A NEW TEST OF INDEPENDENCE AND ITS APPLICATION TO VARIABLE SELECTION

by

Haeun Moon

B.S in Mathematics, KAIST, 2012

M.A in Economics, Seoul National University, 2015

Submitted to the Graduate Faculty of

the Dietrich School of Arts and Sciences in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2022

UNIVERSITY OF PITTSBURGH

DIETRICH SCHOOL OF ARTS AND SCIENCES

This dissertation was presented

by

Haeun Moon

It was defended on

April 25th 2022

and approved by

Kehui Chen, PhD, Department of Statistics

Yu Cheng, PhD, Department of Statistics

Zhao Ren, PhD, Department of Statistics

Ying Ding, PhD, Department of Biostatistics

Dissertation Director: Kehui Chen, PhD, Department of Statistics

A NEW TEST OF INDEPENDENCE AND ITS APPLICATION TO VARIABLE SELECTION

Haeun Moon, PhD

University of Pittsburgh, 2022

In the first part of our research, we propose a new interpoint-ranking sign covariance measure for nonparametric test of independence. The proposed method is applicable to general types of random objects as long as a meaningful similarity measure can be defined, and it is shown to be zero if and only if the two random variables are independent. The test statistic is a U-statistic, whose large sample behavior guarantees that the proposed test is consistent against general types of alternatives. Numerical experiments and data analyses demonstrate the great empirical performance of the proposed method. In the second part, we propose to combine the frequent voting idea with the proposed and existing test of independence methods for model-free variable selection. This research is motivated and illustrated by an application in selecting important genes related to suicidal behaviors. Numerical experiments demonstrate nice empirical performance of the proposed method.

keyword: Consistent; Independence Test; Interpoint distance; Nonparametric; Sign Covariance; Model-free selection.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1.0 INTRODUCTION

The first part of the thesis focuses on developing a rank based statistic for measuring the association and statistical dependence between two outcomes of interest, where the outcomes are not measured by a real number, but measured with multi-dimensional quantities, curves, images or more general types of data. Traditional rank-based statistics such as Kendall's Tau in one-dimensional cases are widely used. However, the extension to multi-dimensional cases and general types of data is a challenging problem, because the ordering is not well defined for data beyond real numbers. This difficulty has been explicitly noted in several recent papers. We have worked out a method based on interpoint distances for this problem. We propose to first compute the interpair rankings based on the similarity matrix, where for each fixed data point, order all the other data points according to the similarity to this particular point, then we compute a rank-based dependence measure with respect to this interpair ranking, and finally summarize over all data points to obtain the final statistic. We are able to prove that the proposed test can detect general types of association/dependency, meaning that the computed statistic will be away from zero if and only if there exists association/dependency between the two random objects. This research is summarized in chapter 2 and chapter 3. The second part of the thesis concerns the application of the independence test and dependence measures in scientific data analysis, especially for the variable selection problem. We combine the resampling procedure with the test of independence and suggest a new selection procedure based on the frequency of voting. In chapter 4, we evaluate the finite sample performance of the proposed method in simulated data. We also illustrate that the method can be applied to select groups of variables with potentially different dimensions through a data application, where we select important genes for suicidal behaviors.

2.0 BACKGROUND AND REVIEW OF INDEPENDENT TESTS

Let X and Y be random variables with marginal distributions P_X on $\mathcal{X}$ and P_Y on $\mathcal{Y}$, respectively, and joint distribution P_{XY} on $\mathcal{X} \times \mathcal{Y}$. Our research aim to test

$$H_0 : P_{XY} = P_X P_Y \text{ versus } H_1 : P_{XY} \neq P_X P_Y$$

based on samples $(X_1, Y_1),\ldots,(X_n, Y_n) \in \mathcal{X} \times \mathcal{Y}$ of size n drawn independently and identically from P_{XY}. This fundamental statistical question has received much attention with a wide range of applications. Here we consider nonparametric independence tests and no distributional assumptions will be made on P_X, P_Y or P_{XY}.

Let $\mathcal{P}_0$, $\mathcal{P}_1$ be collection of distributions of X and Y satisfying H_0 and H_1, respectively. Assume we develop a statistical test $\psi_n : \{\mathcal{X}_n \times \mathcal{Y}_n\} \to \{0, 1\}$ with a test statistic $T^{(n)}$. At a significance level α, we decide whether to accept H_0 or not by comparing $T^{(n)}$ with a critical value C_α, i.e., $\psi_n = \mathbf{1}\{T^{(n)} \geq C_\alpha\}$. A desirable independence test should at least control the type I error, i.e.,

$$\lim_n \sup_{P \in \mathcal{P}_0} E(\Psi_n = 1) \leq \alpha. \tag{2.1}$$

This can be achieved by deriving the null distribution of $T^{(n)}$ and using the upper $(1-\alpha)$-th quantile of the null distribution as the critical value C_α. Alternatively, a brute force solution is to use a permutation test, where the critical value C_α is taken to be the the upper $(1-\alpha)$-th quantile of the $T^{(n)}$ values from the permuted data.

For an independence test to be useful, a further requirement is that the test should have non-trivial power for a general class of $\mathcal{P}_1$. To achieve this, ideally the population counterpart of T^n (or a scaled version $b_n T_n$), denoted by θ, should serve as a meaningful measure of dependence and have an independence-zero equivalence property. The test statistic T^n should

converge to its population version with a desirable rate. We summarize the requirement as follows.

(i) $\theta = 0$ if $P \in \mathcal{P}_0$ and $\theta > 0$ if $P \in \mathcal{P}_1$.

(ii) $b_n T^{(n)}$ converges to θ as $n \to \infty$.

However, this requirement is not easy to achieve for a general collection of $\mathcal{P}_1$. Even in the univariate case, $(\mathcal{X}, \mathcal{Y}) = (\mathbb{R}, \mathbb{R})$, classical measures of association such as Pearson correlation (Pearson, 1895), Kendall's τ (Kendall, 1938), and Spearman's ρ (Spearman, 1904) could be zero even in the presence of an association between random variables, i.e., tests based on these coefficients are not consistent against general types of alternatives. The problem becomes more challenging and remains largely unsolved for multivariate data and more general random objects until some breakthrough work in recent years. In this chapter, we provide a review of recent developments in the test of independence. We draw particular attention to a rank-based nonparametric test, sign covariance test (denoted as τ^*), which was introduced as a modification of Kendall's τ by Bergsma and Dassios (2014). Being based on the concordance and discordance of four points rather than two points as τ does, the test based on τ^* is consistent against general types of alternatives and meanwhile, it enjoys robustness, simplicity and interpretability (Bergsma and Dassios, 2014; Nandy et al., 2016; Dhar et al., 2016; Weihs et al., 2018). However, the original paper (Bergsma and Dassios, 2014) only developed tests for bivariate distributions of (X, Y), and they particularly noted the difficulty in generalizing to multivariate settings.

2.1 RANK BASED TESTS

Let $(X, Y) \in (\mathcal{X}, \mathcal{Y})$ be random variables and (X^i, Y^i) are $i.i.d$ copies of (X, Y). For a simple case, $(\mathcal{X}, \mathcal{Y}) = (\mathbb{R}, \mathbb{R})$, tests based on Kendall's τ (Kendall, 1938) ($:=\mathbb{E}sign(X^1 - X^2)(Y^1 - Y^2)$) and Spearman's ρ (Spearman, 1904) ($:=3\mathbb{E}sign(X^1 - X^2)(Y^1 - Y^3)$) have been used widely. Despite their simplicity and interpretability, it has been shown that they could be zero under the nonlinear type of alternatives i.e., these tests are not consistent.

Two alternative tests based on Hoeffding's D coefficient (Hoeffding, 1948) and Blum-Kiefer-Rosenblatt's R coefficient (Blum et al., 1961) were developed with better consistency guarantee. Denoting marginal distributions of X and Y as F_X and F_Y, respectively, and joint distribution as F_{XY}, Hoeffding's D is given as

$$D = \int (F_{XY}(x,y) - F_X(x)F_Y(y))^2 dF_{XY}(x,y), \tag{2.2}$$

and Blum-Kiefer-Rosenblatt's R is given as

$$R = \int (F_{XY}(x,y) - F_X(x)F_Y(y))^2 dF_X(x)F_Y(y). \tag{2.3}$$

These coefficients basically measure the distance between two cumulative distribution functions, $F_{XY}(x,y)$ and $F_X(x)F_Y(y)$, with respect to some measure on $\mathcal{X} \times \mathcal{Y}$. These are Cramer-von Mises type distances, which are widely used in two-sample testing problems. The Hoeffding's D coefficient satisfies independence-zero property for bivariate continuous distributions, and the BKR-R coefficient satisfies independence-zero property for both continuous and discrete variables. In univariate cases, these coefficients can be estimated by rank-based U-statistics.

A recent paper Bergsma and Dassios (2014) proposed a sign covariance as a new measure of dependence between two random variables.

The population version of sign covariance is defined as

$$\tau^*(X,Y) = \mathbb{E}a(X^1, X^2, X^3, X^4)a(Y^1, Y^2, Y^3, Y^4), \tag{2.4}$$

where

$$a(z_1, z_2, z_3, z_4) = sign(|z_1 - z_2| + |z_3 - z_4| - |z_1 - z_3| - |z_2 - z_4|) \tag{2.5}$$

for $z_1, z_2, z_3, z_4 \in \mathbb{R}$.

The main theorem in Bergsma and Dassios (2014) states that when (X,Y) has a bivariate discrete or continuous distribution, or a mixture of the two, $\tau^*(X,Y) \geq 0$ with equality if and only if X and Y are independent. Moreover, the authors of that paper conjectured that this property holds in general without continuous or discrete conditions. Actually, the

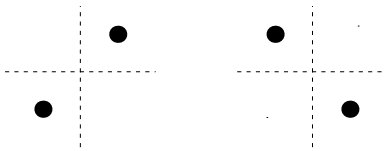


Figure 1: Kendall's τ: Configurations of concordant (left) and discordant (right) quadruples for two observations.

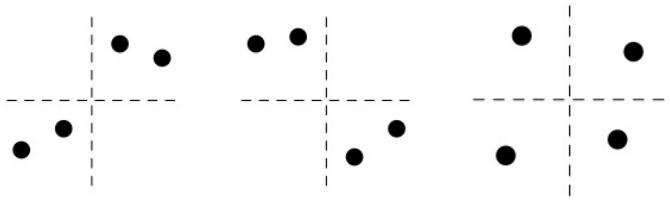
conditions have been eased later, though not completely, to the extent that X and Y have both continuous marginal distributions (Drton et al., 2018).

This sign covariance τ^* can be viewed a modified Kendall's τ. From the definition of Kendall's τ, it can be shown that

$$\tau^2 = \mathbb{E}s(X^1, X^2, X^3, X^4)s(Y^1, Y^2, Y^3, Y^4)$$

where $s(z_1, z_2, z_3, z_4) = sign(|z_1 - z_2|^2 + |z_3 - z_4|^2 - |z_1 - z_3|^2 - |z_2 - z_4|^2)$. This becomes τ^* if we replace the squared distance $|\cdot|^2$ with the absolute value $|\cdot|$. In addition, Kendall's τ has an equivalent formula $\tau = \Pi_{C_2} - \Pi_{D_2}$, where Π_{C_2} and Π_{D_2} denote the probabilities that two randomly chosen points are concordant and discordant, respectively, where the term concordant and discordant for two points is illustrated in Figure 1. The sign covariance τ^* also has an equivalent formula, $\tau^* = (2\Pi_{C_4} - \Pi_{D_4})/3$, where Π_{C_4} and Π_{D_4} denote the probabilities that four randomly chosen points are concordant and discordant, respectively, where the term concordant and discordant for four points is illustrated in Figure 2. Under independence, $\Pi_{C_4} = 1/3$ and $\Pi_{D_4} = 2/3$. If one variable is a strictly monotone function of the other, then $\Pi_{C_4} = 1$ and $\Pi_{D_4} = 0$. Moreover, Drton et al. (2018) proved that for bivariate normal data, the sign covariance is a monotone function of the correlation $|\rho|$.

In the univariate case, the sign covariance τ^* can be estimated by a rank-based U-statistic with order 4. The test of independence based on τ^* is proven to be consistent against general alternatives and enjoy robustness, simplicity and interpretability (Bergsma and Dassios, 2014; Nandy et al., 2016; Dhar et al., 2016; Weihs et al., 2018). However, the generalization to multivariate variables is not easy. The authors mentioned the definition of τ^* can be straightforwardly extended to variables in an arbitrary metric space, by defining



(a) Concordant (b) Discordant

Figure 2: Sign Covariance τ^*: Configurations of concordant (left) and discordant (right) quadruples for four observations.

$a_d(z_1, z_2, z_3, z_4) = sign(d(z_1, z_2) + d(z_3, z_4) - d(z_1, z_3) - d(z_2, z_4))$. But in this case, τ^* might be smaller than zero and the consistency of the τ^* test does not hold any more. In Chapter 3, we extend the sign covariance to work for general types of random objects while preserving a general consistency, and we also give the general form to generalize other coefficients, including the Hoeffding's D coefficient (Hoeffding, 1948) and the Blum-Kiefer-Rosenblatt's R coefficient (Blum et al., 1961).

Generalizing rank-based tests from univariate data to multivariate data is an active research area. A straightforward extension was explored in Leung et al. (2018) and Drton et al. (2018), which derived a family of test statistics for testing mutual independence by collecting all pairwise dependent signals, where univariate measures can be readily computed for each pair of one-dimensional variables. For testing mutual independence between $X^{(1)}, ..., X^{(p)}$, the resulted statistic is represented as the sum or maxima of $\binom{p}{2}$ univariate measure of independence. Another appealing idea is to use the projection approach. Kim et al. (2020) illustrated that using a projection-averaging approach, the sign covariance independence test can be generalized to multivariate data through integrations of the projected univariate τ^* over the unit sphere, as

$$\tau^*_{pq} = \int_{\mathbb{S}^{p-1}} \int_{\mathbb{S}^{q-1}} \mathbb{E}[a(\alpha^T X^1, \alpha^T X^2, \alpha^T X^3 \alpha^T X^4) a(\beta^T Y^1, \beta^T Y^2, \beta^T Y^3, \beta^T Y^4)] d\lambda \alpha \lambda \beta,$$

where a function is defined in Equation (2.5). Similar projection idea has been suggested by Zhu et al. (2017) to generalize Hoeffding's D (Hoeffding, 1948) to multivariate cases. Instead

of projecting to a one-dimensional space, there are also works trying to define multivariate ranks directly. Two recent papers, Deb and Sen (2019) and Shi et al. (2019), developed tests of independence for multivariate variables based on a recent breakthrough in Hallin's multivariate rank (Hallin et al., 2018), where one first "discretize" the unit ball S_d to n grid points with a well-defined ordering and obtain multivariate rank through an optimal coupling between the observed data points and the grid points. These above-mentioned tests have appealing properties in $\mathbb{R}^d$ settings, but in general are not extendable to more general types of data, such as spherical surfaces, planar graph, symmetric positive matrices equipped with Riemannian geometry and manifold-valued functional data. These complex data increasingly arise in practice (Free et al., 2001; Zheng, 2015; Dai et al., 2018; Adriaenssens et al., 2011; Masucci et al., 2009). We also would like to note that there was earlier work using interpoint distances to rank multivariate data or build graphical tests, for example, Mantel (1967); Friedman et al. (1983); Biswas et al. (2016); Sarkar and Ghosh (2018). However, these tests are not consistent against general types of alternatives, and we focus on developing a consistent nonparametric test in this paper.

Finally, two alternatives we would like to mention are Heller et al. (2012) and Pan et al. (2019). The "HHG" method (Heller et al., 2012) transforms the original problem into many aggregated 2×2 contingency tables and use the Pearson's χ^2 test of independence. The ball covariance method (Pan et al., 2019) defines a class of Ball covariance measures by integrating the Hoeffding's dependence measure on the coordinate of radius over poles. Interestingly, the "HHG" method, the ball covariance with their recommended weight functions and our proposed method all depend on the ranking of interpoint distances, although they are derived from three different perspectives.

2.2 NON-RANK BASED TESTS

For multidimensional data, there also exists some non-rank based tests which are consistent against general types of alternatives. Distance covariance ($dCov$) (Székely et al., 2007) was suggested as consistent measure of dependence for multivariate random variables in

$(\mathbb{R}^p, \mathbb{R}^q)$, and later generalized to a metric space with some restrictions (Lyons et al., 2013). Denoting the characteristic function of F_X, F_Y and F_{XY} as ψ_X, ψ_Y and ψ_{XY}, respectively, population $dCov$ is defined as

$$dCov^2(X,Y) = \frac{1}{c_p c_q} \int_{\mathbb{R}^p \times \mathbb{R}^p} \frac{|\psi_{XY}(s,t) - \psi_X(s)\psi_Y(t)|^2}{\|t\|^{1+p}\|s\|^{1+q}} dsdt \tag{2.6}$$

where c_p, c_q are constants. It has an nice empirical counterpart, with $a_{kl} = |X_k - X_l|$, $b_{kl} = |Y_k - Y_l|$,

$$V_n^2(X,Y) := \frac{1}{n^2} \sum_{k,l=1}^{\infty} (a_{kl} - \bar{a_{k\cdot}} - \bar{a_{\cdot l}} + \bar{a_{\cdot\cdot}})(b_{kl} - \bar{b_{k\cdot}} - \bar{b_{\cdot l}} + \bar{b_{\cdot\cdot}}),$$

which is easy to compute and converges to the population coefficient.

With independent copies of (X, Y), the definition of $dCov$ can also be written as

$$dCov = \mathbb{E}\|X^1 - X^2\|\|Y^1 - Y^2\| + \mathbb{E}\|X^1 - X^2\|\mathbb{E}\|Y^1 - Y^2\| - 2\mathbb{E}\|X^1 - X^2\|\|Y^1 - Y^3\|.$$

Then for univariate X and Y, we can write it as

$$dCov = \mathbb{E}b(X^1, X^2, X^3, X^4)b(Y^1, Y^2, Y^3, Y^4)$$

where $b(z_1, z_2, z_3, z_4) = |z_1 - z_2| + |z_3 - z_4| - |z_1 - z_3| - |z_2 - z_4|$. Here the b function differs from the a function in Equation (2.5) only by a sign operator. This reveals the connection between the sign covariance τ^* and the distance covariance measure.

Another popular measure is the Hilbert-Schmidt independence criterion (HSIC) (Gretton et al., 2005, 2008), which is based on the Maximum Mean discrepancy between $P_{XY}(x, y)$ and $P_X(x)P_Y(y)$. According to Sejdinovic et al. (2013), $dCov$ and HSIC are equivalent if the kernel is in the equivalence class that is associated to the distance. The $dCov$ test and the HSIC test are both consistent against general types of alternatives for multivariate data. For more general types of random objects, additional assumptions such as strong negative type spaces and characteristic kernels are needed (Lyons et al., 2013; Sejdinovic et al., 2013).

In Chapter 3, we compare the proposed test of independence with these existing methods, including the distance covariance test (Székely et al., 2007), Hilbert-Schmidt independence criterion (Gretton et al., 2008), two versions of ball covariance tests (Pan et al., 2019) and the HHG method (Heller et al., 2012).

3.0 A NEW TEST BASED ON INTERPOINT-RANKING SIGN COVARIANCE

In this chapter, we introduce a new dependence measure, which we call interpoint-ranking sign covariance (IPR-τ^*), and develop a new test of independence based on IPR-τ^*. The new test inherits good properties from sign covariance τ^*, and is applicable to general types of random objects. We are able to prove the zero-independence equivalence of the proposed IPR-τ^* and to show that the test is consistent against general types of alternatives. We also discuss the general form of the interpoint-ranking based generalization of other coefficients, including the Hoeffding's D coefficient (Hoeffding, 1948) and the Blum-Kiefer-Rosenblatt's R coefficient (Blum et al., 1961).

3.1 DEFINITION AND PROPERTIES OF IPR-τ^*

Let $(\mathcal{X}, \rho)$, $(\mathcal{Y}, \zeta)$ be two separable Banach spaces, where ρ and ζ also represent distances induced by norms. Let θ be a Borel probability measure on $\mathcal{X} \times \mathcal{Y}$ with marginals μ on $\mathcal{X}$ and ν on $\mathcal{Y}$. Let (X, Y) be a pair of random variables where $(X, Y) \sim \theta$, $X \sim \mu$ and $Y \sim \nu$. Let (X^0, Y^0),...,(X^4, Y^4) be $i.i.d$ copies of (X, Y). Then for $i \in \{1, 2, 3, 4\}$, we can refer $\rho(X^0, X^i)$ and $\zeta(Y^0, Y^i)$ as X- and Y-interpoint distance arisen from (X^0, Y^0). The interpoint-ranking sign covariance (IPR-τ^*) collects the signal of dependency between X- and Y-interpoint distances arisen from all anchor points (X^0, Y^0) in $\mathcal{X} \times \mathcal{Y}$.

Definition 1. *The interpoint-ranking sign covariance, or IPR-τ^*, is defined as*

$$IPR\text{-}\tau^*(X,Y) = Ea(\rho(X^0,X^1),\rho(X^0,X^2),\rho(X^0,X^3),\rho(X^0,X^4)) \\ a(\zeta(Y^0,Y^1),\zeta(Y^0,Y^2),\zeta(Y^0,Y^3),\zeta(Y^0,Y^4)),$$

where $a(z_1,z_2,z_3,z_4)$ is defined in Equation (2.5).

