A New Gini Correlation between Quantitative and Qualitative Variables

Xin Dang[a], Dao Nguyen[a],Yixin Chen[b] and Junyin Zhang[c]

[a]Department of Mathematics, University of Mississippi

[b]Department of Computer and Information Science, University of Mississippi

[c]Department of Mathematics, Taiyuan University of Technology

Abstract

We propose a new Gini correlation to measure dependence between a categorical and numerical variables. Analogous to Pearson R^2 in ANOVA model, the Gini correlation is interpreted as the ratio of the between-group variation and the total variation, but it characterizes independence (zero Gini correlation mutually implies independence). Closely related to the distance correlation, the Gini correlation is of simple formulation by considering the nature of categorical variable. As a result, the proposed Gini correlation has a simpler computation implementation than the distance correlation and is more straightforward to perform inference. Simulation and real data applications are conducted to demonstrate the advantages.

Running Title: Categorical Gini Correlation

Keywords: distance correlation, energy distance, Gini mean difference, Gini correlation

1 Introduction

Measuring strength of association or dependence between two variables or two sets of variables is of vital importance in many research fields. Various correlation notions have been developed and studied mainly for two numerical variables (Kendall & Gibbons, 1990; Mari & Kotz, 2001). The widely-used Pearson product correlation measures the linear relationship. Rank based or copula based correlations such as Spearman's ρ (Spearman, 2004) and Kendall's τ (Kendall, 1938) explore monotonic relationships. Gini correlation (Schechtman & Yitzhaki, 1987, 2003) is based on the covariance of one variable and rank of the other. A symmetric version of Gini correlation is proposed by Sang *et al.* (2016). Other robust correlation measures are surveyed in Shevlyakov & Smirnov (2011) and explored in detail in Shevlyakov &

Oja (2016). The celebrated distance correlation proposed by Székely *et al.* (2007) extends to multivariate date and characterizes dependence. Those correlations, however, are defined for numerical and/or ordinal variables. They, except for the distance correlation, can not be directly applied to a categorical variable.

If both variables are nominal, Cramér's V (Cramér, 1946) and Tschuprow's T (Tschuprow, 1939) based on χ^2 test statistic can be used to measure their association. Theoretically based on information theory, mutual information is popular due to its easy computation for two discrete variables. However, mutual information correlation (Ross, 2014; Gao *et al.*, 2017; Beknazaryan *et al.*, 2019) loses the computational attractiveness for measuring dependence between categorical and numerical variables, especially when the numerical variable is in high dimension.

For this case, two approaches are typically used for defining association measures. The first one treats the continuous numerical variable X as the response variable and the categorical variable Y as the predictor. Pearson R^2 of the analysis of variance (ANOVA) or η^2 of MANOVA is then a measure of correlation between them. The second approach considers Y being the response and X as the explanatory variable(s). A pseudo-R^2 of the logistic or other generalized regression model serves a measure of correlation (Hu *et al.*, 2006; Tjur, 2009). If X and Y are independent, those correlation parameters are zero. However, the converse is not true in general. Those correlations do not characterize independence. In this paper, we propose a new Gini correlation (denoted as ρ_g) for measuring dependence between categorical and numerical variables. The contributions of this paper are as follows.

- A new dependence measure between categorical and numerical variables. The proposed Gini correlation characterizes independence: zero correlation mutually implies independence. It also has a nice interpretation as the ratio of between Gini variation and the total Gini variation.
- Limiting distributions of sample Gini correlation obtained under independence and dependence cases.
- Extension of the distance correlation for dependence measure between categorical and numerical variables.
- Connection and comparison of Gini correlation and the distance correlation. Comparing with the distance correlation, Gini correlation has a simpler form, leading a simple computation and easy inference.

The remainder of the paper is organized as follows. Section 2 motivates the dependence measure between one-dimensional numerical variable and a categorical variable. The connection to Gini mean difference leads to a natural generalization and nice interpretation. The properties of the generalized Gini correlation are studied in Section 3. The relationship to the distance correlation is treated in Section 3.2 and three examples are given in Section 3.3. Section 4 is devoted to inferences of the Gini correlation. Asymptotic behavior of the sample Gini correlation is explored. In Section 5, we conduct

experimental studies by simulation and real data applications to demonstrate advantages of the Gini correlation over the distance correlation. We conclude and discuss future works in Section 6. Some detailed derivations of Remarks and Example results are provided in Appendix.

2 Motivation

2.1 Proposed correlation

We consider to measure association between a numerical variable X in $\mathbb{R}$ and a categorical variable Y. Suppose that Y takes values $L_1, ..., L_K$. Assume the categorical distribution P_Y of Y is $P(Y = L_k) = p_k > 0$ and the conditional distribution of X given $Y = L_k$ is F_k. Then the joint distribution of X and Y is $P(X \leq x, Y = k) = p_k F_k(x)$. When the conditional distribution of X given Y is the same as the marginal distribution of X, X and Y are independent. In that case, we say there is no correlation between them. However, when they are dependent, i.e $F(x) \neq F_k(x)$ for some k, we would like to measure this dependence. Intuitively, the larger the difference between the marginal distribution and conditional distribution is, the stronger association should be. With that consideration, a natural correlation measure shall be proportional to

$$D := \mathbb{E} \int_{\mathbb{R}} (F(x|Y) - F(x))^2 \, dx = \sum_{k=1}^{K} p_k \int_{\mathbb{R}} (F_k(x) - F(x))^2 \, dx, \tag{1}$$

the expectation of the integrated squared difference between conditional and marginal distribution functions, if D is finite. In other words, the correlation is proportional to the L_2 distance of marginal and conditional distributions.

Clearly, the corresponding correlation is non-negative, just like Pearson R^2 type of correlations. It, however, has an advantage that the correlation is zero if and only if X and Y are independent, while for Pearson R^2 type of correlation, zero does not mutually imply independence.

Next, we need to find the standardization term so that the corresponding correlation has a range of $[0, 1]$, a desired property for a dependence measure specified in Renyi (1959). Under some condition of F, we would like to obtain $\max D$ among all F_k and p_k, which can be formulated to solve the following optimization problem.

$$\max_{F_k, p_k} D = \max_{F_k, p_k} \sum_{k=1}^{K} p_k \int_{\mathbb{R}} (F_k(x) - F(x))^2 \, dx, \tag{2}$$

$$\text{subject to } p_k > 0, \sum_{k=1}^{K} p_k = 1, \sum_{k=1}^{K} p_k F_k(x) = F(x)$$

$$\text{and } F_k(x) \text{ is a distribution function for } k = 1, ..., K.$$

Note that $\sum_{k=1}^{K} p_k(F_k(x) - F(x))^2 = \sum_{k=1}^{K} p_k F_k^2(x) - F^2(x) \geq 0$ for any x. Since $F_k(x)$ is a cumulative distribution function, we have

$$D = \int_{\mathbb{R}} \sum_{k=1}^{K} p_k F_k^2(x) - F^2(x) dx \leq \int_{\mathbb{R}} F(x) - F^2(x) dx.$$

The equality holds if and only if F_k is a single point mass distribution. In that case, F is a discrete distribution with at most K distinct values. Assuming that $0 < \int_{\mathbb{R}} F(x) - F^2(x)\, dx < \infty$, we propose the correlation between X and Y as

$$\rho_g(X, Y) = \frac{\sum_{k=1}^{K} p_k \int_{\mathbb{R}} (F_k(x) - F(x))^2\, dx}{\int_{\mathbb{R}} F(x) - F^2(x)\, dx}. \tag{3}$$

From the discussion above, we have the following immediate results.

1. $0 \leq \rho_g(X, Y) \leq 1$.
2. $\rho_g(X, Y) = 0$ if and only if X and Y are independent.
3. $\rho_g(X, Y) = 1$ if and only if F_k is a single point mass distribution.

Assumption $\int_{\mathbb{R}} F(x) - F^2(x)\, dx > 0$ implies that F is not a single point mass distribution, meaning that X is non-degenerate. Assumption $\int_{\mathbb{R}} F(x) - F^2(x)\, dx < \infty$ means $\mathbb{E}|X| < \infty$, which we will see in the next subsection. Further, $\rho_g(X, Y)$ can be written as

$$\rho_g(X, Y) = 1 - \frac{2\sum_{k=1}^{K} p_k \int_{\mathbb{R}} F_k(x) - F_k^2(x)\, dx}{2\int_{\mathbb{R}} F(x) - F^2(x)\, dx}. \tag{4}$$

This formulation provides a Gini mean difference representation of the proposed correlation.

2.2 Gini distance representation

Gini mean difference (GMD) was introduced as an alternative measure of variability to the usual standard deviation (Gini, 1914; David, 1968; Yitzhaki & Schechtman, 2013). Let X and X' be independent random variables from a distribution F with finite first moment in $\mathbb{R}$. The GMD of F is

$$\Delta = \Delta(X) = \Delta(F) = \mathbb{E}|X - X'|,$$

the expected distance between two independent random variables. Dorfman (1979) proved that for non-negative random variables,

$$\Delta = 2 \int F(x)(1 - F(x))\, dx. \tag{5}$$

The proof can be easily extended to any random variable with $\mathbb{E}|X| < \infty$ (Yitzhaki & Schechtman, 2013). Note that (5) also holds for discrete random variables. Hence, we can write the correlation of

(4) as

$$\rho_g(X,Y) = 1 - \frac{\sum_{k=1}^K p_k\Delta_k}{\Delta} = \frac{\Delta - \sum_{k=1}^K p_k\Delta_k}{\Delta}, \tag{6}$$

where Δ is the Gini mean difference (GMD) of F and Δ_k is the GMD of F_k. For this reason, we call the proposed correlation $\rho_g(X,Y)$ as the Gini correlation $\mathrm{gCor}(X,Y)$.

The representation of (6) allows another interpretation. Consider that $\sum_{k=1}^K p_k\Delta_k$, the weighted average of Gini mean differences, is a measure of within-group variation and $\Delta - \sum_{k=1}^K p_k\Delta_k$ is the corresponding between group variation. The proposed correlation is the ratio of the between-group Gini variation and the total Gini variation, analogue to the Pearson R^2 correlation in ANOVA (Analysis of Variance). The squared Pearson correlation is defined to be the ratio of between variance and the total variance. Denote μ, σ^2, μ_k, and σ_k^2 as the mean and variance of F and F_k, respectively. The variance of X can be partitioned to the within variation and the between variation as below,

$$\sigma^2 = Var(X) = \mathbb{E}[\mathbb{E}X^2|Y] - (\mathbb{E}[\mathbb{E}X|Y])^2 = \sum_{k=1}^K p_k(\sigma_k^2 + \mu_k^2) - \mu^2 = \sum_{k=1}^K p_k\sigma_k^2 + (\sum_{k=1}^K p_k\mu_k^2 - \mu^2).$$

And Pearson R^2 correlation, denoted as $\rho_p^2(X,Y)$, is

$$\rho_p^2(X,Y) = 1 - \frac{\sum_{k=1}^K p_k\sigma_k^2}{\sigma^2} = \frac{\sum_{k=1}^K p_k\mu_k^2 - \mu^2}{\sigma^2}.$$

Let (X,X'), (X_k,X_k'), (X_l,X_l') be independent pair variables independently from F, F_k and F_l, respectively. It is easy to derive that

$$\Delta = \mathbb{E}|X-X'| = \mathbb{E}\mathbb{E}(|X-X'||Y,Y') = \sum_{k=1}^K p_k^2\Delta_k + 2\sum_{1\le k<l\le K} p_kp_l\Delta_{kl}, \tag{7}$$

where $\Delta_k = \mathbb{E}|X_k - X_k'|$ and $\Delta_{kl} = \mathbb{E}|X_k - X_l|$. Then the between Gini variation, denoted as the Gini distance covariance between X and Y, is

$$\mathrm{gCov}(X,Y) = \Delta - \sum_{k=1}^K p_k\Delta_k = 2\sum_{1\le k<l\le K} p_kp_l\Delta_{kl} - \sum_{k=1}^K p_k(1-p_k)\Delta_k, \tag{8}$$

and the Gini distance correlation between X and Y is

$$\mathrm{gCor}(X,Y) = \rho_g(X,Y) = \frac{\mathrm{gCov}(X,Y)}{\Delta(X)}. \tag{9}$$

The total Gini variation is partitioned to the within and the between Gini variation. Frick *et al.* (2006) consider another decomposition of the Gini variation, which is represented by four components, i.e, within Gini variation, between Gini variation among group means and two effects of overlapping among groups. Although the extra terms provide some insights of the extent of group intertwining,

their decomposition is complicated. Not only our representation of the total Gini variation is simple and easy to interpret, but also it is natural to extend to the multivariate case.

3 Proposed Gini Correlation

3.1 Generalized Gini Correlation

There are two multivariate generalizations for the Gini mean difference. One is the Gini covariance matrix proposed by Dang *et al.* (2019). Along this line, one may extend the Gini correlation based on an analog of Wilk's lambda or Hotelling-Lawley trace in MANOVA. That leaves for future work. Here we explore another generalization defined in Koshevoy & Mosler (1997). That is, the Gini mean difference of a distribution F in $\mathbb{R}^d$ is

$$\Delta = \mathbb{E}\|\boldsymbol{X} - \boldsymbol{X}'\|,$$

or even more generally for some α,

$$\Delta(\alpha) = \mathbb{E}\|\boldsymbol{X} - \boldsymbol{X}'\|^\alpha, \tag{10}$$

where $\|\boldsymbol{x}\|$ is the Euclidean norm of $\boldsymbol{x}$. With this generalized multivariate Gini mean difference (10), we can define the Gini correlation in (4) as follows.

Definition 1 *For a non-degenerate random vector $\boldsymbol{X}$ in $\mathbb{R}^d$ and a categorical variable Y, if $\mathbb{E}\|\boldsymbol{X}\|^\alpha < \infty$ for $\alpha \in (0, 2)$, the Gini correlation of $\boldsymbol{X}$ and Y is defined as*

$$\rho_g(\boldsymbol{X}, Y; \alpha) = 1 - \frac{\sum_{k=1}^K p_k \Delta_k(\alpha)}{\Delta(\alpha)} = \frac{\Delta(\alpha) - \sum_{k=1}^K p_k \Delta_k(\alpha)}{\Delta(\alpha)}, \tag{11}$$

where $\Delta_k(\alpha)$ and $\Delta(\alpha)$ are the generalized Gini differences of F_k and F, respectively.

