^2 - \mathbb{E}\left(\sum_{l=1}^{n}\sigma_{n,l}^2\right)\right| > \varepsilon\operatorname{var}(G_n)\right) \\ &\leq \frac{\operatorname{var}(\sum_{l=1}^{n}\sigma_{n,l}^2)}{\epsilon^2\operatorname{var}^2(G_n)} \to 0.\end{aligned}$$

This completes the proof for Lemma A.4. □

**Lemma A.5.** *Under conditions **C1**–**C2** and independence of $\boldsymbol{X}$ and $Y$, as* $\min\{n_1, n_2, n_3\} \to \infty$, *we have*

$$\frac{\sum_{l=1}^{n} \mathbb{E}(M_{n,l}^4)}{var^2(G_n)} \to 0.$$

*Proof.* Now we compute $\mathbb{E}M_{n,l}^4$ under independence of $\boldsymbol{X}$ and $Y$.

**Case 1**, for $1 \le l \le n_1$, we have

$$\begin{aligned}
\mathbb{E}M_{n,l}^4 &= \mathbb{E}\left\{-\frac{2(n-n_1)}{n(n-1)(n_1-1)} \sum_{j=1}^{l-1} d\big(\boldsymbol{X}_l, \boldsymbol{X}_j^{(1)}\big)\right\}^4 \\
&= \frac{16(n-n_1)^4}{n^4(n-1)^4(n_1-1)^4}\big\{(l-1)\mathbb{E}d^4(\boldsymbol{X}_1, \boldsymbol{X}_l) \\
&\quad + 3(l-1)(l-2)\mathbb{E}d^2(\boldsymbol{X}_1, \boldsymbol{X}_l)d^2(\boldsymbol{X}_2, \boldsymbol{X}_l)\big\} \\
&= \frac{16(n-n_1)^4}{n^4(n-1)^4(n_1-1)^4}\big\{(l-1)\omega^4 + 3(l-1)(l-2)\gamma^4\big\};
\end{aligned}$$

**Case 2**, for $n_1 < l \le n_1 + n_2$, we have

$$\begin{aligned}
\mathbb{E}M_{n,l}^4 &= \mathbb{E}\left\{\frac{-2(n-n_2)}{n(n-1)(n_2-1)} \sum_{j=1}^{l-n_1-1} d\big(\boldsymbol{X}_l, \boldsymbol{X}_j^{(2)}\big) + \binom{n}{2}^{-1} \sum_{i=1}^{n_1} d\big(\boldsymbol{X}_l, \boldsymbol{X}_i^{(1)}\big)\right\}^4 \\
&= \mathbb{E}A^4 + \mathbb{E}B^4 + 6\mathbb{E}A^2B^2,
\end{aligned}$$

where

$$\begin{aligned}
\mathbb{E}A^4 &= \frac{16(n-n_2)^4}{n^4(n-1)^4(n_2-1)^4}\big\{(l-n_1-1)\omega^4 + 3(l-n_1-1)(l-n_1-2)\gamma^4\big\}, \\
\mathbb{E}B^4 &= \binom{n}{2}^{-4}\big\{n_1\omega^4 + 3n_1(n_1-1)\gamma^4\big\}, \\
\mathbb{E}A^2B^2 &= \left(-\frac{2(n-n_2)}{n(n-1)(n_2-1)}\right)^2\binom{n}{2}^{-2}\big\{n_1(l-n_1-1)\big\}\gamma^4.
\end{aligned}$$