The interpoint distances sometimes is called pairwise distances. Empirically, if there are n copies of the data X, one can compute an $n \times n$ pairwise distance matrix $[\rho_{ij}]$ for X, with $n(n-1)/2$ distinct values. For each fixed data point i, one can rank all other points based on the order of ρ_{ij}, and compute the interpoint ranking $R_{i(j)}$. Interpoint distance has been used in various ways to characterize the distribution and geometry of multivariate data, including some independence tests. However, most of them directly build a dependence measure on the two vectors containing $n(n-1)/2$ X-interpoint distances (rankings) and $n(n-1)/2$ Y-interpoint distances (Mantel, 1967; Guo and Modarres, 2020), which oversimplified the structure. In Definition 1, for a fixed reference point (X^0,Y^0), interpoint distance maps the original data to a one dimensional space and the univariate τ^* can be applied. Then we view the reference point (X^0,Y^0) as an extra independent copy in the definition. Here IPR-τ^* is defined using five independent copies of (X,Y) rather than four copies as used in the original sign covariance, which can be considered as the expense of the extended domain. IPR-τ^* remain invariant under the monotone transformation of distances since a represents a coordination structure of interpoint distances.

Theorem 1 states that this new coefficient is nonnegative and becomes zero if and only if X and Y are independent, provided that the joint probability distribution is discrete or continuous, or a mixture of the two.

Theorem 1. *Let $(\mathcal{X},\rho)$, $(\mathcal{Y},\zeta)$ be two separable Banach spaces and θ be a Borel probability measure on $\mathcal{X}\times\mathcal{Y}$ with marginals μ on $\mathcal{X}$ and ν on $\mathcal{Y}$. Assume θ is discrete or continuous, or a mixture of the two, that is, there exists a probability mass function P_{XY} and a density function h such that*

$$\theta(A\times B) = \sum_{x_i\in A, y_i\in B} P_{XY}(x_i,y_i) + \int_{A\times B} h(x,y)G(dx)G(dy),$$

where $A \subset \mathcal{X}$, $B \subset \mathcal{Y}$ are any two open sets and G is the Abstract Wiener measure on $\mathcal{X}$ and $\mathcal{Y}$. In addition, assume $h(x, y)$ is continuous on any continuous point of θ. Then, IPR-$\tau^*(X, Y) \geq 0$ with equality if and only if X and Y are independent.

Our domain, separable Banach space, is chosen so that the proper measure exists with essential properties. Abstract Wiener measure is such measure which is a standardized multivariate Gaussian measure extendable to separable Banach space. It is defined on a Borel σ-algebra generated by open subsets and has a positive measure for any open subset. Due to the absence of the Lebesgue measure in infinite dimensional spaces, we use the Abstract Wiener measure to define the notion of continuous distributions, and it is easy to see that it is equivalent to the conventional definition when restricted to a finite dimensional space.

However, the results may be pushed to a separable metric space with a more mathematically sophisticated definition of continuous distributions in the absence of the Lebesgue measure. The assumptions regarding continuous and discrete distributions are inherited from the original sign covariance paper, which may be relaxed as the authors of the sign covariance paper conjectured. Our numerical experiments show that the method works under various settings within and beyond these requirements. Moreover, well behaved metric spaces are isometric to subspaces of Banach Space; See for example Kuratowski embedding for bounded metric spaces (Kuratowski, 1935), the generalized Banach-Mazur theorem for separable metric spaces (Kleiber and Pervin, 1969), and Nash embedding (Nash, 1956) for Riemannian manifolds. Therefore the independence-zero equivalence property in most applications can be studied in Banach spaces, with a restricted measure support.

We finally note that IPR-τ^* have a nice coverage to work. For example, if we consider $\mathcal{X} = \mathbb{R}^p$, $\mathcal{Y} = \mathbb{R}^q$, with l^d metric, for $p, q, d \geq 3$, the spaces are not strong negative type (Lyons et al., 2013), and the distance covariance test is not consistent. The proposed method can cover these cases. Also, there is no finite moment condition.

3.2 TEST OF INDEPENDENCE

Let $(x_1, y_1), \ldots, (x_n, y_n)$ be *i.i.d.* sample realizations of (X, Y) from a joint distribution θ.

Definition 2. *We propose an empirical IPR-τ^* in the form of a U-Statistic of order 5,*

$$H_n(X,Y) = \frac{1}{\binom{n}{5}} \sum_{1 \le i_1 < i_2 < i_3 < i_4 < i_5 \le n} \phi((x_{i_1}, y_{i_1}), (x_{i_2}, y_{i_2}), (x_{i_3}, y_{i_3}), (x_{i_4}, y_{i_4}), (x_{i_5}, y_{i_5})) \quad (3.1)$$

with the kernel $\phi((x_1, y_1), (x_2, y_2), \ldots, (x_5, y_5))$ defined as

$$\frac{1}{5!} \sum_{(j_1,j_2,j_3,j_4,j_5) \in \mathcal{P}_5} a(\rho(x_{j_1}, x_{j_2}), \rho(x_{j_1}, x_{j_3}), \rho(x_{j_1}, x_{j_4}), \rho(x_{j_1}, x_{j_5}))$$
$$a(\zeta(y_{j_1}, y_{j_2}), \zeta(y_{j_1}, y_{j_3}), \zeta(y_{j_1}, y_{j_4}), \zeta(y_{j_1}, y_{j_5})).$$

To understand this coefficient, it is helpful to consider more straightforward expression of Equation (3.1). Let

$$A_i = \frac{1}{\binom{n-1}{4}} \sum_j \frac{1}{4!} \sum_{\pi \in \mathcal{P}_4} a(\rho(x_i, x_{j_{\pi(1)}}), \ldots, \rho(x_i, x_{j_{\pi(4)}})) a(\zeta(y_i, y_{j_{\pi(1)}}), \ldots, \zeta(y_i, y_{j_{\pi(4)}})),$$

where the outer sum is taken over the set of all ordered subsets j of 4 different integers chosen from $\{1, 2, \ldots, n\}/\{i\}$. Here A_i is a U-statistic of order 4 and gives rise to the τ^* between the interpoint distance-induced random variables, $\rho(x_i, X)$ and $\zeta(y_i, Y)$. Then $\frac{1}{n}\sum_i A_i$ equals the empirical IPR-$\tau^*(X, Y)$. As a result, our empirical IPR-τ^* is the average of τ^*s between X- and Y- interpoint distances anchored in each n data points. Since A_is are not independent to each other, it is easier to see when expressed as a U-statistic formula as Equation (3.1) to drive the asymptotic property of H_n.

Now we drive the asymptotic behavior of H_n to see that H_n is indeed a consistent estimator of IPR-τ^* and test based on H_n is consistent to general alternatives. Let

$$\phi_i((x_1, y_1), \ldots, (x_i, y_i)) = E\phi((x_1, y_1), \ldots, (x_i, y_i), (X_{i+1}, Y_{i+1}), \ldots, (X_5, Y_5))$$

and $\sigma_i = Var\phi_i$ for $i = 1, \ldots, 5$. We first present the general results based on the large-sample theory of the U-statistics (Section 5.5 of Serfling (2009)).

Lemma 1. *If $\sigma_5 < \infty$, we have*

$$n^{1/2}(H_n(X,Y) - \textit{IPR-}\tau^*(X,Y)) \to N(0, 5^2\sigma_1).$$

In the case that $\sigma_1 = 0$, the above Gaussian limit is degenerate, and we refer to a second lemma.

Lemma 2. *If $\sigma_5 < \infty$ and $0 = \sigma_1 < \sigma_2$, we have*

$$n(H_n(X,Y) - \textit{IPR-}\tau^*(X,Y)) \to \binom{5}{2} \sum_{m=1}^{\infty} \lambda_m(\chi^2_{1m} - 1) \tag{3.2}$$

where $\chi^2_{1m}s$ are independent χ^2_1 variables and $\lambda_m s$ are the solutions of the eigen equation

$$\int_{\mathcal{X}\times\mathcal{Y}} \phi_2((x,y),(x',y'))\psi(x',y')d\theta = \lambda\psi(x,y), \psi \in L_2. \tag{3.3}$$

Under the null hypothesis, interpoint distances arising from fixed (x_1, y_1) are still $i.i.d$ random variables with no association. Therefore, $\phi_1((x_1, y_1))$ equals zero (Nandy et al., 2016) and so does σ_1. Then H_n is a degenerate U-statistic, and Lemma 2 applies. Under the alternative hypothesis, $H_n(X,Y)$ converges to IPR-$\tau^*(X,Y)$ and IPR-$\tau^* > 0$ by Theorem 1, so $nH_n(X,Y) \to \infty$ as $n \to \infty$. In both cases, $H_n(X,Y)$ converges to IPR-$\tau^*(X,Y)$, i.e., $H_n(X,Y)$ is a consistent estimator.

We propose to reject the null hypothesis when $nH_n(X,Y) > C_\alpha$, where C_α is the α-level critical value from the null distribution. Combining Theorem 1 and Lemmas 1 - 2, we can obtain the following theorem.

Theorem 2. *If X and Y are jointly distributed as specified in Theorem 1, $nH_n(X,Y)$ can serve as a test statistic for a test of independence which is consistent against the alternatives. Specifically,*

(a) If X and Y are independent,

$$nH_n(X,Y) \xrightarrow{d} \binom{5}{2} \sum_{m=1}^{\infty} \lambda_m(\chi^2_{1m} - 1), \tag{3.4}$$

where $\chi^2_{1m}s$ are independent χ^2_1 variables and $\lambda_m s$ are the solutions of the eigen equation

$$\int_{\mathcal{X}\times\mathcal{Y}} \phi_2((x,y),(x',y'))\psi(x',y')d\theta = \lambda\psi(x,y), \psi \in L_2. \tag{3.5}$$

(b) If X and Y are dependent, $nH_n(X,Y) \xrightarrow{p} \infty$.

In Theorem 1, we proved that IPR-$\tau^*(X,Y) > 0$ under the alternative. It is easy to see that $nH_n(X,Y) \to \infty$ in probability and the power goes to 1 if we consider a fixed dependent signal. As long as IPR-$\tau^*(X,Y)$ does not decrease as the dimension increases, the proposed test should work for large dimensions. However, in the sparse high dimensional case, where $(\mathcal{X},\mathcal{Y}) = (\mathbb{R}^p, \mathbb{R}^q)$, p, q are large relative to n, and the dependence only exists among a small subsets of coordinates, the dependence signal IPR-$\tau^*(X,Y)$ should not be considered as a constant; rather it is a function of $1/p$ and $1/q$, and the power of the test depends on the ratio of p and n.

Spectral approximation To get the critical value, one way is to solve the eigen equation in Equation (3.5). The first step is to get a left-hand side approximation of the equation 3.5 by

$$\frac{1}{n}\sum_{i=1}^{n} \phi_2((x,y),(x_i,y_i))\psi(x_i,y_i).$$

Then we plug $(x_1,y_1),...,(x_n,y_n)$ into (x,y) to obtain a matrix eigenproblem

$$\sum_{i=1}^{n} \phi_2((x_j,y_j),(x_i,y_i))\psi(x_i,y_i) = \hat{v}\psi(x_j,y_j).$$

The solutions for $\hat{v}_m$, m=1,2,.., are the eigenvalues of the Gram matrix

$$\Theta_{ij} = \phi_2((x_i,y_i),(x_j,y_j)) \text{ for } j = 1,..,n,$$

and $\hat{\lambda}_m \simeq \frac{1}{n}\hat{v}_m$. So $\sum\limits_m \hat{\lambda}_m \simeq \frac{1}{n}tr(\Theta)$ and $\sum\limits_m \hat{\lambda}_m^2 \simeq \frac{1}{n^2}tr(\Theta^2)$.

On the other hand, Welch Satterthwaite equation gives an approximation for a χ^2 type mixture $\sum\limits_{m=1}^{\infty} \lambda_m \chi_{1m}^2$ as $\beta\chi_d^2$. By matching the first 2 cumulant, we obtain $\beta = \frac{\sum\lambda_m^2}{\sum\lambda_m}$ and $d = \frac{(\sum\lambda_m)^2}{\sum\lambda_m^2}$. Plugging in the approximation of $\sum\limits_m \hat{\lambda}_m$ and $\sum\limits_m \hat{\lambda}_m^2$, we obtain

$$\beta \simeq \frac{1}{n}\frac{tr(\Theta^2)}{tr(\Theta)},\ d \simeq \frac{tr(\Theta)^2}{tr(\Theta^2)},$$

where $\sum_{m=1}^{\infty} \lambda_m(\chi_{1m}^2 - 1)$ is approximated by $\beta\chi_d^2 - \beta d$.

Permutation approximation An alternative is to approximate the critical value by a permutation distribution. For a class of U-statistic based tests with a continuous asymptotic null distribution, Theorem 2.5 in Kim et al. (2020) establish that the permutation critical

value converge to the oracle critical value under both null and alternative distributions, and the power will be asymptotically the same as the test using the oracle critical values.

Computational Complexity For n sample points, efficient algorithms to compute the sign covariance has been developed in Heller and Heller (2016) with $\mathcal{O}(n^2)$ operations and later in Even-Zohar and Leng (2019) with $\mathcal{O}(n)$ operations. Our statistic is a summation of n different sign covariances of interpoint distances. Our numerical experiments showed that the permutation method (with 1000 permutation samples) is generally faster than the null distribution approximation using the spectral method.

3.3 A GENERAL FORM TO GENERALIZE OTHER STATISTICS

Finally, we introduce a general form to generalize other statistics. The idea developed in this paper can be used to generalize other univariate dependence measures, such as Hoeffding's D, Blum-Kiefer-Rosenblatt's R, the squared Kendall's τ, and the univariate distance covariance. These coefficients can be estimated by a U-statistic; see for example Drton et al. (2018). Let

$$U_m = \frac{1}{\binom{n}{m}} \sum_{1 \le i_1 < i_2, \ldots, i_{m-1} < i_m \le n} \phi((x_{i_1}, y_{i_1}), \ldots, (x_{i_m}, y_{i_m})),$$

where the kernel $\phi((x_1, y_1), \ldots, (x_m, y_m))$ is defined as

$$\frac{1}{(m)!} \sum_{(j_1, \ldots, j_m) \in \mathcal{P}_m} h((x_{j_1}, y_{j_1}), \ldots, (x_{j_m}, y_{j_m})).$$

Then we can generalize U_m to work for multivariate data or more general objects by introducing an extra independent pair,

$$U_{m+1} = \frac{1}{\binom{n}{m+1}} \sum_{1 \le i_1 < i_2, \ldots, i_m < i_{m+1} \le n} \tilde{\phi}((x_{i_1}, y_{i_1}), \ldots, (x_{i_{m+1}}, y_{i_{m+1}})),$$

where the new kernel $\tilde{\phi}((x_1, y_1), \ldots, (x_{m+1}, y_{m+1}))$ is defined as

$$\frac{1}{(m+1)!} \sum_{(j_1, \ldots, j_{m+1}) \in \mathcal{P}_{m+1}} h((\rho(x_{j_1}, x_{j_2}), \zeta(y_{j_1}, y_{j_2})), \ldots, (\rho(x_{j_1}, x_{j_{m+1}}), \zeta(y_{j_1}, y_{j_{m+1}}))).$$

When this generalization is applied to Hoeffiding's D or Blum-Kiefer-Rosenblatt's R, the empirical performance is very similar to that of IPR-τ^*. In this thesis, we focus on generalizing the sign covariance test because it has nice theoretical properties and also the construction of τ^* is simpler than the construction of Hoeffding's D and Blum-Kiefer-Rosenblatt's R. When both X and Y have continuous distributions, Drton et al. (2018) derived an interesting identity between τ^*, Hoeffding's D, and Blum-Kiefer-Rosenblatt's R, which gives $1/18\tau^* = 1/30D + 1/45R$. The same identity holds for IPR-τ^*, IPR-D and IPR-R.

In addition, for $X \in \mathbb{R}$ and $Y \in \mathbb{R}$, the distance covariance (Székely et al., 2007) has the equivalent formula

$$dCov = 1/4Eb(X^1, X^2, X^3, X^4)b(Y^1, Y^2, Y^3, Y^4),$$

with $b(z_1, z_2, z_3, z_4) = |z_1 - z_2| + |z_3 - z_4| - |z_1 - z_3| - |z_2 - z_4|$. Here the b function differs from the a function used in τ^* only by a sign operator (Bergsma and Dassios, 2014). Székely et al. (2007) showed that the distance covariance test works for multivariate data if the absolute distance $|z_1 - z_2|$ is replaced by a Euclidean distance on $\mathbb{R}^p$. If we estimate the univariate distance covariance by U_m as analogous to that for the sign covariance τ^*, the form of U_{m+1} naturally provides another way to generalize the univariate distance covariance to the multivariate case.

3.4 SIMULATIONS

In the first three simulation, we study the empirical performance of our proposed IPR-τ^* test: Simulation I, Multivariate data; Simulation II, Manifold-valued data; and Simulation III, Manifold-valued functional data. We compare the Type-I error and statistical power with several existing tests of independence: the distance covariance test denoted by "dCov" using the R package *energy* (Székely et al., 2007), the test based on the summation of Pearson chi-square statistic denoted by "HHG" using the R package *HHG* (Heller et al., 2012), the Ball covariance test with a constant weight denoted by"BCov1" and a probability weight denoted by "BCov2" using the R package *Ball* (Pan et al., 2019) and the Hilbert-Schmidt

independence criterion with Gaussian kernel denoted by "HSIC" using the R package *HSIC* (Gretton et al., 2008). Our proposed test is implemented in R. The code is currently available upon request and the package will soon be publicly available. The distance covariance test is applicable to metric spaces of strong negative type and the HSIC method requires the kernels to be characteristic. These conditions are in general hard to check for non-Hilbertian data, and we know some of the non-Hilbertian data are not of strong negative type. Also the original "HHG" paper (Heller et al., 2012) only proved consistency of the test for $\mathbb{R}^p$ and $\mathbb{R}^q$. Nevertheless, we still applied these methods to all of the examples since these methods are all based on pairwise distances and can be empirically applied to complex objects as long as an appropriate distance/metric can be defined. In the implementation, we use the same metric for all methods, and p-values are all based on 1000 permutations. In all of the following settings, we report the results with sample sizes 20, 50, 100 and 200. The significance level is 0.05, and powers are based on 1000 simulations. In Simulation IV, we show some results for other generalizations given in Section 3.3 .

Simulation I

We consider multivariate variables $X = (X_1, X_2, X_3, X_4, X_5)$ and $Y = (Y_1, Y_2, Y_3, Y_4, Y_5)$ with the regular Euclidean distance. Examples 3.1-3.2 assess Type-I error rates and Examples 3.3-3.11 compare power performances. Similar settings have been used in Székely et al. (2007), Heller et al. (2012) and Pan et al. (2019).

Example 3.1. *(Type I) X, Y are generated from a multivariate normal distribution with mean* $\mathbf{0}$ *and* $\text{cov}(X_i, X_j) = \text{cov}(Y_i, Y_j) = 0.1$ *for* $i \neq j$, $i, j = 1, 2, 3, 4, 5$. *There is no correlation between X and Y components.*

Example 3.2. *(Type I) X, Y are generated from a multivariate $t(v)$ distribution for $v = 1, 2$.*

Example 3.3. *(Linear) X,Y are generated from a jointly normal distribution with mean* $\mathbf{0}$ *and* $\text{cov}(X_i, X_j) = \text{cov}(Y_i, Y_j) = 0.1$ *for* $i \neq j$, $\text{cov}(X_i, Y_i) = 0.3$ *for* $i, j = 1, 2, 3, 4, 5$.

In Examples 3.4-3.8, X is generated from a multivariate normal distribution with mean $\mathbf{0}$ and $\text{cov}(X_i, X_j) = 0.1$ for $i \neq j$, $i, j = 1, 2, 3, 4, 5$.

Example 3.4. *(Quadratic)* $Y_i = 0.5X_i^2 + \epsilon$ *with* $\epsilon \sim N(0, 1)$ *for* $i = 1, 2, 3, 4, 5$.

Test	Sample size	Normal	t(1)	t(2)
IPR-τ^*	20/50/100/200	0.030/0.026/0.035/0.043	0.064/0.056/0.048/0.032	0.036/0.047/0.045/0.046
dCov	20/50/100/200	0.046/0.045/0.054/0.050	0.054/0.042/0.038/0.049	0.040/0.055/0.060/0.042
HHG	20/50/100/200	0.039/0.048/0.051/0.052	0.059/0.055/0.046/0.034	0.043/0.052/0.054/0.057
BCov1	20/50/100/200	0.040/0.052/0.047/0.058	0.058/0.044/0.045/0.033	0.036/0.069/0.056/0.058
BCov2	20/50/100/200	0.043/0.053/0.054/0.042	0.060/0.043/0.046/0.044	0.041/0.069/0.056/0.048
HSIC	20/50/100/200	0.037/0.051/0.053/0.044	0.054/0.041/0.040/0.051	0.040/0.050/0.057/0.052

Table 1: Empirical Type-I error rates at nominal significance level 0.05 in Simulation I. Results are based on 1,000 simulations.