Remark 1 *Note that a small $\alpha > 0$ provides a weak assumption of $\mathbb{E}\|\boldsymbol{X}\|^\alpha < \infty$ on distributions, which allows applications of the Gini correlation to heavy-tailed distributions.*

Remark 2 *If $\alpha = 2$ and $d = 1$, $\rho_g(X, Y; 2) = \rho_p^2(X, Y)$ because of the fact that $\Delta(2) = \mathbb{E}|X - X'|^2 = 2\,Var(X)$. The requirement of $\alpha \in (0, 2)$ is for desired properties of the Gini correlation.*

The next theorem states the properties of the proposed Gini correlation.

Theorem 1 *For a categorical variable Y and a continuous random vector $\boldsymbol{X}$ in $\mathbb{R}^d$ with $\mathbb{E}\|\boldsymbol{X}\|^\alpha < \infty$ for $0 < \alpha < 2$, $\rho_g(\boldsymbol{X}, Y; \alpha)$ has following properties.*

1. $0 \le \rho_g(\boldsymbol{X}, Y; \alpha) \le 1$.
2. $\rho_g(\boldsymbol{X}, Y; \alpha) = 0$ *if and only if* $\boldsymbol{X}$ *and* Y *are independent.*

3. $\rho_g(\boldsymbol{X}, Y; \alpha) = 1$ *if and only if* F_k *is a single point mass and not identical distribution for* $k = 1, ..., K$.

4. $\rho_g(aO\boldsymbol{X} + \boldsymbol{b}, Y; \alpha) = \rho_g(\boldsymbol{X}, Y; \alpha)$ *for any orthonormal matrix* O $(O^T = O^{-1})$, *nonzero constant* a *and vector* $\boldsymbol{b}$.

Proof of Theorem 1. First of all, $\Delta_k(\alpha) \geq 0$, so we have $\rho_g(\boldsymbol{X}, Y; \alpha) \leq 1$. It is obvious that $\rho_g(\boldsymbol{X}, Y; \alpha) = 1$ if and only if $\Delta_k(\alpha) = 0$ for each k, which mutually implies that F_k is a singleton distribution. Orthogonal invariance of the Property (4) is a result from the Euclidean distance used in the Gini correlation. The proof for the remaining part has two steps. In Step 1, we can write

$$\mathrm{gCov}(\boldsymbol{X}, Y; \alpha) = \Delta(\alpha) - \sum_{k=1}^{K} p_k \Delta_k(\alpha) = \sum_{k=1}^{K} p_k T(\boldsymbol{X}_k, \boldsymbol{X}; \alpha), \tag{12}$$

where $T(\boldsymbol{X}_k, \boldsymbol{X}; \alpha) = 2\mathbb{E}\|\boldsymbol{X}_k - \boldsymbol{X}\|^\alpha - \mathbb{E}\|\boldsymbol{X}_k - \boldsymbol{X}_k'\|^\alpha - \mathbb{E}\|\boldsymbol{X} - \boldsymbol{X}'\|^\alpha$. This is because

$$\begin{aligned}
\sum_{k=1}^{K} p_k T(\boldsymbol{X}_k, \boldsymbol{X}; \alpha) &= \sum_{k=1}^{K} p_k \Big\{ 2p_k \Delta_k(\alpha) + 2 \sum_{l \neq k} p_l \Delta_{kl}(\alpha) - \Delta_k(\alpha) - \Delta(\alpha) \Big\} \\
&= \sum_{k=1}^{K} (2p_k^2 - p_k) \Delta_k(\alpha) + 2 \sum_{k=1}^{K} \sum_{l \neq k} p_k p_l \Delta_{kl}(\alpha) - \Delta(\alpha) \\
&= 2 \sum_{1 \leq k < l \leq K} p_k p_l \Delta_{kl}(\alpha) - \sum_{k=1}^{K} p_k (1 - p_k) \Delta_k(\alpha) \qquad (13) \\
&= \Delta(\alpha) - \sum_{k=1}^{K} p_k \Delta_k(\alpha).
\end{aligned}$$

The third equality (13) is obtained by plugging in (7) and the last equality is due to (8). In Step 2, one recognizes that $T(\boldsymbol{X}_k, \boldsymbol{X}, \alpha)$ is the energy distance between $\boldsymbol{X}$ and $\boldsymbol{X}_k$ defined in Székely & Rizzo (2013, 2017). Applying the Proposition 2 of Székely & Rizzo (2013), for $0 < \alpha < 2$, we have

$$T(\boldsymbol{X}_k, \boldsymbol{X}; \alpha) = c(d, \alpha) \int_{\mathbb{R}^d} \frac{|\psi_k(\boldsymbol{t}) - \psi(\boldsymbol{t})|^2}{\|\boldsymbol{t}\|^{d+\alpha}} d\boldsymbol{t}, \tag{14}$$

where ψ_k and ψ are the characteristic functions of $\boldsymbol{X}_k$ and $\boldsymbol{X}$, respectively, and $c(d, \alpha)$ is a constant only depending on d and α, i.e.,

$$c(d, \alpha) = \frac{\alpha 2^\alpha \Gamma((d+\alpha)/2)}{2\pi^{d/2} \Gamma(1 - \alpha/2)}.$$

Results of (12) and (14) show that for all $0 < \alpha < 2$, we have $\mathrm{gCov}(\boldsymbol{X}, Y; \alpha) \geq 0$ and hence $\rho_g(\boldsymbol{X}, Y, \alpha) \geq 0$ with equality to zero if and only if $\boldsymbol{X}$ and $\boldsymbol{X}_k$ are identically distributed for all $k = 1, ..., K$. □

Below we provide a couple of remarks and their proofs are given in Appendix.

Remark 3 $T(\boldsymbol{X}_k, \boldsymbol{X}; \alpha)$ *is the energy distance of* $\boldsymbol{X}_k$ *and* $\boldsymbol{X}$, *which is the weighted* L_2 *distance of*

characteristic functions of $\boldsymbol{X}_k$ and $\boldsymbol{X}$. For $d = 1$, $T(X_k, X; 1)$ is also the L_2 distance of the distribution function F_k and F multiplying a constant. However, such a relationship does not hold for $d > 1$.

Remark 4 *The Gini covariance of $\boldsymbol{X}$ and Y is the weighted average of energy distance between $\boldsymbol{X}_k$ and $\boldsymbol{X}$. It is also a linear combination of energy distances between $\boldsymbol{X}_k$ and $\boldsymbol{X}_l$. That is, $gCov(\boldsymbol{X}, Y; \alpha) = \sum_{k=1}^{K} p_k T(\boldsymbol{X}_k, \boldsymbol{X}; \alpha) = \sum_{1 \le k < l \le K} p_k p_l T(\boldsymbol{X}_k, \boldsymbol{X}_l; \alpha)$.*

Particularly for $K = 2$, the between variation $\mathrm{gCov}(\boldsymbol{X}, Y; \alpha) = \Delta(\alpha) - p_1\Delta_1(\alpha) - p_2\Delta_2(\alpha)$ is simplified to be

$$p_1 T(\boldsymbol{X}_1, \boldsymbol{X}; \alpha) + p_2 T(\boldsymbol{X}_2, \boldsymbol{X}; \alpha) = p_1 p_2 T(\boldsymbol{X}_1, \boldsymbol{X}_2; \alpha)$$

which is proportional to $T(\boldsymbol{X}_1, \boldsymbol{X}_2; \alpha)$, the energy distance used in Székely & Rizzo (2013, 2017). Székely & Rizzo (2004) considered a special case of the energy distance of $\alpha = 1$ and proposed a test for the equality of two distributions F_1 and F_2, which is also studied in Baringhaus & Franz (2004). The test is equivalent to test $\rho_g(\boldsymbol{X}, Y; \alpha) = 0$. The test of $\rho_g(\boldsymbol{X}, Y; \alpha) = 0$ can be also used for the K-sample problem. In that case, it is equivalent to the test of DISCO (DIStance COmponent) analysis in Rizzo & Székely (2010). The test statistic in DISCO takes the ratio of the between and the within group Gini variations for the K-sample problem. Testing $\rho_g(\boldsymbol{X}, Y; \alpha) = 0$ is equivalent to the one-way DISCO analysis. What we contribute in the dependence test is that our test is able to provide power analysis for a particular alternative if it is specified as $\rho_g(\boldsymbol{X}, Y; \alpha) = \rho_0$ where $\rho_0 > 0$.

Remark 5 *The recent paper of Zhang et al. (2019) proposes an extension of Gini covariance and correlation by considering an energy distance in reproducing kernel Hilbert spaces (RKHS).*

More specifically, let $\kappa : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a Mercer kernel (Mercer, 1909). There is an associated RKHS $\mathcal{H}_\kappa$ of real functions on $\mathbb{R}^d$ with reproducing kernel κ, where the function $D : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ defines a distance in $\mathcal{H}_\kappa$, $D_\kappa(\boldsymbol{x}, \boldsymbol{x}') = \sqrt{\kappa(\boldsymbol{x}, \boldsymbol{x}) + \kappa(\boldsymbol{x}', \boldsymbol{x}') - 2\kappa(\boldsymbol{x}, \boldsymbol{x}')}$. Hence Gini covariance and Gini correlation are generalized to RKHS, $\mathcal{H}_\kappa$, as

$$\begin{aligned}
\mathrm{gCov}(\boldsymbol{X}, Y; \kappa) &= \sum_{k=1}^{K} p_k T(\boldsymbol{X}_k, \boldsymbol{X}; \kappa) = \sum_{k=1}^{K} p_k \left[2\mathbb{E}D_\kappa(\boldsymbol{X}_k, \boldsymbol{X}) - \mathbb{E}D_\kappa(\boldsymbol{X}_k, \boldsymbol{X}_k{}') - \mathbb{E}D_\kappa(\boldsymbol{X}, \boldsymbol{X}') \right], \\
\mathrm{gCor}(\boldsymbol{X}, Y; \kappa) &= \frac{\sum_{k=1}^{K} p_k \left[2\mathbb{E}D_\kappa(\boldsymbol{X}_k, \boldsymbol{X}) - \mathbb{E}D_\kappa(\boldsymbol{X}_k, \boldsymbol{X}_k{}') - \mathbb{E}D_\kappa(\boldsymbol{X}, \boldsymbol{X}') \right]}{\mathbb{E}D_\kappa(\boldsymbol{X}, \boldsymbol{X}')}. \qquad (15)
\end{aligned}$$

Denote $\mathrm{gCor}(\boldsymbol{X}, Y; \kappa)$ as $\rho_\kappa(\boldsymbol{X}, Y)$. With a choice of a bounded kernel, the moment assumption on $\boldsymbol{X}$ can be relaxed for the kernel Gini covariance and kernel Gini correlation.

3.2 Connection to the Distance Correlation

The proposed Gini correlation is closely related to but different from the celebrated distance correlation studied by Székely *et al.* (2007), Székely & Rizzo (2009). Their distance correlation considers correlation between two sets of continuous random variables. Later the distance covariance and distance correlation are extended from Euclidean space to general metric spaces by Lyons (2013). Based on that idea, we define the discrete metric

$$d(y, y') = |y - y'| := I(y \neq y'),$$

where $I(\cdot)$ is the indicator function. Equipped with this set difference metric on the support of Y and Euclidean distance on the support of $\boldsymbol{X}$, the corresponding distance covariance and distance correlation for numerical $\boldsymbol{X}$ and categorical Y variables are as follows.

$$\mathrm{dCov}(\boldsymbol{X}, Y; \alpha) = c(d, \alpha) \sum_{k=1}^{K} \int \frac{(p_k \psi_k(\boldsymbol{t}) - p_k \psi(\boldsymbol{t}))^2}{\|\boldsymbol{t}\|^{d+\alpha}} d\boldsymbol{t}, \tag{16}$$

$$\mathrm{dCov}(\boldsymbol{X}, \boldsymbol{X}; \alpha) = c(d, \alpha)^2 \int \frac{(\psi(\boldsymbol{t}+\boldsymbol{s}) - \psi(\boldsymbol{t})\psi(\boldsymbol{s}))^2}{\|\boldsymbol{t}\|^{d+\alpha}\|\boldsymbol{s}\|^{d+\alpha}} d\boldsymbol{t} d\boldsymbol{s},$$

$$\mathrm{dCov}(Y, Y) = \sum_{k=1}^{K} p_k^2 - 2\sum_{k=1}^{K} p_k^3 + (\sum_{k=1}^{K} p_k^2)^2, \tag{17}$$

$$\rho_d(\mathbf{X}, Y; \alpha) := \ \mathrm{dCor}(\boldsymbol{X}, Y; \alpha) = \frac{\mathrm{dCov}(\boldsymbol{X}, Y, \alpha)}{\sqrt{\mathrm{dCov}(\boldsymbol{X}, \boldsymbol{X}; \alpha)}\sqrt{\mathrm{dCov}(Y, Y)}}.$$

Remark 6 *As expected,* $dCov(\boldsymbol{X}, Y; \alpha) = \mathbb{E}|\boldsymbol{X} - \boldsymbol{X}'|^\alpha |Y - Y'|^\alpha + \mathbb{E}|\boldsymbol{X} - \boldsymbol{X}'|^\alpha \mathbb{E}|Y - Y'|^\alpha - 2\mathbb{E}|\boldsymbol{X} - \boldsymbol{X}'|^\alpha |Y - Y''|^\alpha$, *where* $(\boldsymbol{X}, Y), (\boldsymbol{X}', Y'), (\boldsymbol{X}'', Y'')$ *are i.i.d.*

The proofs of this identity in Remark 6, (17) and the following Remark 9 are given in Appendix.

Comparing (16) with (12) and (14), it is easy to make the following conclusions.

Remark 7 $dCov(\boldsymbol{X}, Y; \alpha) = \sum_{k=1}^{K} p_k^2 T(\boldsymbol{X}_k, \boldsymbol{X}; \alpha)$, *while* $gCov(\boldsymbol{X}, Y; \alpha) = \sum_{k=1}^{K} p_k T(\boldsymbol{X}_k, \boldsymbol{X}; \alpha)$ *is more natural and simple, resulting in a simple form and nice interpretation of gCor.*

Remark 8 *When* $p_1 = p_2 = ... = p_K = \frac{1}{K}$, $gCov(\boldsymbol{X}, Y; \alpha) = K dCov(\boldsymbol{X}, Y; \alpha)$.

Remark 9 *For* $K = 2$, $gCov(\boldsymbol{X}, Y; \alpha) = dCov(\boldsymbol{X}, Y; \alpha)/(2p_1p_2) = dCov(\boldsymbol{X}, Y; \alpha)/\sqrt{dCov(Y, Y)}$, *i.e.,* $\rho_g(\boldsymbol{X}, Y; \alpha)$ *and* $\rho_d(\boldsymbol{X}, Y; \alpha)$ *are only different on a scaling factor that is independent with weights of the two classes.*

Remark 10 *For the case of* $d = 1$, $dCov(X, X; 1)$ *is studied in [11] and*

$$dCov(X, Y; 1) = 2\sum_{k=1}^{K} \int (p_k F_k(x) - p_k F(x))^2 \, dx. \tag{18}$$

Comparison of (18) and (1) explains the difference of our Gini approach and distance correlation approach in the one dimensional case. The distance covariance of X and Y is based on squared difference of the joint distribution $p_k F_k(x)$ and the product of the marginal distributions $p_k F(x)$, while the Gini one is based on the squared difference between the conditional distribution $F_k(x)$ and the marginal distribution $F(x)$. Our Gini dependence measure considers the categorical nature of Y and has a simpler formulation than the distance correlation, leading a simpler inference and computation.

Before we discuss their computation and inference, let us first demonstrate the Gini correlation and distance correlation in several examples.

3.3 Examples

Three examples for $K = 2$, $d = 1$ and $\alpha = 1$ are provided. Denote p_1 as p. The detailed derivations and proofs for the example results are provided in the Appendix.

Example 1. Let $F_1 = \text{Exp}(\theta)$ and $F_2 = \text{Exp}(\beta)$. We have

$$\mu_1 = \sigma_1 = \Delta_1 = \theta, \mu_2 = \sigma_2 = \Delta_2 = \beta, \Delta_{12} = \frac{\theta^2 + \beta^2}{\theta + \beta},$$
$$\text{dCov}(X, X) = 2p^2\theta^2 + 2(1-p)^2\beta^2 + (p^2\theta + (1-p)^2\beta)^2 - \frac{8}{3}p^3\theta^2 - \frac{8}{3}(1-p)^3\beta^2 + \frac{16p(1-p)\theta^2\beta^2}{(\theta+\beta)^2}$$
$$+ \frac{32p^2(1-p)^2\theta^2\beta^2}{(\theta+\beta)^2} + \frac{8p^3(1-p)\theta^2\beta}{\theta+\beta} + \frac{8p(1-p)^3\theta\beta^2}{\theta+\beta} - \frac{8p(1-p)^2\theta\beta^2(5\theta+\beta)}{(2\theta+\beta)(\theta+\beta)}$$
$$- \frac{8p^2(1-p)\theta^2\beta(\theta+5\beta)}{(\theta+2\beta)(\theta+\beta)}.$$

As we see, the formula of $\text{dCov}(X, X)$ is complicated for the 2-component exponential mixture distribution. The correlations are given as follows.

$$\begin{aligned}
\rho_g(X, Y) &= \frac{p(1-p)(\theta-\beta)^2}{(2p-p^2)\theta^2 + (1-p^2)\beta^2 + (1-2p+2p^2)\theta\beta}, \\
\rho_d(X, Y) &= \frac{p(1-p)(\theta-\beta)^2}{2(\theta+\beta)\sqrt{\text{dCov}(X, X)}}, \\
\rho_p^2(X, Y) &= \frac{p(1-p)(\theta-\beta)^2}{p\theta^2 + (1-p)\beta^2 + p(1-p)(\theta-\beta)^2}.
\end{aligned}$$

Figure 1 demonstrates Gini correlation, distance correlation and squared Pearson correlation in the exponential mixtures. The cases of $p = 0$ or $p = 1$ in (a) and $\beta = \theta = 1$ in (b) have zero Gini, zero distance and zero Pearson correlation coefficients, corresponding to the case of independence of X and Y. The value of the Gini correlation is between the squared Pearson correlation and distance correlation. As expected, all correlations increase as the ratio $r = \beta/\theta$ increases for $r \geq 1$. This result (at least for the Gini and Pearson R^2 correlations) can be proved by the positiveness of the first derivative of the correlation with respect to r, which is given in the Appendix.

[Put Figure 1 here]

Example 2. Let $F_1 = \text{Normal}(\mu_1, \sigma^2)$, $F_2 = \text{Normal}(\mu_2, \sigma^2)$ and $a = |\mu_1 - \mu_2|/\sigma$. We have

$$\Delta_1 = \Delta_2 = \frac{2\sigma}{\sqrt{\pi}}, \Delta_{12} = \sigma[2a\Phi(a/\sqrt{2}) + 2\sqrt{2}\phi(a/\sqrt{2}) - a],$$

where $\phi(x)$ and $\Phi(x)$ are the density and cumulative functions of the standard normal distribution, respectively. But it is too complicate to derive formula of $dCov(X, X)$ when X is from a mixture of two normal distributions. In this case, we are only able to derive Gini correlation and the squared Pearson correlation as follows.

$$\begin{aligned} \rho_g(X,Y) &= \frac{p(1-p)[2a\Phi(a/\sqrt{2}) + 2\sqrt{2}\phi(a/\sqrt{2}) - a - 2/\sqrt{\pi}]}{(p^2 + (1-p)^2)/\sqrt{\pi} + p(1-p)[2a\Phi(a/\sqrt{2}) + 2\sqrt{2}\phi(a/\sqrt{2}) - a]}, \\ \rho_p^2(X,Y) &= \frac{p(1-p)a^2}{1 + p(1-p)a^2}. \end{aligned}$$

For a mixture of two normal distributions with a same standard deviation but different means, independence of X and Y is equivalent to either $p = 0, p = 1$ in (a) or $a = 0$ in (b) for both correlations, which is demonstrated in Figure 2. For dependence cases, the squared Pearson correlation is larger than the Gini correlation. With any fixed $a \neq 0$ (i.e., $\mu_1 \neq \mu_2$), the largest correlation is obtained at $p = 0.5$ (the balance case) for both correlations. Also both correlations are monotonically increasing functions of $|a|$ for any $p \neq 0$ or 1.

[Put Figure 2 here]

Example 3. Let $F_1 = \text{Normal}(\mu, \sigma_1^2)$, $F_2 = \text{Normal}(\mu, \sigma_2^2)$ and $r = \sigma_2/\sigma_1$. Again, it is too complicate to derive the formula of $\text{dCov}(X, X; 1)$ in this example. Since two distributions have a same mean, $\rho_p^2(X, Y)$ is always 0 and hence it completely fails to measure the difference of two distributions when $\sigma_1 \neq \sigma_2$. For the Gini correlation, we have

$$\Delta_1 = \frac{2\sigma_1}{\sqrt{\pi}}, \Delta_2 = \frac{2\sigma_2}{\sqrt{\pi}}, \Delta_{12} = \frac{\sqrt{2(\sigma_1^2 + \sigma_2^2)}}{\sqrt{\pi}}.$$

Then

$$\rho_g(X,Y) = \frac{p(1-p)(\sqrt{2(1+r^2)} - 1 - r)}{p^2 + (1-p)^2 r + p(1-p)\sqrt{2(1+r^2)}}.$$

Figure 3 (a) shows Gini correlation changes with p for normal mixture under different ratios of standard deviations. Figure 3 (b) shows the changes of Gini correlation with ratio of standard deviations of normal mixture under different p. In the cases of $p = 0, 1$ and $r = 1$ in (a) and the case of the ratio to be 1 in (b), the Gini correlation is 0, corresponding to the independence of X and Y. The Gini correlation is monotonically increasing in $r > 1$ for each $p \neq 0$ or 1.

[Put Figure 3 here]

3.4 Extension to Categorical Random Vector

The approach can easily extend to deal with multiple factors or multivariate $\mathbf{Y}$ since the total Gini variation can be partitioned in a similar way as multi-way ANOVA, which is studied by Rizzo & Székely (2010). We present here with an emphasis on association measures of the levels of $\boldsymbol{Y}$ and $\boldsymbol{X}$.

Consider a bivariate $\boldsymbol{Y} = (Y_a, Y_b)^T$. The multivariate $\boldsymbol{Y}$ cases can be obtained similarly. Suppose Y_a is a categorical variable taking A levels with distribution $P(Y_a = a_i) = p_i$, $i = 1, ..., A$ and Y_b is a categorical variable taking B levels with the distribution $P(Y_b = b_k) = q_k$, $k = 1, ..., B$. The conditional distribution of $\boldsymbol{X}$ given $Y_a = a_i$ and $Y_b = b_k$ is F_{ik}^{ab}, the conditional distribution of $\boldsymbol{X}$ given $Y_a = a_i$ is F_i^a and the conditional distribution of $\boldsymbol{X}$ given $Y_b = b_k$ is F_k^b. Then the marginal distribution of $\boldsymbol{X}$ is

$$F(\boldsymbol{x}) = \sum_{i=1}^{A} p_i F_i^a(\boldsymbol{x}) = \sum_{k=1}^{B} q_k F_k^b(\boldsymbol{x}) = \sum_{i=1}^{A}\sum_{k=1}^{B} p_i q_k F_{ik}^{ab}(\boldsymbol{x}).$$

Let $(\boldsymbol{X}, \tilde{\boldsymbol{X}})$, $(\boldsymbol{X}_i^a, \tilde{\boldsymbol{X}}_i^a)$, $(\boldsymbol{X}_k^b, \tilde{\boldsymbol{X}}_k^b)$, $(\boldsymbol{X}_{ik}^{ab}, \tilde{\boldsymbol{X}}_{ik}^{ab})$ be independent pairs independently from F, F_i^a, F_k^b and F_{ik}^{ab}, respectively. Define the generalized Gini mean distances as follows.

$$\Delta = \mathbb{E}\|\boldsymbol{X} - \tilde{\boldsymbol{X}}\|^\alpha,\ \Delta_i^a = \mathbb{E}\|\boldsymbol{X}_i^a - \tilde{\boldsymbol{X}}_i^a\|^\alpha,\ \Delta_k^b = \mathbb{E}\|\boldsymbol{X}_k^b - \tilde{\boldsymbol{X}}_k^b\|^\alpha,\ \Delta_{ik}^{ab} = \mathbb{E}\|\boldsymbol{X}_{ik}^{ab} - \tilde{\boldsymbol{X}}_{ik}^{ab}\|^\alpha.$$

Then the total Gini variation can be decomposed as $\Delta = S(a) + S(b) + S(ab) + W(ab)$, where

$$\begin{aligned}
S(a) &= \Delta - \sum_{i=1}^{A} p_i \Delta_i^a, \quad S(b) = \Delta - \sum_{k=1}^{B} q_k \Delta_k^b, \\
S(ab) &= \sum_{i=1}^{A} p_i \Delta_i^a + \sum_{k=1}^{B} q_k \Delta_k^b - \Delta - \sum_{i=1}^{A}\sum_{k=1}^{B} p_i q_k \Delta_{ik}^{ab} \\
W(ab) &= \sum_{i=1}^{A}\sum_{k=1}^{B} p_i q_k \Delta_{ik}^{ab}.
\end{aligned}$$

$W(ab)$ is the within variation. $W(ab) = \Delta$ if and only if $\boldsymbol{Y}$ and $\boldsymbol{X}$ are independent. $S(a)$ and $S(b)$ represent the between variation attributed from Y_a and Y_b, respectively. While $S(ab)$ accounts for the intersection of Y_a and Y_b on $\boldsymbol{X}$. Analogous to ω^2's in two-way ANOVA, we can define the individual ratios $S(a)/\Delta$, $S(b)/\Delta$ and $S(ab)/\Delta$ as association measures. They indicate the proportions of Gini variance in $\boldsymbol{X}$ that are accounted for by the levels of $\boldsymbol{Y}$.

4 Inference

4.1 Estimation

The inference in this section focuses on single categorical variable Y. The extension to infer multivariate $\boldsymbol{Y}$ is straightforward. Suppose a sample data $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}$ for $i = 1, ..., n$ available. We can write $\mathcal{D} = \{\mathcal{D}_1, \mathcal{D}_2, ..., \mathcal{D}_K\}$, where $\mathcal{D}_k = \{\boldsymbol{x}_1^{(k)}, ..., \boldsymbol{x}_{n_k}^{(k)}\}$ is the sample with $y_i = L_k$ and n_k is the number of observations in $\mathcal{D}_k$. p_k is estimated by $\hat{p}_k = n_k/n$. Δ_k and Δ can be estimated as follows.

$$\hat{\Delta}_k(\alpha) = n_k^{-2} \sum_{1 \le i,j \le n_k} \|\boldsymbol{x}_i^{(k)} - \boldsymbol{x}_j^{(k)}\|^\alpha, \quad \hat{\Delta}(\alpha) = n^{-2} \sum_{1 \le i,j \le n} \|\boldsymbol{x}_i - \boldsymbol{x}_j\|^\alpha. \tag{19}$$

Then a corresponding point estimator of $\rho_g(\alpha)$ is

$$\hat{\rho}_g(\alpha) = 1 - \frac{\sum_{k=1}^K \hat{p}_k \hat{\Delta}_k(\alpha)}{\hat{\Delta}(\alpha)} = \frac{\hat{\Delta}(\alpha) - \sum_{k=1}^K \hat{p}_k \hat{\Delta}_k(\alpha)}{\hat{\Delta}(\alpha)}. \tag{20}$$

The estimators in (19) are V Statistics, which are biased. We work with biased sample versions to avoid dealing with complicated constants in the ensuing result of $\hat{\rho}(\alpha)$. We have the following theorems on the asymptotic behavior of the sample Gini correlation.

Theorem 2 *If $\mathbb{E}\|\boldsymbol{X}\|^\alpha < \infty$ and $p_k > 0$ for all $k = 1, ..., K$, then almost surely*

$$\lim_{n \to \infty} \hat{\rho}_g(\alpha) = \rho_g(\alpha).$$

Proof of Theorem 2. By the SLLN, $\hat{p}_k = \frac{1}{n} \sum_{i=1}^n I(y_i = L_k)$ converges to p_k with probability 1. Also by the almost sure behavior of V-statistics (Serfling, 1980), $\hat{\Delta}_k(\alpha)$ and $\hat{\Delta}(\alpha)$ converge with probability 1 to $\Delta_k(\alpha)$ and $\Delta(\alpha)$, respectively. Let g be the function $g(a_1, ..., a_K, b_1, ..., b_K, b) = 1 - \sum_{k=1}^K a_k b_k / b$, which is continuous for $b > 0$. Therefore, the strong consistency of the sample Gini correlation follows by the fact that $\hat{\rho}_g(\alpha) = g(\hat{p}_1, ..., \hat{p}_K, \hat{\Delta}_1(\alpha), ..., \hat{\Delta}_K(\alpha), \hat{\Delta}(\alpha))$. □

Theorem 3 *Suppose that $\mathbb{E}\|\boldsymbol{X}\|^{2\alpha} < \infty$, $p_k > 0$ for all $k = 1, ..., K$ and $\rho_g(\alpha) \neq 0$. We have*

$$\sqrt{n}(\hat{\rho}_g(\alpha) - \rho_g(\alpha)) \xrightarrow{d} N(0, \sigma_g^2(\alpha)),$$

where $\sigma_g^2(\alpha)$ is the asymptotic variance given in the proof.

