Statistica Sinica Preprint No: SS-2023-0070

Title	Frequent-voting Independence Screening for Data of Different Types or Different Dimensions
Manuscript ID	SS-2023-0070
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202023.0070
Complete List of Authors	Haeun Moon and Kehui Chen
Corresponding Authors	Haeun Moon
E-mails	haeunmoon10@gmail.com

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0070

Statistica Sinica

Frequent-voting Independence Screening for Data of Different Types or Different Dimensions

Haeun Moon and Kehui Chen

Carnegie Mellon University and University of Pittsburgh

Abstract: Modern datasets often include different types of variables with complex features, making variable selection particularly challenging. For example, a measure of dependence with the response variable may not be directly comparable among predictor variables of different types and different dimensions. To address this challenge, this work proposes a frequent-voting based independent screening method for variable selection, which avoids a direct comparison of the dependence measure among different variables. Asymptotic analyses show that the proposed method selects all of the active variables with probability converging to one. We also demonstrate its great finite sample performance through numerical experiments and the application to an ADHD study.

Key words and phrases: Model-free; Sure screening; Variable selection; Test of independence

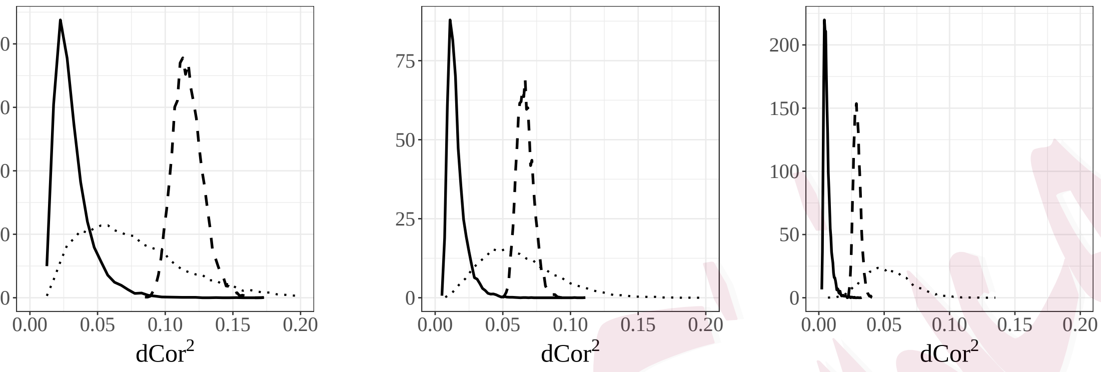
1. Introduction

This work is motivated by collaborative projects with Psychiatrists where the goal of the study is to find risk factors that are predictive of mental disorders such as Attention Deficit Hyperactivity Disorder (ADHD), Major depression, and Suicidal behaviors. This kind of study usually has a large pool of candidate risk factors, consisting of genetic, brain imaging, clinical and demographic variables. Therefore, variable selection plays an important role in these studies. The imaging data such as fMRI data and EEG data are time course data observed at many brain regions. Our data starts with aggregated time course data in each brain region (an average over all voxels in the region). The time course data are then transformed to a multivariate vector using frequency analysis and basis expansion. At this point, we select the multivariate vector as a whole. For the gene data analysis, researchers may be interested in selecting relevant gene pathways consisting a group of genes, where the variables (gene pathways) under consideration are multivariate variables. Meanwhile, many clinical data and demographic data are also available in the form of a continuous variable or a categorical variable. In addition, the response variable could be multivariate as well. For example, we may follow subjects for a few time points and produce a longitudinal outcome. These complex data structures often leads to complex nonlinear relationships as well.

There is a rich body of literature on variable selection research, including penalization/regularization methods (Tibshirani, 1996, 1997; Roth, 2004; Fan and Li, 2001; Zou and Hastie, 2005), partial likelihood function approaches (Xu and Chen, 2014; Yang et al., 2016; Liu et al., 2021), and FDR control techniques (Barber and Candès, 2015; Du et al., 2023), among others. When adopting such approaches, a common prerequisite is to construct a joint model for all variables. However, creating such models is typically challenging when data have different types and dimensions. Considering the diverse types of variables in our applications, we have chosen to first work within a marginal selection framework and employ a test of independence. A joint model then might be built after screening.

With the development of asymptotically consistent methods for the test of independence (Székely et al. (2007); Gretton et al. (2007); Heller et al. (2012); Bergsma and Dassios (2014); Pan et al. (2020)), researchers have naturally considered performing variable selection using dependence measures such as the Distance Correlation measure (Li et al. (2012)) and the Ball Correlation measure (Pan et al. (2019)). In their papers, for each predictor variable in the candidate set, one computes a measure of dependence between the variable under consideration and the response variable, and selects the top N predictor variables based on the magnitude of the dependence measure. These can be viewed as an extension of the Sure Independence Screening method (SIS) proposed by Fan and Lv (2008). The original SIS and its extensions (Fan et al. (2009); Li et al. (2012); Zhao and Li (2012); Barut et al. (2016)) select variables by ordering the marginal correlation coefficients. The new methods based on dependence measures are nice additions to these existing sure screening methods because dependence measures are well defined for multivariate data and beyond, which are particularly suitable in our application.

To apply dependence-measure-based sure screening method, existing works assume that the magnitude of the estimated dependence measure represents the strength of the variable importance, and the variables with larger values of the dependence measure will be selected. We call these *scale-based* methods. However, we encountered a major challenge when trying to apply these scale-based methods to our studies. When the variables are with different dimensions and in different types, it does not always make sense to determine their importance based on the estimated magnitude of the dependence measure. For example, when employing the distance covariance measure, to apply the scale-based method, one has to first standardize the distance covariance measure to the distance correlation measure. The $\mathrm{dCor}^2{}_n(X, Y)$ from 10,000 simulations are shown in Figure 1, for three candidate predictors, where $X_1 \in R$ is a 1-dimensional inactive variable, $X_2 \in R^{20}$ is a 20-dimensional inactive variable, and $X_3 \in R$ is a 1-dimensional active variable. All of them are generated from Gaussian distributions. At moderate sample sizes $n = 100$ and 200, the values for $\mathrm{dCor}^2{}_n(X_2, Y)$ are mostly larger than $\mathrm{dCor}^2{}_n(X_3, Y)$. If we directly compare the estimated dCor^2 for X_2 and X_3, the 20-dimensional inactive variable X_2 will be preferred over the 1-dimensional active variable X_3. In our simulations, the scale-based method always tends to miss the true low dimensional variables and over selects larger-dimensional inactive variables. When the sample size gets larger, the problem will be alleviated, as shown in the last panel of Figure 1 with $n = 500$. When we consider asymptotics with n goes to ∞, the active variables can eventually be separated from the inactive variables, and therefore the sure screening property still holds asymptotically. However, with a moderate sample size, an ordering based on the magnitude of the estimated dCor could be misleading. Scale-based methods based on other dependence measures such as the Ball correlation measure (Pan et al., 2019) and the bias-corrected distance correlation measure (Székely and Rizzo, 2013) might be employed to alleviate the effect of dimension on the null distribution, However, it's still hard to justify that the estimated standardized



(a) n=100 (b) n=200 (c) n=500

Figure 1: Distribution of $\mathrm{dCor}^2{}_n(X_1, Y)$ (——, for a 1-dim inactive variable), $\mathrm{dCor}^2{}_n(X_2, Y)$ (··········, for a 20-dim inactive variable), and $\mathrm{dCor}^2{}_n(X_3, Y)$ (- - - , for 1-dim active variable).

dependence measures can be ordered directly among variables with different dimensions and types.

To address this challenge in data analysis, we propose a *frequent-voting* method that does not directly compare the magnitude of the dependence measure between different types of data. We order variables by their frequency to pass the independence test in sub-samples. We are able to prove that all active variables are selected with probability converging to one, and with appropriate assumptions on the re-sampling ratio, we also show that with probability converging to one, none of the inactive variables will be selected.

Incorporating a resampling procedure into variable selection is intuitively appealing

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0070

and not a new idea. Meinshausen and Bühlmann (2010) provides some formal analysis on this and illustrates cases where this approach can be combined with various kinds of selection methods. Bühlmann and Yu (2002) investigates the gains of bootstrap aggregation (bagging) in the context of decision trees. Meinshausen et al. (2009) aggregates inference for individual variables across multiple random sample splits. The exploration of sample splitting and bagging strategies is also evident in the works of Du et al. (2023), Han et al. (2022) and Dai et al. (2023), where they construct a statistic possessing a global symmetric property. Most of these studies are based on a joint modeling framework and not directly applicable to our case.

In Section 2, we formally propose a Frequent-voting Independence Screening method and establish its asymptotic sure screening properties. In Section 3, we demonstrate the superb finite sample performance of the proposed method using numerical experiments. In Section 4, we apply the proposed method to the ADHD-200 dataset (part of the 1000 human connectome study). The data consists of one-dimensional phenotype variables and time course fMRI data observed at many brain regions. We show that selection based on the scale-based method may result in selecting redundant brain regions and the frequent-voting method provides more parsimonious and sensible selection results. Proofs for the theorems can be found in the appendix.

2. Frequent-voting independence screening method

Consider a response variable Y and p predictor variables $X_1, \ldots, X_p$. Here Y and $X_k, 1 \leq k \leq p$, can have different dimensions. The scale-based independence screening will utilize a dependence measure $\rho(X_k, Y)$ as a marginal utility to rank the importance of X_k.

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 $\mathcal{X} \times \mathcal{Y}$. We say that X and Y are independent if $P_{XY} \neq P_X P_Y$. A good dependence measure ρ is generally believed to satisfy the following consistency condition.

(C1) $\rho = 0$ if and only if X and Y are independent and $\rho > 0$ otherwise.

Meanwhile, there needs to be a computable empirical counterpart $\hat{\rho}_n$ that satisfies the convergence property,

(C2) Given n independent samples, $n\hat{\rho}_n$ converges to a null distribution when X and Y are independent and diverges to ∞ otherwise.

In recent years, there have been active attempts to develop dependence measures that satisfy the above conditions. The distance covariance measure (Székely et al., 2007) is a widely used measure that satisfy (C1) and (C2). This measure is originally defined for $X \in \mathbb{R}^{d_X}$ and $Y \in \mathbb{R}^{d_Y}$ through

$$\mathrm{dCov}^2(X, Y) = \int_{\mathbb{R}^{d_X + d_Y}} ||\phi_{X,Y}(t, s) - \phi_X(t)\phi_Y(s)||^2 w(t, s) dt ds,$$

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0070

where $\phi_X(\cdot)$, $\phi_Y(\cdot)$, $\phi_{X,Y}(\cdot)$ are respective characteristic functions and $w(s,t) = (c_{d_X} c_{d_Y} |t|_{d_X}^{1+d_X} |s|_{d_Y}^{1+d_Y})^{-1}$ with $c_d = \pi^{(1+d)/2}\Gamma((1+d)/2)$. Later it is has been put into a more general formula and works for more general metric spaces (Lyons, 2013).

To use this measure for scale-based independence screening, one need to consider the standardized version

$$\mathrm{dCor}^2(X,Y) = \frac{\mathrm{dCov}^2(X,Y)}{\sqrt{\mathrm{dCov}^2(X,X)\mathrm{dCov}^2(Y,Y)}},$$

ranging from 0 to 1. An empirical version ${\mathrm{dCor}^2}_n$ can be obtained by plugging in ${\mathrm{dCov}^2}_n$. