Estimating Feature-Label Dependence Using Gini Distance Statistics

Silu Zhang, Xin Dang, *Member, IEEE*, Dao Nguyen, Dawn Wilkins, and Yixin Chen, *Member, IEEE*

Abstract—Identifying statistical dependence between the features and the label is a fundamental problem in supervised learning. This paper presents a framework for estimating dependence between numerical features and a categorical label using *generalized Gini distance*, an energy distance in reproducing kernel Hilbert spaces (RKHS). Two Gini distance based dependence measures are explored: *Gini distance covariance* and *Gini distance correlation*. Unlike Pearson covariance and correlation, which do not characterize independence, the above Gini distance based measures define dependence as well as independence of random variables. The test statistics are simple to calculate and do not require probability density estimation. Uniform convergence bounds and asymptotic bounds are derived for the test statistics. Comparisons with distance covariance statistics are provided. It is shown that Gini distance statistics converge faster than distance covariance statistics in the uniform convergence bounds, hence tighter upper bounds on both Type I and Type II errors. Moreover, the probability of Gini distance covariance statistic under-performing the distance covariance statistic in Type II error decreases to 0 exponentially with the increase of the sample size. Extensive experimental results are presented to demonstrate the performance of the proposed method.

Index Terms—Energy distance, feature selection, Gini distance covariance, Gini distance correlation, distance covariance, reproducing kernel Hilbert space, dependence test, supervised learning

1 Introduction

Building a prediction model from observations of features and responses (or labels) is a well-studied problem in machine learning and statistics. The problem becomes particularly challenging in a high dimensional feature space. A common practice in tackling this challenge is to reduce the number of features under consideration, which is in general achieved via feature combination or feature selection.

Feature combination refers to combining high dimensional inputs into a smaller set of features via a linear or nonlinear transformation, e.g., principal component analysis (PCA) [35], independent component analysis (ICA) [16], curvilinear components analysis [21], multidimensional scaling (MDS) [81], nonnegative matrix factorization (NMF) [47], Isomap [80], locally linear embedding (LLE) [63], Laplacian eigenmaps [6], stochastic neighbor embedding (SNE) [33], etc. Feature selection, also known as variable selection, aims at choosing a subset of features that is "relevant" to the response variable [9], [40], [41]. In terms of interpretability, feature selection is more appealing than feature combination because it preserves the physical meaning of the original features.

- *S. Zhang is with the Department of Diagnostic Imaging, St. Jude Children's Research Hospital. 262 Danny Thomas Place, Memphis, TN 38105. E-mail: silu.zhang@stjude.org.*
- *X. Dang and D. Nguyen are with the Department of Mathematics, University of Mississippi, University, MS 38677. E-mail: {xdang, dxnguyen}@olemiss.edu.*
- *D. Wilkins and Y. Chen are with the Department of Computer and Information Science, University of Mississippi, University, MS 38677. E-mail: {dwilkins, yixin}@olemiss.edu.*

Manuscript received 20 Feb. 2019; revised 14 Oct. 2019; accepted 30 Nov. 2019. Date of publication 0 . 0000; date of current version 0 . 0000. (Corresponding author: Silu Zhang.) Recommended for acceptance by P. Ravikumar. Digital Object Identifier no. 10.1109/TPAMI.2019.2960358

Feature selection under supervised setting can further be broadly categorized into filter models, wrapper models and embedded models. Filter models separates the feature selection task from the classification task to avoid increasing learning bias. A common approach is to use correlation to measure feature importance. A wrapper model aims to select a feature subset that achieves optimal classification performance for a predetermined classifier. An embedded model is one that achieves feature selection during the learning process, i.e., feature selection and the training of the classifier are performed simultaneously. For datasets with limited sample size and ultrahigh dimension, both wrapper model and embedded model suffer from over-fitting, whereas filter models are more applicable. In this paper, we present a filter-based feature selection method using new dependence measures—generalized Gini distance covariance and correlation. Unlike the commonly used Pearson correlation, which is only sensitive to linear dependence and does not characterize independence, our method also characterizes independence. Gini distance statistics measures the dependence between a continues random variable/vector and a categorical response, well suited for feature selection in classification tasks. They also have nice interpretations: Gini distance covariance is a measure of between-group variation and Gini distance correlation is the ratio of between group-variation and the total variation. The proposed statistics are closely related to distance covariance and correlation, which measure the dependence between two continuous random variables/vectors. Theoretical results show that Gini distance statistics are likely to perform better in terms of Type II error.

Next, we review work most related to ours. For a more comprehensive survey of this subject, the reader is referred to [30], [31], [50], [90].

TABLE 1
Summary of Related Work on Feature Relevance

Category	Representatives
Pearson correlation (linear model) based	Stoppiglia *et al.* [72], Wei and Billings [88], Fan and Lv [24], Fan *et al.* [25]
Linearization based	Song *et al.* [71], Sun *et al.* [73], Armanfard *et al.* [1], Yao *et al.* [92]
Mutual information (divergence) based	Iannarilli Jr. and Rubin [38], Novovicová *et al.* [58], Javed *et al.* [39], Wang *et al.* [83], Zhai *et al.* [96], Maji and Pal [52], Sindhwani *et al.* [67], Naghibi *et al.* [56]
Mutual information approximation based	Kwak and Choi [44], [45], Peng *et al.* [61], Lefakis and Fleuret [48], Ding *et al.* [22]
Model-free	Székely *et al.* [76], [77], Li *et al.* [49], Cui *et al.* [18]

1.1 Related Work

With a common goal of improving the generalization performance of the prediction model and providing a better interpretation of the underlying process, all feature selection methods are built around the concepts of *feature relevance* and *feature redundancy*.

1.1.1 Feature Relevance

The concept of relevance has been studied in many fields outside machine learning and statistics [34]. In the context of feature selection, John *et al.* [40] defined feature relevance via a probabilistic interpretation where a feature and the response variable are irrelevant if and only if they are conditionally independent given any subset of features. Following this definition, Nilsson *et al.* [57] investigated distributions under which an optimal classifier can be trained over a minimal number of features. Although the above definition of relevance characterizes the statistical dependence, testing the conditional dependence is in general a challenge for continuous random variables.

Significant amount of efforts have been devoted to finding a good trade-off between theoretical rigor and practical feasibility in defining dependence measures. A summary of related work on defining feature relevance is shown in Table 1. Pearson Correlation [60] based methods [24], [25], [72], [88] are among the most popular approaches. Pearson correlation and its variations are in general straightforward to implement, but is sensitive only to linear dependence between two variables. Specifically, Pearson correlation can be zero for dependent random variables. To address the nonlinear dependence, many researchers tackled nonlinear dependence via linearization [1], [71], [73], [92].

Other correlation measures, which treat linear and nonlinear dependence under the same framework, have been developed to address the limitation of Pearson correlation based methods. Among these, mutual information (divergence) based approaches have been investigated extensively [5], [38], [39], [52], [56], [58], [67], [83], [96]. As mutual information is hard to evaluate, several approximations have been suggested [15], [22], [44], [45], [48], [61].

Mutual information relies on the estimation of the probability density functions, which is especially challenging when the sample size is small, e.g., in the medical domain. This motivated the development of model-free approaches [18], [49], [76], [77]. Cui *et al.* [18] defined a new index using the mean variance (MV) of the conditional distribution function of a feature given the class variable. It considers the ranking of the samples in a dependence measure, hence is a robust method for heavy-tailed datasets. The distance covariance and correlation proposed by Székely *et al.* [76], [77] measures the dependence between two numerical random variables of arbitrary dimension. Our approach fits into the model-free category and is closely related to distance covariance and correlation, but aims to measure the dependence between a numerical random variable and a categorical random variable.

1.1.2 Feature Redundancy

Although one may argue that all features dependent on the response variable are informative, redundant features unnecessarily increase the dimensionality of the learning problem, hence may reduce the generalization performance [93]. Eliminating feature redundancy is, therefore, an essential step in feature selection [61].

Several methods were proposed to reduce redundancy explicitly via a feature dependence measure [7], [10], [55], [82], [87], [89]. There are also many methods that formulate feature selection as an optimization problem where redundancy reduction is implicitly achieved via optimizing an objective function, for example, [13], [17], [32], [43], [66], [91], [95]. Particularly, class separation has been widely used as an objective function in redundancy reduction [11], [14], [86], [97]. Many researchers investigated optimal feature subset selection under various optimization formulations, such as using a special class of monotonic feature selection criterion functions [70], or incorporating a regularization term to control the sparsity of the solution [3], [19], [53], [62], [84], [85], [94].

1.2 An Overview of the Proposed Approach

For problems of large scale (large sample size and/or high feature dimension), feature selection is commonly performed in two steps. A subset of candidate features are first identified via a screening [24] (or a filtering [31]) process based upon a predefined "importance" measure that can be calculated efficiently. The final collection of features are then chosen from the candidate set by solving an optimization problem. Usually, this second step is computationally more expensive than the first step. Hence for problems with very high feature dimension, identifying a subset of "good" candidate features, thus reducing the computational cost of the subsequent optimization algorithm, is essential.

The work presented in this article is a model-free approach that aims at improving the feature screening process via a new dependence measure. Székely *et al.* [76], [77] introduced distance covariance and distance correlation, which extended the classical bivariate product-moment covariance and correlation to random vectors of arbitrary dimension. Distance covariance (and distance correlation) characterizes independence: it is zero if and only if the two random vectors are

independent. Moreover, the corresponding statistics are simple to calculate and do not require estimating the distribution function of the random vectors. These properties make distance covariance and distance correlation particularly appealing to the dependence test, which is a crucial component in feature selection [8], [49].

Although distance covariance and distance correlation can be extended to handle categorical variables using a metric space embedding [51], Gini distance covariance and Gini distance correlation [20] provide a natural alternative to measuring dependence between a numerical random vector and a categorical random variable. In this article, we investigate selecting informative features for supervised learning problems with numerical features and a categorical response variable using generalized Gini distance covariance and Gini distance correlation. The contributions of this paper are given as follows:

- *Generalized Gini Distance Covariance and Gini Distance Correlation.* We extend Gini distance covariance and Gini distance correlation to RKHS via positive definite kernels. The choice of kernel not only brings flexibility to the dependence tests, but also makes it easier to derive theoretical performance bounds on the tests.
- *Simple Dependence Tests.* Gini distance statistics are simple to calculate. We prove that when there is dependence between the feature vector and the response variable, the probability of Gini distance covariance statistic under-performing distance covariance statistic approaches 0 with the growth of the sample size.
- *Uniform Convergence Bounds and Asymptotic Analysis.* Under the bounded kernel assumption, we derive uniform convergence bounds for both Type I and Type II errors. Compared with distance covariance and distance correlation statistics, the bounds for Gini distance statistics are tighter. Asymptotic analysis is also presented.

1.3 Outline of the Paper

The remainder of the paper is organized as follows: Section 2 motivates Gini distance covariance and Gini distance correlation from energy distance. We then extend them to RKHS and present a connection between generalized Gini distance covariance and generalized distance covariance. Section 3 provides estimators of Gini distance covariance and Gini distance correlation. Dependence tests are developed using these estimators. We derive uniform convergence bounds for both Type I and Type II errors of the dependence tests. In Section 3.3 we present connections with dependence tests using distance covariance. A connection to maximum mean discrepancy (MMD) [29] is shown in Section 3.4. Asymptotic results are given in Section 3.5. We discuss several algorithmic issues in Section 4. In Section 5, we explain the extensive experimental studies conducted and demonstrate the results. We conclude and discuss the strengths and limitations of the proposed method in Section 6.

2 Gini Distance Covariance and Correlation

In this section, we first present a brief review of the energy distance. As an instance of the energy distance, Gini distance covariance is introduced to measure dependence between numerical and categorical random variables. Gini distance covariance and correlation are then generalized to reproducing kernel Hilbert spaces (RKHS) to facilitate convergence analysis in Section 3. Connections with distance covariance are also discussed.

2.1 Energy Distance

Energy distance was first introduced in [4], [74], [75] as a measure of statistical distance between two probability distributions with finite first order moments. The energy distance between the q-dimensional independent random variables X and Y is defined as [78]

$$\mathcal{E}(X,Y) = 2\mathbb{E}|X-Y|_q - \mathbb{E}|X-X'|_q - \mathbb{E}|Y-Y'|_q, \tag{1}$$

where $|\cdot|_q$ is the euclidean norm in $\mathbb{R}^q$, $\mathbb{E}|X|_q + \mathbb{E}|Y|_q < \infty$, X' is an iid copy of X, and Y' is an iid copy of Y.

Energy distance has many interesting properties. It is scale equivariant: for any $a \in \mathbb{R}$,

$$\mathcal{E}(aX, aY) = |a|\mathcal{E}(X,Y).$$

It is rotation invariant: for any rotation matrix $\mathbf{R} \in \mathbb{R}^{q\times q}$

$$\mathcal{E}(\mathbf{R}X, \mathbf{R}Y) = \mathcal{E}(X,Y).$$

Test statistics of an energy distance are in general relatively simple to calculate and do not require density estimation (Section 3). Most importantly, as shown in [75], if φ_X and φ_Y are the characteristic functions of X and Y, respectively, the energy distance (1) can be equivalently written as

$$\mathcal{E}(X,Y) = c(q)\int_{\mathbb{R}^q} \frac{[\varphi_X(x)-\varphi_Y(x)]^2}{|x|_q^{q+1}}dx, \tag{2}$$

where $c(q) > 0$ is a constant only depending on q. Thus $\mathcal{E} \geq 0$ with equality to zero if and only if X and Y are identically distributed. The above properties make energy distance especially appealing to testing identical distributions (or dependence).

2.2 Gini Distance Covariance and Gini Distance Correlation

Gini distance covariance was proposed in [20] to measure dependence between a numerical random variable $X \in \mathbb{R}^q$ from function F (cumulative distribution function, CDF) and a categorical variable Y with K values $L_1, \ldots, L_K$. If we assume the categorical distribution P_Y of Y is $\Pr(Y = L_k) = p_k$ and the conditional distribution of X given $Y = L_k$ is F_k, the marginal distribution of X is

$$F(x) = \sum_{k=1}^{K} p_k F_k(x).$$

When the conditional distribution of X given Y is the same as the marginal distribution of X, X and Y are independent, i.e., there is no correlation between them. However, when they are dependent, i.e., $F \neq F_k$ for some k, the dependence can be measured through the difference between the marginal distribution F and conditional distribution F_k.