Proof of Theorem 3. For simplicity of presentation, we suppress α in notations in the proof without confusion. Let $\boldsymbol{q}$ be the vector of length $K(K-1)$ with elements $\{p_k p_l\}_{1 \le k \neq l \le K}$. Let $\boldsymbol{h}$ be the kernel functions of form $\boldsymbol{h} = \{h_{kl}\}_{1 \le k \neq l \le K}$, where

$$h_{kl} := h(\boldsymbol{x}_1^{(k)}, \boldsymbol{x}_2^{(k)}; \boldsymbol{x}_1^{(l)}, \boldsymbol{x}_2^{(l)}) = \frac{1}{2}[\|\boldsymbol{x}_1^{(k)} - \boldsymbol{x}_1^{(l)}\|^\alpha + \|\boldsymbol{x}_2^{(k)} - \boldsymbol{x}_2^{(l)}\|^\alpha - \|\boldsymbol{x}_1^{(k)} - \boldsymbol{x}_2^{(k)}\|^\alpha - \|\boldsymbol{x}_1^{(l)} - \boldsymbol{x}_2^{(l)}\|^\alpha].$$

Let $(\boldsymbol{X}_1^{(k)}, \boldsymbol{X}_2^{(k)})$ and $(\boldsymbol{X}_1^{(l)}, \boldsymbol{X}_2^{(l)})$ be independent pairs independently from distributions F_k and F_l, respectively. Then $\boldsymbol{q}^T \mathbb{E}\boldsymbol{h}(\boldsymbol{X}_1^{(1)}, \boldsymbol{X}_2^{(1)}; ...; \boldsymbol{X}_1^{(K)}, \boldsymbol{X}_2^{(K)}) = \sum_{1\le k\neq l\le K} p_k p_l T(\boldsymbol{X}^{(k)}, \boldsymbol{X}^{(l)}; \alpha)$, which is $\mathrm{gCov}(\boldsymbol{X}, Y; \alpha)$ by Remark 4. Let $\hat{T}_{kl}$ be the V-statistic estimator of $T(\boldsymbol{X}^{(k)}, \boldsymbol{X}^{(l)}; \alpha)$. That is,

$$\begin{aligned}\hat{T}_{kl} &= \frac{1}{n_k^2}\frac{1}{n_l^2}\sum_{i_1,i_2=1}^{n_k}\sum_{j_1,j_2=1}^{n_l} h(\boldsymbol{x}_{i_1}^{(k)}, \boldsymbol{x}_{i_2}^{(k)}, \boldsymbol{x}_{j_1}^{(l)}, \boldsymbol{x}_{j_2}^{(l)}) \\ &= \frac{1}{n_k n_l}\sum_{i_1=1}^{n_k}\sum_{j_1=1}^{n_l}\|\boldsymbol{x}_{i_1}^{(k)} - \boldsymbol{x}_{j_1}^{(l)}\|^\alpha - \frac{1}{2n_k^2}\sum_{i_1,i_2=1}^{n_k}\|\boldsymbol{x}_{i_1}^{(k)} - \boldsymbol{x}_{i_2}^{(k)}\|^\alpha - \frac{1}{2n_l^2}\sum_{j_1,j_2=1}^{n_l}\|\boldsymbol{x}_{j_1}^{(l)} - \boldsymbol{x}_{j_2}^{(l)}\|^\alpha.\end{aligned}$$

Then an estimator of $\mathrm{gCov}(\boldsymbol{X}, Y; \alpha)$ given by $\sum_{1\le k\neq l\le K}\hat{p}_k\hat{p}_l\hat{T}_{kl}$ is same as the V-statistic estimator given in (20) because

$$\begin{aligned}&\sum_{1\le k\neq l\le K}\hat{p}_k\hat{p}_l\hat{T}_{kl} = \frac{1}{n^2}\left[\sum_{1\le k\neq l\le K}\sum_{i_1=1}^{n_k}\sum_{j_1=1}^{n_l}\|\boldsymbol{x}_{i_1}^{(k)} - \boldsymbol{x}_{j_1}^{(l)}\|^\alpha - \sum_{k=1}^{K}\frac{n-n_k}{n_k}\sum_{i_1,i_2=1}^{n_k}\|\boldsymbol{x}_{i_1}^{(k)} - \boldsymbol{x}_{i_2}^{(k)}\|^\alpha\right] \\ &= \frac{1}{n^2}\left[\sum_{k=1}^{K}\sum_{i_1,i_2=1}^{n_k}\|\boldsymbol{x}_{i_1}^{(k)} - \boldsymbol{x}_{i_2}^{(k)}\|^\alpha + \sum_{k\neq l}\sum_{i_1=1}^{n_k}\sum_{j_1=1}^{n_l}\|\boldsymbol{x}_{i_1}^{(k)} - \boldsymbol{x}_{j_1}^{(l)}\|^\alpha\right] - \sum_{k=1}^{K}\frac{n_k}{n}\frac{1}{n_k^2}\sum_{i_1,i_2=1}^{n_k}\|\boldsymbol{x}_{i_1}^{(k)} - \boldsymbol{x}_{i_2}^{(k)}\|^\alpha \\ &= \hat{\Delta}(\alpha) - \sum_k \hat{p}_k\hat{\Delta}_k(\alpha).\end{aligned}$$

Consider the centered kernel function $\tilde{h}_{kl} = h_{kl} - \mathbb{E}h_{kl}$ and its first order projections as follows.

$$\begin{aligned}\tilde{h}_{kl}^{10}(\boldsymbol{x}^{(k)}) &= \mathbb{E}\tilde{h}_{kl}(\boldsymbol{x}^{(k)}, \boldsymbol{X}_2^{(k)}; \boldsymbol{X}_1^{(l)}, \boldsymbol{X}_2^{(l)}) = \frac{1}{2}[\mathbb{E}\|\boldsymbol{x}^{(k)} - \boldsymbol{X}_1^{(l)}\|^\alpha - \mathbb{E}\|\boldsymbol{x}^{(k)} - \boldsymbol{X}_2^{(k)}\|^\alpha - \Delta_{kl}(\alpha) + \Delta_k(\alpha)], \\ \tilde{h}_{kl}^{01}(\boldsymbol{x}^{(l)}) &= \mathbb{E}\tilde{h}_{kl}(\boldsymbol{X}_1^{(k)}, \boldsymbol{X}_2^{(k)}; \boldsymbol{x}^{(l)}, \boldsymbol{X}_2^{(l)}) = \frac{1}{2}[\mathbb{E}\|\boldsymbol{X}^{(k)} - \boldsymbol{x}^{(l)}\|^\alpha - \mathbb{E}\|\boldsymbol{x}^{(l)} - \boldsymbol{X}_2^{(l)}\|^\alpha - \Delta_{kl}(\alpha) + \Delta_l(\alpha)].\end{aligned} \tag{21}$$

Denote the first order projection of $\hat{T}_{kl}$ as $\hat{T}_{kl}^{(1)} = 2[\frac{1}{n_k}\sum_{i=1}^{n_k}\tilde{h}_{kl}^{10}(\boldsymbol{x}_i^{(k)}) + \frac{1}{n_l}\sum_{j=1}^{n_l}\tilde{h}_{kl}^{01}(\boldsymbol{x}_j^{(l)})]$. Then

$$\sum_{k\neq l}\hat{p}_k\hat{p}_l\hat{T}_{kl}^{(1)} = \frac{4}{n^2}\sum_{k\neq l}n_l\sum_{i=1}^{n_k}\tilde{h}_{kl}^{10}(\boldsymbol{x}_i^{(k)}).$$

If at least one $\sigma_{kl}^2(\alpha) := var(\tilde{h}_{kl}^{10}(\boldsymbol{X}^{(k)}))$ is not zero, the corresponding V-statistic has an asymptotic normal distribution (Theorem A of Section 6.4, Serfling, 1980). That is,

$$\sqrt{n}(\hat{\Delta}(\alpha) - \sum_k \hat{p}_k\hat{\Delta}_k(\alpha) - \mathrm{gCov}(\boldsymbol{X}, y; \alpha)) \to \mathcal{N}(0, 16\sum_{k\neq l}^{K} p_l^2 p_k \sigma_{kl}^2(\alpha)).$$

Next, we need to show that $\rho_g(\boldsymbol{X}, Y; \alpha) = 0$ if and only if all $\sigma_{kl}^2(\alpha) = 0$. First, $\rho_g(\boldsymbol{X}, Y; \alpha) = 0$ implies that $F_1 = F_2 ... = F_K = F$, $\tilde{h}_{kl}^{10}(\boldsymbol{X}^{(k)})) = 0$ and hence all $\sigma_{kl}^2(\alpha) = 0$. On the other hand, if $\sigma_{kl}^2(\alpha) = 0$, then $\tilde{h}_{kl}^{10}(\boldsymbol{X}^{(k)}))$ is a constant C almost surely. Taking the expectation on $\tilde{h}_{kl}^{10}(\boldsymbol{X}^{(k)}))$ gives $C = 0$. Also

$\tilde{h}_{kl}^{10}(\boldsymbol{X}^{(k)}) = \tilde{h}_{lk}^{10}(\boldsymbol{X}^{(l)}) = 0$ implies that $\Delta_k = \Delta_l = \Delta_{kl}$. Then we have $\rho_g(\boldsymbol{X}, Y; \alpha) = 0$. Hence, if $\rho_g(\boldsymbol{X}, Y; \alpha) > 0$, there is at least one nonzero σ_{kl}^2. Then by Slustky's theorem,

$$\sqrt{n}(\hat{\rho}_g(\alpha) - \rho_g(\boldsymbol{X}, Y; \alpha)) \to \mathcal{N}(0, \sigma_g^2(\alpha)),$$

where $\sigma_g^2(\alpha) = 16 \sum_{k \neq l}^{K} p_l^2 p_k \sigma_{kl}^2(\alpha)/\Delta^2(\alpha)$. □

Although we have a formula of $\sigma_g^2(\alpha)$, it is rarely known in practice because it depends on unknown F_k and p_k. To overcome this difficulty, we can estimate $\sigma_g^2(\alpha)$ by the jackknife method. Let $\hat{\rho}_{(-i)}(\alpha)$ be the Gini correlation estimator based on the sample with the i^{th} observation deleted. The jackknife estimator of standard error $\sigma_g(\alpha)/\sqrt{n}$ is

$$\hat{SE}(\alpha) = \sqrt{\frac{n-1}{n} \sum_{i=1}^{n} (\hat{\rho}_{(-i)}(\alpha) - \bar{\hat{\rho}}_{(\cdot)}(\alpha))^2}, \tag{22}$$

where $\bar{\hat{\rho}}_{(\cdot)}(\alpha) = 1/n \sum_{i=1}^{n} \hat{\rho}_{(-i)}(\alpha)$. See Shao & Tu (1996) for details. Then a $(1-\gamma)100\%$ confidence interval of $\rho_g(\alpha)$ is

$$\hat{\rho}_g(\alpha) \pm z_{\gamma/2} \hat{SE}(\alpha),$$

where $z_{\gamma/2}$ is the $(1-\gamma/2)100\%$ quantile of the standard normal variable.

Theorem 3 states the asymptotic normality of $\hat{\rho}_g(\alpha)$ when $\boldsymbol{X}$ and Y are dependent. However, if $\rho_g(\alpha) = 0$ when $\boldsymbol{X}$ and Y are independent, the behavior of $\hat{\rho}_g(\alpha)$ is quite different since $\sigma_g^2(\alpha) = 0$. In this degenerate case, the limiting distribution of $n(\hat{\Delta}(\alpha) - \sum_k \hat{p}_k \hat{\Delta}_k(\alpha))$ converges in distribution to a quadratic form of Gaussian random variables, similar to the results in Székely & Rizzo (2005, 2013) and Rizzo & Székely (2010). They have proved for balanced cases that S_α, the between sample dispersion by the DISCO decomposition, converges in distribution to a quadratic form of centered Gaussian random variables. We state the following theorem and provide a proof that does not require balanced sizes.