Example 3.5. *(Y=Xϵ)* $Y_i = X_i\epsilon$ *with* $\epsilon \sim N(0,1)$ *for* $i = 1,2,3,4,5$.

Example 3.6. *(Y=Inv(X))* $Y_i = 1/|X_i|$ *for* $i = 1,2,3,4,5$.

Example 3.7. *(Concave)* $Y_i = \pm 1/|X_i|$ *with random signs of equal probability, for* $i = 1,2,3,4,5$.

Example 3.8. *(X-shape)* $Y_i = \pm X_i$ *with random signs of equal probability, for* $i = 1,2,3,4,5$.

In Examples 3.9-3.11, $Z = (Z_1, Z_2, Z_3, Z_4, Z_5)$ is generated from a multivariate normal distribution with mean $\mathbf{0}$ and $\text{cov}(Z_i, Z_j) = 0.1$ for $i \neq j$, $i, j = 1,2,3,4,5$.

Example 3.9. *(Circle)* $X_i = 2logit^{-1}(Z_i) - 1$ *and* $Y_i = \pm(1 - X_i^2)^{1/2}$ *with random signs of equal probability, for* $i = 1,2,3,4,5$.

Example 3.10. *(Diamond)* $X_i = 2logit^{-1}(Z_i) - 1$ *and* $Y_i = \pm(1 - |X_i|)$ *with random signs of equal probability, for* $i = 1,2,3,4,5$.

Example 3.11. *(Two-pieces)* $X_i = 2logit^{-1}(Z_i) - 1$ *and* $Y_i = (0.9I_{\{|X_i|<0.5\}} + 0.1)\epsilon$ *with* $\epsilon \sim N(0, 0.1)$ *for* $i = 1,2,3,4,5$.

Table 1 confirms that the Type-I error rates are well-controlled for all methods. Figure 3 summarizes the empirical powers. The proposed method "IPR-τ^*" shows a good performance, and the powers reach one as sample size increases to 200. In general, we see that "IPR-τ^*" and "HHG" always belong to a group with the highest power except for the linear case. While the distance covariance test has the best power for the linear case, it generally has the lowest power for other non-linear cases. The power of "dCov" is poor for "$Y = Inv(X)$", "Concave", "Circle", "Diamond" and "Two-pieces" even with the sample

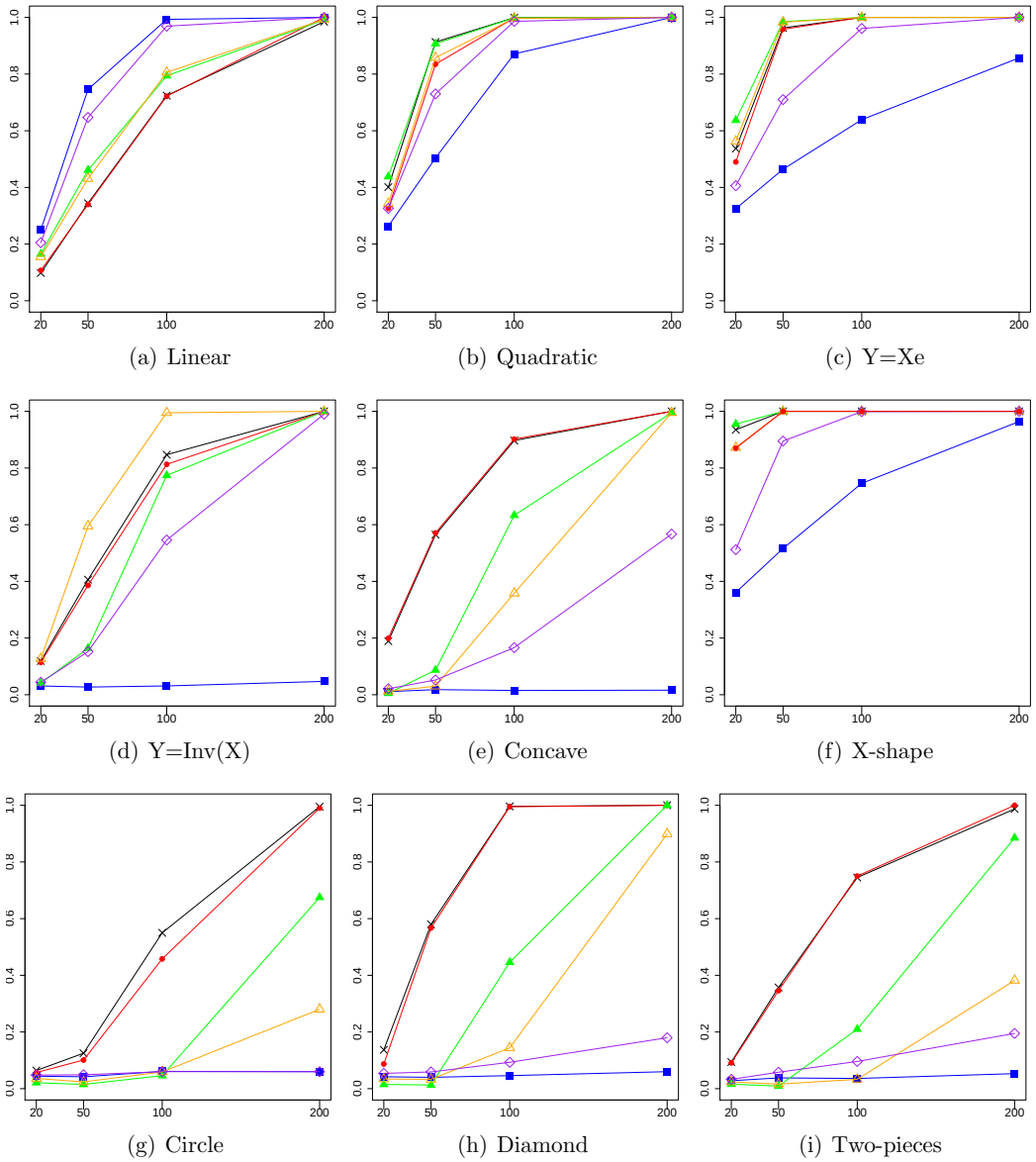


Figure 3: Simulation I: Empirical power of the tests for IPR-τ^* (●, red), dCov (■, blue), HHG (×, black), BCov1 (▲, green), BCov2 (△, orange) and HISC (◇, purple). Power values are computed for each of the sample sizes 20, 50, 100, 200 with 1,000 simulations.

size 200. The performances of "BCov1", "BCov2" and "HSIC" are somewhat in between. The power of "HSIC" with Gaussian kernel is seen to be poor for "Circle", "Diamond" and "Two-pieces".

Simulation II

We consider variables in a non-Euclidean space with Riemannian Metric as a distance. We first consider variables on a spherical coordinate of the unit sphere S^2. In a spherical coordination, each point on a sphere, denoted by $(\theta, \phi)_s$, is uniquely represented with longitude θ and latitude ϕ, where θ specifies the east-west position on a spherical surface and ϕ specifies an angle which range from 0 at the equator to $\pi/2$ at the north, and $-\pi/2$ at the south; so $\theta \in [-\pi, \pi]$ and $\phi \in [-\pi/2, \pi/2]$. We use the great-circle distance, the shortest distance over the surface of a sphere between two points.

Example 3.12. *(Type I)* $X = (X_1, X_2)$ *where* $X_1, X_2 \sim \mathcal{U}(0, 1)$, $Y = (\theta, \phi)_s \in S^2$ *with* $\theta \sim \mathcal{U}(-\pi, \pi)$, $\phi \sim \mathcal{U}(-\pi/2, \pi/2)$.

Example 3.13. *(Sphere 1)* X *is same as Example* 3.12, $Y = (\theta, \phi)_s \in S^2$ *with* $\theta \sim \mathcal{U}(-\pi, \pi)$, $\phi = \pi(X_1 + X_2)\epsilon/2 - \pi/2$, $\epsilon \sim \mathcal{U}(0, 1)$.

Example 3.14. *(Sphere 2)* X *is same as Example* 3.12, $Y = (\theta, \phi)_s \in S^2$ *with* $\theta \sim \mathcal{U}(-\pi, \pi)$, $\phi = \pi|X_1 - X_2|\epsilon - \pi/2$, $\epsilon \sim \mathcal{U}(0, 1)$.

Example 3.15. *(Sphere 3)* X *is same as Example* 3.12, $Y = (\theta, \phi)_s \in S^2$ *with* $\theta \sim \mathcal{U}(-\pi, \pi)$, $\phi = \pi(X_1 + X_2)^2\epsilon/4 - \pi/2$, $\epsilon \sim \mathcal{U}(0, 1)$.

We also consider symmetric positive matrices. Specifically, we consider a 3 by 3 symmetric positive matrix variable whose every non-diagonal element equal to ρ. We use the affine invariant Riemannian metric, $d(A, B) = ||log(A^{-1/2}BA^{-1/2})||_F$, where $log(A)$ is the matrix logarithm of A, and $||A||_F$ is the Frobenius norm of A.

Example 3.16. *(Type I)* $X = \left(\begin{smallmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{smallmatrix}\right)$, $Y = \epsilon$ *with* $\rho \sim \mathcal{U}(0, 0.3)$, $\epsilon \sim N(0, 0.3)$.

Example 3.17. *(PD 1)* $X = \left(\begin{smallmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{smallmatrix}\right)$, $Y = \rho + \epsilon$ *with* $\rho \sim \mathcal{U}(0, 0.3)$, $\epsilon \sim N(0, 0.3)$.

Example 3.18. *(PD 2)* $X = \left(\begin{smallmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{smallmatrix}\right)$, $Y = \epsilon$ *with* $\rho \sim \mathcal{U}(0, 0.3)$, $\epsilon \sim N(0, \rho/3)$.

Example 3.19. *(PD 3)* $X = \left(\begin{smallmatrix} 1 & \rho_1 & \rho_1 \\ \rho_1 & 1 & \rho_1 \\ \rho_1 & \rho_1 & 1 \end{smallmatrix}\right)$, $Y = \left(\begin{smallmatrix} 1 & \rho_2 & \rho_2 \\ \rho_2 & 1 & \rho_2 \\ \rho_2 & \rho_2 & 1 \end{smallmatrix}\right)$, $\rho_1 = 1/(1 + \lambda_1^2)$, $\rho_2 = 1/(1 + \lambda_2^2)$ *with* $\left(\begin{smallmatrix} \lambda_1 \\ \lambda_2 \end{smallmatrix}\right) \sim N(\left(\begin{smallmatrix} 1 \\ 1 \end{smallmatrix}\right), \left(\begin{smallmatrix} 1 & 0.5 \\ 0.5 & 1 \end{smallmatrix}\right))$.

We observed that the Type-I error rates for all methods are well-controlled. The empirical powers are shown in Figure 4. All methods have increasing power towards 1 as the sample size increases. The "dCov" method seems to perform the best in linear relationships and is less competitive in other settings.

Simulation III

We consider manifold-valued functional trajectories $X(t) = (\theta(t), \phi(t))_s$ with $t \in [0, 1]$. For each time point t, $X(t)$ is a point on the unit sphere S^2, so $\theta(t) \in [-\pi, \pi]$ and $\phi(t) \in [-\pi/2, \pi/2]$. We generate data as follows.

$$\theta(t) = (\eta_0 t + \sum_{j=1}^{20} \eta_j sin(j\pi t))(mod 2\pi) - \pi,$$
$$\phi(t) = ((\xi_0 t + \sum_{j=1}^{20} \xi_j sin(j\pi t))/2 \vee \pi/2) \wedge -\pi/2$$

with coefficients ηs and ξs drawn independently from a normal distribution with mean zero. Standard deviations are 1 for η_0, ξ_0 and $j^{-6/5}$ for η_j, ξ_j $(j = 1, \ldots, 20)$. When defining $\phi(t)$, we made upper and lower bounds to ensure $\phi(t) \in [-\pi/2, \pi/2]$, even though $(\xi_0 t + \sum_{j=1}^{20} \xi_j sin(j\pi t))/2$ exceeding $\pi/2$ or below $-\pi/2$ is unlikely in this setting. Examples of sample trajectories $X(t)$ are shown in Figure 5.

The distance between two trajectories is measured by

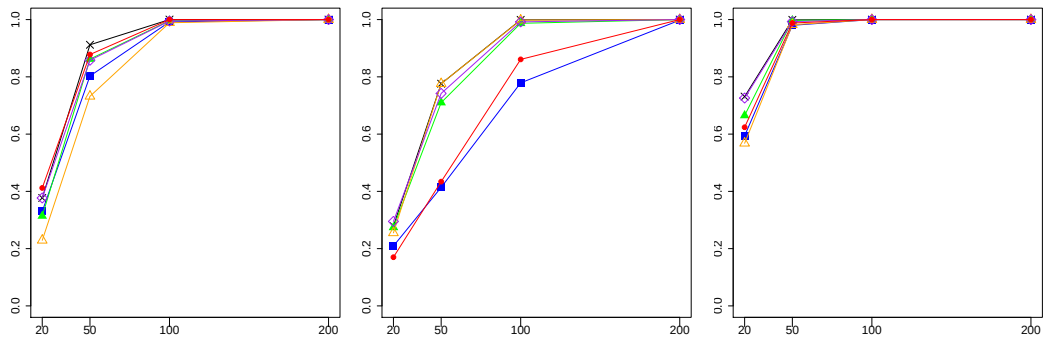
$$d_1(X_i(t), X_{i'}(t)) = sup_{t\in[0,1]} d(X_i(t), X_{i'}(t))$$

where d refers a great-circle distance between two points on a unit sphere. In our simulation, we generate $n = 20, 50, 100, 200$ sample curves with 101 observed points on each trajectory where the observed points are equally spaced between $[0, 1]$.

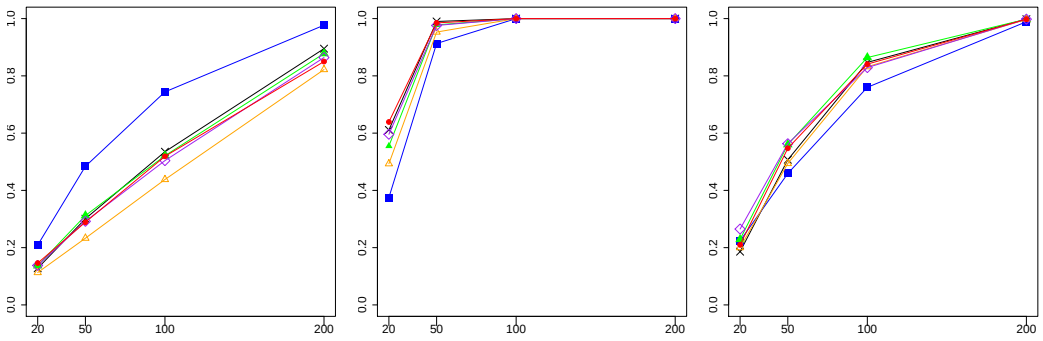
Example 3.20. *(Type I)* $X(t) = (\theta(t), \phi(t))_s$, $Y \sim N(0, 1)$.

Example 3.21. *(FT1)* $X(t) = (\theta(t), \phi(t))_s$, $Y = (1/\int_0^1 |\theta(t)|dt, 1/\int_0^1 |\phi(t)|dt)$.

Example 3.22. *(FT2)* $X(t) = (\theta(t), \phi(t))_s$, $Y = (\int_0^1 |\theta(t)|dt + \epsilon_1, \int_0^1 |\phi(t)|dt + \epsilon_2)$ *with* ϵ_1, $\epsilon_2 \sim N(0, 0.5)$.



(a) Sphere 1 (b) Sphere 2 (c) Sphere 3



(d) PD 1 (e) PD 2 (f) PD 3

Figure 4: Simulation II: Empirical power of the tests for IPR-τ^* (●, red), dCov (■, blue), HHG (×, black), BCov1 (▲, green), BCov2 (△, orange) and HISC (⋄, purple). Power values are computed for each of the sample sizes 20, 50, 100, 200 with 1,000 simulations.

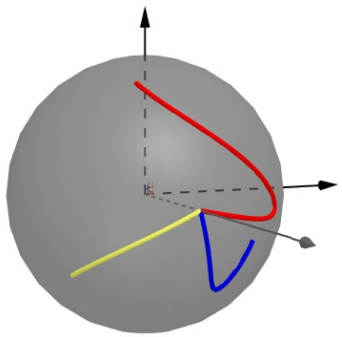
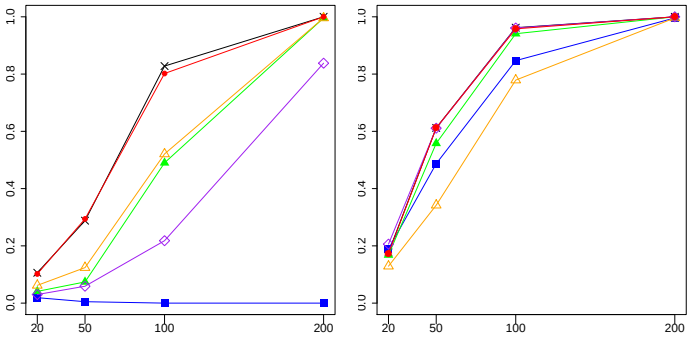


Figure 5: A demonstration of sample curves $X_i(t)$ generated in Simulation III.



(a) FT1 (b) FT2

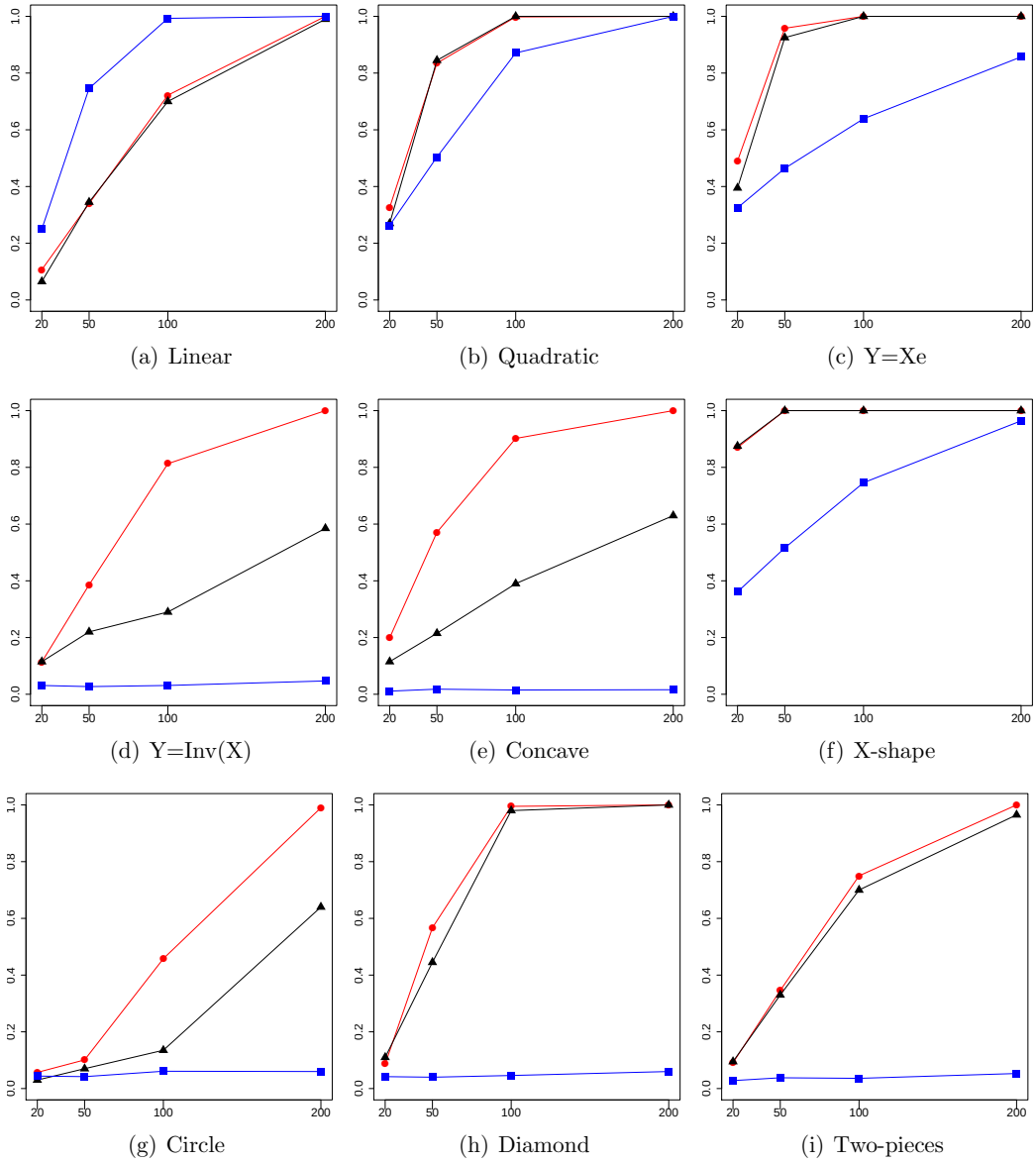
Figure 6: Simulation III: Empirical power of the tests for IPR-τ^* (●,red), dCov (■, blue), HHG (×, black), BCov1 (▲, green), BCov2 (△, orange) and HISC (⋄, purple). Power values are computed for each of the sample sizes 20, 50, 100, 200 with 1,000 simulations.