This difference is measured by Gini distance covariance, $\mathrm{gCov}(X,Y)$, which is defined as the expected weighted L_2 distance between characteristic functions of the conditional and marginal distributions (if the expectation is finite):

$$\mathrm{gCov}(X,Y) := c(q)\sum_{k=1}^{K} p_k \int_{\mathbb{R}^q} \frac{[\varphi_k(x)-\varphi(x)]^2}{|x|_q^{q+1}}\,dx,$$

where $c(q)$ is the same constant as in (2), φ_k and φ are the characteristic functions for the conditional distribution F_k and marginal distribution F, respectively. It follows immediately that $\mathrm{gCov}(X,Y)=0$ mutually implies independence between X and Y. Based on (1) and (2), the Gini distance covariance is clearly a weighted energy distance, hence can be equivalently defined as

$$\mathrm{gCov}(X,Y) = \sum_{k=1}^{K} p_k\Big[2\mathbb{E}|X_k - X|_q - \mathbb{E}|X_k - X_k{}'|_q - \mathbb{E}|X - X'|_q\Big], \tag{3}$$

where $(X_k, X_k{}')$ and (X, X') are independent pair variables from F_k and F, respectively.

Gini distance covariance can be standardized to have a range of [0, 1], a desired property for a correlation measure. The resulting measure is called Gini distance correlation, denoted by $\mathrm{gCor}(X,Y)$, which is defined as

$$\mathrm{gCor}(X,Y) = \frac{\sum_{k=1}^{K} p_k\Big[2\mathbb{E}|X_k - X|_q - \mathbb{E}|X_k - X_k{}'|_q - \mathbb{E}|X - X'|_q\Big]}{\mathbb{E}|X - X'|_q}, \tag{4}$$

provided that $\mathbb{E}|X|_q + \mathbb{E}|X_k|_q < \infty$ and F is not a degenerate distribution. Gini distance correlation satisfies the following properties [20].

1) $0 \le \mathrm{gCor}(X,Y) \le 1$.
2) $\mathrm{gCor}(X,Y)=0$ if and only if X and Y are independent.
3) $\mathrm{gCor}(X,Y)=1$ if and only if F_k is a single point mass distribution almost surely for all $k=1,\ldots,K$.
4) $\mathrm{gCor}(a\mathbf{R}X+b, Y) = \mathrm{gCor}(X,Y)$ for all $a\neq 0$, $b\in\mathbb{R}^q$, and any orthonormal matrix $\mathbf{R}\in\mathbb{R}^{q\times q}$.

Property 2 are especially useful in testing dependence.

2.3 Gini Distance Statistics in RKHS

Energy distance based statistics naturally generalizes from a euclidean space to metric spaces [51]. By using a positive definite kernel (Mercer kernel) [54], distributions are mapped into a RKHS [69] with a kernel induced distance. Hence one can extend energy distances to a much richer family of statistics defined in RKHS [64]. Let $M:\mathbb{R}^q\times\mathbb{R}^q\to\mathbb{R}$ be a Mercer kernel [54]. There is an associated RKHS $\mathcal{H}_M$ of real functions on $\mathbb{R}^q$ with reproducing kernel M, where the function $d:\mathbb{R}^q\times\mathbb{R}^q\to\mathbb{R}$ defines a distance in $\mathcal{H}_M$,

$$d_M(x,x') = \sqrt{M(x,x)+M(x',x')-2M(x,x')}. \tag{5}$$

Hence Gini distance covariance and Gini distance correlation are generalized to RKHS, $\mathcal{H}_M$, as

$$\mathrm{gCov}_M(X,Y) = \sum_{k=1}^{K} p_k[2\mathbb{E}d_M(X_k,X) - \mathbb{E}d_M(X_k,X_k{}') - \mathbb{E}d_M(X,X')], \tag{6}$$

$$\mathrm{gCor}_M(X,Y) = \frac{\sum_{k=1}^{K} p_k[2\mathbb{E}d_M(X_k,X) - \mathbb{E}d_M(X_k,X_k{}') - \mathbb{E}d_M(X,X')]}{\mathbb{E}d_M(X,X')}. \tag{7}$$

The choice of kernels allows one to design various tests. In this paper, we focus on bounded translation and rotation invariant kernels. Our choice is based on the following considerations:

1) The boundedness of a positive definite kernel implies the boundedness of the distance in RKHS, which makes it easier to derive strong (exponential) convergence inequalities based on bounded deviations (discussed in Section 3);
2) Translation and rotation invariance is an important property to have for testing of dependence.

Same as in $\mathbb{R}^q$, Gini distance covariance and Gini distance correlation in RKHS also characterize independence, i.e., $\mathrm{gCov}_M(X,Y)=0$ and $\mathrm{gCor}_M(X,Y)=0$ if and only if X and Y are independent. This is derived as the following from the connection between Gini distance covariance and distance covariance in RKHS. Distance covariance was introduced in [76] as a dependence measure between random variables $X\in\mathbb{R}^p$ and $Y\in\mathbb{R}^q$. If X and Y are embedded into RKHS's induced by M_X and M_Y, respectively, the generalized distance covariance of X and Y is [64]:

$$\begin{aligned}&\mathrm{dCov}_{M_X,M_Y}(X,Y)\\ &= \mathbb{E}d_{M_X}(X,X')d_{M_Y}(Y,Y') + \mathbb{E}d_{M_X}(X,X')\mathbb{E}d_{M_Y}(Y,Y')\\ &\quad - 2\mathbb{E}\big[\mathbb{E}_{X'}d_{M_X}(X,X')\mathbb{E}_{Y'}d_{M_Y}(Y,Y')\big].\end{aligned} \tag{8}$$

In the case of Y being categorical, one may embed it using a set difference kernel M_Y,

$$M_Y(y,y') = \begin{cases}\frac{1}{2} & if\ y=y',\\ 0 & otherwise.\end{cases} \tag{9}$$

This is equivalent to embedding Y as a simplex with edges of unit length [51], i.e., L_k is represented by a K dimensional vector of all zeros except its kth dimension, which has the value $\frac{\sqrt{2}}{2}$. The distance induced by M_Y is called the set distance, i.e., $d_{M_Y}(y,y')=0$ if $y=y'$ and 1 otherwise. Using the set distance, we have the following results on the generalized distance covariance between a numerical and a categorical random variable.

Lemma 1. *Suppose that $X\in\mathbb{R}^q$ is from distribution F and Y is a categorical variable with K values $L_1,\ldots,L_K$. The categorical distribution P_Y of Y is $P(Y=L_k)=p_k$ and the conditional distribution of X given $Y=L_k$ is F_k, the marginal distribution of X is $F(x)=\sum_{k=1}^{K}p_kF_k(x)$. Let $M_X:\mathbb{R}^q\times\mathbb{R}^q\to\mathbb{R}$ be a*

Mercer kernel and M_Y a set difference kernel. The generalized distance covariance $\mathrm{dCov}_{M_X,M_Y}(X,Y)$ *is equivalent to*

$$\mathrm{dCov}_{M_X,M_Y}(X,Y) := \mathrm{dCov}_{M_X}(X,Y) = \sum_{k=1}^{K} p_k^2 \big[2\mathbb{E}d_{M_X}(X_k,X) - \mathbb{E}d_{M_X}(X_k,X_k{}') - \mathbb{E}d_{M_X}(X,X')\big]. \tag{10}$$

From (6) and (10), it is clear that the generalized Gini covariance is always larger than or equal to the generalized distance covariance under the set difference kernel and the same M_X, i.e., [1]

$$\mathrm{gCov}_{M_X}(X,Y) \geq \mathrm{dCov}_{M_X}(X,Y), \tag{11}$$

where they are equal if and only if both are 0, i.e., X and Y are independent. This yields the following theorem. The proof of Lemma 1 is given in Appendix A, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/TPAMI.2019.2960358.

Theorem 2. *For any bounded Mercer kernel $M : \mathbb{R}^q \times \mathbb{R}^q \to \mathbb{R}$, $\mathrm{gCov}_M(X,Y) = 0$ if and only if X and Y are independent. The same result holds for $\mathrm{gCor}_M(X,Y)$ assuming that the marginal distribution of X is not degenerate.*

Proof. The proof of the sufficient part for $\mathrm{gCov}_M(X,Y)$ is immediate from the definition (6). The inequality (11) suggests that $\mathrm{dCov}_M = 0$ when $\mathrm{gCov}_M = 0$. Hence the proof of the necessary part is complete if we show that $\mathrm{dCov}_M = 0$ implies independence. This is proven as the following.

Let $\mathcal{X}$ and $\mathcal{Y}$ be the RKHS induced by M and the set difference kernel (9), respectively, with the associated distance metrics defined according to (5). $\mathcal{X}$ and $\mathcal{Y}$ are both separable Hilbert spaces [2], [12] as they each have a countable set of orthonormal basis [54]. Hence $\mathcal{X}$ and $\mathcal{Y}$ are of strong negative type (Theorem 3.16 in [51]). Because the metrics on $\mathcal{X}$ and $\mathcal{Y}$ are bounded, the marginals of (X,Y) on $\mathcal{X} \times \mathcal{Y}$ have finite first moment in the sense defined in [51]. Therefore, $\mathrm{dCov}_M(X,Y) = 0$ implies that X and Y are independent (Theorem 3.11 [51]).

Finally, the proof for $\mathrm{gCor}_M(X,Y)$ follows from the above and the condition that X is not degenerate. □

In the remainder of the paper, unless noted otherwise, we use the default distance function [2]

$$d_M(x,x') = \sqrt{1 - e^{-\frac{|x-x'|_q^2}{\sigma^2}}},$$

induced by a weighted Gaussian kernel, $M(x,x') = \frac{1}{2}e^{-\frac{|x-x'|_q^2}{\sigma^2}}$. It is immediate that the above distance function is translation and rotation invariant and is bounded with the range [0, 1). Moreover, using Taylor expansion, it is not difficult to show that gCor_M approaches gCor when the kernel parameter σ approaches ∞.

1. The inequality holds for Gini covariance and distance covariance as well, i.e., $\mathrm{gCov}(X,Y) \geq \mathrm{dCov}(X,Y)$ where $X \in \mathbb{R}^d$ and Y is categorical. The notations of $\mathrm{gCov}_{M_X}(X,Y)$ and $\mathrm{gCov}_M(X,Y)$ are used 'interchangeably with both M_X and M representing a Mercer kernel.

2. Since any bounded translation and rotation invariant kernels can be normalized to define a distance function with the maximum value no greater than 1, the results in Sections 2.3 and 3 hold for these kernels as well.

3 Dependence Tests

We first present an unbiased estimator of the generalized Gini distance covariance. Probabilistic bounds for large deviations of the empirical generalized Gini distance covariance are then derived. These bounds lead directly to two dependence tests. We also provide discussions on connections with the dependence test using generalized distance covariance and connection with maximum mean discrepancy (MMD) [29]. Finally, asymptotic analysis of the test statistics is presented.

3.1 Estimation

In Section 2.2, Gini distance covariance and Gini distance correlation were introduced from an energy distance point of view. An alternative interpretation based on Gini mean difference was given in [20]. This definition yields simple point estimators.

Let $X \in \mathbb{R}^q$ be a random variable from distribution F. Let $Y \in \mathbb{Y} = \{L_1, \ldots, L_K\}$ be a categorical random variable with K values and $\Pr(Y = L_k) = p_k \in (0,1)$. The conditional distribution of X given $Y = L_k$ is F_k. Let (X,X') and $(X_k,X_k{}')$ be independent pair variables from F and F_k, respectively. The Gini distance covariance (3) and Gini distance correlation (4) can be equivalently written as

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

$$\mathrm{gCor}(X,Y) = \frac{\Delta - \sum_{k=1}^{K} p_k \Delta_k}{\Delta}, \tag{13}$$

where $\Delta = \mathbb{E}|X - X'|_q$ and $\Delta_k = \mathbb{E}|X_k - X_k{}'|_q$ are the Gini mean difference (GMD) of F and F_k in $\mathbb{R}^q$ [26], [27], [42], respectively. This suggests that Gini distance covariance is a measure of between-group variation and Gini distance correlation is the ratio of between-group variation and the total Gini variation. Replacing $|\cdot|_q$ with $d_M(\cdot,\cdot)$ in (12) and (13) yields the GMD version of (6) and (7).

Given an iid sample data $\mathcal{D} = \{(x_i, y_i) \in \mathbb{R}^q \times \mathbb{Y} : i = 1, \ldots, n\}$, let $\mathcal{I}_k$ be the index set of sample points with $y_i = L_k$. The probability p_k is estimated by the sample proportion of category k, i.e., $\hat{p}_k = \frac{n_k}{n}$ where $n_k = |\mathcal{I}_k| > 2$. The point estimators of the generalized Gini distance covariance and Gini distance correlation for a given kernel M are

$$\mathrm{gCov}_M^n := \hat{\Delta} - \sum_{k=1}^{K} \hat{p}_k \hat{\Delta}_k, \tag{14}$$

$$\mathrm{gCor}_M^n := \frac{\hat{\Delta} - \sum_{k=1}^{K} \hat{p}_k \hat{\Delta}_k}{\hat{\Delta}}, \tag{15}$$

where

$$\hat{\Delta}_k = \binom{n_k}{2}^{-1} \sum_{i<j \in \mathcal{I}_k} d_M(x_i, x_j), \tag{16}$$

$$\hat{\Delta} = \binom{n}{2}^{-1} \sum_{i<j} d_M(x_i, x_j). \tag{17}$$

Theorem 3. *The point estimator (14) of the generalized Gini distance covariance is unbiased.*

Proof. Clearly, $\hat{\Delta}_k$ and $\hat{\Delta}$ are unbiased because they are U-statistics of size 2. Also $\hat{p}_k\hat{\Delta}_k$ is unbiased since $\mathbb{E}[\hat{p}_k\hat{\Delta}_k] = \mathbb{E}\mathbb{E}[\hat{p}_k\hat{\Delta}_k|n_k] = \mathbb{E}[\frac{n_k}{n}\Delta_k] = p_k\Delta_k$. This leads to the unbiasedness of gCov_M^n. □

3.2 Uniform Convergence Bounds

We derive two probabilistic inequalities, from which dependence tests using point estimators (14) and (15) are established.