Theorem 4 *If $\rho_g(\boldsymbol{X}, Y; \alpha) = 0$, $\mathbb{E}\|\boldsymbol{X}\|^{2\alpha} < \infty$ and $p_k > 0$ for $k = 1, ..., K$, then*

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

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

Proof of Theorem 4. Under $\rho_g(\boldsymbol{X}, Y; \alpha) = 0$, $F_1 = F_2 = ... = F_K = F$ implies all $\tilde{h}_{kl}^{10}(\boldsymbol{x}^{(k)})$'s in (21)

are zero almost surely. In this degenerate case, we need the second order projections of $\tilde{h}_{kl}$.

$$\begin{aligned}
\tilde{h}^{20}_{kl}(\boldsymbol{x}^{(k)}_1, \boldsymbol{x}^{(k)}_2) &= \mathbb{E}\tilde{h}_{kl}(\boldsymbol{x}^{(k)}_1, \boldsymbol{x}^{(k)}_2; \boldsymbol{X}^{(l)}_1, \boldsymbol{X}^{(l)}_2) \\
&= \frac{1}{2}[\mathbb{E}\|\boldsymbol{x}^{(k)}_1 - \boldsymbol{X}^{(l)}_1\|^\alpha + \mathbb{E}\|\boldsymbol{x}^{(k)}_2 - \boldsymbol{X}^{(l)}_2\|^\alpha - \|\boldsymbol{x}^{(k)}_1 - \boldsymbol{x}^{(k)}_2\|^\alpha - 2\Delta_{kl}(\alpha) + \Delta_k(\alpha)] \\
\tilde{h}^{02}_{kl}(\boldsymbol{x}^{(l)}_1, \boldsymbol{x}^{(l)}_2) &= \mathbb{E}\tilde{h}_{kl}(\boldsymbol{X}^{(k)}_1, \boldsymbol{X}^{(k)}_2; \boldsymbol{x}^{(l)}_1, \boldsymbol{x}^{(l)}_2) \\
&= \frac{1}{2}[\mathbb{E}\|\boldsymbol{X}^{(k)}_1 - \boldsymbol{x}^{(l)}_1\|^\alpha + \mathbb{E}\|\boldsymbol{X}^{(k)}_2 - \boldsymbol{x}^{(l)}_2\|^\alpha - \|\boldsymbol{x}^{(l)}_1 - \boldsymbol{x}^{(l)}_2\|^\alpha - 2\Delta_{kl}(\alpha) + \Delta_l(\alpha)], \\
\tilde{h}^{11}_{kl}(\boldsymbol{x}^{(k)}_1, \boldsymbol{x}^{(l)}_1) &= \mathbb{E}\tilde{h}_{kl}(\boldsymbol{x}^{(k)}_1, \boldsymbol{X}^{(k)}_2; \boldsymbol{x}^{(l)}_1, \boldsymbol{X}^{(l)}_2) \\
&= \frac{1}{2}[2\|\boldsymbol{x}^{(k)}_1 - \boldsymbol{x}^{(l)}_1\|^\alpha - \mathbb{E}\|\boldsymbol{x}^{(k)}_1 - \boldsymbol{X}^{(k)}_2\|^\alpha - \mathbb{E}\|\boldsymbol{x}^{(l)}_1 - \boldsymbol{X}^{(l)}_2\|^\alpha - 2\Delta_{kl}(\alpha) + \Delta_k(\alpha) + \Delta_l(\alpha)].
\end{aligned}$$

If $\rho_g(\boldsymbol{X}, Y; \alpha) = 0$, we have $\tilde{h}^{20}_{kl} = \tilde{h}^{02}_{kl} = \tilde{h}^{11}_{kl}$ for all $k \neq l$ and denote them as h_2. Let $h_2(\boldsymbol{x}_1, \boldsymbol{x}_2) = \sum_{s=1}^{\infty} \lambda_s \phi_s(\boldsymbol{x}_1)\phi_s(\boldsymbol{x}_2)$, where

$$\int_{\mathbb{R}^d} h_2(\boldsymbol{x}_1, x_2)\phi_s(\boldsymbol{x}_2)dF(\boldsymbol{x}_2) = \lambda_s\phi_s(\boldsymbol{x}_1).$$

Under the assumption of $\mathbb{E}\|\boldsymbol{X}\|^{2\alpha} < \infty$, we have $\sum_{s=1}^{\infty} \lambda_s < \infty$.

Denote the second order projection of $\hat{T}_{kl}$ as $\hat{T}^{(2)}_{kl}$ that is given by

$$\hat{T}^{(2)}_{kl} = \frac{2}{n_k^2} \sum_{1 \leq i,j \leq n_k} h_2(\boldsymbol{x}^{(k)}_i, \boldsymbol{x}^{(k)}_j) + \frac{2}{n_l^2} \sum_{1 \leq i,j \leq n_l} h_2(\boldsymbol{x}^{(l)}_i, \boldsymbol{x}^{(l)}_j) + \frac{4}{n_k n_l} \sum_{i=1}^{n_k} \sum_{j=1}^{n_l} h_2(\boldsymbol{x}^{(k)}_i, \boldsymbol{x}^{(l)}_j).$$

We obtain the second order projection of the estimator to be

$$\sum_{k \neq l} \hat{p}_k \hat{p}_l \hat{T}^{(2)}_{kl} = \frac{1}{n}[\sum_{k=1}^{K} 4(1 - p_k)\frac{1}{n_k} \sum_{1 \leq i,j \leq n_k} h_2(\boldsymbol{x}^{(k)}_i, \boldsymbol{x}^{(k)}_j) + \sum_{k \neq l} 4 \sum_{i=1}^{n_k} \sum_{j=1}^{n_l} h_2(\boldsymbol{x}^{(k)}_i, \boldsymbol{x}^{(l)}_j)].$$

By the V-statistic theorem (Theorem B of Section 6.4, Serfling,1980), we have

$$n(\hat{\Delta} - \sum_k \hat{p}_k \hat{\Delta}_k) \xrightarrow{d} \sum_{s=1}^{\infty} \sum_{k=1}^{K} 4(1 - p_k)\lambda_s Z^2_{s,k} + \sum_{s=1}^{\infty} \sum_{1 \leq k < l \leq K} 4\sqrt{p_k p_l}\lambda_s Z_{s,k} Z_{s,l},$$

where $Z_{s,k} (k = 1, ..., K, s = 1, 2, ...)$ are independent standard normal variates. An immediate application of Slustky's Theorem completes the proof. □

4.2 Testing Dependence

The Gini correlation is zero if and only $\boldsymbol{X}$ and Y are independent. Hence, for a given $0 < \alpha < 2$, the independence test can be stated as

$$\mathcal{H}_0 : \rho_g(\boldsymbol{X}, Y; \alpha) = 0, \quad \text{vs} \quad \mathcal{H}_1 : \rho_g(\boldsymbol{X}, Y; \alpha) = \rho_0 > 0. \tag{23}$$

Note that the null hypothesis of the test in (23) is equivalent to the null of the K-sample test

$$\mathcal{H}_0' : F_1 = F_2 = ... = F_K = F. \tag{24}$$

Reject $\mathcal{H}_0$ or $\mathcal{H}_0'$ if $\hat{\rho}_g$ is large. The critical value of the test of significance level γ, however, is difficult to obtain from Theorem 4 by two reasons. Firstly λ_l's depend on unknown distribution F. Secondly, it is a mixture of infinitely many distributions. To overcome this difficulty, a permutation procedure is used to estimate the critical value and p-value.

Let $\nu = 1 : n$ be the vector of original sample indices of the sample for Y labels and $\hat{\rho}_g(\alpha) = \hat{\rho}(\nu; \alpha)$. Let $\pi(\nu)$ denote a permutation of the elements of ν and the corresponding $\hat{\rho}_g(\pi; \alpha)$ is computed. Under the $\mathcal{H}_0$, $\hat{\rho}_g(\nu)$ and $\hat{\rho}_g(\pi; \alpha)$ are identically distributed for every permutation π of ν. Hence, based on M permutations, the critical value q_γ is estimated by the $(1-\gamma)100\%$ sample quantile of $\hat{\rho}_g(\pi_m; \alpha)$, $m = 1, ..., M$ and the p-value is estimated by the proportion of $\hat{\rho}_g(\pi_m; \alpha)$ greater than $\hat{\rho}_g(\nu; \alpha)$. Usually $100 \leq M \leq 1000$ is sufficient for a good estimation on the critical value or p-value. In the simulation next section, we use $M = 200$. Further, if ρ_0 is specified, the power of the test can be estimated by

$$\text{power} = 1 - \Phi\left(\frac{\hat{q}_\gamma - \rho_0}{\hat{\sigma}_g(\alpha)/\sqrt{n}}\right),$$

where $\Phi(x)$ is the cdf of the standard normal random variable.

4.3 Computation issues

The computation of the sample Gini correlation $\hat{\rho}_g(\alpha)$ in (20) is straightforward. In general, it has a computational complexity $O(n^2)$ since all distinct pair distances need to calculate. However, when the numerical variable is univariate and $\alpha = 1$, we have a much faster algorithm that only costs $O(n \log n)$ computation. This is because the univariate Gini mean distance can be written as a linear combination of order statistics (Schechtman & Yitzhaki, 1987). Suppose that $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$ are the order statistics of $x_1, x_2, ..., x_n$. Then

$$\hat{\Delta}(X; 1) = \binom{n}{2}^{-1} \sum_{1 \leq i < j \leq n} |x_i - x_j| = \binom{n}{2}^{-1} \sum_{i=1}^{n} (2i - n - 1) x_{(i)}.$$

This fast algorithm is crucial for Gini correlation in application of feature screening. For the classification problem with ultrahigh-dimensional data, the first step is to screen out unimportant predictors. We can rank features by their Gini correlations with the class label and screen out less correlated predictors, analogue to the sure independent screening procedures (Fan & Lv, 2008; Li *et al.*, 2012) in which they consider a numerical response variable and use Pearson correlation or distance correlation to do feature selection.

Note that for the sample counterpart of the kernelized Gini correlation of (15), it does not exist a fast algorithm for the univariate case. All pairwise distances need to compute for $\hat{\rho}_\kappa$.

For sample distance correlation $\hat{\rho}_d(\alpha)$, its computation follows as the average of the element-wise product of two centered pairwise distance matrices, which is described in Szekely & Rizzo (2004, 2007). With small adjustments in centering, an unbiased estimator is provided in Székely & Rizzo (2004). More specifically, let $A = (a_{ij})$ be a symmetric, $n \times n$, centered distance matrix of sample $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n$. The (i, j)-th entry of A is

$$A_{ij} = \begin{cases} a_{ij} - \frac{1}{n-2} a_{i\cdot} - \frac{1}{n-2} a_{\cdot j} + \frac{1}{(n-1)(n-2)} a_{\cdot\cdot}, & i \neq j; \\ 0, & i = j, \end{cases}$$

where $a_{ij} = \|\boldsymbol{x}_i - \boldsymbol{x}_j\|^\alpha$, $a_{i\cdot} = \sum_{j=1}^n a_{ij}$, $a_{\cdot j} = \sum_{i=1}^n a_{ij}$, and $a_{\cdot\cdot} = \sum_{i,j=1}^n a_{ij}$. Similarly, using the set difference metric, a symmetric, $n \times n$, centered distance matrix is calculated for samples $y_1, \cdots, y_n$ and denoted by $B = (b_{ij})$. Unbiased estimators of $dCov(\boldsymbol{X}, Y; \alpha)$, $dCov(\boldsymbol{X}, \boldsymbol{X}; \alpha)$ and $dCov(Y, Y; \alpha)$are given respectively as,

$$\frac{1}{n(n-3)} \sum_{i \neq j} A_{ij} B_{ij}, \qquad \frac{1}{n(n-3)} \sum_{i \neq j} A_{ij}^2, \qquad \frac{1}{n(n-3)} \sum_{i \neq j} B_{ij}^2.$$

Note that for univariate X, a fast $O(n \log n)$ algorithm for sample distance correlation is available (Huo & Székely, 2016) and is implemented in "distcor" from "dcortools" package on GitHub. We compare its computation time with those of the Gini correlation and the kernel Gini correlation implemented in "RcppgCor" and "RcppKgCor" from R package "GiniDistance". We simulate standard normal random samples in $\mathbb{R}^d (d = 1, d = 100)$ of different sizes. Half of sample points are randomly assigned to Class 1 and the other half forms Class 2. The process is repeated 30 times. The code is run on a MacBook Pro with I7 2.8 Ghz CPU. The mean and standard deviation of the computation time for the Gini correlation, distance correlation and kernel Gini correlation are recorded in Table 1.

[Put Table 1 here]

From Table 1, it is clear to see that the kernel Gini correlation $\hat{\rho}_\kappa$ has no fast algorithm in $d = 1$. For $n = 100,000$, $\hat{\rho}_\kappa$ takes 85 seconds to be computed. Although they have a same computation complexity, $\hat{\rho}_g$ is about 5 times faster than $\hat{\rho}_d$ in $d = 1$ and about 3 times faster in $d = 100$. The slower computation speed in the distance correlation can be interpreted by its implementation in dealing with the centering process.

Another computation issue is the choice of α, the parameter of distance metric in $\mathbb{R}^d$. A natural choice is $\alpha = 1$, which corresponds to the Euclidean distance and leads to fast algorithms for the univariate case. However, if outliers appear in data, we want to choose a small α value so that the Gini and distance correlations are insensitive to the outliers, as mentioned in Remark 1. Another possible

way is to choose the α value to maximize the correlations. The idea is similar to the approach in the maximal correlation (Sarmanov, 1958; Renyi, 1959). They choose the transformation of the data to achieve the largest association. One may select the metric on the original data so that the correlation is the greatest. It is worthwhile to continue the research in this directions in the future. In the next section, we use $\alpha = 1$ in the first two simulation studies and in the real data application. And $\alpha = 0.5, 0.75, 1$ are used in the last simulation for demonstration of different α values for a heavy-tailed distribution.

5 Experiments

5.1 Simulation

Three simulations are conducted to demonstrate the performance of Gini correlation. The first one is to check the coverage probabilities of the confidence intervals based on the asymptotic normality with the asymptotic standard error estimated by the Jackknife method of (22). The second simulation is to compare dependence tests based on Gini correlation and the distance correlation, and the last simulation is to illustrate that a small α value is more proper for data from heavy-tailed distributions.

For the first simulation on confidence intervals, we consider examples studied in the previous section. The coverage probabilities of confidence intervals in Table 2 are computed based on 10000 repetitions. Two confidence levels 0.90 and 0.95 are considered under sample sizes of $n = 60$ and $n = 120$. Comparison with confidence intervals of ρ_d is only available for Example 1 where random samples are generated from the mixture of two exponential distributions because the true values of ρ_d are unknown in the other two cases. From Table 2, we observe that the coverage probabilities of confidence intervals of ρ_d are unacceptable. One possible reason is the double-centering procedure in the computation of the sample distance correlation, which makes its finite sample performance undesirable. The coverage probabilities of confidence intervals for ρ_g are satisfying. They are reasonably close to the nominal levels even under small sample size of $n = 60$.

[Put Table 2 here]

For the second simulation on dependence test, the following scenarios with balanced $\boldsymbol{p} = (p_1, p_2, p_3) = (1/3, 1/3, 1/3)$, lightly unbalanced $\boldsymbol{p} = (5/12, 4/12, 3/12)$ and heavily unbalanced $\boldsymbol{p} = (0.6, 0.3, 0.1)$ of the total sample sizes of $(n = 60, n = 120)$ are considered.

- $X \sim p_1 \exp(1) + p_2 \exp(\theta_1) + p_3 \exp(\theta_2)$;
- $X \sim p_1 \mathcal{N}(0, 1) + p_2 \mathcal{N}(\mu_1, \sigma_1^2) + p_3 \mathcal{N}(\mu_2, \sigma_2^2)$;

The size and power of each test based on 1000 repetitions are reported in Table 3. The cases of $\theta_1 = \theta_2 = 1$, $\mu_1 = \mu_2 = 0$ and $\sigma_1 = \sigma_2 = 1$ imply independence of X and Y. Two tests maintain the test level 0.05 well. In the balanced cases ($\boldsymbol{p} = (1/3, 1/3, 1/3)$), two permutation tests perform

similarly. The powers of the ρ_g method are close to the powers of the ρ_d test. Indeed, two permutation tests are asymptotically equivalent for the balanced scenarios since the population Gini correlation is a multiple of the population distance correlation (Remark 8). When the weights are unequal, the Gini correlation method performs better than the distance correlation. The power gain of the ρ_g test over the ρ_d test in the lightly unbalanced cases ($\boldsymbol{p} = (5/12, 4/12, 3/12)$) is about 4-5% for the exponential mixtures and 3-6% for the normal mixtures. For the heavily unbalance cases ($\boldsymbol{p} = (0.6, 0.3, 0.1)$), the Gini method performs much better than the distance method, with the power gain reaching as large as 20%. This phenomenon can be explained from the formulation of two correlations in Remark 7. Both Gini and distance covariance are weighted energy distances, however, due to the weight p_k^2 in the distance covariance, the contribution portion from the third class with $p_3 = 0.1$ will be reduced largely, resulting in less power for the distance test.

[Put Table 3 here]

Last, to illustrate robustness of the method with a small α value, we generate random variables from balanced 2-class Cauchy mixtures with Class 1 centered at 0 and Class 2 centered at δ changing from 0 to 1 . We take three values of $\alpha = 0.5, 0.75, 1$. Table 4 reports the level and power of the permutation tests based on each α value. The sizes of three tests are similar to each other. As expected, $\alpha = 1$ performs inferior to the other two since the Cauchy distribution has no first moment. The test with $\alpha = 0.5$ yields the highest power among the three.

[Put Table 4 here]

5.2 Real Data Application

Four data sets are studied for dependence of categorical variable and numerical variables. The first data set is the famous Iris data set with the measurements in centimeters on sepal length and width and petal length and width, for 50 flowers from each of 3 species of iris. Table 5 lists the Gini and distance correlations between Species and each of measurements, also between Species and all measurements. The standard deviations in parentheses are computed through the Jackknife procedure. Note that values in each column in Table 5 are not comparable because they estimate different quantities. Across each row, we can conclude that iris species have higher correlation with petal size than sepal size. Species account for 77.3% of the total variation of petal length and 75.3% of total variation of petal width. Overall, 62.4% of variation on the all measurements can be explained by species. Gini correlation estimators have smaller standard errors than the distance correlation. In other words, Gini correlation estimators are more statistically efficient. Consequently, they lead to shorter confidence intervals than the distance correlation estimators do.

[Put Table 5 here]

The second data analysis is to illustrate a simple application on multi-factors. We use the R built-in data set ToothGrowth. It contains data from a study evaluating the effect of vitamin C on tooth growth in Guinea pigs. The experiment has been performed on 60 pigs, where each animal received one of three dose levels of vitamin C (0.5, 1, and 2 mg/day) by one of two delivery methods, orange juice or ascorbic acid. The overall Gini variation of tooth length is 8.839, in which dose accounts for 45.1%, supplement type accounts for only 3.2%, and dose and supplement type jointly account for 5.4%.

The third data is LSVT Voice Rehabilitation dataset from UCI machine learning repository [10]. After speech rehabilitation treatments in Parkinson's disease, phonations of 126 patients were evaluated as 'acceptable' or 'unacceptable' based on 309 attributes. Refer to Tsanas *et al.* (2014) for details of the data set and dysphonia measure attributes. This data set has the dimension larger than the sample size. Our goal is to demonstrate a simple feature selection based on the highest correlated variables with the response class so that the selected subset of attributes is able to effectively predict a phonation as 'acceptable' or 'unacceptable'. We evaluate the selection method by its performance of the selected feature set in classification. We use Random Forest as the classifier due to its simplicity, popularity and effectiveness. Once d features are selected, the random forest classifier (R package randomForest) with default parameters is applied and out of bag (oob) miss-classfication error is used as the performance criterion. As a benchmark, the oob error of the random forest classifier that uses all 309 predictors has a median of 0.167. The boxplots of oob errors based on 50 repetitions of random forest classifiers on each of d values are provided in Figure 4.

[Put Figure 4 here]

The top one, two and three features selected by the Gini correlation coincide with those by the distance correlation. The feature 60 is the fourth highest correlated variable in terms of Gini correlation, while the feature 151 is the 4^{th} ranked according to the distance correlation. Based on the top 4 features selected by Gini correlation, oob errors has a median of 0.103, significantly better than the median of 0.127 for the distance method and the Pearson R^2 method. It is worthwhile to mention that the variation of oob errors for the model selected by Gini method is extremely small with a standard deviation of 0.003 and a median absolute deviation of 0, indicating stability of the selected model. The attribute 80 ranks the fifth in three methods. However, its inclusion degrades performance. The error medians increase to 0.111, 0.143 and 0.143 respectively in three models. The model with top 6 correlated features selected by Gini and the distance methods is identical, and hence skipped in the boxplot. The differences of the top 7 and 8 features are the attribute 82 in the Gini method and the attribute 154 in the distance method. The Gini method yields better performance. However, when considering the top 9 and the top 10 features, the distance and Pearson R^2 methods are better since they select the feature 155. For $d = 13$, the median error rate of the Gini selected model is about 0.8% smaller than the one based on the other two methods. For $d = 50$, Gini and distance methods produce a model with a similar performance

in classification. In general, the distance correlation and Pearson R^2 selection methods perform similarly and Gini feature selection performs better than the other two. The model with 4 features selected by Gini correlation has the best performance.

The fourth dataset is the TCGA breast cancer microarray dataset from the UCSC Xena database (Goldman *et al.*, 2018). This data contains expression levels of 17278 genes from 506 patients and each patient has a breast cancer subtype label (luminal A, luminal B, HER2-enriched, or basal-like). PAM50 is a gene signature consisting of 50 genes derived from microarray data and is considered as the gold-standard for breast cancer subtype prognosis and prediction (Parker*et al.*, 2009). We compare each method by comparing the top d genes with PAM50 gene signature. Also we compare classification performance of each selected model. We randomly hold-out 20% as test data, select top d genes by each method and build random forest classifier based on the remaining 80% data, then evaluate performance of the selected model by the test classification accuracy. We repeat the random forest classification 30 times.

[Put Table 6 here]

Table 6 shows the classification performance using selected top d genes by different correlations. $\hat{\rho}_g$ and $\hat{\rho}_\kappa$ perform similar. Among the four selection methods under comparison, they have the best overall performance, even outperforms PAM50 with $d = 40$ and $d = 50$. This suggests that they are able to select a smaller number of genes and the predictions are better than the gold standard. Although $\hat{\rho}_d$ perform the best for $d = 20$, but it is not able to exceed PAM50 within 50 genes, neither do the Pearson R^2.

[Put Figure 5 here]

Figure 5 shows the number of PAM50 genes appear in the top d selected genes for each selection method. It is obvious to see that Pearson R^2 selects the smallest number of PAM50. Among the top 1000 genes selected by Pearson R^2, only 40% of the PAM50 genes are included, while 80% overlap the 1000 genes selected by the kernel Gini correlation. Gini correlation and kernel Gini correlation are able to select more PAM50 genes than the distance correlation as d increases. The small ratio of PAM50 included in the selected genes by any method is because of the high correlation among genes. PAM50 is derived by not only selecting most subtype dependent genes, but also less mutually dependent genes to obtain a smaller set of genes for the same prediction accuracy. Even though any of the selection methods under comparison does not take the feature-feature dependence into consideration, $\hat{\rho}_g$ and $\hat{\rho}_\kappa$ are able to select a better gene set than PAM50 for classification.

6 Conclusion and Future Work

Dependence between categorical and numerical variables is an importance topic, which is relevant in practice, but generally receives little attention. In this paper, we have proposed the Gini correlation, which takes advantages of the nature of the categorical variable and hence has a simpler formulation than the celebrated distance correlation. As a result, the sample Gini correlation is more computationally and statistically efficient than the sample distance correlation. As shown in the experiment, the standard deviations of Gini correlation estimators are usually smaller than those of distance correlation estimators. Like Pearson R^2 correlation, Gini correlation has a nice interpretation as the ratio of the between variation and the total variation. Unlike Pearson R^2, Gini correlation characterizes independence. Testing zero Gini correlation or distance correlation is equivalent to testing the distribution equality of K-samples. In the balanced case, the permutation test based on the Gini correlation is asymptotically equivalent to the distance test. However, the Gini test is more powerful than the distance test for the unbalance cases, which are more common in real applications.

Although the proposed Gini correlation has advantages over others, it has some limitations. It is only orthogonal invariant but not affine invariant in general. One way to make it affine invariant is to consider the standardized samples $\boldsymbol{z}_i$ defined by $\boldsymbol{z}_i = S^{-1/2}\boldsymbol{x}_i$, where S is the sample covariance matrix of $\mathcal{D} = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_n\}$. Then an affine Gini correlation estimator can be defined as

$$\hat{\rho}_G(\boldsymbol{X}, Y; \alpha) = \hat{\rho}_g(\boldsymbol{Z}, Y; \alpha) = \frac{\hat{\Delta}(\boldsymbol{Z}; \alpha) - \sum_k \hat{p}_k \hat{\Delta}_k(\boldsymbol{Z}, \alpha)}{\hat{\Delta}(\boldsymbol{Z}; \alpha)}.$$

For the purpose of robustness, S can be chosen to be some robust shape matrix estimators such as M-estimator and S-estimator (Shevlyakov & Oja, 2016).

Affine invariance property preserves the equivalance of statistical inference under linear transformations of data. A more desired property for a dependence measure is invariant under monotone transformations (Renyi, 1959). We would like to have a dependence Gini measure such that

$$\rho_g(\boldsymbol{X}, Y) = \rho_g(\boldsymbol{h}(\boldsymbol{X}), Y),$$

where $\boldsymbol{h}$ is a one-to-one function. If X is one-dimensional, one option shall be rank-based Gini correlation. Rather than using values of X, its ranks are used in calculation of Gini correlation. The rank-based approach preserves the monotonicity relationships and is robust against outliers. It turns out that the rank-based Gini correlation is 6 times of the mean variance index (MV) proposed by Cui *et al.* (2015) and Cui & Zhang (2019). That is,

$$\rho_g(R(X), Y) = \frac{\sum_{k=1}^K p_k \int_{\mathbb{R}} (F_k(x) - F(x))^2 \, dF(x)}{\int_{\mathbb{R}} F(x) - F^2(x) \, dF(x)} = 6MV(X, Y)$$

The dependence test based on MV is distribution-free and hence no permutation procedure is required. However, it may lose too much statistical efficiency. Also extensions to the multivariate case is not direct. Continuities of the research in this direction are worthwhile.

In the simulation, we empirically demonstrated that the ρ_g test performs better than the ρ_d test under heavily unbalanced weights. Such empirical evidences immediately call for a theoretical justification. How does the weight play a role in distinguishing gCov from dCov? What are the asymptotic distributions of gCov and dCov when n and K both go to infinity. Developing theory to answer these questions is one focus of research in the near future.

References

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

[2] Beknazaryan, A., Dang, X. and Sang, H. (2019). On mutual information estimation for mixed-pair random variables. *Statist. Probab. Lett.* **148**, 9-16.

[3] Cramér, H. (1946). *Mathematical Methods of Statistics*. Princeton Press, NJ.

[4] Cui, H., Li, R. and Zhong, W. (2015). Model-free feature screening for ultrahigh dimensional discriminant analysis, *J. Amer. Statist. Assoc.* **110**, 630-641.

[5] Cui, H. and Zhong, W. (2019). A distribution-free test of independence based on mean variance index, *Comput. Statist. Data Anal.* **139**, 117-133.

[6] Dang, X., Sang, H. and Weatherall, L. (2019). Gini covariance matrix and its affine equivariant version. *Statist. Papers* **60** (3), 291-316.

[7] David, H. A. (1968). Gini's mean difference rediscovered. *Biometrika* **55**, 573-575.

[8] Devlin, S.J., Gnanadesikan, R. and Kettering, J.R. (1975). Robust estimation and outlier detection with correlation coefficients. *Biometrika* **62**, 531-545.

[9] Dorfman, R. (1979). A formula for the Gini coefficient. *Review of Economics and Statistics*, **61**, 146-149.

[10] 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.

[11] Edelmann, D., Richards, D. and Vogel, D. (2017). The distance standard deviation. arXiv:1705.05777v1

[12] Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space (with discussion). *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **70**, 849-911.

[13] Frick, R., Goebel, J., Schechtman, E., Wagner, G. and Yitzhaki, S. (2006). Using analysis of Gini (ANOGI) for detecting whether two sub-samples represent the same universe: The German Socio-Economic Panel study (SOEP) experience. *Sociological Methods and Research*, **34**, 427-468.

[14] Gao, W., Kannan, S., Oh, S. and Viswanath, P. (2017). Estimating mutual information for discrete-continuous mixtures. In proceedings of 31^{th} *Conference on Neural Information Processing Systems (NIPS)*, Long Beach, CA.

[15] Gini, C. (1914). Sulla misura della concentrazione e della variabilità dei caratteri. *Atti del Reale Istituto Veneto di Scienze, Lettere ed Aeti* **62**, 1203-1248. English Translation: On the measurement of concentration and variability of characters (2005). *Metron* LXIII (1), 3-38.

[16] Goldman, M., Craft, B., Brooks, A.N., Zhu, J. and Haussler, D. (2018). The ucsc xena platform for cancer genomics data visualization and interpretation. *bioRxiv*.

[17] Hu, B., Shao, J. and Palta, M. (2006). Pseudo-R^2 logistic regression model. *Statist. Sinica* **16**, 847-860.

[18] Huo, X. and Székely, G. (2016). Fast computing for distance covariance. *Technometrics* **58** (4), 435-447.

[19] Kendall, M.G. and Gibbons, J.D. (1990). *Rank Correlation Methods, 5^{th} edn.* Griffin, London.

[20] Kendall, M.G. (1938). A new measure of rank correlation. *Biometrika* **60** (2), 185-196.

[21] Koshevoy, G. and Mosler, K. (1997). Multivariate Gini indices. *J. Multivariate Anal.* **60**, 252-276.

[22] Li, R., Zhong, W. and Zhu, L. (2012). Feature screening via distance correlation learning. *J. Amer. Statist. Assoc.* **107**, 1129-1139.

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

[24] Mari, D.D. and Kotz, S. (2001). *Correlation and dependence*. Imperial College Press, London.

[25] 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.

[26] Parker, J.S. et al. (2009). Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of Clinical Oncology* **27** (8), 1160-1167.

[27] Renyi, A. (1959). On measures of dependence. *Acta Math. Acad. Sci. Hung.* **10**, 441-451.