In Simulation III, the Type-I errors are well-controlled for all the methods. Figure 6 show that the powers of all methods increase as the sample size increases, except for the "dCov" method in Example 3.21. The proposed test "IPR-τ^*" and "HHG" maintain good power in both cases.

Overall, numerical experiments have confirmed that the proposed test is consistent against general alternatives. The empirical powers for "IPR-τ^*" and "HHG" are similar and within the highest power group, although the test statistics are derived from very different perspectives. The ball covariance methods are powerful in most cases, and we find that different weights lead to different performances with no obvious winner. The ball covariance is proved to be asymptotically equivalent to the "HHG" test if a chi-square type weight is used (Pan et al., 2019), and we only focus on finite sample performance here. The HSIC method with Gaussian kernel is usually in the middle group, but has poor power for some cases such as Examples 3.9 and 3.10. The performance of "dCov" is somehow divided. The method tends to perform better than others in normal and linear cases, but clearly less competitive in terms of power in all other cases.

Simulation IV

Referring Section 3.3, general U_{m+1} form gives rise to IPR-D or IPR-R, for Hoeffiding's



(a) Linear (b) Quadratic (c) Y=Xe

(d) Y=Inv(X) (e) Concave (f) X-shape

(g) Circle (h) Diamond (i) Two-pieces

Figure 7: Empirical power of the tests for IPR-τ^* (●, red), IP-dCov (▲, black) and dCov(■, blue). Power values are computed for each of the sample sizes 20, 50, 100, 200 with 200 simulations.

D and Blum-Kiefer-Rosenblatt's R respectively. We conducted simulations to compare IPR-τ^*, IPR-D and IPR-R using examples as in Simulation I. The empirical performances of these methods were very similar.

Next, we compared IPR-τ^*, IP-dCov and dCov using the same examples as in Simulation I. We found that IP-dCov generally had a better empirical performance than dCov (except for the linear case). The performance of IP-dCov was not as good as IPR-τ^* in $Y = Inv(X)$, concave and circle. Results are shown in Figure 7.

3.5 DATA EXAMPLE

DLBCL Data

We first apply the proposed test of independence to study the relationship between gene expression and survival outcomes. We use the data provided by Rosenwald et al. (2002), in which the survival time of patients with diffuse large B-cell lymphoma after chemotherapy is recorded as well as the related gene expression profiles. As the survival time is believed to be influenced by the molecular features of tumor, previous papers, including Bair and Tibshirani (2004), Bair et al. (2006), Bøvelstad et al. (2007), Chen et al. (2011) and Chen et al. (2017), have tried to build predictive models for survival time of a patient, using gene-expression patterns as predictors.

In the study, 240 patients were examined for 7399 gene expression profiles with the use of DNA microarrays. Following the same approach as the above-cited authors, we pre-screen the genes and use only 240 most relevant ones. The subset selection is performed by fitting a univariate Cox regression model of each gene expression value on survival one-by-one and ranking the obtained Cox scores from largest to smallest (Chen et al., 2011). We apply the proposed method "IPR-τ^*" as well as "dCov", "HHG", "BCov1", "BCov2", "HSIC" to test independence between 240 gene expressions and the survival time. Euclidean distance is used to measure the distance between two gene expressions for all the methods.

All of the methods detect the dependency with 0.05 significance level. More efforts can be put into building predictive models after the nonparametric tests show statistically

significant results. As a comparison, we test with the 240 genes that have the lowest Cox scores and repeat the same sets of tests. No method concludes the dependency.

Farm Data

We investigate the dependency between the annual crop yields and the temperature using the dataset described in Wong et al. (2019). In this dataset, the annual yield of two major crops, corn and soybean, is recorded in bushels per acre from 105 counties of Kansas from 1999 and 2011, provided by the National Agricultural Statistics Agency at https://quickstats.nass.usda.gov/. The weather data is from the National Climatic Data Center at https://www.ncdc.noaa.gov/data-access and contains the daily minimum, maximum temperature aggregated at the county level.

Following the source paper, we let Y be the annual corn or soybean yield for a specific year and county, respectively, $X_1(t)$ and $X_2(t)$ be the daily maximum and minimum temperatures for the same year and county, and $X(t) = (X_1(t), X_2(t))$. We aim at testing independence between the annual crop yields and temperature trajectories. For this purpose, we introduce the daily heat unit accumulation defined as, $HU(X,t) = [(X_1(t) + X_2(t))/2 - T_{base}]_+$, based on which phenological development of plants occurs according to the EPIC plant growth model (Williams et al., 1989), i.e., no growth occurs at or below T_{base}. In the formula, T_{base} is a crop-specific base temperature. Following the reference paper, we use $T_{base} = 8°C$ for corn and $10°C$ for soybean. c_+ denotes the positive part of c. The dissimilarity between temperatures is measured by

$$d_X(X(t), X'(t)) = \Big(\int_0^{365} \Big([\frac{X_1(t) + X_2(t)}{2} - T_{base}]_+ - [\frac{X_1'(t) + X_2'(t)}{2} - T_{base}]_+\Big)^2 dt\Big)^{1/2},$$

and Euclidean distance is used for Y.

All six methods have the same statistical conclusion at the 0.05 significance level. Significant dependence between temperature and annual crop yields are found for both soybean data and the corn data. In this data example, we find that the largest sign covariance values arise from anchor points in year 2004, which indicates that the X and Y-interpoint distances computed for a data point in year 2004 tend to be highly correlated. Further interpretation and visualization of the dependence could be a topic for future research.

4.0 APPLICATION TO VARIABLE SELECTION

4.1 BACKGROUND

High-dimensional data provide rich ground for studying relationships between variables. However, it also posts new challenges in building and estimating a model. Variable selection is a crucial procedure to reduce the number of variables and make the analysis concise and interpretable. In this chapter, we investigate a way to use a measure of independence to select important variables with an application to gene selection related to suicidal behaviors.

There exist a vast amount of old and new methods to select variables in a high - dimensional setting. The L_1 penalized approach "LASSO" (Tibshirani, 1996) is probably one of the most widely-used methods in practice. Assuming that response variable Y can be expressed as a linear combination of explanatory variables $\{X_j\}_{j=1}^{p}$ plus an error, that is,

$$Y_i = \sum_{j=1}^{\infty} \beta_j X_{ij} + \epsilon_i,$$

LASSO jointly achieves low prediction error and sparse estimation of $\{\beta_j\}_{j=1}^{p}$ by introducing L_1 penalty term to an optimization function and provides a natural way to select variables. It has a computationally feasible algorithm and nice regime to decide the size of selection through cross-validation which enhances its applicability. Many LASSO variants have been proposed in order to extend a linear model assumption (Tibshirani, 1997; Roth, 2004; Krishnapuram et al., 2005) and to improve prediction accuracy (Zou and Hastie, 2005) of the original method, but they are still in a framework of penalizing coefficients and minimizing model-based residuals. However, as the data collection method evolves, it gets harder to know in advance how the variables are related and whether the assumed model fits the data.

There is no systematic results on the control of false-positive and false-negative errors if a misspecified model is used.

Recently, Li et al. (2012) and Pan et al. (2018) suggested to use a measure of independence ("dCov" and "BCov") to select important variables. In their papers, one computes a measure of dependence with the response variable for each independent variable in the candidate set, and selects the top N of them where the variables are ordered by the magnitude of the dependence measure (scale-based approach). With the development of conditional independence measures (Wang et al., 2015), there are also some recent work proposed to perform an analogous variable selection using conditional dependence measures instead of dependence measures (Lu and Lin, 2020; Liu and Wang, 2018). These can be viewed as an extension of the Sure Independence Screening method (SIS) proposed by Fan and Lv (2008). The original SIS approach assumes a linear model and selects variables by ordering the marginal Pearson correlation coefficients. The authors showed that this process achieves sure screening property which means that all active (truly important) variables are selected with probability converging to 1 as the sample size goes to infinity with an assumption of Gaussian errors. Later their work was extended to generalized linear models (Fan et al., 2009), Cox models (Zhao and Li, 2012) and linear model with preselected variables (Barut et al., 2016). The new methods based on independence measures are nice additions to the existing selection methods as they offer a completely model-free approach without assuming any kind of relationships in advance. However, these methods are all scale-based approaches, which possess several problems in practice. First, they rely on the assumption that the magnitude of the estimated dependence measure represents the strength of variable importance. In finite sample experiments, the ordering based on the estimated dependence measure is not always stable due to natural estimation variability. In particular, when there exist various kinds of relationships, the magnitude of the dependence measure does not necessarily have a interpretable scale as in a linear measure. Second, although the theory says the true active variables will all be retained with a large enough model size, determining the cutoff (model size) is practically difficult. Third, it is particularly hard to compare groups of variables with different dimensions. For example, one may want to cluster genes and select them at all or not if they are in the same cluster. One of the LASSO variants, group-LASSO (Zou

and Hastie, 2005) provides a way to perform such selection where they enforce each cluster a different penalty depending on the size of clusters to make a fair comparison. However, it is hard to make such adjustments for a measure of independence because we do not exactly understand the effect of dimensions on the magnitude of the measure.

Motivated by the need to address these practical challenges in data analysis, we propose a frequent voting method that uses but do not completely rely on the absolute ordering of the dependence measure. Instead of the magnitude of the dependence measure, we rather order variables by their frequency to pass the independence test in bootstrapping samples. The size of a model is decided by the amount of stability we want to achieve (70, 80, or 90% of total bootstrapping samples). The work is motivated by a project selecting genes related to suicidal behaviors. The number of genes is greater than the sample size, and we don't have any knowledge about how genes are related to the probability of committing suicidal behaviors. Longitudinal observations are available for some subjects, and we wish to conduct a gene cluster selection where each cluster has different sizes. We will illustrate that the proposed method is proper to conduct these analyses.

Incorporating a resampling procedure into variable selection is intuitively appealing and certainly not a new idea. Meinshausen and Bühlmann (2010) provides some formal framework on this and illustrates cases where this approach can be combined with various kinds of selection methods. They mention that selecting variables based on "a stability measure", the probability for each variable to be selected in a random resampling of the data, can provide better separation of relevant variables from irrelevant ones, proper guidance to decide an amount of regularization (size of the model), and sometimes achieves a consistent selection with settings where the original methods fail. We found that this framework is particularly well-suited when combined with a test of independence, as it opens up the possibility to convert a mere test to a new selection method with a completely model-free setting while properly addressing the problems of the scale-based approach.

The remaining of this chapter is organized as follows. In Section 4.2, we introduce a new method based on a frequent voting. In Section 4.3, we assess the finite sample performance of the proposed method with comparison to the scale-based method and LASSO. In Section 4.4, we apply the method to select important genes related to suicidal behaviors.

We present various selection results including unconditional selection, multivariate selection, and conditional selection to illustrate possible extensions of our proposed method.

4.2 FREQUENT VOTING METHOD

Let Y be a response vector and $X_1, ..., X_p$ be vectors of predictors. A proper test of independence needs to be chosen depending on the dimension of the data. For illustration, in this chapter, we use distance covariance measure and the IPR-τ^* measure for $(X_j, Y) \in (\mathbb{R}^p, \mathbb{R}^q)$. We use distance covariance measure and τ^* for $(X_j, Y) \in (\mathbb{R}, \mathbb{R})$.

To start, we create a set of bootstrapping samples $(X_1^*, ..., X_p^*, Y^*)$ and perform a test of independence between X_j^* and Y^* for $j = 1, ..., p$, separately. If the test rejects the null hypothesis for X_j, then X_j gets one vote. We repeat this procedure B times, a predefined number of bootstraps, and collect the vote for each variable X_j throughout the B times repeated procedure. We finally order $X_1, ..., X_p$ by the vote they have, whose possible range is between 0 and B. We decide the cutoff by an amount of stability we want to achieve. We may select every X_j whose vote exceeds $0.8B$.

The suggested method is simple and intuitive but possesses advantages over the scale-based method. First, the method is free from a direct comparison of measures between predictors. The method only uses whether the measure of independence exceeds the α-level critical value of the null distribution, so the decision we make for the procedure is supported by the large sample theory of the independence test. Second, the method has a natural way of selecting a cutoff; rather than deciding the size of the model, one chooses a level of stability to achieve. Third, the method allows predictors to have different dimensions. The independence measure are not compared between groups of predictors with different dimensions, but they are compared with their critical value.

This simple method is mainly aimed at screening and retaining important variables. A second step model-based approach with the selected variables may be employed to assess the prediction error. Some other extensions such as iterative screening based on residuals as discussed in Pan et al. (2018) can also be considered.

4.3 SIMULATION

In this section, we conduct a Monte-Carlo simulation to assess the empirical performance of the proposed frequent voting method, compared to the scale-based method and LASSO. Two measures of dependency, distance covariance ("dCov") and sign covariance (τ^*) are employed for illustration. A random tie-break is used when τ^* is employed. The results are based on the implementation in R packages "dcov", "TauStar" and "glmnet", respectively.

We generate $X = (X_1, ..X_p)^T$ from a multivariate normal distribution with zero mean and covariance matrix $\Sigma = (\sigma_{ij})_{p\times p}$ where $\sigma_{ij} = \rho^{|i-j|}$. We consider three levels of intercorrelation (a) $\rho = 0$, (b) $\rho = 0.2$ and (c) $\rho = 0.8$. We fix the dimension p to be 200, and consider sample size $n = 100, 150, 200$ for the mixture examples in 4.1-4.3 and $n = 200, 400$ for logistic examples in 4.4-4.5.

For the frequent voting method, the results are based on $B = 100$ of bootstrapped samples with a significance level of 0.05 for the independence test. Critical values are approximated by a permutation procedure with data generated under each setting and they are averaged over 5 times. For the scale-based method, normalized version of measures are used for a comparison which are

$$dCor_j = \frac{dCov(X_j, Y)}{\sqrt{dCov(X_j, X_j)dCov(Y, Y)}} \text{ and } \tau^*_{b,j} = \frac{\tau^*(X_j, Y)}{\sqrt{\tau^*(X_j, X_j)\tau^*(Y, Y)}}$$

for $j = 1, ..., p$. A linear LASSO model is used in Examples 4.1-4.3, and a logistic LASSO model is used in Example 4.4-4.5.

We evaluate the performance through $\mathcal{P}_1$, $\mathcal{P}_{11}$, $\mathcal{P}_{21}$, $\mathcal{P}_{31}$ and $\mathcal{P}_{\text{all}}$ which are the proportions that each or all active variables (X_1, X_{11}, X_{21}, X_{31}) are selected in 200 replications, for a given size of the model. Here we do not really have a model, the size of the model is the total number of variables selected. For the frequent voting method, the size of the model is determined by a frequency cutoff; variables are selected if they achieve 90%, 80%, or 70% of the vote from bootstrapped data. Since the scale-based method doesn't have a common way to determine a cutoff and for comparison purpose, their size of selection is matched with the frequent voting method. For LASSO, the size of the model is determined to achieve a

minimum mean squared error through cross-validation. We hope to retain the true variables with high probability with a reasonable size of the model.

Example 4.1-4.3 are designed to illustrate the various mixture relationships. Response variable Y is generated from the following models which include a linear term, quadratic term, interaction term, and other nonlinear relationships.

Example 4.1. $Y = 0.8X_1 + X_{11} + 0.8X_{21}^2 + X_{31}^2 + \epsilon,\ \epsilon \sim N(0,1)$

Example 4.2. $Y = X_1 + 1.5X_{11}^2 + 4.5I(X_{21}^2 > 0.675^2)I(X_{31}^2 > 0.675^2) + \epsilon,\ \epsilon \sim N(0,1)$

Example 4.3. $Y = X_1 + 4X_{11}I(X_{21} < 0) + \exp(|X_{31}|) + \epsilon,\ \epsilon \sim N(0,1)$

The results are summarized in Tables 2 - 10. In each table, the nine columns represent the results under three different frequency cutoffs ($d = 90\%, 80\%, 70\%$) and with three levels of intercorrelation between covariates ($\rho = 0, 0.2, 0.8$). As expected, we observe that the average size of the model increases as the intercorrelation becomes bigger as well as the cutoff becomes lower. For each example, we present results for $n = 100$, $n = 150$ and $n = 200$ in three separate tables. The performance of the frequent voting method and the scale-based method improves as the sample size increases, but the LASSO method does not improve much due to the lack of ability to detect nonlinear relationships. Varying the model size for LASSO method does not improve the performance much for the same reason. We include the LASSO method as a benchmark for baseline comparison. In the following, we discuss the performance of the frequent voting method and the scale-based method.

Overall, the performance of the frequent voting method is very satisfactory with a moderate sample size $n = 150$ or 200, in the sense that the probability of retaining all important variable is close to 1 with a reasonable model size. In all of the settings, the frequent voting method is always better than the scale-based method, when τ^* is used as an independence measure, in the sense that the probability of retaining all true variables is much higher with the frequent voting method when the size of the model is kept the same. Using "dCov" as a dependence measure, the performance of the frequent voting method and the scale-based method is similar in Example 4.1 as shown in Tables 2-4. The frequent voting method is clearly better in Example 4.2 as shown in Tables 5-7. The frequent voting method is mod-

erately better than the scale-based method in Example 4.3 as shown in Tables 8-10. This is reasonable. When the chosen dependence is suitable and has a clear scale ordering for detecting the true relationship, the scale-based method will be relatively stable in practice. Overall, the frequent voting method largely stablize the performance and works well in all cases with a moderate sample size $n = 150$.

Example 4.4 and Example 4.5 are designed to assess the performance of methods when Y is binary. In both examples, Y follows the Bernoulli distribution with probability $1/1 + e^p$ and p is generated as follows. Example 4.4 is a logistic version of Example 4.1 and Example 4.4 illustrate various relationships without linear terms.

Example 4.4. $p = 0.8X_1 - X_{11} + 0.8X_{21}^2 - X_{31}^2$

Example 4.5. $p = X_1^2 - X_{11}^2 + 2I(X_{21} > 0.675) - 2I(X_{31}^2 > 0.675^2)$

The simulations results are shown in Tables 11-14. As expected, we observe that the average size of the selected model increases as the intercorrelation becomes bigger as well as the cutoff becomes lower. For each example, we present results for $n = 200$, and $n = 400$ in separate tables. The performance of the frequent voting method and the scale-based method improves as the sample size increases, but the LASSO method does not improve much due to the lack of ability to detect nonlinear relationships. Overall, the performance of the frequent voting method is very satisfactory with a sample size $n = 400$, in the sense that the probability of retaining all important variable is close to 1 with a moderate model size. In all of the settings, the frequent voting method is better than the scale-based method, in the sense that the probability of retaining all true variables is higher with the frequent voting method when the size of the model is kept the same.

			4.1(a)			4.1(b)			4.1(c)		
			d1	d2	d3	d1	d2	d3	d1	d2	d3
dCov	Freq	$\mathcal{P}_1$	0.72	0.87	0.93	0.73	0.89	0.95	0.87	0.95	0.96
		$\mathcal{P}_{11}$	0.96	1.00	1.00	0.95	0.99	1.00	0.98	1.00	1.00
		$\mathcal{P}_{21}$	0.46	0.62	0.77	0.48	0.68	0.81	0.73	0.89	0.95
		$\mathcal{P}_{31}$	0.75	0.87	0.93	0.81	0.91	0.94	0.85	0.95	0.97
		$\mathcal{P}_{all}$	**0.29**	**0.50**	**0.67**	**0.28**	**0.53**	**0.72**	**0.52**	**0.80**	**0.87**
	Scale	$\mathcal{P}_1$	0.78	0.90	0.95	0.79	0.90	0.95	0.88	0.96	0.99
		$\mathcal{P}_{11}$	0.98	0.99	1.00	0.98	1.00	1.00	0.99	1.00	1.00
		$\mathcal{P}_{21}$	0.39	0.61	0.77	0.45	0.70	0.84	0.60	0.83	0.91
		$\mathcal{P}_{31}$	0.70	0.83	0.95	0.74	0.91	0.93	0.84	0.93	0.98
		$\mathcal{P}_{all}$	**0.25**	**0.50**	**0.72**	**0.26**	**0.56**	**0.73**	**0.47**	**0.75**	**0.87**
	Avg.Size		4.9	10.8	21.5	5.7	12.8	24.3	17.1	30.6	47.8
τ^*	Freq	$\mathcal{P}_1$	0.81	0.91	0.95	0.84	0.92	0.97	0.87	0.94	0.97
		$\mathcal{P}_{11}$	0.99	1.00	1.00	0.97	1.00	1.00	1.00	1.00	1.00
		$\mathcal{P}_{21}$	0.42	0.63	0.76	0.43	0.64	0.74	0.50	0.72	0.86
		$\mathcal{P}_{31}$	0.78	0.90	0.98	0.76	0.88	0.93	0.76	0.87	0.93
		$\mathcal{P}_{all}$	**0.25**	**0.50**	**0.70**	**0.24**	**0.49**	**0.66**	**0.32**	**0.57**	**0.77**
	Scale	$\mathcal{P}_1$	0.85	0.92	0.98	0.86	0.92	0.96	0.90	0.96	0.98
		$\mathcal{P}_{11}$	1.00	1.00	1.00	0.99	1.00	1.00	1.00	1.00	1.00
		$\mathcal{P}_{21}$	0.20	0.49	0.67	0.24	0.48	0.69	0.23	0.52	0.73
		$\mathcal{P}_{31}$	0.49	0.80	0.91	0.47	0.76	0.89	0.49	0.77	0.90
		$\mathcal{P}_{all}$	**0.08**	**0.36**	**0.59**	**0.06**	**0.31**	**0.57**	**0.08**	**0.37**	**0.63**
	Avg.Size		6.2	13.2	24.8	6.4	13.4	24.9	14.5	24.1	36.8
Lasso		$\mathcal{P}_1$		0.92			0.87			0.91	
		$\mathcal{P}_{11}$		0.99			0.95			0.94	
		$\mathcal{P}_{21}$		0.13			0.17			0.12	
		$\mathcal{P}_{31}$		0.18			0.21			0.13	
		$\mathcal{P}_{all}$		**0.03**			**0.03**			**0.01**	
	Avg.size			13.3			15.6			12.0	

Table 2: Simulation result of Example 4.1 for n=100. $\mathcal{P}_1$, $\mathcal{P}_{11}$, $\mathcal{P}_{21}$, $\mathcal{P}_{31}$ and $\mathcal{P}_{all}$ are proportions that each or every active variables $(X_1, X_{11}, X_{21}, X_{31})$ are included in selection with frequency cutoff 90%(d_1), 80%(d_2) and 70%(d_1) over 200 repetitions. Avg.Size means an average size of selection with a given cutoff.