Theorem 4. *Let $\mathcal{D} = \{(x_i, y_i) \in \mathbb{R}^q \times \mathbb{Y} : i = 1, \ldots, n\}$ be an iid sample of (X, Y) and M a Mercer kernel over $\mathbb{R}^q \times \mathbb{R}^q$ that induces a distance function $d_M(\cdot,\cdot)$ with bounded range [0, 1). For every $\epsilon > 0$,*

$$\Pr\big[\mathrm{gCov}_M^n - \mathrm{gCov}_M(X,Y) \geq \epsilon\big] \leq \exp\left(\frac{-n\epsilon^2}{12.5}\right), \text{ and}$$
$$\Pr\big[\mathrm{gCov}_M(X,Y) - \mathrm{gCov}_M^n \geq \epsilon\big] \leq \exp\left(\frac{-n\epsilon^2}{12.5}\right).$$

Theorem 5. *Under the condition of Theorem 4, for every $\epsilon > 0$*

$$\Pr[\hat{\Delta} - \Delta \geq \epsilon] \leq \exp\left(\frac{-n\epsilon^2}{2}\right), \text{ and}$$
$$\Pr[\Delta - \hat{\Delta} \geq \epsilon] \leq \exp\left(\frac{-n\epsilon^2}{2}\right).$$

Proofs of Theorem 4 and Theorem 5 are given in Appendix B, available in the online supplemental material. Next we consider a dependence test based on gCov_M^n. Theorem 2 shows that $\mathrm{gCov}_M(X,Y) = 0$ mutually implies that X and Y are independent. This suggests the following null and alternative hypotheses:

$$H_0 : \mathrm{gCov}_M(X,Y) = 0,$$
$$H_1 : \mathrm{gCov}_M(X,Y) \geq 2cn^{-t},\ c > 0 \text{ and } t > 0.$$

The null hypothesis is rejected when $\mathrm{gCov}_M^n \geq cn^{-t}$ where $c > 0$ and $t \in (0, \frac{1}{2})$. Next we establish upper bounds for the Type I and Type II errors of the above dependence test.

Corollary 6. *Under the conditions of Theorem 4, the following inequalities hold for any $c > 0$ and $t \in (0, \frac{1}{2})$:*

$$\text{Type I} : \Pr\big[\mathrm{gCov}_M^n \geq cn^{-t}|H_0\big] \leq \exp\left(-\frac{c^2n^{1-2t}}{12.5}\right), \tag{18}$$

$$\text{Type II} : \Pr\big[\mathrm{gCov}_M^n \leq cn^{-t}|H_1\big] \leq \exp\left(-\frac{c^2n^{1-2t}}{12.5}\right). \tag{19}$$

Proof. Let $\epsilon = cn^{-t}$. The Type I bound is immediate from Theorem 4. The Type II bound is derived from the following inequality and Theorem 4.

$$\begin{aligned}&\Pr\big[\mathrm{gCov}_M^n \leq cn^{-t}|H_1\big]\\ &\leq \Pr\big[cn^{-t} - \mathrm{gCov}_M^n + \mathrm{gCov}_M(X,Y) - 2cn^{-t} \geq 0|H_1\big]\\ &= \Pr\big[\mathrm{gCov}_M(X,Y) - \mathrm{gCov}_M^n \geq cn^{-t}|H_1\big].\end{aligned}$$

□

A dependence test can also be performed using the empirical Gini distance correlation under the above null and alternative hypotheses with gCor_M replacing gCov_M. The null hypothesis is rejected when $\mathrm{gCor}_M^n \geq cn^{-t}$ where $c > 0$ and $t \in (0, \frac{1}{4})$. Type I and Type II bounds are presented as below.

Corollary 7. *Under the conditions of Theorem 4 and Theorem 5 where additionally $\Delta \geq 2n^{-t}$, the following inequalities hold for any $c > 0$ and $t \in (0, \frac{1}{4})$:*

$$\begin{aligned}\text{Type I} : &\Pr\big[\mathrm{gCor}_M^n \geq cn^{-t}|H_0\big]\\ &\leq \exp\left(-\frac{c^2n^{1-4t}}{12.5}\right) + \exp\left(-\frac{n^{1-2t}}{2}\right),\end{aligned} \tag{20}$$

$$\text{Type II} : \Pr\big[\mathrm{gCor}_M^n \leq cn^{-t}|H_1\big] \leq \exp\left(-\frac{c^2n^{1-2t}}{12.5}\right). \tag{21}$$

Proof. From (15), we have

$$\begin{aligned}&\Pr\big[\mathrm{gCor}_M^n \geq cn^{-t}|H_0\big]\\ &\leq \Pr\Big[\mathrm{gCov}_M^n \geq cn^{-2t} \ \text{ OR } \ \hat{\Delta} \leq n^{-t}|H_0\Big]\\ &\leq \Pr\big[\mathrm{gCov}_M^n \geq cn^{-2t}|H_0\big] + \Pr\Big[\hat{\Delta} \leq n^{-t}|H_0\Big]\\ &\leq \Pr\big[\mathrm{gCov}_M^n \geq cn^{-2t}|H_0\big] + \Pr\Big[\Delta - \hat{\Delta} \geq n^{-t}|H_0\Big].\end{aligned}$$

Let $\epsilon_1 = cn^{-2t}$ and $\epsilon_2 = n^{-t}$. The Type I bound is derived from Theorem 4 and Theorem 5. The boundedness of $d_M(\cdot,\cdot)$ implies that $\hat{\Delta} < 1$. Therefore,

$$\begin{aligned}&\Pr\big[\mathrm{gCor}_M^n \leq cn^{-t}|H_1\big]\\ &\leq \Pr\big[\mathrm{gCov}_M^n \leq cn^{-t}|H_1\big]\\ &\leq \Pr\big[\mathrm{gCov}_M(X,Y) - \mathrm{gCov}_M^n \geq cn^{-t}|H_1\big].\end{aligned}$$

Hence the Type II bound is given by Theorem 4 with $\epsilon = cn^{-t}$. □

3.3 Connections to Generalized Distance Covariance

In Section 2.3, generalized Gini distance covariance is related to generalized distance covariance through (11). Under the conditions of Lemma 1, $\mathrm{dCov}_{M_X,M_Y}(X,Y) = 0$ if and only if X and Y are independent. Hence dependence tests similar to those in Section 3.2 can be developed using empirical estimates of $\mathrm{dCov}_{M_X,M_Y}(X,Y)$. Next, we establish a result similar to Theorem 4 for generalized distance covariance. We demonstrate that generalized Gini distance covariance has a tighter probabilistic bound for large deviations than its generalized distance covariance counterpart.

Using the unbiased estimator for distance covariance developed in [79], we generalize it to an unbiased estimator for $\mathrm{dCov}_{M_X,M_Y}(X,Y)$ defined in (8). Let $\mathcal{D} = \{(x_i, y_i) \in \mathbb{R}^q \times \mathbb{R}^p : i = 1, \ldots, n\}$ be an iid sample from the joint distribution of X and Y. Let $A = (a_{ij})$ be a symmetric, $n \times n$, centered kernel distance matrix of sample $x_1, \ldots, x_n$. 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 $a_{ij} = d_{M_X}(x_i, x_j)$, $a_{i\cdot} = \sum_{j=1}^{n} a_{ij}$, $a_{\cdot j} = \sum_{i=1}^{n} a_{ij}$, and $a_{\cdot\cdot} = \sum_{i,j=1}^{n} a_{ij}$. Similarly, using $d_{M_Y}(y_i, y_j)$, a symmetric, $n \times n$, centered kernel distance matrix is calculated for samples $y_1, \ldots, y_n$ and denoted by $B = (b_{ij})$. An unbiased estimator of $\mathrm{dCov}_{M_X,M_Y}(X,Y)$ is given as

$$\mathrm{dCov}^n_{M_X,M_Y} = \frac{1}{n(n-3)} \sum_{i \neq j} A_{ij} B_{ij}. \tag{22}$$

We have the following result on the concentration of $\mathrm{dCov}^n_{M_X,M_Y}$ around $\mathrm{dCov}_{M_X,M_Y}(X,Y)$.

Theorem 8. *Let $\mathcal{D} = \{(x_i, y_i) \in \mathbb{R}^q \times \mathbb{R}^p : i = 1, \ldots, n\}$ be an iid sample of (X,Y). Let $M_X : \mathbb{R}^q \times \mathbb{R}^q \to \mathbb{R}$ and $M_Y : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ be Mercer kernels. $d_{M_X}(\cdot,\cdot)$ and $d_{M_Y}(\cdot,\cdot)$ are distance functions induced by M_X and M_Y, respectively. Both distance functions have a bounded range [0, 1). For every $\epsilon > 0$,*

$$\Pr\Big[\mathrm{dCov}^n_{M_X,M_Y} - \mathrm{dCov}_{M_X,M_Y}(X,Y) \geq \epsilon\Big] \leq \exp\left(\frac{-n\epsilon^2}{512}\right),$$

and

$$\Pr\Big[\mathrm{dCov}_{M_X,M_Y}(X,Y) - \mathrm{dCov}^n_{M_X,M_Y} \geq \epsilon\Big] \leq \exp\left(\frac{-n\epsilon^2}{512}\right).$$

The proof is provided in Appendix C, available in the online supplemental material. Note that the above result is established for both X and Y being numerical. When Y is categorical, it can be embedded into $\mathbb{R}^K$ using the set difference kernel (9). Therefore, in the following discussion, we use the simpler notation introduced in Lemma 1 where dCov_{M_X,M_Y} is denoted by dCov_{M_X}.

The upper bounds for generalized Gini distance covariance is clearly tighter than those for generalized distance covariance. Replacing $\mathrm{gCov}_M(X,Y)$ in H_0 and H_1 with $\mathrm{dCov}_{M_X}(X,Y)$, one may develop dependence tests parallel to those in Section 3.2: reject the null hypothesis when $\mathrm{dCov}^n_{M_X} \geq cn^{-t}$ where $c > 0$ and $t \in (0, \frac{1}{2})$. Upper bounds on Type I and Type II errors can be established in a result similar to Corollary 6 with the only difference being replacing the constant 12.5 with 512. Hence the bounds on the generalized Gini distance covariance based dependence test are tighter than those on the generalized distance covariance based dependence test.

To further compare the two dependence tests, we consider the following null and alternative hypotheses:

$$H_0 : S(X,Y) = 0,$$
$$H_1 : S(X,Y) \geq \mathcal{T},\ \mathcal{T} > 0,$$

where $S(X,Y) = \mathrm{gCov}_{M_X}(X,Y)$ or $\mathrm{dCov}_{M_X}(X,Y)$ with the corresponding test statistics $S_n = \mathrm{gCov}^n_{M_X}$ or $\mathrm{dCov}^n_{M_X}$, respectively. The null hypothesis is rejected when $S_n \geq \tau$ where $0 < \tau \leq \mathcal{T}$. Note that this test is more general than the dependence test discussed in Section 3.2, which is a special case with $\mathcal{T} = 2cn^{-t}$ and $\tau = cn^{-t}$. Upper bounds on Type I errors follow immediately from (18) by replacing cn^{-t} with τ. Type II error bounds, however, are more difficult to derive due to the fact that $\tau = \mathcal{T}$ would make deviation nonexistent. Next, we take a different approach by establishing which one of $\mathrm{gCov}^n_{M_X}$ and $\mathrm{dCov}^n_{M_X}$ is less likely to underperform in terms of Type II errors.

Under the alternative hypothesis

$$H_1{}' : \mathrm{dCov}_{M_X}(X,Y) \geq \mathcal{T},\ \mathcal{T} > 0,$$

we compare two dependence tests:

- accepting $H_1{}'$ when $\mathrm{gCov}^n_{M_X} \geq \tau, 0 < \tau \leq \mathcal{T}$;
- accepting $H_1{}'$ when $\mathrm{dCov}^n_{M_X} \geq \tau, 0 < \tau \leq \mathcal{T}$.

We call that "$\mathrm{gCov}^n_{M_X}$ *underperforms* $\mathrm{dCov}^n_{M_X}$" if and only if

$$\mathrm{gCov}^n_{M_X} < \tau \leq \mathrm{dCov}^n_{M_X},$$

i.e., the dependence between X and Y is detected by $\mathrm{dCov}^n_{M_X}$ but not by $\mathrm{gCov}^n_{M_X}$. The following theorem demonstrates an upper bound on the probability that $\mathrm{gCov}^n_{M_X}$ underperforms $\mathrm{dCov}^n_{M_X}$.