**Case 3**, for $n_1 + n_2 < l \le n$, we have

$$\begin{aligned}
\mathbb{E}M_{n,l}^4 &= \mathbb{E}\left\{-\frac{2(n-n_3)}{n(n-1)(n_3-1)} \sum_{j=1}^{l-n_1-n_2-1} d\big(\boldsymbol{X}_l, \boldsymbol{X}_j^{(3)}\big)\right. \\
&\quad \left. + \binom{n}{2}^{-1} \sum_{i=1}^{n_1} d\big(\boldsymbol{X}_l, \boldsymbol{X}_i^{(1)}\big) + \binom{n}{2}^{-1} \sum_{i=1}^{n_2} d\big(\boldsymbol{X}_l, \boldsymbol{X}_i^{(2)}\big)\right\}^4 \\
&= \mathbb{E}A^4 + \mathbb{E}B^4 + \mathbb{E}C^4 + 6\big(\mathbb{E}A^2B^2 + \mathbb{E}A^2C^2 + \mathbb{E}B^2C^2\big),
\end{aligned}$$

where

$$\begin{aligned}
\mathbb{E}A^4 &= \frac{16(n-n_3)^4}{n^4(n-1)^4(n_3-1)^4}\big\{(l-n_1-n_2-1)\omega^4+3(l-n_1-n_2-1)\\
&\quad\times(l-n_1-n_2-2)\gamma^4\big\},\\
\mathbb{E}B^4 &= \binom{n}{2}^{-4}\big\{n_1\omega^4+3n_1(n_1-1)\gamma^4\big\},\\
\mathbb{E}C^4 &= \binom{n}{2}^{-4}\big\{n_2\omega^4+3n_2(n_2-1)\gamma^4\big\},\\
\mathbb{E}A^2B^2 &= \left(-\frac{2(n-n_3)}{n(n-1)(n_3-1)}\right)^2\binom{n}{2}^{-2}\big\{n_1(l-n_1-n_2-1)\big\}\gamma^4,\\
\mathbb{E}A^2C^2 &= \left(-\frac{2(n-n_3)}{n(n-1)(n_3-1)}\right)^2\binom{n}{2}^{-2}\big\{n_2(l-n_1-n_2-1)\big\}\gamma^4,\\
\mathbb{E}B^2C^2 &= \binom{n}{2}^{-4}n_1n_2\gamma^4.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\sum_{l=1}^{n}\mathbb{E}M_{n,l}^4 &= \sum_{l=1}^{n_1}\frac{16(n-n_1)^4}{n^4(n-1)^4(n_1-1)^4}\big\{(l-1)\omega^4+3(l-1)(l-2)\gamma^4\big\}\\
&\quad+\sum_{l=n_1+1}^{n_2}\bigg\{\frac{16(n-n_1)^4}{n^4(n-1)^4(n_1-1)^4}\big\{(l-n_1-1)\omega^4\\
&\quad+3(l-n_1-1)(l-n_1-2)\gamma^4\big\}\\
&\quad+\binom{n}{2}^{-4}\big\{n_1\omega^4+3n_1(n_1-1)\sigma^4\big\}\\
&\quad+6\left(\frac{2(n-n_2)}{n(n-1)(n_2-1)}\right)^2\binom{n}{2}^{-2}\big\{n_1(l-n_1-1)\big\}\gamma^4\bigg\}\\
&\quad+\sum_{l=n_1+n_2+1}^{n}\{\frac{16(n-n_3)^4}{n^4(n-1)^4(n_3-1)^4}\big\{(l-n_1-n_2-1)\omega^4\\
&\quad+3(l-n_1-n_2-1)(l-n_1-n_2-2)\gamma^4\big\}\\
&\quad+\binom{n}{2}^{-4}\big\{n_1\omega^4+3n_1(n_1-1)\sigma^4\big\}+\binom{n}{2}^{-4}\big\{n_2\omega^4+3n_2(n_2-1)\gamma^4\big\}\\
&\quad+6\left(-\frac{2(n-n_3)}{n(n-1)(n_3-1)}\right)^2\binom{n}{2}^{-2}\big\{n_1(l-n_1-n_2-1)\big\}\gamma^4\\
&\quad+6\left(\frac{2(n-n_3)}{n(n-1)(n_3-1)}\right)^2\binom{n}{2}^{-2}\big\{n_2(l-n_1-n_2-1)\big\}\gamma^4\\
&\quad+6\binom{n}{2}^{-4}n_1n_2\gamma^4
\end{aligned}$$

$$= O\big(n^{-5}\big)\gamma^4 + O\big(n^{-5}\big)\omega^4$$
$$= o\big(n^{-4}\big)\sigma^4.$$