			4.1(a)			4.1(b)			4.1(c)		
			d1	d2	d3	d1	d2	d3	d1	d2	d3
dCov	Freq	$\mathcal{P}_1$	0.93	0.96	0.97	0.94	0.97	0.99	1.00	1.00	1.00
		$\mathcal{P}_{11}$	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		$\mathcal{P}_{21}$	0.83	0.93	0.96	0.81	0.90	0.97	0.83	0.94	0.98
		$\mathcal{P}_{31}$	0.97	0.98	1.00	0.98	1.00	1.00	0.95	1.00	1.00
		$\mathcal{P}_{all}$	**0.75**	**0.87**	**0.93**	**0.75**	**0.87**	**0.96**	**0.78**	**0.94**	**0.98**
	Scale	$\mathcal{P}_1$	0.96	0.98	0.99	0.96	0.99	1.00	1.00	1.00	1.00
		$\mathcal{P}_{11}$	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		$\mathcal{P}_{21}$	0.79	0.93	0.97	0.80	0.94	1.00	0.75	0.88	0.97
		$\mathcal{P}_{31}$	0.94	0.98	0.99	0.97	0.99	1.00	0.96	0.99	1.00
		$\mathcal{P}_{all}$	**0.71**	**0.91**	**0.96**	**0.73**	**0.92**	**1.00**	**0.73**	**0.87**	**0.97**
	Avg.Size		7.9	17.6	33.8	9.3	21.4	40.0	19.6	30.5	44.8
τ^*	Freq	$\mathcal{P}_1$	0.94	0.97	0.99	0.95	0.97	0.99	0.99	1.00	1.00
		$\mathcal{P}_{11}$	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		$\mathcal{P}_{21}$	0.77	0.89	0.97	0.77	0.89	0.95	0.79	0.88	0.94
		$\mathcal{P}_{31}$	0.94	0.98	1.00	0.98	1.00	1.00	0.96	0.99	1.00
		$\mathcal{P}_{all}$	**0.67**	**0.85**	**0.95**	**0.71**	**0.86**	**0.94**	**0.75**	**0.87**	**0.93**
	Scale	$\mathcal{P}_1$	0.97	0.97	0.99	0.95	0.98	1.00	1.00	1.00	1.00
		$\mathcal{P}_{11}$	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		$\mathcal{P}_{21}$	0.44	0.75	0.93	0.49	0.73	0.90	0.52	0.76	0.88
		$\mathcal{P}_{31}$	0.86	0.94	0.99	0.83	0.96	1.00	0.83	0.96	0.99
		$\mathcal{P}_{all}$	**0.36**	**0.68**	**0.90**	**0.35**	**0.68**	**0.89**	**0.40**	**0.73**	**0.87**
	Avg.Size		7.1	14.7	26.6	7.5	15.1	28.0	18.9	28.6	42.0
Lasso		$\mathcal{P}_1$		0.98			0.96			0.97	
		$\mathcal{P}_{11}$		1.00			1.00			0.99	
		$\mathcal{P}_{21}$		0.20			0.13			0.09	
		$\mathcal{P}_{31}$		0.22			0.20			0.19	
		$\mathcal{P}_{all}$		**0.07**			**0.03**			**0.02**	
	Avg.size			15.1			13.8			13.0	

Table 3: Simulation result of Example 4.1 for n=150. $\mathcal{P}_1$, $\mathcal{P}_{11}$, $\mathcal{P}_{21}$, $\mathcal{P}_{31}$ and $\mathcal{P}_{all}$ are proportions that each or every active variables $(X_1, X_{11}, X_{21}, X_{31})$ are included in selection with frequency cutoff 90%(d_1), 80%(d_2) and 70%(d_1) over 200 repetitions. Avg.size means an average size of selection with a given cutoff.

			4.1(a)			4.1(b)			4.1(c)		
			d1	d2	d3	d1	d2	d3	d1	d2	d3
dCov	Freq	$\mathcal{P}_1$	0.99	1.00	1.00	0.97	1.00	1.00	1.00	1.00	1.00
		$\mathcal{P}_{11}$	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		$\mathcal{P}_{21}$	0.98	1.00	1.00	1.00	1.00	1.00	0.97	0.99	1.00
		$\mathcal{P}_{31}$	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		$\mathcal{P}_{all}$	**0.96**	**0.99**	**1.00**	**0.96**	**1.00**	**1.00**	**0.96**	**0.99**	**1.00**
	Scale	$\mathcal{P}_1$	1.00	1.00	1.00	0.98	1.00	1.00	1.00	1.00	1.00
		$\mathcal{P}_{11}$	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		$\mathcal{P}_{21}$	0.96	0.99	1.00	0.93	1.00	1.00	0.91	0.99	1.00
		$\mathcal{P}_{31}$	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		$\mathcal{P}_{all}$	**0.95**	**0.99**	**1.00**	**0.91**	**0.99**	**1.00**	**0.91**	**0.99**	**1.00**
	Avg.Size		7.9	16.9	31.7	8.0	16.9	31.6	22.7	33.1	47.4
τ^*	Freq	$\mathcal{P}_1$	0.99	1.00	1.00	0.98	0.99	1.00	1.00	1.00	1.00
		$\mathcal{P}_{11}$	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		$\mathcal{P}_{21}$	0.93	0.98	1.00	0.94	1.00	1.00	0.95	0.99	1.00
		$\mathcal{P}_{31}$	0.99	1.00	1.00	1.00	1.00	1.00	0.99	1.00	1.00
		$\mathcal{P}_{all}$	**0.91**	**0.98**	**1.00**	**0.92**	**0.99**	**1.00**	**0.94**	**0.99**	**1.00**
	Scale	$\mathcal{P}_1$	1.00	1.00	1.00	0.98	0.99	1.00	1.00	1.00	1.00
		$\mathcal{P}_{11}$	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		$\mathcal{P}_{21}$	0.72	0.92	0.98	0.70	0.93	1.00	0.77	0.93	0.99
		$\mathcal{P}_{31}$	0.97	1.00	1.00	0.98	1.00	1.00	0.97	0.99	1.00
		$\mathcal{P}_{all}$	**0.69**	**0.92**	**0.98**	**0.68**	**0.92**	**1.00**	**0.74**	**0.92**	**0.99**
	Avg.Size		7.5	15.2	27.9	8.0	15.9	28.9	21.8	31.5	44.9
Lasso		$\mathcal{P}_1$		1.00			0.99			1.00	
		$\mathcal{P}_{11}$		1.00			1.00			0.99	
		$\mathcal{P}_{21}$		0.16			0.13			0.11	
		$\mathcal{P}_{31}$		0.18			0.18			0.20	
		$\mathcal{P}_{all}$		**0.04**			**0.03**			**0.02**	
	Avg.size			13.8			12.7			13.0	

Table 4: Simulation result of Example 4.1 for n=200. $\mathcal{P}_1$, $\mathcal{P}_{11}$, $\mathcal{P}_{21}$, $\mathcal{P}_{31}$ and $\mathcal{P}_{all}$ are proportions that each or every active variables $(X_1, X_{11}, X_{21}, X_{31})$ are included in selection with frequency cutoff 90%(d_1), 80%(d_2) and 70%(d_1) over 200 repetitions. Avg.size means an average size of selection with a given cutoff.

			4.2(a)			4.2(b)			4.2(c)		
			d1	d2	d3	d1	d2	d3	d1	d2	d3
dCov	Freq	$\mathcal{P}_1$	0.72	0.83	0.91	0.72	0.89	0.93	0.78	0.91	0.95
		$\mathcal{P}_{11}$	0.84	0.94	0.97	0.85	0.94	0.98	0.91	0.97	0.98
		$\mathcal{P}_{21}$	0.34	0.54	0.69	0.29	0.49	0.66	0.46	0.68	0.81
		$\mathcal{P}_{31}$	0.34	0.55	0.66	0.41	0.60	0.74	0.47	0.64	0.76
		$\mathcal{P}_{all}$	**0.11**	**0.29**	**0.46**	**0.10**	**0.26**	**0.46**	**0.20**	**0.40**	**0.61**
	Scale	$\mathcal{P}_1$	0.74	0.86	0.93	0.76	0.89	0.95	0.84	0.92	0.96
		$\mathcal{P}_{11}$	0.84	0.96	0.98	0.84	0.95	0.99	0.91	0.98	1.00
		$\mathcal{P}_{21}$	0.18	0.43	0.59	0.19	0.40	0.62	0.31	0.54	0.73
		$\mathcal{P}_{31}$	0.22	0.39	0.55	0.19	0.38	0.60	0.33	0.58	0.77
		$\mathcal{P}_{all}$	**0.04**	**0.16**	**0.36**	**0.02**	**0.12**	**0.37**	**0.08**	**0.32**	**0.54**
	Avg.Size		3.8	8.4	16.2	4.0	9.3	18.3	8.6	19.2	34.3
τ^*	Freq	$\mathcal{P}_1$	0.84	0.96	0.99	0.92	0.97	1.00	0.86	0.94	0.98
		$\mathcal{P}_{11}$	0.93	0.98	0.99	0.94	0.98	0.99	0.84	0.94	0.98
		$\mathcal{P}_{21}$	0.34	0.55	0.72	0.34	0.55	0.69	0.34	0.49	0.67
		$\mathcal{P}_{31}$	0.38	0.60	0.73	0.34	0.62	0.74	0.36	0.55	0.74
		$\mathcal{P}_{all}$	**0.08**	**0.29**	**0.52**	**0.08**	**0.31**	**0.46**	**0.08**	**0.24**	**0.46**
	Scale	$\mathcal{P}_1$	0.87	0.97	0.98	0.91	0.96	0.98	0.86	0.93	0.98
		$\mathcal{P}_{11}$	0.72	0.91	0.98	0.70	0.89	0.97	0.74	0.90	0.98
		$\mathcal{P}_{21}$	0.15	0.37	0.62	0.16	0.42	0.63	0.18	0.39	0.61
		$\mathcal{P}_{31}$	0.19	0.43	0.62	0.15	0.38	0.60	0.22	0.43	0.68
		$\mathcal{P}_{all}$	**0.01**	**0.12**	**0.37**	**0.01**	**0.09**	**0.34**	**0.02**	**0.09**	**0.38**
	Avg.Size		5.7	12.9	24.3	5.8	13.3	24.4	7.6	16.2	29.4
Lasso		$\mathcal{P}_1$		0.72			0.70			0.71	
		$\mathcal{P}_{11}$		0.18			0.11			0.10	
		$\mathcal{P}_{21}$		0.06			0.05			0.04	
		$\mathcal{P}_{31}$		0.08			0.05			0.05	
		$\mathcal{P}_{all}$		**0.01**			**0.00**			**0.01**	
	Avg.size			10.5			8.9			8.0	

Table 5: Simulation result of Example 4.2 for $n = 100$. $\mathcal{P}_1$, $\mathcal{P}_{11}$, $\mathcal{P}_{21}$, $\mathcal{P}_{31}$ and $\mathcal{P}_{all}$ are proportions that each or every active variables $(X_1, X_{11}, X_{21}, X_{31})$ are included in selection with frequency cutoff 90%(d_1), 80%(d_2) and 70%(d_1) over 200 repetitions. Avg.size means an average size of selection with a given cutoff.

			4.2(a)			4.2(b)			4.2(c)		
			d1	d2	d3	d1	d2	d3	d1	d2	d3
dCov	Freq	$\mathcal{P}_1$	0.98	1.00	1.00	0.97	0.99	1.00	0.93	0.98	0.99
		$\mathcal{P}_{11}$	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		$\mathcal{P}_{21}$	0.72	0.86	0.92	0.83	0.91	0.98	0.66	0.82	0.90
		$\mathcal{P}_{31}$	0.80	0.92	0.97	0.78	0.92	0.96	0.76	0.94	0.97
		$\mathcal{P}_{all}$	**0.58**	**0.80**	**0.89**	**0.66**	**0.84**	**0.94**	**0.52**	**0.75**	**0.87**
	Scale	$\mathcal{P}_1$	0.96	1.00	1.00	0.98	1.00	1.00	0.94	0.99	0.99
		$\mathcal{P}_{11}$	0.99	0.99	1.00	0.99	1.00	1.00	1.00	1.00	1.00
		$\mathcal{P}_{21}$	0.53	0.77	0.90	0.61	0.82	0.94	0.55	0.76	0.89
		$\mathcal{P}_{31}$	0.59	0.82	0.93	0.54	0.79	0.94	0.57	0.79	0.92
		$\mathcal{P}_{all}$	**0.31**	**0.65**	**0.83**	**0.34**	**0.64**	**0.88**	**0.32**	**0.60**	**0.81**
	Avg.Size		6.7	14.6	27.8	7.0	14.7	27.7	9.6	17.6	29.1
τ^*	Freq	$\mathcal{P}_1$	1.00	1.00	1.00	0.99	0.99	1.00	0.98	0.99	1.00
		$\mathcal{P}_{11}$	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		$\mathcal{P}_{21}$	0.62	0.81	0.89	0.71	0.87	0.94	0.72	0.88	0.94
		$\mathcal{P}_{31}$	0.71	0.85	0.93	0.75	0.88	0.95	0.76	0.94	0.98
		$\mathcal{P}_{all}$	**0.42**	**0.70**	**0.83**	**0.53**	**0.78**	**0.90**	**0.53**	**0.81**	**0.92**
	Scale	$\mathcal{P}_1$	0.99	1.00	1.00	0.98	1.00	1.00	0.98	0.99	1.00
		$\mathcal{P}_{11}$	0.96	0.99	1.00	0.96	1.00	1.00	0.98	1.00	1.00
		$\mathcal{P}_{21}$	0.40	0.70	0.85	0.48	0.75	0.91	0.49	0.78	0.91
		$\mathcal{P}_{31}$	0.45	0.71	0.86	0.41	0.71	0.88	0.52	0.75	0.89
		$\mathcal{P}_{all}$	**0.15**	**0.48**	**0.73**	**0.18**	**0.54**	**0.80**	**0.25**	**0.57**	**0.80**
	Avg.Size		6.7	14.4	26.5	7.1	14.7	26.3	10.4	19.5	32.8
Lasso		$\mathcal{P}_1$		0.87			0.84			0.88	
		$\mathcal{P}_{11}$		0.17			0.16			0.12	
		$\mathcal{P}_{21}$		0.04			0.07			0.03	
		$\mathcal{P}_{31}$		0.08			0.06			0.06	
		$\mathcal{P}_{all}$		**0.00**			**0.01**			**0.00**	
	Avg.size			10.3			9.9			7.6	

Table 6: Simulation result of Example 4.2 for $n = 150$. $\mathcal{P}_1$, $\mathcal{P}_{11}$, $\mathcal{P}_{21}$, $\mathcal{P}_{31}$ and $\mathcal{P}_{all}$ are proportions that each or every active variables $(X_1, X_{11}, X_{21}, X_{31})$ are included in selection with frequency cutoff 90%(d_1), 80%(d_2) and 70%(d_1) over 200 repetitions. Avg.size means an average size of selection with a given cutoff.

			4.2(a)			4.2(b)			4.2(c)		
			d1	d2	d3	d1	d2	d3	d1	d2	d3
dCov	Freq	$\mathcal{P}_1$	1.00	1.00	1.00	1.00	1.00	1.00	0.99	1.00	1.00
		$\mathcal{P}_{11}$	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		$\mathcal{P}_{21}$	0.90	0.98	1.00	0.93	0.97	0.99	0.91	0.97	0.99
		$\mathcal{P}_{31}$	0.94	0.98	0.99	0.86	0.94	0.99	0.91	0.96	0.97
		$\mathcal{P}_{all}$	**0.83**	**0.96**	**0.98**	**0.81**	**0.92**	**0.98**	**0.81**	**0.92**	**0.96**
	Scale	$\mathcal{P}_1$	1.00	1.00	1.00	1.00	1.00	1.00	0.98	0.99	1.00
		$\mathcal{P}_{11}$	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		$\mathcal{P}_{21}$	0.79	0.94	0.99	0.77	0.93	0.97	0.72	0.93	0.97
		$\mathcal{P}_{31}$	0.83	0.94	0.97	0.77	0.90	0.97	0.84	0.93	0.97
		$\mathcal{P}_{all}$	**0.64**	**0.89**	**0.96**	**0.61**	**0.84**	**0.95**	**0.59**	**0.85**	**0.94**
	Avg.Size		7.6	16.1	30.5	7.3	14.7	27.3	12.7	21.7	34.7
τ^*	Freq	$\mathcal{P}_1$	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		$\mathcal{P}_{11}$	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		$\mathcal{P}_{21}$	0.91	0.97	0.99	0.89	0.96	0.98	0.89	0.98	0.99
		$\mathcal{P}_{31}$	0.91	0.97	0.99	0.87	0.97	0.98	0.89	0.95	0.99
		$\mathcal{P}_{all}$	**0.83**	**0.94**	**0.97**	**0.78**	**0.93**	**0.96**	**0.79**	**0.93**	**0.98**
	Scale	$\mathcal{P}_1$	1.00	1.00	1.00	1.00	1.00	1.00	0.99	1.00	1.00
		$\mathcal{P}_{11}$	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		$\mathcal{P}_{21}$	0.68	0.89	0.98	0.65	0.86	0.96	0.68	0.90	0.98
		$\mathcal{P}_{31}$	0.70	0.91	0.97	0.62	0.86	0.96	0.75	0.91	0.97
		$\mathcal{P}_{all}$	**0.43**	**0.81**	**0.95**	**0.41**	**0.75**	**0.93**	**0.50**	**0.82**	**0.95**
	Avg.Size		7.2	14.8	26.9	7.4	15.2	27.6	12.9	23.1	37.4
Lasso		$\mathcal{P}_1$		0.92			0.96			0.91	
		$\mathcal{P}_{11}$		0.17			0.16			0.15	
		$\mathcal{P}_{21}$		0.03			0.05			0.03	
		$\mathcal{P}_{31}$		0.06			0.08			0.05	
		$\mathcal{P}_{all}$		**0.00**			**0.00**			**0.01**	
	Avg.size			8.8			9.0			7.0	

Table 7: Simulation result of Example 4.2 for $n = 200$. $\mathcal{P}_1$, $\mathcal{P}_{11}$, $\mathcal{P}_{21}$, $\mathcal{P}_{31}$ and $\mathcal{P}_{all}$ are proportions that each or every active variables $(X_1, X_{11}, X_{21}, X_{31})$ are included in selection with frequency cutoff 90%(d_1), 80%(d_2) and 70%(d_1) over 200 repetitions. Avg.size means an average size of selection with a given cutoff.