Theorem 9. *Under $H_1{}'$ and conditions of Theorem 8, there exists $\gamma > 0$ such that the following inequality holds for any $\mathcal{T} > 0$ and $0 < \tau \leq \mathcal{T}$:*

$$\Pr\Big[\mathrm{gCov}^n_{M_X} \text{ underperforms } \mathrm{dCov}^n_{M_X} | H_1{}'\Big] \leq 2e^{-n\gamma^2}.$$

Proof. Lemma 1 implies that $\mathrm{gCov}_{M_X}(X,Y) \geq \mathrm{dCov}_{M_X}(X,Y)$ where the equality holds if and only if both are 0, i.e., X and Y are independent. Therefore, under $H_1{}'$, for any $\mathcal{T} > 0$ and $0 < \tau \leq \mathcal{T}$, we define

$$\gamma = \frac{\mathrm{gCov}_{M_X}(X,Y) - \mathrm{dCov}_{M_X}(X,Y)}{\sqrt{12.5} + \sqrt{512}} > 0.$$

It follows that

$$\begin{aligned}
&\Pr\Big[\mathrm{gCov}^n_{M_X} \text{ underperforms } \mathrm{dCov}^n_{M_X} | H_1{}'\Big] \\
&= \Pr\Big[\mathrm{gCov}^n_{M_X} < \tau \leq \mathrm{dCov}^n_{M_X} | H_1{}'\Big] \\
&\leq \Pr\Big[\mathrm{gCov}^n_{M_X} < \mathrm{dCov}^n_{M_X} | H_1{}'\Big] \\
&\leq \Pr\Big[\mathrm{gCov}_{M_X}(X,Y) - \mathrm{gCov}^n_{M_X} \geq \sqrt{12.5}\gamma \ \text{ OR} \\
&\quad \mathrm{dCov}^n_{M_X} - \mathrm{dCov}_{M_X}(X,Y) \geq \sqrt{512}\gamma | H_1{}'\Big] \\
&\leq \Pr\Big[\mathrm{gCov}_{M_X}(X,Y) - \mathrm{gCov}^n_{M_X} \geq \sqrt{12.5}\gamma | H_1{}'\Big] \\
&+ \Pr\Big[\mathrm{dCov}^n_{M_X} - \mathrm{dCov}_{M_X}(X,Y) \geq \sqrt{512}\gamma | H_1{}'\Big] \\
&\leq 2e^{-n\gamma^2},
\end{aligned}$$

where the last step is from Theorems 4 and 8. □

3.4 Connections to Maximum Mean Discrepancy

In [29], Gretton *et al.* proposed a method in testing if two samples are drawn from different distributions based on maximum mean discrepancy (MMD), defined as the largest difference in expectations over functions in a RKHS. Sejdinovic *et al.* [64] showed the equivalence of distance-based and RKHS-based methods in hypothesis testing. In particular, it was shown that distance covariance and HSIC are equivalent, and MMD is equivalent to energy distance when the distance is computed with a semimetric of negative type.

The Gini distance statistics was generalized to RKHS via a kernel induced energy distance while MMD measures the difference between two distributions in RKHS. The following result shows a close connection between Gini distance covariance in RKHS and the average of squared MMD between the margin distribution F and conditional distributions F_k's.

Corollary 10. *Suppose that $X \in \mathbb{R}^q$ is from distribution F and Y is a categorical variable with K values $L_1, \ldots, L_K$. The categorical distribution P_Y of Y is $P(Y = L_k) = p_k$ and the conditional distribution of X given $Y = L_k$ is F_k, the marginal distribution of X is $F(x) = \sum_{k=1}^{K} p_k F_k(x)$. For any Mercer kernel $M : \mathbb{R}^q \times \mathbb{R}^q \to \mathbb{R}$, there exists a Mercer kernel $\widehat{M} : \mathbb{R}^q \times \mathbb{R}^q \to \mathbb{R}$ such that*

$$\mathrm{gCov}_M(X, Y) = \sum_{k=1}^{K} 2p_k \delta^2_{\widehat{M}}(F, F_k),$$

where $\delta^2_{\widehat{M}}(F, F_k)$ is the squared MMD between F and F_k in RKHS of $\widehat{M}$.

Proof. Let $d_M(x, x')$ be a distance induced by M as defined in (5). We construct a distance induced kernel $\widehat{M}$ centered at x_0 as

$$\widehat{M}(x, x') = \frac{1}{2}[d_M(x, x_0) + d_M(x', x_0) - d_M(x, x')].$$

$\widehat{M}$ is positive definite (Lemma 12, [64]). From Theorem 22 of [64], we have

$$2\mathbb{E}d_M(X_k, X) - \mathbb{E}d_M(X_k, X_k') - \mathbb{E}d_M(X, X') = 2\delta^2_{\widehat{M}}(F, F_k).$$

The proof follows (6). □

The result above means that for any Mercer kernel M, one can construct another Mercer kernel $\widehat{M}$ such that the Gini covariance in M is equivalent to a weighted average of squared MMD in $\widehat{M}$.

3.5 Asymptotic Analysis

We now present asymptotic distributions for the proposed Gini covariance and Gini correlation.

Theorem 11. *Assume $\mathbb{E}(d^2_M(X, X')) < \infty$ and $p_k > 0$ for $k = 1, \ldots, K$. Under dependence of X and Y, $\mathrm{gCov}^n_{M_X}$ and $\mathrm{gCor}^n_{M_X}$ have the asymptotic normality property. That is,*

$$\sqrt{n}(\mathrm{gCov}^n_{M_X} - \mathrm{gCov}_{M_X}(X, Y)) \xrightarrow{D} \mathcal{N}(0, \sigma_v^2), \tag{23}$$

$$\sqrt{n}(\mathrm{gCor}^n_{M_X} - \mathrm{gCor}_{M_X}(X, Y)) \xrightarrow{D} \mathcal{N}(0, \frac{\sigma_v^2}{\Delta^2}), \tag{24}$$

where σ_v^2 is given in the proof.

Under independence of X and Y, $\mathrm{gCov}^n_{M_X}$ and $\mathrm{gCor}^n_{M_X}$ converge in distribution, respectively, according to

$$n(\mathrm{gCov}^n_{M_X}) \xrightarrow{D} \sum_{l=1}^{\infty} \lambda_l(\chi^2_{1l} - 1), \tag{25}$$

$$n(\mathrm{gCor}^n_{M_X}) \xrightarrow{D} \frac{1}{\Delta} \sum_{l=1}^{\infty} \lambda_l(\chi^2_{1l} - 1), \tag{26}$$

where $\lambda_1, \ldots$ are non-negative constants dependent on F and $\chi^2_{11}, \chi^2_{12}, \ldots,$ are independent χ^2_1 variates.

Note that the boundedness of the positive definite kernel M implies the condition of $\mathbb{E}(d^2_M(X, X')) < \infty$.

Proof. We focus on a proof for the generalized Gini distance covariance and results for the correlation follow immediately from Slutsky's theorem [68] and the fact that $\hat{\Delta}$ is a consistent estimator of Δ.

Let $g(x) = \mathbb{E}d_M(x, X') - \mathbb{E}d_M(X, X')$. With the U-statistic theorem, we have

$$\sqrt{n}(\hat{\Delta} - \Delta) \xrightarrow{D} N(0, v^2),$$

where $v^2 = 4\mathbb{E}g^2(X) = 4\sum_k p_k \mathbb{E}g^2(X_k)$. Similarly, let $g_k(x) = \mathbb{E}d_M(x, X_k') - \mathbb{E}d_M(X_k, X_k')$ for $k = 1, 2, \ldots, K$ and $v_k^2 = 4\mathbb{E}g_k^2(X_k)/p_k$. We have

$$\sqrt{n}(\hat{\Delta}_k - \Delta_k) \xrightarrow{D} N(0, v_k^2).$$

Let $\boldsymbol{\Sigma}$ be the variance and covariance matrix for $\tilde{\boldsymbol{g}} = 2(g_1(X_1), \ldots, g_K(X_K), g(X))^T$, where $X = X_k$ with probability p_k. In other words, $\boldsymbol{\Sigma} = \mathbb{E}\tilde{\boldsymbol{g}}\tilde{\boldsymbol{g}}^T$. Denote $(\hat{\Delta}_1, \ldots, \hat{\Delta}_K, \hat{\Delta})^T$ as $\hat{\boldsymbol{\delta}}$ and $(\Delta_1, \ldots, \Delta_K, \Delta)^T$ as $\boldsymbol{\delta}$. From the U-statistic theorem [36], we have $\sqrt{n}(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}) \xrightarrow{D} N(\mathbf{0}, \boldsymbol{\Sigma})$. Let $\boldsymbol{b} = (-p_1, \ldots, -p_K, 1)^T$ be the gradient vector of $\mathrm{gCov}_{M_X}(X, Y)$ with respect to $\boldsymbol{\delta}$. Then $\sigma_v^2 = \boldsymbol{b}^T\boldsymbol{\Sigma}\boldsymbol{b} > 0$ under the assumption of dependence of X and Y, since

$$\begin{aligned} h(x) &:= \boldsymbol{b}^T\tilde{\boldsymbol{g}}(x) = 2\sum_k p_k(g(x_k) - g_k(x_k)) \\ &= 2\sum_k p_k(\mathbb{E}d_M(x_k, X) - \mathbb{E}d_M(x_k, X_k)) - 2(\Delta - \sum_k p_k\Delta_k) \\ &\neq 0, \end{aligned}$$

and $\sigma_v^2 = \sum_k p_k \mathbb{E}[h(X_k)^2]$. In this case, by the Delta method, $\sqrt{n}\boldsymbol{b}^T(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta})$ is asymptotically normally distributed with 0 mean and variance σ_v^2. With the result of $\hat{\boldsymbol{b}} = (-\hat{p}_1, \ldots, -\hat{p}_K, 1)^T$ being a consistent estimator of $\boldsymbol{b}$ and by the Slutsky's theorem, we have the same limiting normal distribution for $\mathrm{gCov}^n_{M_X} = \hat{\boldsymbol{b}}^T\hat{\boldsymbol{\delta}}$ as that of $\boldsymbol{b}^T\hat{\boldsymbol{\delta}}$. Therefore, the result of (23) is proved.

However, under the independence assumption, $\sigma_v^2 = 0$ because $h(x) = 0$, resulting from the same distribution of X and X_k. This corresponds to the degenerate case of U-statistics and $\boldsymbol{b}^T\hat{\boldsymbol{\delta}}$ has a mixture of χ^2 distributions [65]. Hence the result of (25) holds. □

One way to use the results of (23) and (24) is to test H_0 based on the confidence interval approach. More specifically, an asymptotically $(1 - \alpha)100$ percent confidence interval for $\mathrm{gCov}_{M_X}(X, Y)$ is

$$\mathrm{gCov}^n_{M_X}(X, Y) \pm Z_{1-\alpha/2}\frac{\hat{\sigma}_v^2}{\sqrt{n}},$$

where $\hat{\sigma}_v^2$ is a consistent estimator of σ_v^2 and $Z_{1-\alpha/2}$ is the $1 - \alpha/2$ quantile of the standard normal random variable. If this interval does not contain 0, we can reject H_0 at significance level $\alpha/2$. This test controls Type II error to be $\alpha/2$.

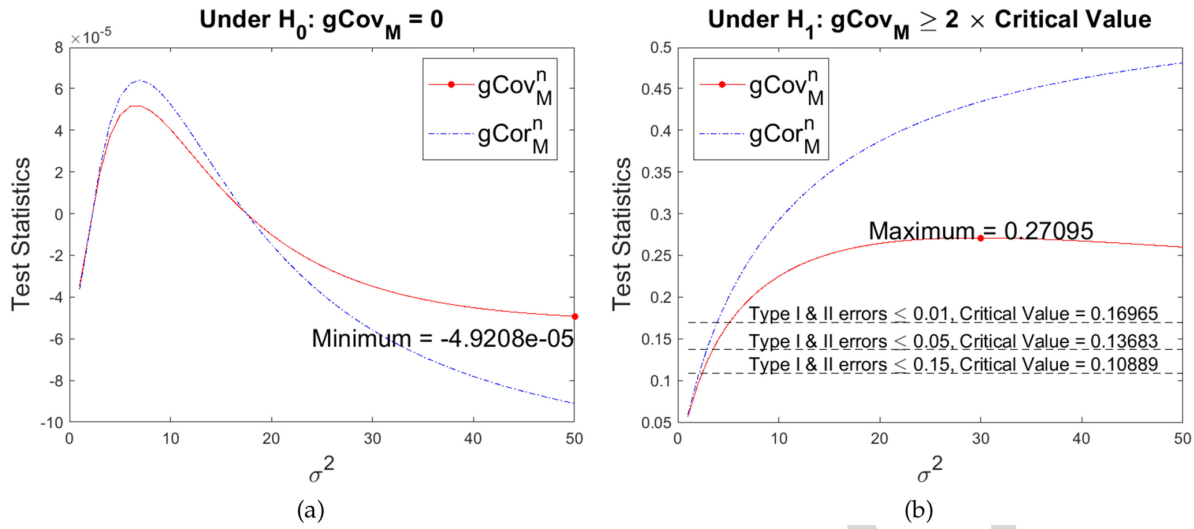


Fig. 1. Estimates of the generalized Gini distance covariance and generalized Gini distance correlation for different kernel parameters using 2000 iid samples: (a) independent case; (b) dependent case. Three critical values are shown in (b). They are calculated for significance levels 0.01, 0.05, and 0.15, respectively. In terms of the uniform convergence bounds, the optimal value of the kernel parameter σ is defined by the minimizer (or maximizer) of the test statistics under H_0 (or H_1).

On the other hand, if a test to control Type I error is preferred, we usually need to rely on a permutation test rather than the results of (25) and (26) since λ's depend on the distribution F, which is unknown. Details of the permutation test are in the next section.

4 An Algorithmic View

Although the uniform convergence bounds for generalized Gini distance covariance and generalized Gini distance correlation in Sections 3.2 and 3.3 are established upon the bounded kernel assumption, all the results also hold for Gini distance covariance and Gini distance correlation if the features are bounded. This is because when the features are bounded, they can be normalized so that $\sup_{x,x'}|x-x'|_q = 1$.

The calculation of test statistics (14) and (15) requires evaluating distances between all unique pairs of samples. Its time complexity is therefore $\Theta(n^2)$, where n is the sample size. In the one dimension case, i.e., $q = 1$, Gini distance statistics can be calculated in $\Theta(n\log n)$ time [20].[3] Note that distance covariance and distance correlation can also be calculated in $\Theta(n\log n)$ time [37]. Nevertheless, the implementation for Gini distance statistics is much simpler as it does not require the centering process.