The last equality is due to Condition **C**2 and Lemma A.3. This completes the proof of this lemma. □

Lemma A.5 implies that the Lindeberg's condition holds. Along with Lemma A.4, an application of the martingale CLT completes the proof of Theorem 2.1.

### *A.3. Proof of Theorem 2.3*

As $\hat{\sigma}_0$ in (2.7) is a ratio consistent estimator for $\sigma_0$, it is sufficient to show that $\frac{\mathrm{gCov}_n(X,Y)}{\sigma_0} > C$ for any arbitrarily large constant $C > 0$ under $\mathcal{H}_1$.

$$\begin{aligned}
&\mathbb{E}\big(\mathrm{gCov}_n(\boldsymbol{X},Y) - \mathrm{gCov}(\boldsymbol{X},Y)\big)^2 \\
&= \mathbb{E}\left((U_n - \mathbb{E}U_n) - \sum_{k=1}^{K}\hat{p}_k(U_{n_k} - \mathbb{E}U_{n_k}) + \sum_{k=1}^{K}(p_k - \hat{p}_k)\mathbb{E}U_{n_k}\right)^2 \\
&\le (2K+1)\left(\mathbb{E}(U_n - \mathbb{E}U_n)^2 + \sum_{k=1}^{K}\hat{p}_k^2\mathbb{E}(U_{n_k} - \mathbb{E}U_{n_k})^2\right. \\
&\left. + \sum_{k=1}^{K}\mathbb{E}(p_k - \hat{p}_k)^2(\mathbb{E}U_{n_k})^2\right) \\
&\le (2K+1)\left(\frac{C_1}{n}\mathbb{E}\|\boldsymbol{X}_i - \boldsymbol{X}_j\|^2 + C_2\sum_{k=1}^{K}\frac{\hat{p}_k^2}{n_k}\mathbb{E}\|\boldsymbol{X}_i^{(k)} - \boldsymbol{X}_j^{(k)}\|^2\right. \\
&\left. + \sum_{k=1}^{K}\frac{p_k(1-p_k)\Delta_k^2}{n_k}\right) \qquad \text{(A.6)} \\
&= \frac{C(2K+1)}{n}\left(\mathbb{E}\|\boldsymbol{X}_i - \boldsymbol{X}_j\|^2 + \sum_{k=1}^{K}\hat{p}_k\mathbb{E}\|\boldsymbol{X}_i^{(k)} - \boldsymbol{X}_j^{(k)}\|^2 + O(1)\right).
\end{aligned}$$

The inequality (A.6) is obtained by applying the moment inequality of $U$-statistics from [18] (p. 72) and conditional Jensen's inequality. Hence,

$$|\mathrm{gCov}_n(\boldsymbol{X},Y) - \mathrm{gCov}(\boldsymbol{X},Y)| = O_p\big(n^{-1/2}\big).$$

With the equation (A.3), we have

$$\left|\frac{\mathrm{gCov}_n(\boldsymbol{X},Y)}{\sigma_0} - \frac{\mathrm{gCov}(\boldsymbol{X},Y)}{\sigma_0}\right| = O_p\big(n^{1/2}\big) \to \infty. \qquad \text{(A.7)}$$

Under condition **C**4, $\sqrt{n}\mathrm{gCov}(\boldsymbol{X},Y) \to \infty$, we have

$$\left|\frac{\mathrm{gCov}_n(\boldsymbol{X},Y) - \mathrm{gCov}(\boldsymbol{X},Y)}{\mathrm{gCov}(\boldsymbol{X},Y)}\right| \to 0 \ \text{ in probability.} \qquad \text{(A.8)}$$

With (A.7) and (A.8) together, we can conclude that $\frac{\text{gCov}_n(\boldsymbol{X},Y)}{\sigma_0} \to \infty$ in probability. Therefore, $P(\text{gCov}_n(\boldsymbol{X},Y) > Z_\alpha \hat{\sigma}_0) \to 1$. We have completed the proof.