			4.3(a)			4.3(b)			4.3(c)		
			d1	d2	d3	d1	d2	d3	d1	d2	d3
dCov	Freq	$\mathcal{P}_1$	0.62	0.79	0.84	0.57	0.73	0.84	0.77	0.88	0.96
		$\mathcal{P}_{11}$	1.00	1.00	1.00	0.98	0.99	1.00	1.00	1.00	1.00
		$\mathcal{P}_{21}$	0.55	0.73	0.81	0.37	0.60	0.74	0.64	0.81	0.90
		$\mathcal{P}_{31}$	0.78	0.92	0.97	0.80	0.90	0.94	0.87	0.94	0.97
		$\mathcal{P}_{all}$	**0.29**	**0.52**	**0.66**	**0.17**	**0.43**	**0.62**	**0.47**	**0.70**	**0.85**
	Scale	$\mathcal{P}_1$	0.66	0.81	0.87	0.58	0.73	0.85	0.86	0.96	0.99
		$\mathcal{P}_{11}$	1.00	1.00	1.00	0.98	0.99	1.00	1.00	1.00	1.00
		$\mathcal{P}_{21}$	0.38	0.64	0.83	0.24	0.53	0.75	0.48	0.74	0.88
		$\mathcal{P}_{31}$	0.78	0.92	0.97	0.78	0.90	0.94	0.85	0.93	0.98
		$\mathcal{P}_{all}$	**0.22**	**0.48**	**0.70**	**0.10**	**0.38**	**0.63**	**0.41**	**0.68**	**0.86**
	Avg.Size		6.9	16.4	31.3	5.0	11.5	22.7	19.1	33.2	51.8
τ^*	Freq	$\mathcal{P}_1$	0.71	0.82	0.88	0.67	0.81	0.89	0.79	0.94	0.98
		$\mathcal{P}_{11}$	1.00	1.00	1.00	0.98	1.00	1.00	1.00	1.00	1.00
		$\mathcal{P}_{21}$	0.49	0.68	0.85	0.39	0.59	0.76	0.47	0.66	0.85
		$\mathcal{P}_{31}$	0.77	0.90	0.97	0.88	0.94	0.96	0.80	0.90	0.94
		$\mathcal{P}_{all}$	**0.23**	**0.49**	**0.71**	**0.21**	**0.45**	**0.66**	**0.30**	**0.55**	**0.77**
	Scale	$\mathcal{P}_1$	0.74	0.85	0.91	0.71	0.84	0.91	0.91	0.96	0.98
		$\mathcal{P}_{11}$	1.00	1.00	1.00	0.99	1.00	1.00	1.00	1.00	1.00
		$\mathcal{P}_{21}$	0.24	0.50	0.72	0.17	0.41	0.67	0.25	0.50	0.74
		$\mathcal{P}_{31}$	0.59	0.81	0.92	0.58	0.83	0.94	0.64	0.86	0.93
		$\mathcal{P}_{all}$	**0.09**	**0.33**	**0.58**	**0.07**	**0.26**	**0.57**	**0.15**	**0.42**	**0.67**
	Avg.Size		6.3	13.4	25.3	6.1	13.1	24.4	15.7	25.8	38.2
Lasso		$\mathcal{P}_1$		0.71			0.66			0.70	
		$\mathcal{P}_{11}$		0.96			0.97			0.96	
		$\mathcal{P}_{21}$		0.05			0.04			0.02	
		$\mathcal{P}_{31}$		0.22			0.24			0.18	
		$\mathcal{P}_{all}$		**0.02**			**0.02**			**0.01**	
	Avg.size			13.5			11.4			10.9	

Table 8: Simulation result of Example 4.3 for $n = 100$. $\mathcal{P}_1$, $\mathcal{P}_{11}$, $\mathcal{P}_{21}$, $\mathcal{P}_{31}$ and $\mathcal{P}_{all}$ are proportions that each or every active variables $(X_1, X_{11}, X_{21}, X_{31})$ are included in selection with frequency cutoff 90%(d_1), 80%(d_2) and 70%(d_1) over 200 repetitions. Avg.size means an average size of selection with a given cutoff.

			4.3(a)			4.3(b)			4.3(c)		
			d1	d2	d3	d1	d2	d3	d1	d2	d3
dCov	Freq	$\mathcal{P}_1$	0.84	0.96	0.99	0.91	0.95	0.98	0.96	0.99	1.00
		$\mathcal{P}_{11}$	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		$\mathcal{P}_{21}$	0.86	0.95	0.98	0.87	0.98	0.99	0.89	0.97	0.98
		$\mathcal{P}_{31}$	0.98	1.00	1.00	0.99	1.00	1.00	0.97	0.99	1.00
		$\mathcal{P}_{all}$	**0.72**	**0.91**	**0.96**	**0.79**	**0.93**	**0.96**	**0.83**	**0.95**	**0.97**
	Scale	$\mathcal{P}_1$	0.90	0.97	0.98	0.90	0.97	0.99	0.98	0.99	1.00
		$\mathcal{P}_{11}$	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		$\mathcal{P}_{21}$	0.74	0.88	0.96	0.79	0.94	0.98	0.82	0.96	0.98
		$\mathcal{P}_{31}$	0.98	1.00	1.00	0.98	1.00	1.00	0.97	0.99	1.00
		$\mathcal{P}_{all}$	**0.66**	**0.86**	**0.93**	**0.71**	**0.90**	**0.97**	**0.78**	**0.94**	**0.97**
	Avg.Size		8.9	21.2	40.2	12.3	29.9	54.3	23.0	36.1	52.9
τ^*	Freq	$\mathcal{P}_1$	0.86	0.96	0.99	0.90	0.96	0.98	0.99	1.00	1.00
		$\mathcal{P}_{11}$	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		$\mathcal{P}_{21}$	0.77	0.88	0.93	0.81	0.93	0.98	0.85	0.95	0.97
		$\mathcal{P}_{31}$	0.99	1.00	1.00	1.00	1.00	1.00	0.97	1.00	1.00
		$\mathcal{P}_{all}$	**0.65**	**0.85**	**0.91**	**0.72**	**0.88**	**0.96**	**0.82**	**0.94**	**0.97**
	Scale	$\mathcal{P}_1$	0.93	0.98	0.99	0.91	0.95	0.99	0.99	1.00	1.00
		$\mathcal{P}_{11}$	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		$\mathcal{P}_{21}$	0.50	0.77	0.87	0.48	0.73	0.90	0.65	0.88	0.96
		$\mathcal{P}_{31}$	0.91	0.98	1.00	0.91	0.99	1.00	0.87	0.98	1.00
		$\mathcal{P}_{all}$	**0.39**	**0.73**	**0.85**	**0.40**	**0.67**	**0.88**	**0.56**	**0.86**	**0.95**
	Avg.Size		7.0	14.0	25.5	7.5	15.1	27.8	20.8	30.8	44.3
Lasso		$\mathcal{P}_1$		0.88			0.87			0.91	
		$\mathcal{P}_{11}$		0.99			1.00			1.00	
		$\mathcal{P}_{21}$		0.06			0.07			0.05	
		$\mathcal{P}_{31}$		0.28			0.28			0.20	
		$\mathcal{P}_{all}$		**0.03**			**0.02**			**0.00**	
	Avg.size			12.6			12.4			12.2	

Table 9: Simulation result of Example 4.3 for $n = 150$. $\mathcal{P}_1$, $\mathcal{P}_{11}$, $\mathcal{P}_{21}$, $\mathcal{P}_{31}$ and $\mathcal{P}_{all}$ are proportions that each or every active variables $(X_1, X_{11}, X_{21}, X_{31})$ are included in selection with frequency cutoff 90%(d_1), 80%(d_2) and 70%(d_1) over 200 repetitions. Avg.size means an average size of selection with a given cutoff.

			4.3(a)			4.3(b)			4.3(c)		
			d1	d2	d3	d1	d2	d3	d1	d2	d3
dCov	Freq	$\mathcal{P}_1$	0.97	0.99	1.00	0.93	0.98	1.00	1.00	1.00	1.00
		$\mathcal{P}_{11}$	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		$\mathcal{P}_{21}$	0.98	1.00	1.00	0.94	1.00	1.00	0.99	1.00	1.00
		$\mathcal{P}_{31}$	0.99	1.00	1.00	0.99	0.99	1.00	1.00	1.00	1.00
		$\mathcal{P}_{all}$	**0.94**	**0.98**	**1.00**	**0.86**	**0.97**	**1.00**	**0.99**	**1.00**	**1.00**
	Scale	$\mathcal{P}_1$	0.96	0.99	1.00	0.96	1.00	1.00	1.00	1.00	1.00
		$\mathcal{P}_{11}$	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		$\mathcal{P}_{21}$	0.95	1.00	1.00	0.87	0.97	1.00	0.97	1.00	1.00
		$\mathcal{P}_{31}$	0.99	1.00	1.00	0.99	1.00	1.00	1.00	1.00	1.00
		$\mathcal{P}_{all}$	**0.90**	**0.98**	**1.00**	**0.83**	**0.97**	**1.00**	**0.96**	**1.00**	**1.00**
	Avg.Size		11.1	26.7	49.8	8.2	18.1	34.5	31.2	48.4	69.6
τ^*	Freq	$\mathcal{P}_1$	0.97	0.99	1.00	0.97	1.00	1.00	0.99	1.00	1.00
		$\mathcal{P}_{11}$	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		$\mathcal{P}_{21}$	0.93	0.98	1.00	0.94	0.97	1.00	0.96	0.99	1.00
		$\mathcal{P}_{31}$	1.00	1.00	1.00	0.99	1.00	1.00	1.00	1.00	1.00
		$\mathcal{P}_{all}$	**0.90**	**0.97**	**1.00**	**0.90**	**0.97**	**1.00**	**0.95**	**0.98**	**1.00**
	Scale	$\mathcal{P}_1$	0.97	0.99	1.00	0.98	1.00	1.00	1.00	1.00	1.00
		$\mathcal{P}_{11}$	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		$\mathcal{P}_{21}$	0.76	0.94	1.00	0.70	0.94	0.98	0.83	0.97	0.99
		$\mathcal{P}_{31}$	0.97	1.00	1.00	0.97	0.99	1.00	0.98	0.99	1.00
		$\mathcal{P}_{all}$	**0.70**	**0.93**	**1.00**	**0.67**	**0.92**	**0.98**	**0.81**	**0.96**	**0.99**
	Avg.Size		7.4	15.2	27.6	7.9	16.1	28.9	23.9	33.9	47.2
Lasso		$\mathcal{P}_1$		0.94			0.93			0.94	
		$\mathcal{P}_{11}$		1.00			1.00			1.00	
		$\mathcal{P}_{21}$		0.07			0.04			0.04	
		$\mathcal{P}_{31}$		0.31			0.29			0.19	
		$\mathcal{P}_{all}$		**0.01**			**0.03**			**0.00**	
	Avg.size			12.4			13.4			11.9	

Table 10: Simulation result of Example 4.3 for $n = 200$. $\mathcal{P}_1$, $\mathcal{P}_{11}$, $\mathcal{P}_{21}$, $\mathcal{P}_{31}$ and $\mathcal{P}_{all}$ are proportions that each or every active variables $(X_1, X_{11}, X_{21}, X_{31})$ are included in selection with frequency cutoff 90%(d_1), 80%(d_2) and 70%(d_1) over 200 repetitions. Avg.size means an average size of selection with a given cutoff.

			4.4(a)			4.4(b)			4.4(c)		
			d1	d2	d3	d1	d2	d3	d1	d2	d3
dCov	Freq	$\mathcal{P}_1$	0.64	0.81	0.90	0.70	0.84	0.90	0.59	0.73	0.81
		$\mathcal{P}_{11}$	0.92	0.96	0.99	0.93	0.97	0.99	0.85	0.96	0.98
		$\mathcal{P}_{21}$	0.30	0.51	0.63	0.41	0.60	0.76	0.44	0.64	0.72
		$\mathcal{P}_{31}$	0.60	0.80	0.88	0.64	0.80	0.85	0.60	0.78	0.88
		$\mathcal{P}_{all}$	**0.10**	**0.33**	**0.48**	**0.13**	**0.38**	**0.59**	**0.07**	**0.28**	**0.47**
	Scale	$\mathcal{P}_1$	0.69	0.83	0.91	0.75	0.85	0.92	0.63	0.74	0.82
		$\mathcal{P}_{11}$	0.92	0.97	1.00	0.95	0.97	0.99	0.89	0.96	0.98
		$\mathcal{P}_{21}$	0.19	0.38	0.56	0.26	0.42	0.58	0.22	0.42	0.62
		$\mathcal{P}_{31}$	0.43	0.64	0.80	0.42	0.65	0.80	0.35	0.60	0.79
		$\mathcal{P}_{all}$	**0.04**	**0.19**	**0.42**	**0.05**	**0.21**	**0.43**	**0.03**	**0.15**	**0.39**
	Avg.Size		3.0	5.1	8.5	3.4	5.7	9.1	5.8	9.9	14.8
τ^*	Freq	$\mathcal{P}_1$	0.58	0.77	0.90	0.63	0.80	0.90	0.55	0.69	0.80
		$\mathcal{P}_{11}$	0.90	0.95	0.97	0.89	0.97	0.99	0.81	0.94	0.98
		$\mathcal{P}_{21}$	0.15	0.29	0.42	0.17	0.33	0.51	0.20	0.39	0.58
		$\mathcal{P}_{31}$	0.45	0.64	0.79	0.40	0.61	0.74	0.37	0.59	0.72
		$\mathcal{P}_{all}$	**0.03**	**0.14**	**0.28**	**0.02**	**0.13**	**0.34**	**0.01**	**0.11**	**0.31**
	Scale	$\mathcal{P}_1$	0.64	0.81	0.90	0.69	0.85	0.92	0.58	0.72	0.80
		$\mathcal{P}_{11}$	0.92	0.96	0.98	0.93	0.97	0.98	0.85	0.96	0.98
		$\mathcal{P}_{21}$	0.09	0.18	0.30	0.08	0.19	0.36	0.04	0.18	0.32
		$\mathcal{P}_{31}$	0.15	0.36	0.59	0.16	0.33	0.55	0.11	0.32	0.55
		$\mathcal{P}_{all}$	**0.01**	**0.05**	**0.15**	**0.01**	**0.04**	**0.20**	**0.00**	**0.02**	**0.13**
	Avg.Size		2.5	4.5	8.0	2.6	4.8	8.5	4.5	8.2	13.2
Lasso		$\mathcal{P}_1$		0.92			0.94			0.87	
		$\mathcal{P}_{11}$		0.98			0.99			0.95	
		$\mathcal{P}_{21}$		0.05			0.05			0.05	
		$\mathcal{P}_{31}$		0.05			0.05			0.01	
		$\mathcal{P}_{all}$		**0.01**			**0.02**			**0.01**	
	Avg.size			12.4			12.2			12.7	

Table 11: Simulation result of Example 4.4 for $n = 200$. $\mathcal{P}_1$, $\mathcal{P}_{11}$, $\mathcal{P}_{21}$, $\mathcal{P}_{31}$ and $\mathcal{P}_{all}$ are proportions that each or every active variables $(X_1, X_{11}, X_{21}, X_{31})$ are included in selection with frequency cutoff 90%(d_1), 80%(d_2) and 70%(d_1) over 200 repetitions. Avg.size means an average size of selection with a given cutoff.

			4.4(a)			4.4(b)			4.4(c)		
			d1	d2	d3	d1	d2	d3	d1	d2	d3
dCov	Freq	$\mathcal{P}_1$	0.97	1.00	1.00	0.98	0.99	1.00	0.90	0.95	0.97
		$\mathcal{P}_{11}$	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		$\mathcal{P}_{21}$	0.93	0.99 1	0.00	0.86	0.92	0.95	0.90	0.96	0.98
		$\mathcal{P}_{31}$	1.00	1.00	1.00	1.00	1.00	1.00	0.99	1.00	1.00
		$\mathcal{P}_{all}$	**0.91**	**0.98**	**1.00**	**0.84**	**0.92**	**0.95**	**0.80**	**0.91**	**0.95**
	Scale	$\mathcal{P}_1$	0.97	1.00	1.00	0.99	1.00	1.00	0.94	0.97	0.98
		$\mathcal{P}_{11}$	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		$\mathcal{P}_{21}$	0.79	0.92	0.97	0.76	0.88	0.94	0.78	0.91	0.97
		$\mathcal{P}_{31}$	0.95	0.99	1.00	0.97	1.00	1.00	0.91	0.99	1.00
		$\mathcal{P}_{all}$	**0.73**	**0.91**	**0.97**	**0.72**	**0.88**	**0.94**	**0.66**	**0.87**	**0.95**
	Avg.Size		4.5	6.3	9.3	4.5	6.4	9.7	11.1	15.3	20.6
τ^*	Freq	$\mathcal{P}_1$	0.96	0.99	1.00	0.97	1.00	1.00	0.90	0.94	0.97
		$\mathcal{P}_{11}$	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		$\mathcal{P}_{21}$	0.77	0.89	0.97	0.70	0.83	0.90	0.74	0.86	0.93
		$\mathcal{P}_{31}$	0.96	0.99	0.99	0.96	1.00	1.00	0.91	0.97	1.00
		$\mathcal{P}_{all}$	**0.71**	**0.88**	**0.96**	**0.65**	**0.83**	**0.90**	**0.60**	**0.78**	**0.90**
	Scale	$\mathcal{P}_1$	0.97	1.00	1.00	0.99	1.00	1.00	0.92	0.95	0.97
		$\mathcal{P}_{11}$	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		$\mathcal{P}_{21}$	0.45	0.67	0.86	0.42	0.64	0.81	0.43	0.68	0.86
		$\mathcal{P}_{31}$	0.82	0.92	0.96	0.79	0.91	0.97	0.68	0.86	0.97
		$\mathcal{P}_{all}$	**0.35**	**0.61**	**0.83**	**0.31**	**0.57**	**0.78**	**0.25**	**0.55**	**0.80**
	Avg.Size		4.1	5.9	9.1	4.2	5.9	9.3	9.2	13.3	18.5
Lasso		$\mathcal{P}_1$		1.00			1.00			1.00	
		$\mathcal{P}_{11}$		1.00			1.00			1.00	
		$\mathcal{P}_{21}$		0.06			0.08			0.06	
		$\mathcal{P}_{31}$		0.08			0.08			0.08	
		$\mathcal{P}_{all}$		**0.01**			**0.02**			**0.01**	
	Avg.size			13.4			12.6			13.0	

Table 12: Simulation result of Example 4.4 for $n = 400$. $\mathcal{P}_1$, $\mathcal{P}_{11}$, $\mathcal{P}_{21}$, $\mathcal{P}_{31}$ and $\mathcal{P}_{all}$ are proportions that each or every active variables $(X_1, X_{11}, X_{21}, X_{31})$ are included in selection with frequency cutoff 90%(d_1), 80%(d_2) and 70%(d_1) over 200 repetitions. Avg.size means an average size of selection with a given cutoff.

			4.5(a)			4.5(b)			4.5(c)		
			d1	d2	d3	d1	d2	d3	d1	d2	d3
dCov	Freq	$\mathcal{P}_1$	0.64	0.78	0.87	0.73	0.85	0.91	0.65	0.78	0.86
		$\mathcal{P}_{11}$	0.49	0.65	0.77	0.52	0.67	0.80	0.41	0.60	0.74
		$\mathcal{P}_{21}$	0.51	0.68	0.78	0.51	0.66	0.78	0.49	0.70	0.79
		$\mathcal{P}_{31}$	0.37	0.55	0.67	0.41	0.60	0.78	0.34	0.45	0.59
		$\mathcal{P}_{all}$	**0.04**	**0.19**	**0.35**	**0.06**	**0.20**	**0.43**	**0.03**	**0.11**	**0.25**
	Scale	$\mathcal{P}_1$	0.56	0.71	0.82	0.55	0.69	0.81	0.48	0.68	0.78
		$\mathcal{P}_{11}$	0.30	0.49	0.65	0.28	0.46	0.64	0.27	0.49	0.67
		$\mathcal{P}_{21}$	0.58	0.71	0.80	0.62	0.71	0.79	0.57	0.73	0.81
		$\mathcal{P}_{31}$	0.20	0.37	0.51	0.24	0.42	0.60	0.19	0.37	0.52
		$\mathcal{P}_{all}$	**0.02**	**0.08**	**0.20**	**0.01**	**0.05**	**0.21**	**0.00**	**0.05**	**0.19**
	Avg.Size		2.6	4.9	8.2	2.7	4.9	8.5	3.5	7.1	11.6
τ^*	Freq	$\mathcal{P}_1$	0.39	0.59	0.70	0.45	0.67	0.79	0.39	0.61	0.70
		$\mathcal{P}_{11}$	0.25	0.47	0.62	0.29	0.51	0.66	0.22	0.41	0.61
		$\mathcal{P}_{21}$	0.45	0.63	0.77	0.47	0.66	0.76	0.45	0.63	0.79
		$\mathcal{P}_{31}$	0.26	0.44	0.63	0.30	0.52	0.71	0.25	0.43	0.57
		$\mathcal{P}_{all}$	**0.00**	**0.06**	**0.21**	**0.01**	**0.08**	**0.23**	**0.00**	**0.05**	**0.16**
	Scale	$\mathcal{P}_1$	0.23	0.44	0.63	0.23	0.41	0.61	0.17	0.36	0.57
		$\mathcal{P}_{11}$	0.12	0.25	0.39	0.12	0.23	0.41	0.10	0.25	0.44
		$\mathcal{P}_{21}$	0.50	0.67	0.77	0.54	0.70	0.76	0.49	0.66	0.80
		$\mathcal{P}_{31}$	0.14	0.25	0.41	0.18	0.30	0.49	0.14	0.30	0.43
		$\mathcal{P}_{all}$	**0.00**	**0.02**	**0.08**	**0.00**	**0.01**	**0.06**	**0.00**	**0.00**	**0.06**
	Avg.Size		1.8	3.9	7.4	2.0	4.1	7.6	2.5	5.8	10.5
Lasso		$\mathcal{P}_1$		0.05			0.06			0.04	
		$\mathcal{P}_{11}$		0.02			0.02			0.03	
		$\mathcal{P}_{21}$		0.51			0.54			0.48	
		$\mathcal{P}_{31}$		0.02			0.04			0.02	
		$\mathcal{P}_{all}$		**0.00**			**0.00**			**0.00**	
	Avg.size			7.7			7.6			6.2	

Table 13: Simulation result of Example 4.5 for n=200. $\mathcal{P}_1$, $\mathcal{P}_{11}$, $\mathcal{P}_{21}$, $\mathcal{P}_{31}$ and $\mathcal{P}_{all}$ are proportions that each or every active variables $(X_1, X_{11}, X_{21}, X_{31})$ are included in selection with frequency cutoff 90%(d_1), 80%(d_2) and 70%(d_1) over 200 repetitions. Avg.size means an average size of selection with a given cutoff.