Generalized Gini distance statistics are functions of the kernel parameter σ. Fig. 1a shows gCov_M^n and gCor_M^n of X_1 and Y_1 for $n = 2000$. The numerical random variable X_1 is generate from a mixture of two dimensional normal distributions: $N_1 \sim \mathcal{N}([1,2]^T, \mathrm{diag}[2,.5])$, $N_2 \sim \mathcal{N}([-3,-5]^T,\ \mathrm{diag}[1,1])$, and $N_3 \sim \mathcal{N}([-1,2]^T, \mathrm{diag}[2,2])$. The three components have equal mixing proportions. The categorical variable $Y_1 \in \{y_1, y_2, y_3\}$ is independent of X_1. The results in Fig. 1b are calculated from X_2 and Y_2 for $n = 2000$. The numerical random variable X_2 is generated by N_i if and only if $Y_2 = y_i$, $i = 1, 2, 3$. The categorical distribution of Y_2 is $\Pr(Y_2 = y_i) = \frac{1}{3}$. It is clear that X_2 and Y_2 are dependent on each other. Fig. 1 shows the impact of kernel parameter σ on the estimated generalized Gini distance covariance and Gini distance correlation. As a result, this affects the Type I and Type II error bounds given in Section 3.2. In this example, under H_0 (or H_1), the minimum (or maximum) gCov_M^n is achieved at $\sigma^2 = 50$ (or $\sigma^2 = 29$). These extremes yield tightest bounds in (18) and (19).[4] Note that gCov_M^n is an unbiased estimate of gCov_M. Although gCov_M can never be negative, gCov_M^n can be negative, especially under H_0.

This example also suggests that in addition to the theoretical importance, the inequalities in (18) and (19) may be directly applied to dependence tests. Given a desired bound (or significance level), α, on Type I and Type II errors, we call the value that determines whether H_0 should be rejected (hence to accept H_1) the *critical value* of the test statistic. Based on (18) and (19), the critical value for gCov_M^n, $\mathrm{cv}(\alpha, n)$, which is a function of α and the sample size n, is calculated as

$$\mathrm{cv}(\alpha, n) = \sqrt{\frac{12.5\log\frac{1}{\alpha}}{n}}.$$

The three horizontal dashed lines in Fig. 1b illustrate the critical values for $\alpha = 0.01$, $\alpha = 0.05$, and $\alpha = 0.15$, respectively. The population Gini distance covariance estimated using 20,000 iid samples is not included in the figure because of its closeness to gCov_M^n. With a proper choice of σ, H_1 should be accepted based on the 2000 samples of (X_2, Y_2) with both Type I and Type II errors no greater than 0.05. Note that we could not really accept H_1 at the level $\alpha = 0.01$ because the estimated maximum gCov_M is around 0.28, which is smaller than 0.3393 (two times the critical value at $\alpha = 0.01$).

The above test, although simple, has two limitations:

- Choosing an optimal σ is still an open problem. Numerical search is computationally expensive even if it is in one dimension;

3. When the inner produce kernel $M(x,x') = x^Tx'$ is chosen, generalized Gini distance statistics reduces to Gini distance statistics.

4. The kernel parameter σ also affects the Type I and Type II error bounds for gCor_M^n in (20) and (21). The Type I error bound for gCov_M^n is significantly tighter than that for gCor_M^n.

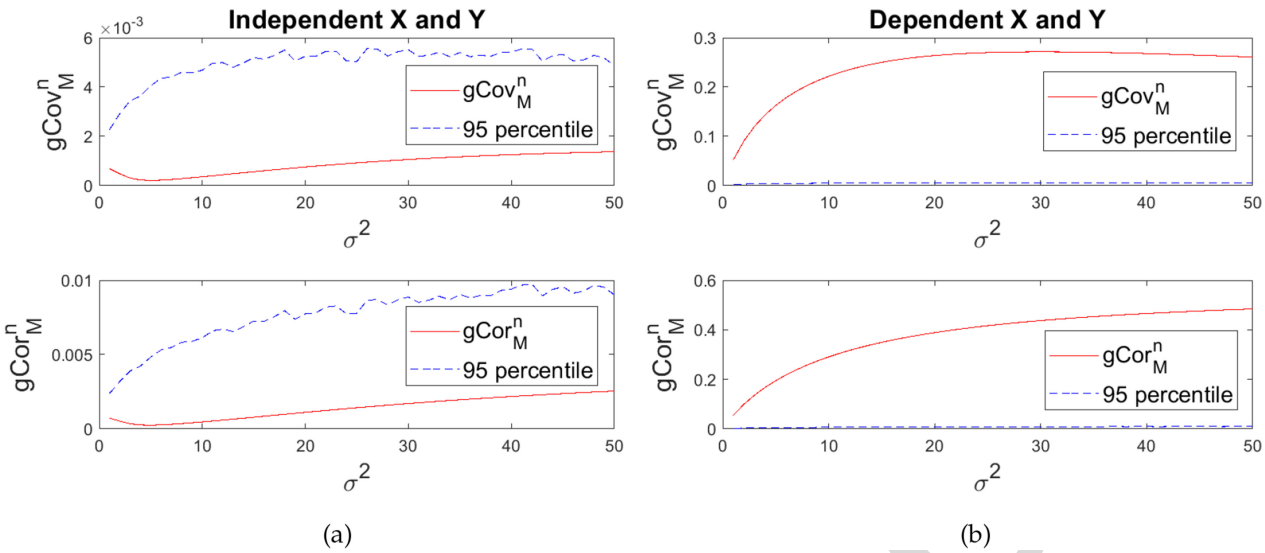


Fig. 2. Permutation tests of the generalized Gini distance covariance and generalized Gini distance correlation for different kernel parameters using 200 iid samples and 5000 random permutations: (a) independent case; (b) dependent case. The 95 percentile curves define the critical values for $\alpha = 0.05$. They are calculated from the permuted data. Test statistics higher (lower) than the critical value suggest accepting H_1 (H_0).

- The simplicity of the distribution free critical value $\mathrm{cv}(\alpha, n)$ comes with a price: it might not be tight enough for many distributions, especially when n is small.

Therefore, we apply permutation test [23], a commonly used statistical tool for constructing sampling distributions, to handle scenarios that the test based on $\mathrm{cv}(\alpha, n)$ is not feasible. We randomly shuffle the samples of X and keep the samples of Y untouched. We expect the generalized Gini distance statistics of the shuffled data should have values close to 0 because the random permutation breaks the dependence between samples of X and samples of Y. Repeating the random permutation many times, we may estimate the critical value for a given significance level α based on the statistics of the permuted data. A simple approach is to use the percentile defined by α, e.g., when $\alpha = 0.05$ the critical value is 95th percentile of the test statistic of the permuted data. The null hypothesis H_0 is rejected when the test statistic is larger than the critical value.

Fig. 2 shows the permutation test results of data generated from the same distributions used in Fig. 1. The plots on the top are gCov^n_M and the critical values. The plots at the bottom are gCor^n_M and the critical values. Test statistics are calculated from 500 samples. The critical values are estimated from 5000 random permutations at $\alpha = 0.05$. As illustrated in Fig. 2a, when X and Y are independent, the permutation tests do not reject H_0 at significance level 0.05. Fig. 2b shows that when X and Y are dependent, H_0 is rejected at significance level 0.05 by the permutation tests. It is interesting to note that the decision to reject or accept H_0 is not influenced by the value of the kernel parameter σ. [5]

5 Experiments

We first compare Gini distance statistics with distance statistics on artificial datasets where the dependent features are known. We then provide comparisons on real world datasets including the MNIST dataset, a breast cancer dataset and 19 publicly available datasets. For these real world datasets, we also include another three baseline methods: Pearson R^2, mean variance (MV) [18] and a direct average of squared MMD (avgMMD^2), i.e., $\frac{1}{K}\sum_{k=1}^{K} {\delta_M}^2(F, F_k)$, where M is the same Gaussian kernel used for Gini and distance statistics.

5.1 Simulation Results

In this experiment, we compare dependence tests using four statistics, dCov^n_M, dCor^n_M, gCov^n_M, and gCor^n_M, on artificial datasets. The kernel parameter was fixed at $\sigma^2 = 10$. The data were generated from three distribution families: normal, exponential, and Gamma distributions under both H_0 (X and Y are independent) and H_1 (X and Y are dependent). Given a distribution family, we first randomly choose a distribution F_0 and generate n iid samples of X. Samples of the categorical Y are then produced independent of X. Repeating the process, we create a total of m independent datasets under H_0. In the dependent case, X is produced by $F = \sum_{k=1}^{K} p_k F_k$, a mixture of K distributions where K is the number of different values that Y takes, p_k is the probability that $Y = y_k$, and F_k is a distribution from the same family that yields the data under H_0. The dependence between X and Y is established by the data generating process: a sample of X is created by F_k if $Y = y_k$. The mixture model is randomly generated, i.e., p_k and F_k are both randomly chosen. For each $Y = y_k$ ($k = 1, \ldots, K$), $n_k = n \cdot p_k$ iid samples of X are produced following F_k. This results in one data set of size $n = \sum_{k=1}^{K} n_k$ under H_1. Following the same procedure, we obtain m independent data sets under H_1 each corresponding to a randomly selected K-component mixture model F. In our experiments, $n = 100$ and $m = 10,000$.

Table 2 summarizes the model parameters of the three distribution families. $I(\cdot)$ is the indicator function. $\Gamma(\cdot)$ is the gamma function. $\mathcal{N}(\mu, \sigma)$ denotes the normal distribution with mean μ and standard deviation σ. $\Gamma(\alpha, \beta)$ denotes the gamma distribution with shape α and rate β. $\mathcal{U}(a, b)$ denotes the uniform distribution over interval $[a, b]$. $\mathrm{Dir}(\alpha)$ denotes the Dirichlet distribution with concentration α. A

5. Although the decision to reject or accept H_0 is not affected by σ, the p-value of the test does vary with respect to σ.

TABLE 2
Models of Different Distribution Families

	$p(x\|\theta)$	θ
Normal	$\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{\sigma^2}}$	$\mu \sim \mathcal{N}(0,5)$ $\sigma^2 \sim 1/\Gamma(1,1)$
Exponential	$\lambda e^{-\lambda x}\mathrm{I}(x \geq 0)$	$\lambda \sim \mathcal{U}(0,5)$
Gamma	$\frac{\beta^\alpha x^{\alpha-1}e^{-\beta x}}{\Gamma(\alpha)}\mathrm{I}(x \geq 0)$	$\alpha \sim \mathcal{U}(0,10)$ $\beta \sim \mathcal{U}(0,10)$
Proportions	$p_k \sim \mathrm{Dir}(1)$	

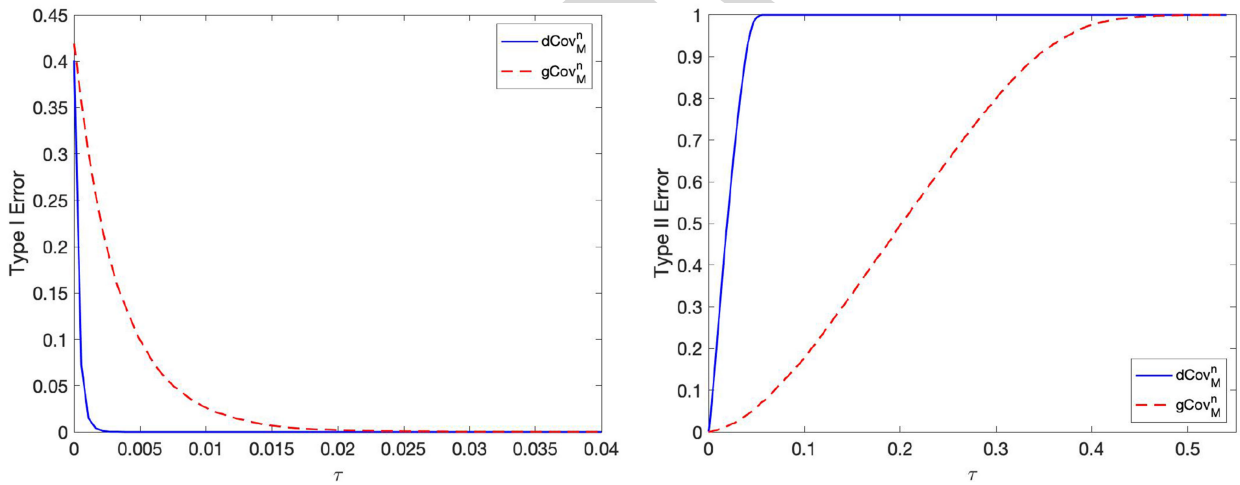
distribution (F_0 or F_k) is randomly chosen via its parameter (s). The unbiasedness of gCov_M^n requires that there are at least two data points for each value of Y. Therefore, random proportions that do not meet this requirement are removed.

Fig. 3 shows the performance of dCov_M^n and gCov_M^n in terms of type I and type II errors with the critical value τ set to different values. As Theorem 9 suggests, for any τ, gCov_M^n outperforms dCov_M^n in type II error. However, with the same value of τ, gCov_M^n underperforms dCov_M^n in terms of type I error. The results presented by Fig. 3 motivates us to compare Gini and distance statistics using power (with type I error α controlled at 0.05) and area under the curve (AUC). Both measures have values between 0 and 1 with a value closer to 1 indicating better performance. Table 3 illustrates the performance of the four test statistics under different values of K for the three distribution families. The highest power and AUC among the four test statistics are shown in bold and the second highest are underlined. In this experiment, gCov_M^n appears to be the most competitive test statistics in terms of power and ROC at all values of K. In addition, both gCov_M^n and gCor_M^n outperform dCov_M^n and dCor_M^n in most of the cases. We also tested the influence of σ^2 and observed stable results (figures are provided in supplementary materials, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/TPAMI.2019.2960358.).

5.2 The MNIST Dataset

We first tested feature selection methods using different test statistics on the MNIST data. The advantage of using an image dataset like MNIST is that we can visualize the selected pixels. We expect useful/dependent pixels to appear in the center part of the image. Some descriptions of the MNIST data are listed in Table 4.

The 7 test statistics under comparisons are: Pearson R^2, MV, avgMMD^2, dCov_M^n, dCor_M^n, gCov_M^n, and gCor_M^n. For each method, the top k features were selected by ranking



(a) (b)

Fig. 3. Simulation results using normal distribution with K = 3: (a) Type I error; (b) Type II error.

TABLE 3
Power ($\alpha = 0.05$) and AUC

		Power				AUC			
		dCov_M^n	dCor_M^n	gCov_M^n	gCor_M^n	dCov_M^n	dCor_M^n	gCov_M^n	gCor_M^n
$K=3$	Normal	0.991	0.993	**0.996**	**0.996**	0.998	0.998	**0.999**	**0.999**
	Exponential	0.666	0.669	**0.701**	0.681	0.871	0.875	0.880	**0.881**
	Gamma	0.956	0.960	**0.974**	0.971	0.988	0.989	**0.992**	**0.992**
$K=4$	Normal	0.999	0.999	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
	Exponential	0.737	0.734	**0.774**	0.756	0.908	0.909	**0.920**	0.918
	Gamma	0.987	0.987	**0.994**	0.992	0.997	0.997	**0.998**	**0.998**
$K=5$	Normal	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
	Exponential	0.790	0.776	**0.823**	0.805	0.931	0.930	**0.941**	0.939
	Gamma	0.995	0.993	**0.998**	0.996	**0.999**	**0.999**	**0.999**	**0.999**

TABLE 4
Data set summary

Data Set	Train/Test Size	Features	Classes
MNIST	60000/10000	784	10
Breast Cancer	405/101	17278	4
GDC PANCAN	2076/519	24981	12
Head and Neck Cancer	2010/502	23686	4
Pancreatic Cancer	77/19	18278	4
Medulloblastoma	228/57	33297	3
Gene Expression (UCI)	641/160	20531	5
Gastrointestinal Lesions	61/15	1396	3
Satellite	4435/2000	36	6
Ecoli	269/67	7	8
Glass	171/43	9	6
Urban Land Cover	168/507	147	9
Wine	142/36	13	3
Anuran Calls	5756/1439	22	4
Breast Tissue	85/21	9	4
Cardiotocography	1701/425	21	10
Leaf	272/68	14	30
Mice Protein Expression	864/216	77	8
HAR	4252/1492	561	6
UJIndoorLoc	19937/1111	520	13
Forest Types	198/325	27	4

the test statistics in descending order. Then the selected feature set was used to train the same classifier and the test accuracies were compared. The classifier used was a random forest consisting of 100 trees. We used the training and test set provided by [46] for training and testing. To reduce computation cost, we randomly selected 5000 samples to calculate test statistics for all methods. For avgMMD^2, Gini and distance statistics, each feature was standardized by subtracting the mean and dividing by the standard deviation. Pearson R^2 and MV are not affected by data standardization. The kernel parameter σ^2 was set to be 10. Because of the randomness involved in training random forest, each experiment was repeated 10 times and the average test accuracy was used for comparison.

The results of the MNIST data are summarized in Fig. 4. From Fig. 4a we can see the clear increasing trend in accuracy as more features are selected, as expected. Among all methods, Pearson R^2 performs the poorest. The discrepancy in accuracy between Pearson R^2 and the other six test statistics comes from the ability of characterizing non-linear dependence. The other five methods behave similar in terms of classification accuracy. Specifically, we expect dCov_M^n and gCov_M^n to be very similar because MNIST is a balanced dataset. Under the following two scenarios $\mathrm{dCov}_M(X,Y)$ and $\mathrm{gCov}_M(X,Y)$ will give the same ranking of the the features because their ratio is a constant (Remark 2.8 & 2.9 of [20]):

1) When the data is balanced, i.e., $p_1 = p_2 = ... = p_K = \frac{1}{K}$, $\mathrm{dCov}_M(X,Y) = \frac{1}{K}\mathrm{gCov}_M(X,Y)$;
2) When the data has only 2 classes, i.e., $K = 2$, $\mathrm{dCov}_M(X,Y) = 2p_1p_2\mathrm{gCov}_M(X,Y)$.

Hence, when n is sufficiently large, dCov_M^n and gCov_M^n will have the same ranking for the features.

The difference between Gini and distance statistics is more observable in the value range, as shown in Fig. 4b. Both dCor_M^n and gCor_M^n are bounded between 0 and 1, but clearly gCor_M^n takes a much wider range than dCor_M^n. Therefore, gCor_M^n is a more sensitive measure of dependence than dCor_M^n. gCov_M^n is also more sensitive than dCov_M^n as shown both empirically in Fig. 4b and theoretically by (11).

Fig. 4c shows the visualization of the selected pixels as white. Pearson R^2 and avgMMD^2 are not able to select some of the pixels in the center part even when k is greater than 400. Other four methods behave similar in this graph.

5.3 The Breast Cancer Dataset

We then compared the 7 feature selection methods on a gene selection task. The dataset used in this experiment was the TCGA breast cancer microarray dataset from the UCSC Xena database [28]. 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 [59]. In this experiment, we randomly hold-out 20 percent as test data, used each method to select top k genes, then evaluated the classification performance and compared the selected genes with the PAM50 gene signature. Because this dataset has a relatively small sample size, all training examples were used to calculate test statistics and train the classifier, and we repeated each classification test 30 times. Other experiment setups were kept the same as previous.

The results are shown in Fig. 5. Fig. 5a shows the classification performance using selected top k genes using different test statistics as well as using all PAM50 genes (shown as a green dotted line). Note that the accuracy of PAM50 is the averaged value from 30 runs. We plot it as a line across the entire k range for easier comparison with other methods. Among the 7 selection methods under comparison, gCov_M^n has the best overall performance, even outperforms PAM50 with $k = 30$. This suggests that gCov_M^n is able to select a smaller number of genes and the prediction is better than the gold standard. We also observe that gCor_M^n outperforms PAM50 with $k = 40$ and $k = 50$. Pearson R^2, MV, avgMMD^2, dCov_M^n, and dCor_M^n are not able to exceed PAM50 within 50 genes. Fig. 5b shows the number of PAM50 genes appear in the top k selected genes for each selection method. It is obvious to see that Pearson R^2 and avgMMD^2 select much smaller number of PAM50 compared to others. Gini statistics are able to select more PAM50 genes than distance statistics as k increases. The small ratio of PAM50 included in the selected genes by any method is because of the high correlation between genes. PAM50 was 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, both gCov_M^n and gCor_M^n are able to select a better gene set than PAM50 for classification.

5.4 Other Publicly Available Datasets

We further tested the 7 feature selection methods on a total of 19 publicly available datasets of classification tasks. The 19 datasets cover a wide range of data type, sample size, and feature set size. Specifically, we avoided binary-class and

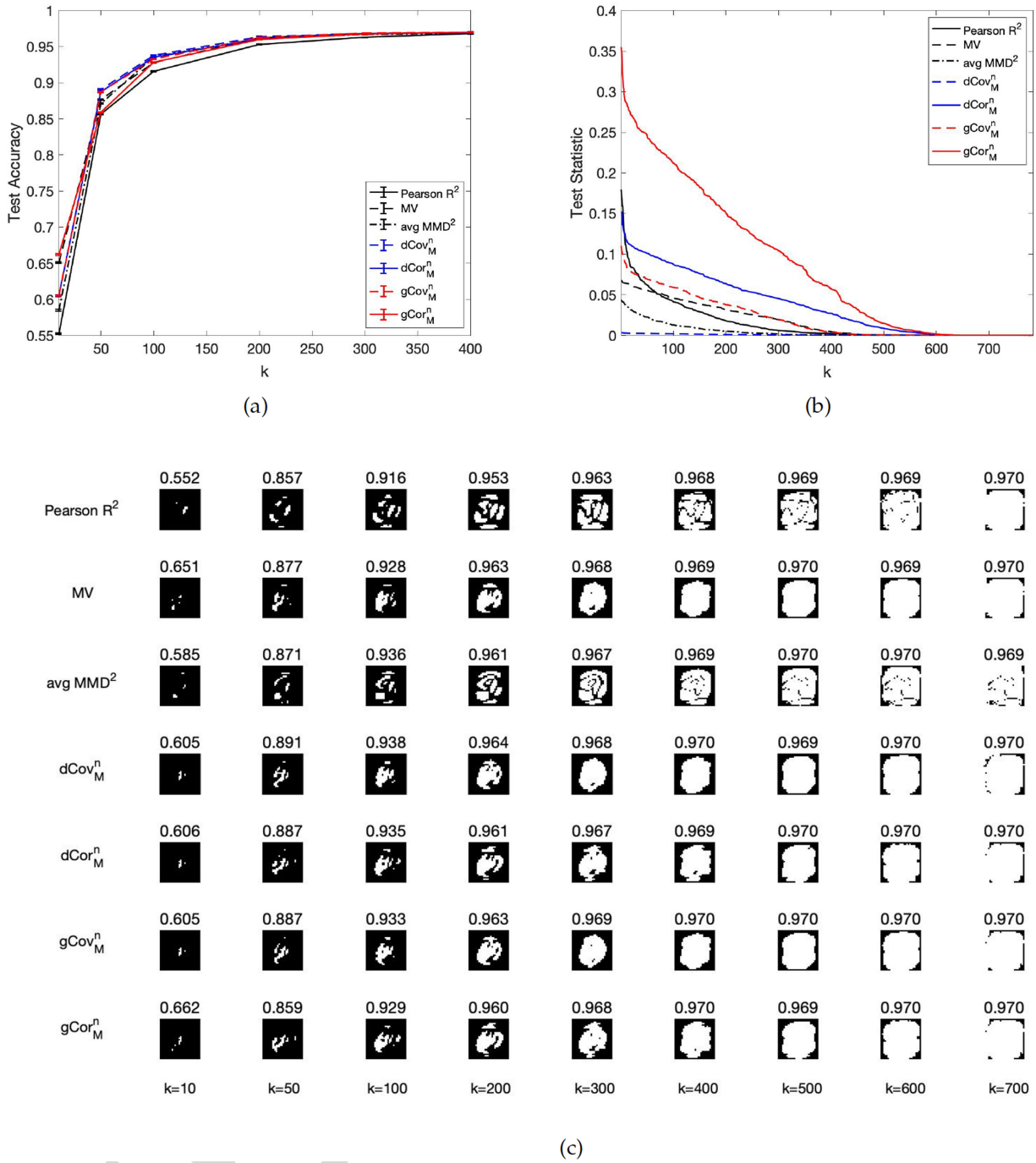


Fig. 4. The MNIST dataset. (a) Test accuracy using the top k selected features. (b) Test statistics of features in descending order. (c) Visualization of the top k pixels selected. White: selected. Black: not selected. Test accuracy using the selected pixels is labeled on the top of each image.

balanced datasets because dCov_M^n and gCov_M^n give the same ranking on these datasets when sufficient training samples are given. For datasets without training and test sets provided, we randomly hold out 20 percent as the test set. The descriptions of these datasets used are summarized in Table 4. GDC PANCAN, Head and Neck Cancer, Pancreatic Cancer and Medulloblastoma are gene datasets from the UCSC Xena database [28]. GDC PANCAN used DNA methylation features, Pancreatic Cancer used RNA-seq features, Head and Neck used single-cell RNA-seq features and Medulloblastoma used microarray features. The Gene Expression from UCI is also a gene data for PANCAN analysis but used RNA-seq features. The remaining datasets are all from UCI. For the UJIndoorLoc datasets, we randomly selected 5000 samples from the training set to calculate test statistics. For the remaining datasets, all training samples were used. For each dataset, each method was used to select top k features for training with three different values of k. Each classification test was repeated 10 times.

As we do not have the ground truth of the dependent features, only classification accuracy was used for evaluation. The average test accuracy (from 10 runs) of the 7 methods under comparison with different values of k on the 19 datasets are summarized in Table 5. Of all methods, the highest accuracy is shown in bold and the second highest one is underlined. The top 1 (top 2) statistic is determined by the number of times a method is shown in bold (bold or

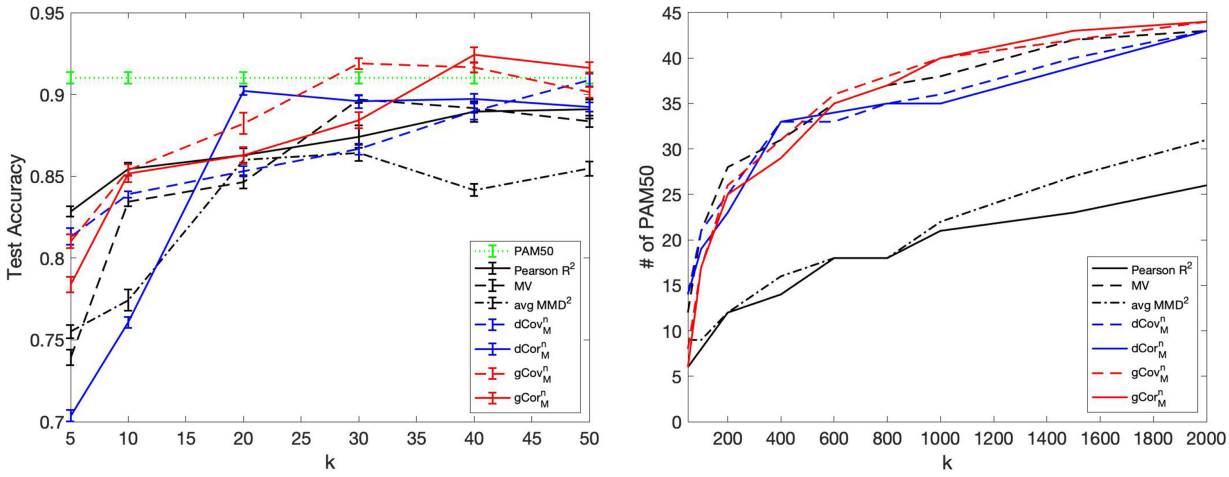


Fig. 5. The breast cancer dataset. (a) Test accuracy using the top k selected genes. (b) Number of PAM50 genes in the selected top k genes.

underlined). Among all methods, gCor_M^n appears 19 times as top 1 and 33 times in top 2, outperforming all other methods. MV, dCov_M^n, and gCov_M^n have similar performance, followed by avgMMD^2 and then dCor_M^n. Pearson R^2 is ranked the last, with 8 times as top 1 and 14 times in top 2. We observed that the performance of gCor_M^n is more superior on the gene datasets evaluated, namely, Breast Cancer, GDC PANCAN, Head and Neck Cancer, Pancreatic Cancer, Medulloblastoma, and Gene Expression (UCI). Gene datasets typically have a large number of features, which is usually an order of magnitude greater than the sample size, making selecting a small set of good features necessary yet challenging. The empirical results on five gene datasets suggests gCor_M^n to be a more competitive feature selection method than other methods under comparison.