[28] Rizzo, M. L. and Székely, G. J. (2010). DISCO analysis: A nonparametric extension of analysis of variance. *Ann. Appl. Stat.* **4** (2), 1034-1055.

[29] Ross, B.C. (2014). Mutual information between discrete and continuous data sets. *PLoS ONE* **9** (2), e87357. doi:10.1371/journal.pone.0087357

[30] Sang, Y., Dang, X. and Sang, H. (2016). Symmetric Gini covariance and correlation. *Canad. J. Statist.* **44** (3), 323-342.

[31] Sarmanov, O.V. (1958). Maximum correlation coefficient (symmetric case). *Doklady Akad. Nauk SSSR* **120**, 715-718.

[32] Schechtman, E. and Yitzhaki, S. (1987). A measure of association based on Gini's mean difference. *Comm. Statist. Theory Methods* **16** (1), 207-231.

[33] Schechtman, E. and Yitzhaki, S. (2003). A Family of correlation coefficients based on the extended Gini index. *J. Econ. Inequality* **1**(2), 129-146.

[34] Serfling, R. (1980). *Approximation theorems of mathematical statistics*. Wiley, New York.

[35] Shao, J. and Tu, D. (1996). *The jackknife and bootstrap*. Springer, New York.

[36] Shevlyakov G.L. and Smirnov P.O. (2011). Robust estimation of the correlation coefficient: an attempt of survey. *Austrian J. Stat.* **40**, 147-156.

[37] Shevlyakov G.L. and Oja, H. (2016). *Robust correlation: theory and applications*. Wiley, Chichester, UK.

[38] Spearman, C. (1904). General intelligence objectively determined and measured. *Amer. J. Psychol.* **15**, 201-293.

[39] Székely, G.J. and Rizzo, M.L. (2004). Testing for equal distributions in high dimension. *InterStat* Nov. (5).

[40] Székely, G. J. and Rizzo, M. L. (2005). Hierarchical clustering via joint between-within distances: Extending Ward's minimum variance method, *J. Classification* **22** (2), 151-183.

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

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

[43] Székely, G. J. and Rizzo, M. L. (2013). Energy statistics: A class of statistics based on distances, *J. Statist. Plann. Inference* **143** (8), 1249-1272.

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

[45] Tsanas, A., Little, M.A., Fox, C. and Ramig, L.O. (2014). Objective automatic assessment of rehabilitative speech treatment in Parkinson's diseases, *IEEE Transactions on Neural Systems and Rehabilitation Engineering* **22**, 181-191.

[46] Tschuprow, A. (1939). *Principles of the Mathematical Theory of Correlation*, W. Hodge & Co.

[47] Tjur, T. (2009). Coefficients of determination in logistic regression models? A new proposal: the coefficient of discrimination, *Amer. Statist.* **63** (4), 366-372.

[48] Yitzhaki, S. and Schechtman, E. (2013). *The Gini Methodology*, Springer, New York.

[49] Zhang, S., Dang, X., Nguyen, D., Wilkins, D. and Chen, Y. (2019). Estimating feature - label dependence using Gini distance statistics, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, accepted. 10.1109/TPAMI.2019.2960358.

Corresponding Author:
XIN DANG
Department of Mathematics, University of Mississippi
315 Hume Hall, University, MS, 38677
Email: xdang@olemiss.edu

Appendix

Proof of Remark 3. For $d = 1$ and $\alpha = 1$, we would like to show that

$$T(X_k, X; 1) = \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{|\psi_k(t) - \psi(t)|^2}{t^2} dt = 2\int_{-\infty}^{\infty} (F_k(x) - F(x))^2 dx. \tag{25}$$

This is true because of the Parseval-Plancherel identity. Let the Fourier transforms of $F_k(x)$ and $F(x)$ are $\Psi_k(t)$ and $\Psi(t)$, respectively. By the Parseval-Plancheral formula,

$$2\pi\int_{-\infty}^{\infty} (F_k(x) - F(x))^2 dx = \int_{-\infty}^{\infty} |\Psi_k(t) - \Psi(t)|^2 dt. \tag{26}$$

Also taking derivatives of $\Psi(t)$ and $\Psi_k(t)$ with respect to x, we obtain that $\Psi(t) = -\psi(t)/(it)$ and $\Psi_k(t) = -\psi_k(t)/(it)$. Plugging them to (26) proves (25).

For $d > 1$, although the Parseval-Plancheral identity still holds in $\mathbb{R}^d$, but there is no implicit relationship betweeen Ψ and ψ. Hence, $T(\boldsymbol{X}_k, \boldsymbol{X}; 1)$ has no interpretation as the L_2 distance between F_k and F. □

Proof of Remark 4. It is sufficient to prove $\sum_{k=1}^{K} p_k|\psi_k - \psi|^2 = \sum_{1\le k\le l\le K} p_k p_l|\psi_k - \psi_l|^2$. With the fact that $\psi = \sum_{l=1}^{K} p_l\psi_l$, we have

$$\begin{aligned}
&\sum_{k=1}^{K} p_k|\psi_k - \psi|^2 = \sum_{k=1}^{K} p_k|\psi_k - \sum_{l=1}^{K} p_l\psi_l|^2 = \sum_{k=1}^{K} p_k|\sum_{l=1}^{K} p_l(\psi_k - \psi_l)|^2 \\
&= \sum_{k=1}^{K} p_k \left[\sum_{l=1}^{K} p_l^2|\psi_k - \psi_l|^2 + \sum_{l\neq m} p_l p_m(\psi_k - \psi_l)\overline{(\psi_k - \psi_m)}\right] \\
&= \sum_{1\neq k\le l\neq K} \left(p_k p_l|\psi_k - \psi_l|^2(\sum_{m=1}^{K} p_m)\right)
\end{aligned}$$

The last equation is due to combining the terms $p_k p_l p_m(\psi_k - \psi_l)\overline{(\psi_k - \psi_m)}$ and $p_k p_l p_m(\psi_k - \psi_l)\overline{(\psi_m - \psi_l)}$ together. Since $\sum p_m = 1$, the remark is proved. □

Proof of Remark 6. On one hand,

$$\begin{aligned}
&\mathbb{E}|\boldsymbol{X} - \boldsymbol{X}'|^\alpha|Y - Y'|^\alpha + \mathbb{E}|\boldsymbol{X} - \boldsymbol{X}'|^\alpha\mathbb{E}|Y - Y'|^\alpha - 2\mathbb{E}|\boldsymbol{X} - \boldsymbol{X}'|^\alpha|Y - Y''|^\alpha \\
&= \sum_{k\neq l} p_k p_l\Delta_{kl}(\alpha) + \Delta(\alpha)(1 - \sum_j p_j^2) - 2\sum_k p_k(1 - p_k)\mathbb{E}|\boldsymbol{X} - \boldsymbol{X}_k|^\alpha \\
&= \sum_{k\neq l} p_k p_l\Delta_{kl}(\alpha) + \sum_{k\neq l} p_k p_l\Delta_{kl}(\alpha)(1 - \sum_j p_j^2) + \sum_k p_k^2\Delta_k(\alpha)((1 - \sum_j p_j^2) \\
&- 2\sum_k p_k(1 - p_k)(p_k\Delta_k(\alpha) + \sum_{k\neq l} p_l\Delta_{kl}(\alpha)) \\
&= \sum_{k\neq l} p_k p_l(2p_k - \sum p_j^2)\Delta_{kl}(\alpha) + \sum_k p_k^2(2p_k - 1 - \sum_j p_j^2)\Delta_k(\alpha).
\end{aligned}$$

On the other hand, by (14)

$$\begin{aligned}
dCov(\boldsymbol{X}, Y; \alpha) &= \sum_k p_k^2 T(\boldsymbol{X}_k, \boldsymbol{X}; \alpha) \\
&= \sum_k p_k^2(2\mathbb{E}|\boldsymbol{X}_k - \boldsymbol{X}|^\alpha - \mathbb{E}|\boldsymbol{X}_k - \boldsymbol{X}_k'|^\alpha - \mathbb{E}|\boldsymbol{X} - \boldsymbol{X}'|^\alpha) \\
&= \sum_k p_k^2[2(\sum_{l\neq k} p_l\Delta_{kl}(\alpha) + p_k\Delta_k(\alpha))] - \sum_k p_k^2\Delta_k(\alpha) - (\sum_k p_k^2)\Delta(\alpha) \\
&= \sum_{k\neq l}(2p_k - \sum_j p_j^2)\Delta_{kl}(\alpha) + \sum_k p_k^2(2p_k - 1 - \sum_j p_j^2)\Delta_k(\alpha).
\end{aligned}$$

The result of Remark 6 is proved. □

Proof of Equation (17). We have

$$\begin{aligned} dCov(Y,Y) &= \mathbb{E}|Y-Y'|^2 + (\mathbb{E}|Y-Y'|)^2 - 2\mathbb{E}|Y-Y'||Y-Y''| \\ &= \sum_k p_k(1-p_k) + (\sum_k p_k(1-p_k))^2 - 2\sum_k p_k(1-p_k)^2 \\ &= 1 - \sum_k p_k^2 + (1-\sum_k p_k^2)^2 - 2 + 4\sum_k p_k^2 - 2\sum_k p_k^3 \\ &= \sum p_k^2 + (\sum_k p_k^2)^2 - 2\sum_k p_k^3. \end{aligned}$$

□

Proof of Remark 9. For $K = 2$ with $p_1 + p_2 = 1$, we have

$$dCov(Y,Y) = p_1^2 + p_2^2 + (p_1^2+p_2^2)^2 - 2(p_1^3+p_2^3) = 4p_1^2p_2^2;$$
$$p_1^2|\psi_1-\psi|^2 + p_2^2|\psi_2-\psi|^2 = p_1^2|\psi_1 - p_1\psi_1 - p_2\psi_2|^2 + p_2^2|\psi_2 - p_1\psi_1 - p_2\psi_2|^2 = 2p_1^2p_2^2|\psi_1-\psi_2|^2;$$
$$p_1|\psi_1-\psi|^2 + p_2|\psi_2-\psi|^2 = p_1p_2^2|\psi_1-\psi_2|^2 + p_1^2p_2|\psi_1-\psi_2|^2 = p_1p_2|\psi_1-\psi_2|^2.$$

Together with the definitions of $gCov(\boldsymbol{X}, Y; \alpha)$ and $dCov(\boldsymbol{X}, Y; \alpha)$, Remark 9 is proved. □

Detailed Derivations for Examples. We use the Theorem 3.2 result of Edelmann *et al.* (2017), which states that

$$dCov(X,X) = 8\int\int_{-\infty<x<z<\infty} F^2(x)(1-F(z))^2 dzdx. \tag{27}$$

Due to the difficulty to evaluate (27), we only provide the distance correlation formula in Example 1 where $X \sim pExp(\theta) + (1-p)Exp(\beta)$.

Example 1. $X_1 \sim Exp(\theta)$ and $X_2 \sim Exp(\beta)$ are independent. Then,

$$\Delta_{12} = \mathbb{E}|X_1 - X_2| = \int_0^\infty \int_0^\infty |x_1 - x_2| \frac{1}{\theta} e^{-x_1/\theta} \frac{1}{\beta} e^{-x_2/\beta} dx_1 dx_2 = \frac{\theta^2+\beta^2}{\theta+\beta}.$$

Plugging $F(x) = p(1-e^{-x/\theta}) + (1-p)(1-e^{-x/\beta}) = 1 - pe^{-x/\theta} - (1-p)e^{-x/\beta}$ in (27) and following a tedious evaluation of the integral, we have the result of $dCov(X,X)$.

To prove that the squared Pearson correlation and Gini correlation are increasing with $r > 1$, we obtain their derivatives with respect to r as follows.

$$\frac{\partial \rho_g}{\partial r} = \frac{p(1-p)(r-1)[4p - p^2 + (1-2p+2p^2)(r-1)]}{(p+(1-p)r^2 + p(1-p)(1-r)^2)^2} > 0$$
$$\frac{\partial \rho_p^2}{\partial r} = \frac{2p(1-p)(r-1)[p + (1-p)r^2 + p(1-p)(1-r)^2 + (1-p)(r-1)((1+p)r-p)]}{(p+(1-p)r^2 + p(1-p)(1-r)^2)^2} > 0$$

The positiveness of those derivatives proves the claim. □

Example 2. $X_1 \sim N(\mu_1, \sigma^2$ and $X_2 \sim N(\mu_2, \sigma^2)$ are independent. Let $a = |\mu_1 - \mu_2|/\sigma$. Then, $Z = \frac{X_1 - X_2}{\sqrt{2}\sigma} \sim N(a/\sqrt{2}, 1)$ and

$$\Delta_{12} = \mathbb{E}|X_1 - X_2| = \sqrt{2}\sigma\mathbb{E}|Z| = \sqrt{2}\sigma\left(\sqrt{2}a\Phi(\frac{a}{\sqrt{2}}) + 2\phi(\frac{a}{\sqrt{2}}) - \frac{a}{\sqrt{2}}\right)$$
$$= \sigma\left(2a\Phi(\frac{a}{\sqrt{2}}) + 2\sqrt{2}\phi(\frac{a}{\sqrt{2}}) - a\right), \tag{28}$$

where Φ and ϕ are the cumulative distribution function and probability density distribution function of $N(0,1)$, respectively. Also $\Delta_1 = \Delta_2 = 2\sigma/\sqrt{\pi}$ is obtained by (28) with $a = 0$.

To prove the monotonecity property of ρ_g in a for any p, it is sufficient to prove that $g(a) := 2a\Phi(a/\sqrt{2}) + 2\sqrt{2}\phi(a/\sqrt{2}) - a$ is increasing in a. This can be done by showing that

$$\frac{\partial g(a)}{\partial a} = 2\Phi(a/\sqrt{2}) - 1 > 0.$$

To obtain the maximum correlations with respect to p, we have

$$\frac{\partial \rho_g}{\partial p} = \frac{(1-2p)(g(a) - 2/\sqrt{\pi})}{[(p^2 + (1-p)^2 + \sqrt{\pi}p(1-p)g(a)]^2} := 0$$

With an additional check of $\frac{\partial^2 \rho_g}{\partial^2 p}|_{p=0.5} < 0$, we conclude that the maximum Gini correlation is achieved at $p = 0.5$. □

Example 3. $X_1 \sim N(\mu, \sigma_1^2)$ and $X_2 \sim N(\mu_2, \sigma_2^2)$ are independent. Then, $X_1 - X_2 \sim N(0, \sigma_1^2 + \sigma_2^2)$. By (28) with $a = 0$ and $\sqrt{2}\sigma = \sqrt{\sigma_1^2 + \sigma_2^2}$, we have $\Delta_{12} = \sqrt{2(\sigma_1^2 + \sigma_2^2)}/\sqrt{\pi}$.

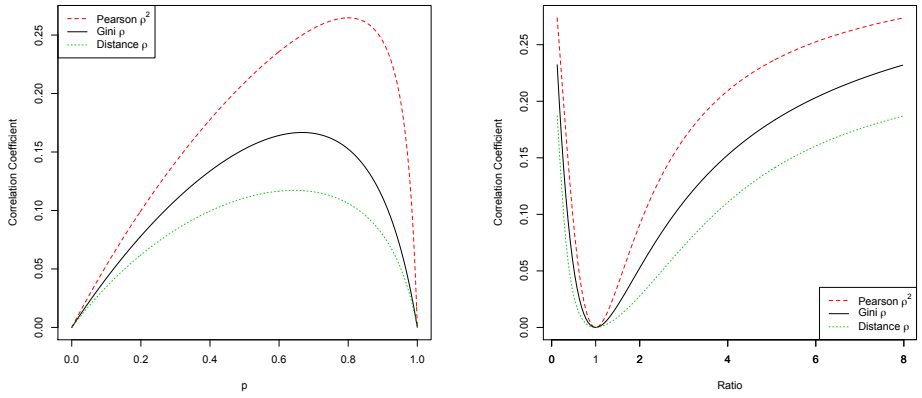
The first derivatives of the correlations with respective to r are

$$\frac{\partial \rho_g}{\partial r} = \frac{2p(1-p)(r-1)[2p(2-p) + (1-2p+2p^2)(r-1)]}{[(2p-p)^2 + (1-p^2)r^2 + (1-2p+2p^2)r]^2} > 0$$
$$\frac{\partial \rho_p^2}{\partial r} = \frac{2p(1-p)(r-1)[p + (1-p)r]}{[p + (1-p)r^2 + p(1-p)(1-r)^2]^2} > 0.$$

This completes the claim that the correlations are increasing in $r > 1$. □

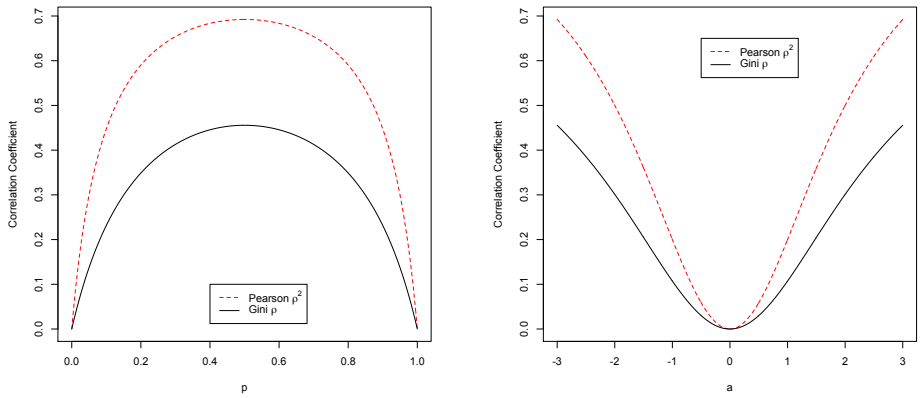
Table 1: Mean computation time in seconds of three correlations with standard errors in parentheses based on 30 repetitions.

	$d = 1$			$d = 100$		
	$n = 1,000$	$n = 10,000$	$n = 100,000$	$n = 100$	$n = 1,000$	$n = 10,000$
$\hat{\rho}_g$	.000(.000)	.001(.000)	0.012(.000)	.001(.000)	.084(.000)	9.934(.020)
$\hat{\rho}_d$	.001(.000)	.008(.000)	0.063(.013)	.003(.000)	.253(.000)	35.18(.149)
$\hat{\rho}_\kappa$	.009(.000)	.854(.001)	85.62(.046)	.002(.000)	.104(.000)	12.09(.022)



(a) (b)

Figure 1: (a) Correlation coefficients vs p in the mixture exponential distribution with $\theta = 1$ and $\beta = 4$; (b) Correlation coefficients vs $r = \beta/\theta$ in the mixture exponential distribution with $p = 0.5$.

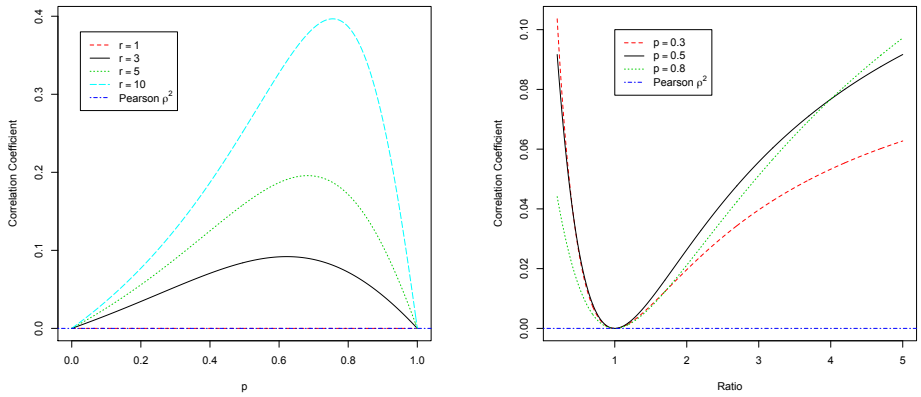


(a) (b)

Figure 2: (a) Correlation coefficient vs p in the mixture normal distribution with $a = |\mu_1 - \mu_2|/\sigma = 3$; (b) Correlation coefficient vs a with $p = 0.5$.

Table 2: Coverage probabilities of the confidence intervals.

Distribution	Parameter	Level = 0.90		Level = 0.95	
		$n = 60$	$n = 120$	$n = 60$	$n = 120$
$0.5\text{Exp}(1) + 0.5\text{Exp}(4)$	$\rho_g = 0.1525$	0.9031	0.9007	0.9442	0.9472
	$\rho_d = 0.1191$	0.5059	0.7690	0.6771	0.8802
$0.5\mathcal{N}(0,1) + 0.5\mathcal{N}(3,1)$	$\rho_g = 0.4556$	0.8898	0.8934	0.9323	0.9437
$0.5\mathcal{N}(0,1) + 0.5\mathcal{N}(0,3^2)$	$\rho_g = 0.0557$	0.9201	0.9093	0.9531	0.9524



(a) (b)

Figure 3: (a) Correlation coefficient vs p in the mixture normal distribution with $\mu_1 = \mu_2$ for different $r = \sigma_2/\sigma_1$; (b) Correlation coefficient vs $r = \sigma_2/\sigma_1$ in the mixture normal distribution with $\mu_1 = \mu_2$ for different p.

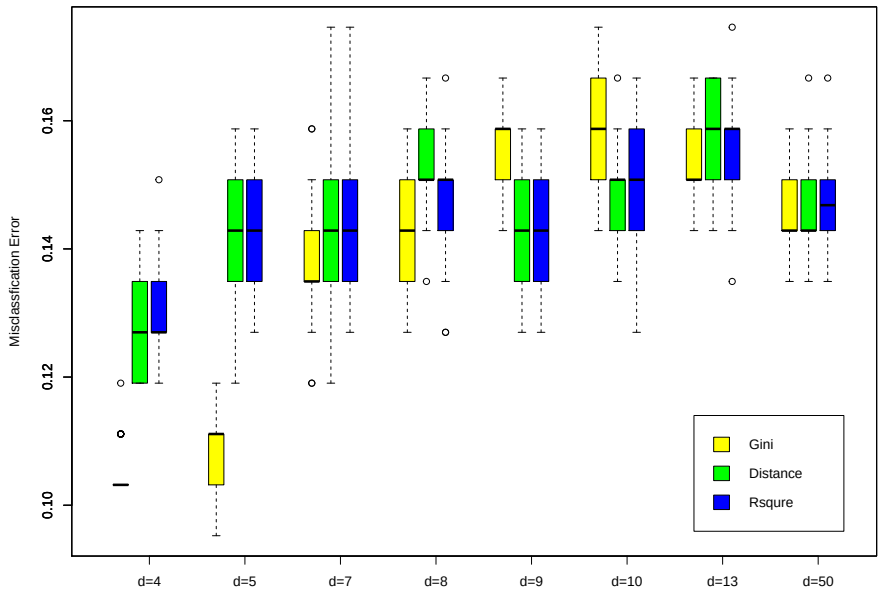


Figure 4: Boxplot of out-of-bag classification errors of random forest classifier based on d highest correlated features selected by Gini, distance and Pearson R^2 correlations.

Table 3: Size and power of dependence tests at 0.05 significance level.

Dist	n	$\boldsymbol{p}$	Method	(θ_1,θ_2) (1,1)	(1.1,1.8)	(1.2,2.6)	(1.3,3.4)	(1.4,4.2)	(1.5,5)
Exp	60	$(\frac{1}{3},\frac{1}{3},\frac{1}{3})$	ρ_g	.056	.345	.799	.922	.976	.992
			ρ_d	.045	.327	.770	.915	.967	.991
		$(\frac{5}{12},\frac{4}{12},\frac{3}{12})$	ρ_g	.056	.319	.742	.897	.948	.976
			ρ_d	.060	..271	.689	.853	.929	.968
		$(.6,.3,.1)$	ρ_g	.050	.206	.458	.653	.759	.867
			ρ_d	.044	.117	.314	.487	.656	.753
	120	$(\frac{1}{3},\frac{1}{3},\frac{1}{3})$	ρ_g	.049	.672	.971	1.00	1.00	1.00
			ρ_d	.041	.667	.974	.999	1.00	1.00
		$(\frac{5}{12},\frac{1}{3},\frac{1}{4})$	ρ_g	.055	.647	.969	.999	1.00	1.00
			ρ_d	.049	.625	.943	.998	1.00	1.00
		$(.6,.3,.1)$	ρ_g	.046	.389	.755	.928	.958	.991
			ρ_d	.050	.212	.549	.799	.920	.970
				(μ_1,μ_2) (0,0)	(0.1,0.5)	(0.2,1)	(0.3,1.5)	(0.4,2)	(0.5,2.5)
				(σ_1^2,σ_2^2) (1,1)	(1.1,1.5)	(1.2,2)	(1.3,2.5)	(1.4,3)	(1.5,3.5)
Norm	60	$(\frac{1}{3},\frac{1}{3},\frac{1}{3})$	ρ_g	.049	.250	.662	.913	.982	1.00
			ρ_d	.051	.222	.621	.907	.981	1.00
		$(\frac{5}{12},\frac{4}{12},\frac{3}{12})$	ρ_g	.046	.191	.639	.875	.982	.998
			ρ_d	.042	.162	.540	.825	.962	.993
		$(.6,.3,.1)$	ρ_g	.043	.141	.385	.607	.840	.924
			ρ_d	.053	.090	.220	.435	.658	.819
	120	$(\frac{1}{3},\frac{1}{3},\frac{1}{3})$	ρ_g	.045	.425	.921	.999	1.00	1.00
			ρ_d	.044	.416	.923	.997	1.00	1.00
		$(\frac{5}{12},\frac{4}{12},\frac{3}{12})$	ρ_g	.051	.398	.913	.999	1.00	1.00
			ρ_d	.051	.339	.858	.996	1.00	1.00
		$(.6,.3,.1)$	ρ_g	.046	.227	.644	.903	.986	.998
			ρ_d	.044	.133	.440	.754	.946	.993

Table 4: Size and power of the Gini correlation permutation test based on different values of α under the Cauchy mixtures.

n	α	$\delta=0$	$\delta=0.25$	$\delta=0.5$	$\delta=0.75$	$\delta=1$
60	0.5	0.059	0.092	0.171	0.362	0.565
	0.75	0.064	0.092	0.139	0.307	0.475
	1	0.055	0.076	0.120	0.227	0.382
120	0.5	0.070	0.118	0.334	0.647	0.876
	0.75	0.069	0.110	0.269	0.545	0.791
	1	0.070	0.086	0.199	0.393	0.633

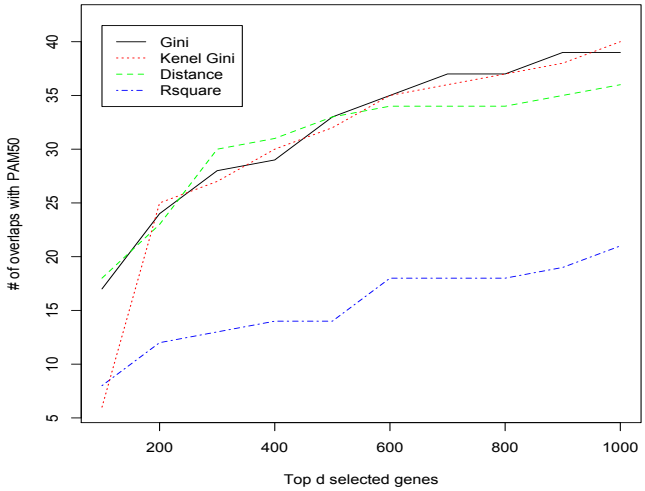


Figure 5: Number of PAM50 genes in the selected top d genes by each of correlations.

Table 5: Correlations between Species and all variables, between Species and each of variables for the Iris data. Standard deviations are in parentheses.

	All	Sepal.L	Sepal.W	Petal.L	Petal.W
$\hat{\rho}_g$	0.624 (.019)	0.398 (.035)	0.223 (.039)	0.773 (.018)	0.753 (.019)
$\hat{\rho}_d$	0.879 (.026)	0.684 (.045)	0.509 (.053)	0.882 (.026)	0.899 (.020)

Table 6: Test accuracy of random forest classifiers based on the top d features selected by each correlation. The standard errors are in parentheses. As a baseline, the test accuracy based on PAM50 is 0.910 (.003).

Method	$d = 5$	$d = 10$	$d = 20$	$d = 30$	$d = 40$	$d = 50$
ρ_g	.779 (.014)	.855 (.011)	.866 (.012)	.887 (.016)	.926 (.010)	.923 (.008)
ρ_κ	.780 (.014)	.852 (.012)	.861 (.016)	.882 (.014)	.923 (.013)	.919 (.008)
ρ_d	.752 (.011)	.759 (.012)	.903 (.006)	.896 (.009)	.890 (.010)	.894 (.008)
R^2	.832 (.008)	.848 (.014)	.861 (.015)	.871 (.016)	.892 (.016)	.878 (.016)