			4.5(a)			4.5(b)			4.5(c)		
			d1	d2	d3	d1	d2	d3	d1	d2	d3
dCov	Freq	$\mathcal{P}_1$	0.99	1.00	1.00	0.99	1.00	1.00	0.99	1.00	1.00
		$\mathcal{P}_{11}$	0.94	0.97	1.00	0.94	0.98	0.99	0.96	0.99	0.99
		$\mathcal{P}_{21}$	0.96	0.98	1.00	0.94	0.98	0.99	0.94	0.98	0.99
		$\mathcal{P}_{31}$	0.93	0.97	1.00	0.91	0.96	0.98	0.86	0.95	0.98
		$\mathcal{P}_{all}$	**0.81**	**0.92**	**1.00**	**0.78**	**0.91**	**0.96**	**0.74**	**0.91**	**0.96**
	Scale	$\mathcal{P}_1$	0.96	0.99	1.00	0.98	0.99	1.00	0.98	0.99	1.00
		$\mathcal{P}_{11}$	0.86	0.93	0.98	0.84	0.93	0.97	0.81	0.94	0.99
		$\mathcal{P}_{21}$	0.96	0.98	0.99	0.95	0.97	0.99	0.96	0.99	1.00
		$\mathcal{P}_{31}$	0.80	0.90	0.96	0.78	0.88	0.93	0.70	0.87	0.95
		$\mathcal{P}_{all}$	**0.60**	**0.81**	**0.93**	**0.62**	**0.77**	**0.89**	**0.51**	**0.80**	**0.92**
	Avg.Size		4.4	6.2	9.4	4.4	6.2	9.4	7.6	12.0	17.1
τ^*	Freq	$\mathcal{P}_1$	0.93	0.97	0.99	0.95	0.99	0.99	0.92	0.98	0.98
		$\mathcal{P}_{11}$	0.84	0.92	0.96	0.80	0.92	0.97	0.84	0.95	0.97
		$\mathcal{P}_{21}$	0.94	0.97	0.99	0.92	0.96	0.99	0.91	0.97	0.99
		$\mathcal{P}_{31}$	0.91	0.96	1.00	0.85	0.95	0.98	0.81	0.92	0.98
		$\mathcal{P}_{all}$	**0.66**	**0.82**	**0.93**	**0.57**	**0.83**	**0.92**	**0.57**	**0.82**	**0.92**
	Scale	$\mathcal{P}_1$	0.75	0.90	0.97	0.78	0.91	0.97	0.75	0.90	0.98
		$\mathcal{P}_{11}$	0.62	0.77	0.90	0.60	0.73	0.90	0.50	0.76	0.91
		$\mathcal{P}_{21}$	0.94	0.97	0.99	0.94	0.96	0.99	0.94	0.98	1.00
		$\mathcal{P}_{31}$	0.71	0.80	0.92	0.70	0.80	0.91	0.56	0.79	0.93
		$\mathcal{P}_{all}$	**0.26**	**0.52**	**0.79**	**0.27**	**0.49**	**0.78**	**0.19**	**0.49**	**0.81**
	Avg.Size		4.0	5.7	8.9	4.0	5.7	9.0	6.2	10.1	15.4
Lasso		$\mathcal{P}_1$		0.05			0.06			0.06	
		$\mathcal{P}_{11}$		0.02			0.02			0.02	
		$\mathcal{P}_{21}$		0.87			0.89			0.81	
		$\mathcal{P}_{31}$		0.03			0.04			0.03	
		$\mathcal{P}_{all}$		**0.00**			**0.00**			**0.00**	
	Avg.size			8.3			8.6			7.6	

Table 14: Simulation result of Example 4.5 for $n = 400$. $\mathcal{P}_1$, $\mathcal{P}_{11}$, $\mathcal{P}_{21}$, $\mathcal{P}_{31}$ and $\mathcal{P}_{all}$ are proportions that each or every active variables $(X_1, X_{11}, X_{21}, X_{31})$ are included in selection with frequency cutoff 90%(d_1), 80%(d_2) and 70%(d_1) over 200 repetitions. Avg.size means an average size of selection with a given cutoff.

4.4 DATA APPLICATION

We apply the proposed method to select important genes related to suicidal behaviors. Suicide ranks among the 10 leading causes of death in the US and is the 2nd leading cause in youth. Suicidal behavior occurs in the context of many psychiatric disorders; however, relatively few subjects with a psychiatric disorder attempt suicide. Current diagnosis largely depends on psychosocial symptoms and doesn't successfully identify individuals at risk (May et al., 2012). One of the most challenging and critical tasks is to identify objective biological predictors for suicidal behavior. The data is a part of ongoing research and its collection is not completed yet. Any result of this section is based on available samples and is only for an illustrative purpose. We acknowledge Prof. Nadine Melhem from the Department of Psychiatry at the University of Pittsburgh for providing this dataset to study our new selection method.

The data combines gene activation levels and EHR records which track suicidal behaviors of a subject. There exists multiple levels of suicidal behaviors which ranging from a suicidal ideation to an successful suicide. For identifying them, the Columbia Suicide Severity Rating Scale (C-SSRS)349 is used. It is a widely used clinical and research tool to assess current and lifetime suicidal ideation and behavior, including intent and lethality rating for attempts at follow-up; and Suicidal Ideation Questionnaire (SIQ)350,351 to assess the severity of suicidal ideation. In this study, we use a binary response variable; $Y = 1$ if one commits any kind of suicidal behaviors within 6 months and 0 otherwise. For gene expression data, Illumina's HumanHT-12 v4 Expression BeadChip Kit is used, which is a microarray providing genome-wide transcriptional coverage of more than 47,000 well-characterized genes, gene candidates, and splice variants. Samples are prepared using the Illumina®Totalprep RNA Amplification Kit. Only genes that passed our quality control (QC) analysis and demonstrated a detection p-value which was significant in at least 50% of the sample were included in our analysis. It has been established in the literature that HPA axis dysregulation is associated with increased risk for suicide. We will focus on examine gene expressions in the HPA axis and inflammatory pathways in a large sample of male and female psychiatric patients at high-risk but with no prior history of suicidal behavior; following the same subjects over time to

examine the temporal sequence between alterations in these pathways and suicide attempts. Finally we have 134 samples and 20 of them are $Y = 1$ cases. At this point, only 46 subjects have longitudinal observations of gene activation levels. The number of genes is 2423.

In the first part, we perform an unconditional selection to discover important genes related to Y. We begin with a plain individual gene selection on their single observation. Then we try to select genes based on a cluster selection, where the genes are selected together or not if they belong to the same cluster. In the second part, we select genes conditional on gender. We compare results between the frequent voting method, scale-based method, and logistic LASSO. For the frequent voting method, 500 bootstrapped samples are used to vote based on a significance level of 0.05. Normalized measures (distance correlation instead of distance covariance) are used for the scale-based method.

4.4.1 Unconditional Selection

We first apply our method to select individual genes based on their single observation. Two measures of independence, "dCov" and τ^*, are used. As seen in Table 15, when we apply frequency cutoff 90, 80, 70%, a total of 1, 6, and 22 genes are selected respectively with "dCov" and 0, 6 and 25 genes are selected with τ^*. For the frequent voting method, the top 6 genes are exactly overlapped between two measures. Figure 8 illustrate the distribution of gene expression levels for the $Y = 1$ and $Y = 0$ groups. The difference between two distributions is prominent in the most relevant gene, 2406. This gene is called "YWHAE" and is known to be related to schizophrenia, a mental disease that can cause suicidal behaviors. The other 5 genes also show some different shapes in distribution between two groups.

The number of genes selected in the scale-based method is matched with the frequent voting method. The top 6 selections for the scale-based method are slightly different depending on which measure is used. We found that 4 of the top 6 genes in the frequent voting method are also selected in the scale-based method for both "dCov" and τ^*. One of the top 6 genes in the frequent voting method, gene 1671, is not selected in the top 25 genes using the scale-based method with "dCov". LASSO didn't select any genes when applying CV to minimize the mean squared error, so we list the top 6 genes in their solution path. Only 2 of

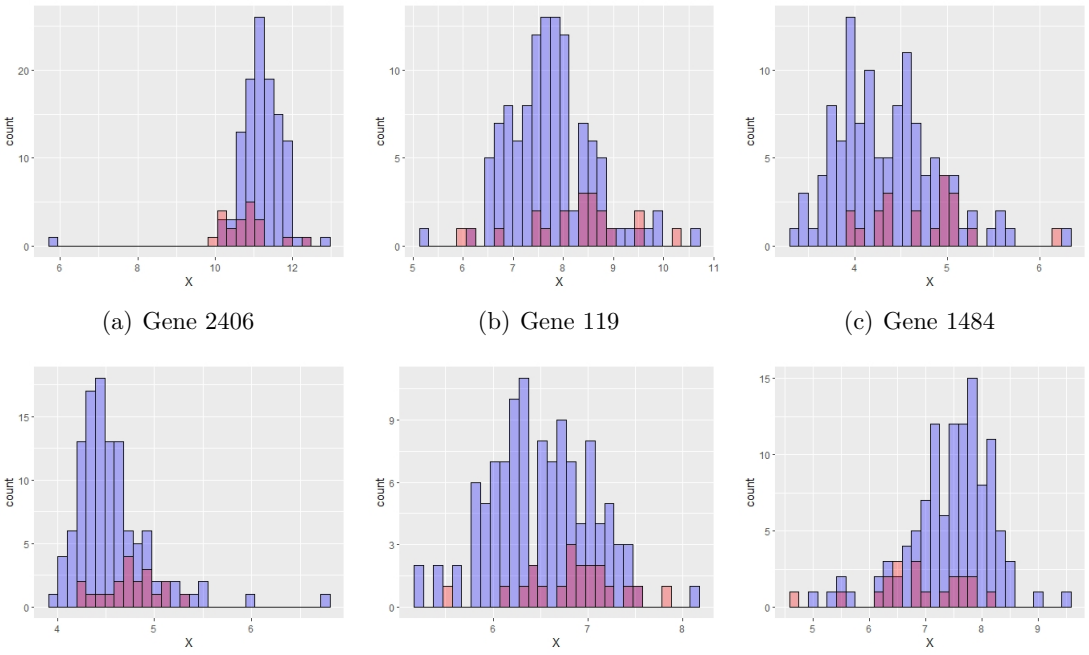
		90%	80%	70%
dCov	Freq	2406 (1)	119 1484 1649 1671 1986 2406 (6)	119 224 296 371 350 378 501 901 999 1235 1484 1530 1538 1649 1671 1916 1958 1974 2014 1986 2060 2406 (22)
	Scale	2406	371 1484 1530 1649 1986 2406	119 296 350 371 378 901 999 1242 1339 1484 1530 1538 1608 1673 1649 1916 1958 1965 1974 1986 2060 2406
τ^*	Freq	(0)	119 1484 1649 1671 1986 2406 (6)	80 119 122 155 296 705 999 1484 1608 1668 1530 1538 1958 1649 1671 1673 1965 1974 1986 2014 2060 2178 2345 2406 2409 (25)
	Scale		1484 1530 1538 1649 1986 2406	119 154 155 296 350 378 501 999 1242 1484 1530 1538 1608 1649 1671 1673 1958 1965 1974 1986 2014 2060 2178 2406 2409
Lasso		296 901 1484 1530 1986 2241		

Table 15: Result of unconditional selection. Each numbers in parentheses refer a size of selection with a given cutoff for the frequent based method. Size of selection for scale-based method and LASSO are matched with that of the frequent voting method.

the Top 6 selected by the frequent voting method are also selected in LASSO. It is noticeable that the most relevant gene in the other two methods, 2406, is not selected in LASSO.

Due to a delay in the sample collection, currently only 46 samples have longitudinal observations. The proposed method can be used to selection genes with longitudinal observations by employing distance covariance or "IPR$-\tau^*$" as dependent measures as they work for multivariate dimensional data.

We now move to the cluster selection which aims to select closely related genes together. Genes are believed to interact and work together. Recently clustering analysis based on gene networks has been studied widely. In the first step, we cluster genes based on their co-expression patterns. We apply a hierarchical clustering algorithm, called WGCNA (a weighted gene co-expression network analysis) (Langfelder and Horvath, 2008). Using R package "WGCNA" and with some tuning parameters (softpower=5, minimum module size=3), we identified 135 clusters whose size range from 3 to 538. There are 24 unassigned genes which we still include as a cluster of size 1. There are different methods for gene network analysis and clustering. We use WGCNA just for illustration purposes. We then apply



(a) Gene 2406 (b) Gene 119 (c) Gene 1484

(d) Gene 1649 (e) Gene 1671 (f) Gene 1986

Figure 8: Different distributions of each gene's activation level between two groups; Y=1 (red) Y=0 (blue). The genes are from the top 6 selected by the frequent voting method.

two multivariate measures of independence, "dCov" and "IPR-τ^*" to test the independence between each cluster vs the response variable. The frequent voting method is then used to decide which clusters to select. For reference, we also present the result of group-LASSO selection (Zou and Hastie, 2005) for logistic regression conducted using R package "gglasso", where the penalty for each cluster is proportional to the square root of the corresponding size of each group (cluster). We still include the results of the scale-based method but note that this method may not be suitable for variables with different dimensions.

		90%	80%	70%
dCov	Freq	(0)	1 (1)	1 2 4 5 8 9 12 14 15 17 33 51 56 122 (14)
	Scale	(0)	1 (1)	1 2 5 8 12 14 17 30 33 51 56 94 122 154
IPR-τ^*	Freq	(0)	24 33 94 116 (4)	15 17 24 28 33 51 79 56 94 116 (10)
	Scale	(0)	24 137 145 154 (4)	15 17 24 33 94 116 137 145 149 154 (10)
g-LASSO			20 21 24 25 80 81 129 112 120 122	

Table 16: Result of a cluster selection. The numbers denote a cluster number. Each numbers in parentheses refer a size of selection with a given cutoff for the frequent based method. Size of selection for group LASSO is matched with that of "IPR-τ^*" selection

Table 16 summarize the selection result. For the frequent voting method, 0,1,14 clusters are selected with "dCov" and 0,4,10 clusters are selected with "IPR-τ^*". Since group LASSO didn't select any clusters with a minimum mean square error criteria, the top 10 results are listed. For the frequent voting method, clusters 15, 17, 33, 51, 56 are commonly included in "dCov" and "IPR-τ^*" selection. Cluster 24 and 122 are commonly selected by group LASSO and one of the two frequent voting selections. There exists some overlaps in selection between the frequent voting method and the scale-based method.

4.4.2 Conditional Selection

In some applications, researchers believe that some covariates Z need to be controlled when studying the effect of X on Y. If we ignore this relationship, the marginal selection may give us an inconsistent result even under a model-free approach. That is, we may drop

important variables or include false variables (Fan and Lv, 2008). A simple linear model can illustrate this point. If two variables are highly correlated, and their effects on Y are in an opposite direction, their marginal correlation with Y can be very low or even zero. Therefore, selection based on a conditional relationship is sometimes very important. Conditional selection provides a way to recruit additional variables based on previous knowledge about relationships.

Formally speaking, X and Y are said to be independent conditional on Z if they satisfies

$$H_0 : P_{XY|Z} = P_{X|Z}P_{Y|Z} \text{ versus } H_1 : P_{XY|Z} \neq P_{X|Z}P_{Y|Z}.$$

Unlike the unconditional independence problem, it is known to be very challenging to identify a conditional relationship (Shah et al., 2020), especially when Z is continuous or has multiple dimensions. One alternative method is to transform Z to discrete data by binning, but then Type 1 error will not be controlled anymore. Nevertheless, there exist some attempts to consider a conditional relationship to select variables. As a popular classical method, forward selection employs an iterative algorithm; at each step, one selects the most relevant variables which minimize the residual sum of squares (RSS) in a linear regression setting. Wang (2009) extends this method to high dimensional data (when p exceeds n) based on a partial correlation coefficient. In a residual-based approach, we first make an unconditional selection, estimate a model to get residuals and then conduct another unconditional selection with these residuals as new response variables (see iterative versions of Fan and Lv (2008); Pan et al. (2018)). As another model-based approach, Barut et al. (2016) suggests a marginal selection method based on a generalized linear model. In their method, when a set of variables are pre-selected, one recruits additional variables if the scale of their conditional marginal likelihood estimator exceeds a certain threshold while every variable is standardized.

Recently, conditional version of distance covariance ("c-dCov") was suggested by Wang et al. (2015). Its population version is quantified by

$$\mathcal{S}_a = \mathbb{E}[\mathcal{D}^2(X, Y|Z)a(Z)],$$

where $\mathcal{D}(X, Y|Z)$ is a measure of distance covariance evaluated at fixed Z and $a(Z)$ is some weight function. The corresponding empirical statistic was also suggested using a kernel-based approach ; based on a distance between Z and impose weights. It provides a natural way to extend our model-free method to conduct a conditional selection, compensating for the possible inconsistency caused by a marginal selection.

Although the measure was successfully developed, implementing the measure with real data has some challenges. Since the measure is an average of multiple unconditional measures, the required sample size for a reliable result is larger than an unconditional method. In other words, the convergence of test statistic is slow for a conditional measure. If we consider a multidimensional Z, the rate of convergence will be even slower; it is a function of $h^{r/2}$ where h is a kernel bandwidth used to approximate a density of Z and r is the dimension of Z. Here we have a relatively small sample and we only illustrate the conditional selection idea using gender as a conditioning variable. We first provide unconditional selection results separately conducted for males and females. Overall, the selection result is similar between the frequent voting method and the scale-based method. However, the selection results are very different between males and females, which means that the important genes may be different in two genders. This difference is possibly caused by a small sample size (71 for males and 63 for females); nevertheless, we proceed with a selection conditioning on gender.

Here R package ("cdcsis") is used to calculate a conditional distance covariance and its normalized version is used for the scale-based method. We also provide a result with conditional LASSO which can be performed by enforcing "gender" to stay in a model.

As seen in Table 18, with 90%, 80%, and 70% of frequency cutoff, a group of 0, 2 and 8 genes are selected respectively with the frequent voting method. It includes two common selections, gene 1484 and gene 1986, with LASSO, and one common selection, gene 2406, with the scale-based method. Selection based on the scale-based method and LASSO have no overlaps.

Compared to the separate results in Table 17, we found that 6 of the top 8 genes originated from the male, and 2 of them are from the female for the scale-based selection. This is probably because there are more samples for males than females. Although the difference is not large, it can largely affect the selection because the empirical distance covariance is

			80%
Male	dCov	Freq	80 624 677 999 1283 1393 1420 1538 1673 1916 (10)
		Scale	80 157 677 999 1283 1393 1420 1538 1673 1916
	τ^*	Freq	1538 1673 (2)
		Scale	1393 1538
Female	dCov	Freq	119 675 811 898 1262 1891 1906 1986 2029 2110 2406 (11)
		Scale	119 675 811 1218 1262 1608 1891 1906 1986 2110 2406
	τ^*	Freq	1262 (1)
		Scale	1262

Table 17: Unconditional selection result for male (71 subjects) and female (63 subjects)

.

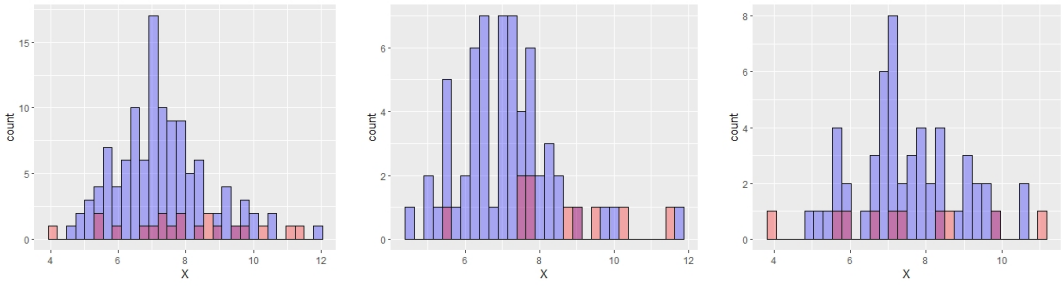
averaged over the fourth of the density function of Z. That is, $a(Z) = 12f^4(Z)$ when f is a density function of Z. Then the relative weight of males over females is $(71/63)^4 = 1.6$. For the frequent voting selection, 3 of the top 8 selections are from females.