6 Conclusion

We proposed a feature selection framework based on a new dependence measure between a numerical feature X and a categorical label Y using generalized Gini distance statistics: Gini distance covariance $\mathrm{gCov}(X,Y)$ and Gini distance correlation $\mathrm{gCor}(X,Y)$. We presented estimators of $\mathrm{gCov}(X,Y)$ and $\mathrm{gCor}(X,Y)$ using n iid samples, i.e., gCov_M^n and gCor_M^n, and derived uniform convergence bounds. We showed that gCov_M^n converge faster than its distance statistic counterpart dCov_M^n, and the probability of gCov_M^n under-performing dCov_M^n in terms of Type II error decreases to 0 exponentially as the sample size increases. gCov_M^n and gCor_M^n are also simpler to calculate than dCov_M^n and dCor_M^n. Extensive experiments were performed to compare gCov_M^n and gCor_M^n with other dependence measures in feature selection tasks using artifical and real world datasets, including MNIST, breast cancer and 19 publicly available daatsets. For simulated data, gCov_M^n and gCor_M^n perform better in terms of power and AUC. For real world datasets, on average, gCor_M^n is able to select more meaningful features and has better classification performances. Notice that the advantage of Gini statistics over distance statistics is less observable in real world datasets than in simulation settings. This is because for real world datasets the ground truth is unavailable and the difference is more difficult to see using classification accuracy as the performance measure. However, when the data dimension is sufficiently large, the difference between methods under comparison is more significant. As we see on the gene datasets, gCor_M^n is significantly better than the baseline methods. Therefore, we would recommend the use of gCor_M^n for high dimension data. In spite of the equivalence between gCov_M and a weighted average of squared MMD in $\widehat{M}$, gCov_M^n is superior to a direct average of squared MMD using the same kernel M in many settings, suggesting the importance of using weighted average and the choice of kernel.

The proposed feature selection method using generalized Gini distance statistics have several limitations:

- Choosing an optimal σ is still an open problem. In our experiments we used $\sigma^2 = 10$ after data standardization;
- The computation cost for gCov_M^n and gCor_M^n is $O(n^2)$, which is same as avgMMD^2, dCov_M^n and dCor_M^n, but more expensive than MV ($O(n\log n)$) and Pearson R^2 ($O(n)$). For large datasets, a sampling of data is desired.
- Features selected by Gini distance statistics, as well as other dependence measure based methods, can be redundant, hence a subsequent feature elimination may be needed for the sake of feature subset selection.

Acknowledgments

The authors would like to thank the reviewers and the associate editor for the suggestions to improve the quality of the paper.

References

[1] N. Armanfard, J. P. Reilly, and M. Komeili, "Local feature selection for data classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 6, pp. 1217–1227, Jun. 2016.
[2] N. Aronszajn, "Theory of reproducing kernels," *Trans. Amer. Soc.*, vol. 68, no. 3, pp. 337–404, 1950.
[3] A. Barbu, Y. She, L. Ding, and G. Gramajo, "Feature selection with annealing for computer vision and big data learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 2, pp. 272–286, Feb. 2017.
[4] L. Baringhaus and C. Franz, "On a new multivariate two-sample test," *J. Multivariate Anal.*, vol. 88, no. 1, pp. 190–206, 2004.
[5] R. Battiti, "Using mutual information for selecting features in supervised neural net learning," *IEEE Trans. Neural Netw.*, vol. 5, no. 4, pp. 537–550, Jul. 1994.

TABLE 5
Classification Accuracies Using Top k Dependent Features

	GDC PANCAN			Head and Neck Cancer			Pancreatic Cancer			Medulloblastoma		
k	5	10	15	3	5	7	10	30	50	5	7	10
Pearson R^2	0.600	0.791	0.841	0.886	0.936	0.959	0.611	0.711	0.774	**0.963**	0.930	0.951
MV	0.792	**0.874**	**0.892**	0.638	0.953	0.957	0.721	0.632	0.768	0.954	0.958	**0.970**
avg MMD^2	0.517	0.639	0.734	0.886	0.938	0.949	0.389	0.553	0.568	0.858	0.872	0.835
dCov_M^n	0.783	0.853	0.888	**0.907**	0.951	0.970	0.658	0.679	0.753	0.932	0.956	0.946
dCor_M^n	0.745	0.852	0.877	0.894	0.923	0.954	0.616	0.700	0.774	0.946	0.939	0.951
gCov_M^n	0.807	0.857	0.882	0.893	0.937	0.955	**0.847**	**0.842**	0.837	0.933	**0.965**	0.965
gCor_M^n	**0.826**	0.863	0.881	0.894	**0.982**	**0.982**	0.800	**0.842**	**0.874**	0.854	0.870	0.916

	Gene Expression (UCI)			Gastrointestinal Lesions			Satellite			Ecoli		
k	5	7	10	10	100	200	5	10	15	2	3	4
Pearson R^2	0.909	0.917	0.934	0.747	**0.700**	0.660	0.598	0.830	0.869	0.624	**0.772**	0.776
MV	**0.936**	0.961	**0.982**	0.660	0.660	0.673	0.831	0.885	0.900	0.713	0.760	0.781
avg MMD^2	0.711	0.696	0.819	0.547	0.680	**0.693**	0.770	0.834	0.860	0.624	**0.772**	**0.845**
dCov_M^n	0.889	0.907	0.963	0.700	0.607	0.673	0.627	0.860	0.896	0.712	0.766	0.788
dCor_M^n	0.851	0.928	0.937	**0.787**	0.633	0.680	0.808	**0.887**	**0.903**	0.713	0.764	0.787
gCov_M^n	0.923	0.933	0.948	0.560	0.647	**0.693**	0.834	0.870	0.881	**0.718**	0.766	0.781
gCor_M^n	0.741	**0.980**	0.972	0.607	**0.700**	0.653	**0.838**	0.873	0.898	0.634	0.758	0.803

	Glass			Urban Land Cover			Wine			Anuran Calls		
k	3	5	7	30	60	90	2	4	6	5	10	15
Pearson R^2	0.702	**0.730**	0.707	0.764	0.778	0.798	0.753	0.969	0.972	0.927	0.957	0.975
MV	0.702	0.667	**0.758**	0.782	0.799	0.805	**0.917**	0.992	0.997	0.933	**0.958**	**0.980**
avg MMD^2	**0.707**	0.670	0.744	0.778	0.796	0.799	0.758	0.903	0.942	0.934	**0.958**	0.979
dCov_M^n	0.702	0.691	0.744	0.786	0.795	**0.809**	0.897	**0.997**	**1.000**	**0.938**	0.957	0.979
dCor_M^n	**0.707**	0.672	0.751	0.781	**0.817**	**0.809**	0.881	0.992	0.997	0.937	0.956	0.978
gCov_M^n	0.705	0.674	0.663	0.784	0.790	0.804	0.881	**0.997**	0.997	**0.938**	0.956	0.979
gCor_M^n	0.693	0.665	0.670	**0.787**	0.804	0.805	0.900	**0.997**	**1.000**	0.937	**0.958**	**0.980**

	Breast Tissue			Cardiotocography			Leaf			Mice Protein Expression		
k	3	5	7	5	10	15	4	7	10	10	20	30
Pearson R^2	0.810	**0.857**	0.833	**0.831**	**0.890**	0.894	0.437	0.637	0.647	**0.978**	0.942	0.980
MV	0.810	**0.857**	0.843	0.805	0.880	0.893	0.494	0.576	0.676	0.868	0.964	0.982
avg MMD^2	0.810	**0.857**	**0.857**	0.675	0.860	0.894	0.443	0.601	0.676	0.977	**0.970**	0.980
dCov_M^n	**0.819**	**0.857**	0.852	0.773	0.874	0.891	0.471	**0.690**	0.663	0.893	0.950	**0.985**
dCor_M^n	0.810	**0.857**	0.838	0.816	0.849	0.894	**0.506**	0.606	0.704	0.890	0.952	0.977
gCov_M^n	0.810	**0.857**	0.852	0.817	0.878	**0.895**	0.468	0.688	0.654	0.888	0.945	0.983
gCor_M^n	0.810	**0.857**	**0.857**	0.766	0.880	0.887	0.503	0.594	**0.706**	0.888	0.943	0.982

	HAR			UJIndoorLoc			Forest Types			Top 1	Top 2
k	100	200	300	100	200	300	5	10	15	(times)	(times)
Pearson R^2	0.756	0.780	0.856	0.695	0.811	0.847	0.751	0.756	0.795	8	14
MV	0.719	0.778	0.775	0.753	0.864	**0.873**	0.699	0.758	0.806	11	25
avg MMD^2	0.781	0.832	**0.863**	0.731	**0.867**	0.869	**0.764**	0.757	0.794	11	18
dCov_M^n	0.765	0.779	0.859	0.768	0.866	**0.873**	0.652	0.809	**0.807**	11	22
dCor_M^n	0.767	0.780	0.853	0.793	0.863	0.871	0.715	0.752	0.802	8	19
gCov_M^n	0.766	**0.853**	0.858	0.763	0.864	0.872	0.651	**0.818**	0.804	11	26
gCor_M^n	**0.821**	**0.853**	0.853	**0.797**	0.864	0.872	0.696	0.734	0.806	**19**	**33**

[6] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2002, vol. 14, pp. 585–591.

[7] K. Benabdeslem and M. Hindawi, "Efficient semi-supervised feature selection: Constraint, relevance, and redundancy," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 5, pp. 1131–1143, May 2014.

[8] J. R. Berrendero, A. Cuevas, and J. L. Torrecilla, "Variable selection in functional data classification: A maxima-hunting proposal," *Statistica Sinica*, vol. 26, no. 2, pp. 619–638, 2016.

[9] A. L. Blum and P. Langley, "Selection of relevant features and examples in machine learning," *Artif. Intell.*, vol. 97, no. 12, pp. 245–271, 1997.

[10] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[11] M. Bressan and J. Vitrià, "On the selection and classification of independent features," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 10, pp. 1312–1317, Oct. 2003.

[12] S. Canu, "Functional learning through kernels," *Advances in Learning Theory: Methods, Models and Application*, J. Suykens, G. Horvath, S. Basu, C. Micchelli, J. Vandewalle (Ed.), Amsterdam, The Netherlands: IOS Press, 2003, pp. 89–110.

[13] R. Chakraborty and N. R. Pal, "Feature selection using a neural framework with controlled redundancy," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 1, pp. 35–50, Jan. 2015.

[14] Q. Cheng, H. Zhou, and J. Cheng, "The fisher-markov selector: Fast selecting maximally separable feature subset for multiclass classification with applications to high-dimensional data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 6, pp. 1217–1233, Jun. 2011.

[15] T. W. S. Chow and D. Huang, "Estimating optimal feature subsets using efficient estimation of high-dimensional mutual information," *IEEE Trans. Neural Netw.*, vol. 16, no. 1, pp. 213–224, Jan. 2005.

[16] P. Comon, "Independent component analysis, A new concept?" *Signal Process.*, vo. 36, no. 3, pp. 287–314, 1994.

[17] C. Constantinopoulos, M. K. Titsias, and A. Likas, "Bayesian feature and model selection for Gaussian mixture models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 6, pp. 1013–1018, Jun. 2006.

[18] H. Cui, R. Li, and W. Zhong,"Model-free feature screening for ultrahigh dimensional discriminant analysis," *J. Amer. Statistical Assoc.*, vol. 110, no. 510, pp. 630–641, 2015.

[19] B. B. Damodaran, N. Courty, and S. Lefèvre, "Sparse Hilbert Schmidt independence criterion and surrogate-kernel-based feature selection for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 4, pp. 2385–2398, Apr. 2017.

[20] X. Dang, D. Nguyen, Y. Chen, and J. Zhang "A new Gini correlation between quantitative and qualitative variables," 2018, *arXiv: 1809.09793*.

[21] P. Demartines and J. Hérault, "Curvilinear component analysis: A self-organizing neural network for nonlinear mapping of data sets," *IEEE Trans. Neural Netw.*, vol. 8, no. 1, pp. 148–154, Jan. 1997.

[22] A. A. Ding, J. G. Dy, Y. Li, and Y. Chang, "A robust-equitable measure for feature ranking and selection," *J. Mach. Learn. Res.*, vol. 18, pp. 1–46, 2017.

[23] E. Edgington and P. Onghena, *Randomization Tests*, 4th ed., London, UK/Boca Raton, FL, USA: Chapman and Hall/CRC, 2007.

[24] J. Fan and J. Lv, "Sure independence screen for ultrahigh dimensional feature space," *J. Roy. Statist. Soc., B*, vol. 70, no. 5, pp. 849–911, 2008.

[25] J. Fan, R. Samworth, and Y. Wu, "Ultrahigh dimensional feature selection: Beyond the linear model," *J. Mach. Learn. Res.*, vol. 10, pp. 2013–2038, 2009.

[26] C. Gini, *Variabilità e Mutabilità: Contributo Allo Studio Delle Distribuzioni e Relazioni Statistiche*, Vol. III (part II), Bologna, Italy: Tipografia di Paolo Cuppin, 1912.

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

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

[29] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *J. Mach. Learn. Res.*, vol. 13, pp. 723–773, 2012.

[30] J. Gui, Z. Sun, S. Ji, D. Tao, and T. Tan, "Feature selection based on structured sparsity: A comprehensive study," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 7, pp. 1490–1507, Jul. 2017.

[31] I. Guyon, and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, 2003.