## Acknowledgments

Thanks Jun Li for sharing R codes on two-sample tests with us.

## References

[1] Anderson, N.H., Hall, P. and Titterington, D.M. (1994). Two-sample test statistics for measuring discrepancies between two multivariate probability density functions using kernel-based density estimates. *J. Multivariate Anal.* **50**, 41–54. MR1292607

[2] Baringhaus, L. and Franz, C. (2004). On a new multivariate two-sample test. *J. Mult. Anal.* **88**, 190–206. MR2021870

[3] Biswas, M. and Ghosh, A.K. (2014). A nonparametric two-sample test applicable to high-dimensional data. *J. Mult. Anal.* **123**, 160–171. MR3130427

[4] Cai, S., Chen, J. and Zidek, J. (2017). Hypothesis testing in the presence of multiple samples under density ratio models. *Statist. Sinica* **27**(2), 761–783. MR3674695

[5] Chen, H. and Friedman, J.H. (2017). A new graph-based two-sample test for multivariate and object data. *J. Am. Statist. Assoc.* **112**, 397–409. MR3646580

[6] Curry, J., Dang, X. and Sang, H. (2019). A rank-based Cramér-von-Mises-type test for two samples. *Braz. J. Probab. Stat.* **33**(3), 425–454. MR3960270

[7] Dang, X., Nguyen, D., Chen, X. and Zhang, J. (2021). A new Gini correlation between quantitative and qualitative variables. *Scand. J. Stat.* **48**(4), 1314–1343. MR4377359

[8] Darling, D.A. (1957). The Kolomogorov-Smirnov, Cramér-von Mises tests. *Ann. Math. Stat.* **28**(4), 823–838. MR0093870

[9] Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [`http://archive.ics.uci.edu/ml`]. Irvine, CA: University of California, School of Information and Computer Science.

[10] Fernández, V., Jimènez Gamerro, M. and Muñoz Garcìa, J. (2008). A test for the two-sample problem based on empirical characteristic functions. *Comput. Statist. Data Anal.* **52**, 3730–3748. MR2427377

[11] Gao, L., Fan, Y., Lv, J. and Shao, Q. (2021). Asymptotic distributions of high-dimensional distance correlation inference. *Ann. Stat.*, accepted. DOI: 10.1214/20-AOS2024. MR4319239

[12] Guyon, I., Gunn, S., Ben-Hur, A., and Dror, G. (2004). Result analysis of the NIPS 2003 feature selection challenge.

[13] Heller, R., Heller, Y. and Gorfine, M. (2013). A consistent multivariate test

of association based on ranks of distances. *Biometrika* **100**(2), 503–510. MR3068450

[14] Heller, R., Heller, Y., Kaofman, S., Brill, B. and Gorfine, M. (2016). Consistent distribution-free $K$-sample and independence test for univariate random variables. *J. Mach. Learn. Res.* **17**(29), 1–54. MR3491123

[15] Huo, X. and Székely, G.J. (2016). Fast computing for distance covariance. *Technometrics* **58**(4), 435–447. MR3556612

[16] Jiang, B., Ye, C. and Liu, J. (2015). Nonparametric $K$-sample tests via dynamic slicing. *J. Amer. Statist. Assoc.* **110**, 642–653. MR3367254

[17] Kiefer, J. (1959). $k$-sample analogues of the Kolmogorov-Simirnov, Cramér-von Mises tests. *Ann. Math. Statist.* **30**, 420–447. MR0102882

[18] Koroljuk, V.S. and Borovskich, Y.V. (1994). *Theory of U-statistics. Mathematics and its applications 273*. Kluwer Academic Publishers Group, Dordrecht Translated from the 1989 Russian original by P.V. Malyshev and D.V. Malyshev and revised by the authors. MR1472486

[19] Li, J. (2018). Asymptotic normality of interpoint distances for high-dimensional data with applications to the two-sample problem. *Biometrika* **105**, 529–546. MR3842883

[20] Lyons, R. (2013). Distance covariance in metric spaces. *Ann. Probab.* **41**(5), 3284–3305. MR3127883

[21] Martìnez-Camblor, P. and de Uña-Álvarez, J. (2009). Non-parametric $k$-sample tests: Density functions vs distribution functions. *Comput. Statist. Data Anal.* **53**, 3344–3357. MR2751143

[22] Mukhopadhyay, S. and Wang, K. (2020). A nonparametric approach to high-dimensional $k$-sample comparison problem. *Biometrika* **107**(3), 555–572. MR4138976

[23] Rizzo, M.L. and Székely, G.J. (2010). Disco analysis: A nonparametric extension of analysis of variance. *Ann. Appl. Stat.* **4**, 1034–1055. MR2758432

[24] Sang, Y., Dang, X. and Zhao, Y. (2020). Jackknife empirical likelihood approach for $K$-sample tests via energy distance. *Canad. J. Statist.*, accepted. DOI: 10.1002/cjs.11611. MR4349638

[25] Scholz, F. W. and Stpephens, M.A. (1987). $K$-sample Anderson-Darling tests. *J. Amer. Statist. Assoc.* **82**(399), 918–924. MR0910001

[26] Székely, G.J. and Rizzo, M.L. (2004). Testing for equal distributions in high dimension. *InterStat* **5**(16.10), 1249–1272.

[27] Székely, G.J., Rizzo, M.L. and Bakirov, N. (2007). Measuring and testing dependence by correlation of distances. *Ann. Statist.* **35**(6), 2769–2794. MR2382665

[28] Székely, G. J. and Rizzo, M. L. (2009). Brownian distance covariance. *Ann. Appl. Stat.* **3**(4), 1233–1303. MR2752135

[29] Székely, G.J. and Rizzo, M.L. (2013). Energy statistics: A class of statistics based on distances. *J. Stat. Plan. Infer.* **143**, 1249–1272. MR3055745

[30] Székely, G.J. and Rizzo, M.L. (2013). The distance of correlation $t$-test of independence in high dimension. *J. Mult. Anal.* **117**, 193–213. MR3053543

[31] Székely, G.J. and Rizzo, M.L. (2017). The energy of data. *Ann. Rev. Stat. Appl.* **4**(1), 447–479.

[32] Tsanas, A., Little, M.A., Fox, C. and Ramig, L.O. (2014). Objective automatic assessment of rehabilitative speech treatment in Parkinson's diseases. *IEEE Trans. Neural Syst. Rehabilitation Eng.* **22**, 181–191.

[33] Wang, C., Marriott, P and Li, P. (2017). Testing homogeneity of multiple nonnegative distributions with excess zero observations. *Comput. Statist. Data Anal.* **114**, 146–157. MR3660845

[34] Zhang, S., Dang, X., Nguyen, D., Wilkins, D. and Chen, Y. (2021). Estimating feature-label dependence using Gini distance statistics. *IEEE Trans. Pattern Anal. Mach. Intell.* **43**(6), 1947–1963.

[35] Zhu, C., Zhang, X., Yao, S. and Shao, X. (2020). Distance-based and RKHS-based dependence metrics in high dimension. *Ann. Stat.* **48**(6), 3366–3394. MR4185812

[36] Zhu, C. and Shao, X. (2021). Interpoint distance based two sample tests in high dimension. *Bernoulli* **27**, 1189–1211. MR4255231