Compared to the unconditional selection in Table 15, the frequent voting method results show that 6 of the top 8 genes are included in the top 22 selection of the unconditional method. Gene 1958 is not selected by the unconditional "dCov" method but can be found in the top 25 of the τ^*-based unconditional selection. Gene 157 is not selected in any of the unconditional selection. We plot histograms of the activation level of gene 157 in Figure 9. Its overall distribution is not very different between $Y = 1$ and $Y = 0$ cases, but when we look at female samples, the suicidal cases are more likely to have high expression values. This illustrates the difference between the conditional selection and the unconditional selection. The scale-based selection does not overlap with unconditional selection as much as the frequent voting method. Only 4 of the top 8 are found in the top 22 selection. The top 2 genes, 1538 and 1916 are not found in the top 6 genes in unconditional selection. For LASSO, since minimizing mean squared error doesn't select any genes, we list the top

		90%	80%	70%
c-dCov	Freq	(0)	119 2406 (2)	119 157 296 371 1484 1958 1986 2406 (8)
	Scale		1538 1916	80 999 1393 1538 1673 1906 1916 2406
LASSO			901 1986	709 901 963 1484 1530 1671 1986 2241

Table 18: Selection result conditional on gender. Each numbers in parentheses refer a size of selection with a given cutoff for the frequent based method. Size of selection for scale-based method and LASSO are matched with that of the frequent voting method.

genes with a matched size to the frequent voting method. LASSO result didn't change much from an unconditional selection to conditional selection, but we note that we only control gender in the model without considering any interaction effects between gender and genes. The conditional dependence test is conceptually different from a model based approach controlling Z in the model.



(a) All (b) female (c) male

Figure 9: Histogram of gene 157's activation level between two groups; Y=1 (red), Y=0 (blue).

5.0 DISCUSSION

In this paper, we have introduced a new measure of independence that works for general types of data. A corresponding empirical measure was suggested as a U-statistic. A test of independence is derived based on a large sample theory. The simulation study illustrates that the proposed measure have nice empirical power in detecting various kind of dependency. We applied the proposed test of independence to gene expression data and crop yields data to study the relationship between variables.

We also have proposed a new model-free variable selection method based on the suggested and existing measures of independence. Combining with a resampling procedure, the new selection method gives a way to reduce the number of independent variables to an adjustable size while considering the stability of their relationships with a response variable. The method was extended to a conditional variable selection which controls the effect of confounding variables. We applied this method to select important genes related to suicidal behaviors. Due to the delay in sample collection, the analysis was conducted based on only available samples. We may further analyze this data when the collection is finished. We may also apply this method using other conditional tests of independence and compare the results with a current selection using a conditional distance covariance. Once the selection is confirmed, we may build a model to assess the real-data prediction performance of the method.

APPENDIX

PROOFS

Lemma 3. *Assume θ is discrete or continuous, or a mixture of the two, that is, assume there exists a probability mass function P_{XY} and a density function h such that*

$$\theta(A \times B) = \sum_{x_i, y_i} P_{XY}(x_i, y_i) + \int_{A \times B} h(x, y) G(dx) G(dy),$$

where $A \subset \mathcal{X}, B \subset \mathcal{Y}$ are any two open sets and G is the abstract Wiener measure on $\mathcal{X}$ and $\mathcal{Y}$. Then, induced random variables defined by $U_x = \rho(x, X)$, $V_y = \zeta(y, Y)$ with $U_x : \mathcal{X} \to \mathbb{R}$, $V_y : \mathcal{Y} \to \mathbb{R}$ are well-defined and has a jointly discrete or continuous distribution, or a mixture of the two.

Proof. Denote the closed ball with center x and radius r_1 in $\mathcal{X}$ as $\bar{B}_\rho(x, r_1)$ or $\bar{B}(x, r_1)$, and the closed ball with center y and radius r_2 in $\mathcal{Y}$ as $\bar{B}_\zeta(y, r_2)$ or $\bar{B}(y, r_2)$.

For U_x,

$$Pr(U_x \in [0, a]) = Pr(X \in \bar{B}(x, a)) = \mu(\bar{B}(x, a)).$$

So U_x is a well-defined Borel probability measure on $\mathbb{R}$. Same applied for V_y and the joint variable.

Now we show that U_x and V_y has a jointly discrete, continuous or a mixture of two distributions.

$$\begin{aligned} Pr(U_x \le a, V_y \le b) &= Pr(\bar{B}(x,a) \times \bar{B}(y,b)) \\ &= \sum_{a' \in [0,a], b' \in [0,b]} \sum_{x_i \in \partial B(x,a'), y_i \in \partial B(y,b')} P_{XY}(x_i, y_i) \\ &+ \int_0^a \int_0^b \int_{\partial B(x,a') \times \partial B(y,b')} h(x,y) G(dx) G(dy) db' da' \end{aligned}$$

Here, $P_{U_x V_y}(a', b') = \sum_{x_i \in \partial \bar{B}(x,a'), y_i \in \partial \bar{B}(y,b')} P_{X,Y}(x_i, y_i)$ is a probability mass function and $\int_{\partial B(x,a') \times \partial B(y,b')} h(x,y) G(dx) G(dy)$ is a density function. □

Proof of Theorem 1

Proof. Let's define $\eta(x,y)$ for $(x,y) \in \mathcal{X} \times \mathcal{Y}$ as

$$\eta(x,y) = Ea(\rho(x,X^1), \rho(x,X^2), \rho(x,X^3), \rho(x,X^4)) a(\zeta(y,Y^1), \zeta(y,Y^2), \zeta(y,Y^3), \zeta(y,Y^4)),$$

where $(X^1,Y^1), \ldots, (X^4,Y^4)$ are independent copies of (X,Y). If we define new random variables induced by x, y as $U_x = \rho(x,X)$, $V_y = \zeta(y,Y)$, respectively, which are well-defined by Lemma 3, $\eta(x,y)$ becomes $\tau^*(U_x, V_y)$. Therefore, under the consistency condition of τ^* met by Lemma 3, $\eta(x,y)$ is zero if U_x and V_y are independent and positive otherwise. Then IPR-$\tau^*(X,Y) = E_{(x,y)\sim\theta}\eta(x,y) \ge 0$ is derived.

If X and Y are independent, then U_x and V_y are independent for every $(x,y) \in \mathcal{X} \times \mathcal{Y}$. So IPR-$\tau^*(X,Y) = 0$. In the following, we will show that IPR-$\tau^*(X,Y) = 0$ only if X and Y are independent.

Step 1) We claim that if $\theta \neq \mu \times \nu$, there exist a point of dependence in the support set of θ. Since $h(x,y)$ is continuous on any continuous point of θ, we have marginal density funtion $f(x)$ and $g(y)$ for any continuous point (x,y) of θ. Denote support sets of θ, μ and ν as S_θ, S_μ and S_ν.

Then, since $S_\theta \subset S_\mu \times S_\nu$, we have

$$\begin{aligned}
1 &= \sum_{S_\mu \times S_\nu} P_{X,Y}(x,y) + \int_{S_\mu \times S_\nu} h(x,y) G(dx \times dy) \\
&= \sum_{S_\theta} P_X(x) P_Y(y) + \sum_{S_\mu \times S_\nu / S_\theta} P_X(x) P_Y(y) \\
&+ \int_{S_\theta} f(x) g(y) G(dx \times dy) + \int_{S_\mu \times S_\nu / S_\theta} f(x) g(y) G(dx \times dy).
\end{aligned}$$

If the claim is not true, we must either have a discrete point $(x,y) \in S_\theta^c$ such that $P_{X,Y}(x,y) \neq P_X(x)P_Y(y)$, and/or sets A and B such that $\int_{A\times B} h(x,y)G(dx\times dy) \neq \int_A f(x)$ $G(dx) \int_B g(x)G(dy)$. In the first case, $P_{X,Y}(x,y) \neq P_X(x)P_Y(y)$ implies $P_X(x)P_Y(y) > 0$, so $(x,y) \in S_\mu \times S_\nu / S_\theta$. Then there should exists another point $(x',y') \in S_\theta$ such that $P_{X,Y}(x',y') > P_X(x')P_Y(y')$ to balance the above equation. The same argument applies to the later case.

Step 2) With the results of *Step 1)*, if (x,y) is a discrete point and $P_{X,Y}(x,y) \neq P_X(x)P_Y(y)$, we can find balls $B_\rho(x,r_1)$ and $B_\zeta(y,r_2)$ such that

$$\theta(B_\rho(x,r_1) \times B_\zeta(y,r_2)) \neq \mu(B_\rho(x,r_1))\nu(B_\zeta(y,r_2)),$$

with small enough r_1 and r_2. Since $\{U_x < r_1\} = B_\rho(x,r_1)$ and $\{V_y < r_2\} = B_\zeta(y,r_y)$, this equation is reduced to

$$P_{U_x,V_y}(r_1,r_2) \neq P_{U_x}(r_1)P_{V_y}(r_2).$$

So U_x and V_y are not independent, i.e. $\eta(x,y) > 0$. Then,

$$\text{IPR-}\tau^*(X,Y) \geq \eta(x,y)\theta(x,y) > 0.$$

If (x,y) is a continuous point, we can say that $h(x,y) > f(x)g(y)$ without a loss of generosity. Since $h(x,y)$ is continuous, so are $f(x)$, $g(y)$. We can find an area A of nonzero measure such that there exist balls $B_\rho(v,r_v)$, $r_v > 0$ and $B_\zeta(w,r_w)$, $r_w > 0$ for every $(v,w) \in A$ where $h(v',w') > f(v')g(w')$ for $v' \in B_\rho(v,r_v)$ and $w' \in B_\zeta(w,r_w)$.

Then,

$$\begin{aligned}\theta(B_\rho(v,r_v)\times B_\zeta(w,r_w)) &= \int_{B_\rho(v,r_v)\times B_\zeta(w,r_w)} h(v',w')G(dv'\times dw') \\ &> \int_{B_\rho(v,r_v)} f(w)G(dw)\int_{B_\zeta(w,r_w)} g(w')G(dw') \\ &= \mu(B_\rho(v,r_v))\nu(B_\zeta(w,r_w))\end{aligned}$$

Same as the discrete case, the inequality is reduced to $P_{U_v,V_y}((-\infty,r_v)\times(-\infty,r_w)) > P_{U_v}((-\infty,r_v))P_{V_w}((-\infty,r_w))$. So U_v and V_w are not independent and $\eta(v,w)>0$ for every $(v,w)\in A$. Then,

$$\text{IPR-}\tau^*(X,Y) \geq \int_A \eta(v,w)h(v,w)G(dv\times dw) > 0,$$

and IPR-$\tau^*(X,Y)=0$ only if X and Y are independent.

□

BIBLIOGRAPHY

Adriaenssens, N., Coenen, S., Versporten, A., Muller, A., Minalu, G., Faes, C., Vankerckhoven, V., Aerts, M., Hens, N., Molenberghs, G., et al. (2011), "European Surveillance of Antimicrobial Consumption (ESAC): outpatient quinolone use in Europe (1997–2009)," *Journal of antimicrobial chemotherapy*, 66, vi47–vi56.

Bair, E., Hastie, T., Paul, D., and Tibshirani, R. (2006), "Prediction by supervised principal components," *Journal of the American Statistical Association*, 101, 119–137.

Bair, E. and Tibshirani, R. (2004), "Semi-supervised methods to predict patient survival from gene expression data," *PLoS Biol*, 2, e108.

Barut, E., Fan, J., and Verhasselt, A. (2016), "Conditional sure independence screening," *Journal of the American Statistical Association*, 111, 1266–1277.

Bergsma, W. and Dassios, A. (2014), "A consistent test of independence based on a sign covariance related to Kendall's tau," *Bernoulli*, 20, 1006–1028.

Biswas, M., Sarkar, S., and Ghosh, A. K. (2016), "On some exact distribution-free tests of independence between two random vectors of arbitrary dimensions," *Journal of Statistical Planning and Inference*, 175, 78–86.

Blum, J. R., Kiefer, J., and Rosenblatt, M. (1961), "Distribution free tests of independence based on the sample distribution function," *The annals of mathematical statistics*, 485–498.

Bøvelstad, H. M., Nygård, S., Størvold, H. L., Aldrin, M., Borgan, Ø., Frigessi, A., and Lingjærde, O. C. (2007), "Predicting survival from microarray data—a comparative study," *Bioinformatics*, 23, 2080–2087.

Chen, K., Chen, K., Müller, H.-G., and Wang, J.-L. (2011), "Stringing high-dimensional data for functional analysis," *Journal of the American Statistical Association*, 106, 275–284.

Chen, K., Zhang, X., Petersen, A., and Müller, H.-G. (2017), "Quantifying infinite-dimensional data: Functional data analysis in action," *Statistics in Biosciences*, 9, 582–604.

Dai, X., Müller, H.-G., et al. (2018), "Principal component analysis for functional data on riemannian manifolds and spheres," *The Annals of Statistics*, 46, 3334–3361.

Deb, N. and Sen, B. (2019), "Multivariate rank-based distribution-free nonparametric testing using measure transportation," *arXiv preprint arXiv:1909.08733*.

Dhar, S. S., Dassios, A., Bergsma, W., et al. (2016), "A study of the power and robustness of a new test for independence against contiguous alternatives," *Electronic Journal of Statistics*, 10, 330–351.

Drton, M., Han, F., and Shi, H. (2018), "High dimensional independence testing with maxima of rank correlations," *arXiv preprint arXiv:1812.06189*.

Even-Zohar, C. and Leng, C. (2019), "Counting Small Permutation Patterns," *arXiv preprint arXiv:1911.01414*.

Fan, J. and Lv, J. (2008), "Sure independence screening for ultrahigh dimensional feature space," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70, 849–911.

Fan, J., Samworth, R., and Wu, Y. (2009), "Ultrahigh dimensional feature selection: beyond the linear model," *The Journal of Machine Learning Research*, 10, 2013–2038.

Free, S., O'Higgins, P., Maudgil, D., Dryden, I., Lemieux, L., Fish, D., and Shorvon, S. (2001), "Landmark-based morphometrics of the normal adult brain using MRI," *Neuroimage*, 13, 801–813.

Friedman, J. H., Rafsky, L. C., et al. (1983), "Graph-theoretic measures of multivariate association and prediction," *The Annals of Statistics*, 11, 377–391.

Gretton, A., Bousquet, O., Smola, A., and Schölkopf, B. (2005), "Measuring statistical dependence with Hilbert-Schmidt norms," in *International conference on algorithmic learning theory*, Springer, pp. 63–77.

Gretton, A., Fukumizu, K., Teo, C. H., Song, L., Schölkopf, B., and Smola, A. J. (2008), "A kernel statistical test of independence," in *Advances in neural information processing systems*, pp. 585–592.

Guo, L. and Modarres, R. (2020), "Nonparametric tests of independence based on interpoint distances," *Journal of Nonparametric Statistics*, 32, 225–245.

Hallin, M., del Barrio, E., Albertos, J. C., and Matrán, C. (2018), "Distribution and Quantile Functions, Ranks, and Signs in dimension d: a measure transportation approach," *arXiv preprint arXiv:1806.01238*.

Heller, R., Heller, Y., and Gorfine, M. (2012), "A consistent multivariate test of association based on ranks of distances," *Biometrika*, 100, 503–510.

Heller, Y. and Heller, R. (2016), "Computing the Bergsma Dassios sign-covariance," *arXiv preprint arXiv:1605.08732*.

Hoeffding, W. (1948), "A non-parametric test of independence," *The annals of mathematical statistics*, 546–557.

Kendall, M. G. (1938), "A new measure of rank correlation," *Biometrika*, 30, 81–93.

Kim, I., Balakrishnan, S., and Wasserman, L. (2020), "Robust Multivariate Nonparametric Tests via Projection-Averaging," *Annals of Statistics*, 1–34.

Kleiber, M. and Pervin, W. (1969), "A generalized Banach-Mazur theorem," *Bulletin of The Australian Mathematical Society*, 1, 169–173.

Krishnapuram, B., Carin, L., Figueiredo, M. A., and Hartemink, A. J. (2005), "Sparse multinomial logistic regression: Fast algorithms and generalization bounds," *IEEE transactions on pattern analysis and machine intelligence*, 27, 957–968.

Kuratowski, C. (1935), "Quelques problèmes concernant les espaces métriques non-séparables," *Fundamenta Mathematicae*, 25, 534–545.

Langfelder, P. and Horvath, S. (2008), "WGCNA: an R package for weighted correlation network analysis," *BMC bioinformatics*, 9, 1–13.

Leung, D., Drton, M., et al. (2018), "Testing independence in high dimensions with sums of rank correlations," *The Annals of Statistics*, 46, 280–307.

Li, R., Zhong, W., and Zhu, L. (2012), "Feature screening via distance correlation learning," *Journal of the American Statistical Association*, 107, 1129–1139.

Liu, Y. and Wang, Q. (2018), "Model-free feature screening for ultrahigh-dimensional data conditional on some variables," *Annals of the Institute of Statistical Mathematics*, 70, 283–301.

Lu, J. and Lin, L. (2020), "Model-free conditional screening via conditional distance correlation," *Statistical Papers*, 61, 225–244.

Lyons, R. et al. (2013), "Distance covariance in metric spaces," *The Annals of Probability*, 41, 3284–3305.

Mantel, N. (1967), "The detection of disease clustering and a generalized regression approach," *Cancer research*, 27, 209–220.

Masucci, A. P., Smith, D., Crooks, A., and Batty, M. (2009), "Random planar graphs and the London street network," *The European Physical Journal B*, 71, 259–271.

May, A. M., Klonsky, E. D., and Klein, D. N. (2012), "Predicting future suicide attempts among depressed suicide ideators: a 10-year longitudinal study," *Journal of psychiatric research*, 46, 946–952.

Meinshausen, N. and Bühlmann, P. (2010), "Stability selection," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72, 417–473.

Nandy, P., Weihs, L., Drton, M., et al. (2016), "Large-sample theory for the Bergsma-Dassios sign covariance," *Electronic Journal of Statistics*, 10, 2287–2311.

Nash, J. (1956), "The imbedding problem for Riemannian manifolds," *Annals of mathematics*, 20–63.

Pan, W., Wang, X., Xiao, W., and Zhu, H. (2018), "A generic sure independence screening procedure," *Journal of the American Statistical Association*.

Pan, W., Wang, X., Zhang, H., Zhu, H., and Zhu, J. (2019), "Ball Covariance: A Generic Measure of Dependence in Banach Space," *Journal of the American Statistical Association*, 1–24.

Pearson, K. (1895), "Notes on Regression and Inheritance in the Case of Two Parents Proceedings of the Royal Society of London, 58, 240-242," .

Rosenwald, A., Wright, G., Chan, W. C., Connors, J. M., Campo, E., Fisher, R. I., Gascoyne, R. D., Muller-Hermelink, H. K., Smeland, E. B., Giltnane, J. M., et al. (2002), "The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma," *New England Journal of Medicine*, 346, 1937–1947.

Roth, V. (2004), "The generalized LASSO," *IEEE transactions on neural networks*, 15, 16–28.

Sarkar, S. and Ghosh, A. K. (2018), "Some multivariate tests of independence based on ranks of nearest neighbors," *Technometrics*, 60, 101–111.

Sejdinovic, D., Sriperumbudur, B., Gretton, A., Fukumizu, K., et al. (2013), "Equivalence of distance-based and RKHS-based statistics in hypothesis testing," *The Annals of Statistics*, 41, 2263–2291.

Serfling, R. J. (2009), *Approximation theorems of mathematical statistics*, vol. 162, John Wiley & Sons.

Shah, R. D., Peters, J., et al. (2020), "The hardness of conditional independence testing and the generalised covariance measure," *Annals of Statistics*, 48, 1514–1538.

Shi, H., Drton, M., and Han, F. (2019), "Distribution-free consistent independence tests via Hallin's multivariate rank," *arXiv preprint arXiv:1909.10024*.

Spearman, C. (1904), "The proof and measurement of association between two things," *The American journal of psychology*, 15, 72–101.

Székely, G. J., Rizzo, M. L., Bakirov, N. K., et al. (2007), "Measuring and testing dependence by correlation of distances," *The annals of statistics*, 35, 2769–2794.

Tibshirani, R. (1996), "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, 58, 267–288.

— (1997), "The lasso method for variable selection in the Cox model," *Statistics in medicine*, 16, 385–395.

Wang, H. (2009), "Forward regression for ultra-high dimensional variable screening," *Journal of the American Statistical Association*, 104, 1512–1524.

Wang, X., Pan, W., Hu, W., Tian, Y., and Zhang, H. (2015), "Conditional distance correlation," *Journal of the American Statistical Association*, 110, 1726–1734.

Weihs, L., Drton, M., and Meinshausen, N. (2018), "Symmetric rank covariances: a generalized framework for nonparametric measures of dependence," *Biometrika*, 105, 547–562.

Williams, J., Jones, C., Kiniry, J., and Spanel, D. A. (1989), "The EPIC crop growth model," *Transactions of the ASAE*, 32, 497–0511.

Wong, R. K., Li, Y., and Zhu, Z. (2019), "Partially linear functional additive models for multivariate functional data," *Journal of the American Statistical Association*, 114, 406–418.

Zhao, S. D. and Li, Y. (2012), "Principled sure independence screening for Cox models with ultra-high-dimensional covariates," *Journal of multivariate analysis*, 105, 397–411.

Zheng, Y. (2015), "Trajectory data mining: an overview," *ACM Transactions on Intelligent Systems and Technology (TIST)*, 6, 1–41.

Zhu, L., Xu, K., Li, R., and Zhong, W. (2017), "Projection correlation between two random vectors," *Biometrika*, 104, 829–843.

Zou, H. and Hastie, T. (2005), "Regularization and variable selection via the elastic net," *Journal of the royal statistical society: series B (statistical methodology)*, 67, 301–320.