[32] X. He, M. Ji, C. Zhang, and H. Bao, "A variance minimization criterion to feature selection using laplacian regularization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 10, pp. 2013–2025, Oct. 2011.

[33] G. E. Hinton and S. T. Roweis, "Stochastic neighbor embedding," in *Proc. Advances Neural Inf. Process. Syst.*, 2002, vol. 15, pp. 833–840.

[34] B. Hjørland, "The foundation of the concept of relevance," *J. Amer. Soc. Inf. Sci. Technol.*, vol. 61, no. 2, pp. 217–237, 2010.

[35] H. Hotelling, "Analysis of a complex of statistical variables into principal components," *J. Educ. Psychol.*, vol. 24, no. 6, pp. 417–441, 1933.

[36] W. Hoeffding, "Class of statistics with asymptotically normal distribution," *Ann. Math. Statist.*, vol. 19, pp. 293–325, 1948.

[37] X. Hou and G. Székely, "Fast computing for distance covariance," *Technometrics*, vol. 58, no. 4, pp. 435–447, 2016.

[38] F. J. Iannarilli Jr, and P. A. Rubin, "Feature selection for multiclass discrimination via mixed-integer linear programming," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 6, pp. 779–783, Jun. 2003.

[39] K. Javed, H. A. Babri, and M. Saeed, "Feature selection based on class-dependent densities for high-dimensional binary data," *IEEE Trans. Knowl. Data Eng.*, vol. 24, no. 3, pp. 465–477, Mar. 2012.

[40] G. H. John, R. Kohavi, and K. Pfleger, "Irrelevant feature and the subset selection problem," in *Proc. 11th Int. Conf. Mach. Learn.*, 1994, pp. 121–129.

[41] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artif. Intell.*, vol. 97, nos. 1–2, pp. 273–324, 1997.

[42] G. Koshevoy and K. Mosler, "Multivariate gini indices," *J. Multivariate Anal.* vol. 60, no. 2, pp. 252–276, 1997.

[43] B. Krishnapuram, A. J. Hartemink, L. Carin, and M. A. T. Figueiredo, "A Bayesian approach to joint feature selection and classifier design," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 9, pp. 1105–1111, Sep. 2004.

[44] N. Kwak and C. Choi, "Input feature selection for classification problems," *IEEE Trans. Neural Netw.*, vol. 13, no. 1, pp. 143–159, Jan. 2002.

[45] N. Kwak and C. Choi, "Input feature selection by mutual information based on Parzen window," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 12, pp. 1667–1671, Dec. 2002.

[46] Y. LeCun and C. Cortes, "MNIST handwritten digit database," 2010.

[47] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999.

[48] L. Lefakis and F. Fleuret, "Jointly informative feature selection made tractable by Gaussian modeling," *J. Mach. Learn. Res.*, vol. 17, pp. 1–39, 2016.

[49] R. Li, W. Zhong, and L. Zhu, "Feature screening via distance correlation learning," *J. J. Amer. Statistical Assoc.*, vol. 107, no. 499, pp. 1129–1139, 2012.

[50] H. Liu and L. Yu, "Toward integrating feature selection algorithms for classification and clustering," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 4, pp. 491–502, Apr. 2005.

[51] R. Lyons, "Distance covariance in metric spaces," *Ann. Probability*, vol. 41, no. 5, pp. 3284–3305, 2013.

[52] P. Maji and S. K. Pal, "Feature selection using f-information measures in fuzzy approximation spaces," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 6, pp. 854–867, Jun. 2010.

[53] Q. Mao and I. W. Tsang, "A feature selection method for multivariate performance measures," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 9, pp. 2051–2063, Sep. 2013.

[54] J. Mercer, "Functions of positive and negative type, and their connection the theory of integral equations," *Philos. Trans. Roy. Soc. A*, vol. 209, pp. 415–446, 1909.

[55] P. Mitra, C. A. Murthy, and K. Pal, "Unsupervised feature selection using feature similarity," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 3, pp. 301–312, Mar. 2002.

[56] T. Naghibi, S. Hoffmann, and B. Pfister, "A semidefinite programming based search strategy for feature selection with mutual information measure," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 8, pp. 1529–1541, Aug. 2015.

[57] R. Nilsson, J. M. Peña, J. Björkegren, and J. Tegnér, "Consistent feature selection for pattern recognition in polynomial time," *J. Mach. Learn. Res.*, vol. 8, pp. 589–612, 2007.

[58] J. Novovicová, P. Pudil, and J. Kittler, "Divergence based feature selection for multimodal class densities," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 18, no. 2, pp. 218–223, Feb. 1996.

[59] J. S. Parker *et al.*, "Supervised risk predictor of breast cancer based on intrinsic subtypes," *J. Clin. Oncology*, vol. 27, no. 8, pp. 1160–1167, 2009.

[60] K. Pearson, "Notes on regression and inheritance in the case of two parents," *Proc. Roy. Soc. London*, vol. 58, pp. 240–242, 1895.

[61] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005.

[62] S. Ren, S. Huang, J. Ye, and X. Qian, "Safe feature screening for generalized LASSO," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol., 40, no. 12, pp. 2992–3006, Dec. 2018.

[63] S. T. Roweis and L. L. Saul, "Locally linear embedding," *Sci.*, vol. 290, pp. 2323–2326, 2000.

[64] D. Sejdinovic, B. Sriperumbudur, A. Gretton, and K. Fukumizu, "Equivalence of distance-based and RKHS-based statistics in hypothesis testing," *Ann. Statist.*, vol. 41, no. 5, pp. 2263–2291, 2013.

[65] R. Serfling, *Approximation Theorems of Mathematical Statistics*," New York, NY, USA: Wiley, 1980.

[66] M. Shah, M. Marchand, and J. Corbeil, "Feature selection with conjunctions of decision stumps and learning from microarray data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 1, pp. 174–186, Jan. 2012.

[67] V. Sindhwani, S. Rakshit, D. Deodhare, D. Erdogmus, J. C. Principe, and P. Niyogi, "Feature selection in MLPs and SVMs based on maximum output information," *IEEE Trans. Neural Netw.*, vol. 15, no. 4, pp. 937–948, Jul. 2004.

[68] E. Slutsky, "Über Stochastische Asymptoten und Grenzwerte," *Metron* (in German), vol. 5, no. 3, pp. 3–89, 1925.

[69] A. J. Smola, A. Gretton, L. Song, and B. Schölkopf, "A Hilbert space embedding for distributions," in *Proc. Conf. Algorithmic Learn. Theory*, 2007, vol. 4754, pp. 13–31.

[70] P. Somol, P. Pudil, and J. Kittler, "Fast branch & bound algorithms for optimal feature selection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 7, pp. 900–912, Jul. 2004.

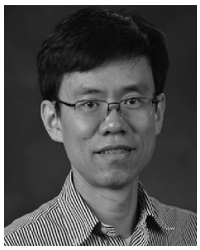
[71] L. Song, A. Smola, A. Gretton, J. Bedo, and K. Borgwardt, "Feature selection via dependence maximization," *J. Mach. Learn. Res.*, vol. 13, pp. 1393–1434, 2012.
[72] H. Stoppiglia, G. Dreyfus, R. Dubois, and Y. Oussar, "Ranking a sandom feature for variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1399–1414, 2003.
[73] Y. Sun, S. Todorovic, and S. Goodison, "Local-learning-based feature selection for high-dimensional data analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1610–1626, Sep. 2010.
[74] G. J. Székely and M. L. Rizzo, "Testing for equal distributions in high dimension," *InterStat*, vol. 5, pp. 1249–1272, 2004.
[75] G. J. Székely and M. L. Rizzo, "A new test for multivariate normality," *J. Multivariate Anal.*, vol. 93, no. 1, pp. 58–80, 2005.
[76] G. J. Székely, M. L. Rizzo, and N. K. Bakirov, "Measuring and testing dependence by correlation of distances," *Ann. Statist.*, vol. 35, no. 6, pp. 2769–2794, 2007.
[77] G. J. Székely and M. L. Rizzo, "Brownian distance covariance," *Ann. Appl. Statist.*, vol. 3, no. 4, pp. 1233–1303, 2009.
[78] G. J. Székely and M. L. Rizzo, "Energy statistics: A class of statistics based on distances," *J. Statist. Planning Inference*, vol. 143, no. 8, pp. 1249–1272, 2013.
[79] G. J. Székely and M. L. Rizzo, "Partial distance correlation with methods for dissimilarities," *Ann. Statist.*, vol. 42, no. 6, pp. 2382–2412, 2014.
[80] J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Sci.*, vol. 290, pp. 2319–2323, 2000.
[81] W. S. Torgerson, "Multidimensional scaling I: Theory and method," *Psychometrika*, vol. 17, no. 4, pp. 401–419, 1952.
[82] E. Tuv, A. Borisov, G. Runger, and K. Torkkola, "Feature selection with ensembles, artificial variables, and redundancy elimination," *J. Mach. Learn. Res.*, vol. 10, pp. 1341–1366, 2009.
[83] H. Wang, D. Bell, and F. Murtagh, "Axiomatic approach to feature subset selection based on relevance," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 21, no. 3, pp. 271–277, Mar. 1999.
[84] J. Wang, P. Zhao, S. C. H. Hoi, and R. Jin, "Online feature selection and its applications," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 3, pp. 698–710, Mar. 2014.
[85] K. Wang, R. He, L. Wang, W. Wang, and T. Tan, "Joint feature selection and subspace learning for cross-modal retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 10, pp. 2010–2023, Oct. 2016.
[86] L. Wang, "Feature selection with kernel class separability," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 9, pp. 1534–1546, Sep. 2008.
[87] L. Wang, N. Zhou, and F. Chu, "A general wrapper approach to selection of class-dependent features," *IEEE Trans. Neural Netw.*, vol. 19, no. 7, pp. 1267–1278, Jul. 2008.
[88] H. Wei and S. A. Billings, "Feature Subset Selection and Ranking for Data Dimensionality Reduction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 1, pp. 162–166, Jan. 2007.
[89] X. Wu, K. Yu, W. Ding, H. Wang, and X. Zhu, "Online feature selection with streaming features," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 5, pp. 1178–1192, May 2013.
[90] Z. J. Xiang, Y. Wang, and P. J. Ramadge, "Screeningtests for Lasso problems," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 5, pp. 1008–1027, May 2017.
[91] S. Yang and B. Hu, "Discriminative feature selection by nonparametric Bayes error minimization," *IEEE Trans. Knowl. Data Eng.*, vol. 24, no. 8, pp. 1422–1434, Aug. 2012.
[92] C. Yao, Y. Liu, B. Jang, J. Han, and J. Han, "LLE score: A new filter-based unsupervised feature selection method based on nonlinear manifold embedding and its application to image recognition," *IEEE Trans. Image Process.*, vol. 26, no. 11, pp. 5257–5269, Nov. 2017.
[93] L. Yu and H. Liu, "Efficient feature selection via analysis of relevance and redundancy," *J. Mach. Learn. Res.*, vol. 5, pp. 1205–1224, 2004.
[94] H. Zeng and Y. Cheung, "Feature selection and kernel learning for local learning-based clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 8, pp. 1532–1547, Aug. 2011.
[95] Z. Zhao, L. Wang, H. Liu, and J. Ye, "On similarity preserving feature selection," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 3, pp. 619–632, Mar. 2013.
[96] Y. Zhai, Y. Ong, and I. Tsang, "Making trillion correlations feasible in feature grouping and selection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 12, pp. 2472–2486, Dec. 2016.
[97] L. Zhou, L. Wang, and C. Shen, "Feature selection with redundancy-constrained class separability," *IEEE Trans. Neural Netw.*, vol. 21, no. 5, pp. 853–858, May 2010.

Silu Zhang received the BS degree in bioengineering from Zhejiang University, China, the MS degree in chemical engineering from North Carolina State University, both in 2011, and the second MS and the PhD degrees in computer science from the University of Mississippi, in 2017 and 2019. She is currently a diagnostic imaging research scientist with St. Jude children's research hospital, working on feature extraction and selection from MRI images and clustering analysis of brain tumor subtypes.

Xin Dang received the BS degree in applied mathematics from Chongqing University, China, in 1991, the master's and the PhD degrees in statistics from the University of Texas at Dallas, in 2003, and 2005. Currently, she is a professor of the Department of Mathematics, University of Mississippi. Her research interests include robust and nonparametric statistics, statistical and numerical computing, and multivariate data analysis. In particular, she has focused on data depth and applications, bioinformatics, machine learning, and robust procedure computation. She is a member of the IMS, ASA, ICSA, and the IEEE.

Dao Nguyen received the bachelor of computer science degree from the University of Wollongong, Australia, in 1997, the PhD degree in electrical engineering from the University of Science and Technology, Korea, in 2010, and the second PhD degree in statistics from the University of Michigan, Ann Arbor, in 2016. He was a postdoc scholar at the University of California, Berkeley for more than a year before becoming an assistant professor of mathematics at the University of Mississippi, in 2017. His research interests include machine learning, dynamics modeling, stochastic optimization, Bayesian analysis. He is a member of the ASA, ISBA.

Dawn Wilkins received the BA and MA degrees in mathematical systems from Sangamon State University (now the University of Illinois–Springfield), in 1981 and 1983, respectively, and the PhD degree in computer science from Vanderbilt University, in 1995. She joined the faculty of the University of Mississippi, where she is currently the professor of computer and information science. Her research interests are primarily in machine learning, computational biology, bioinformatics and database systems.



Yixin Chen received the BS degree from the Department of Automation, Beijing Polytechnic University, in 1995, the MS degree in control theory and application from Tsinghua University, in 1998, the MS and PhD degrees in electrical engineering from the University of Wyoming, in 1999 and 2001, and the another PhD degree in computer science from the Pennsylvania State University, in 2003. He had been an assistant professor of computer science at the University of New Orleans. He is currently a professor of computer and information science with the University of Mississippi. His research interests include machine learning, data mining, computer vision, bioinformatics, and robotics and control. He is a member of the ACM, IEEE, and IEEE Computer Society.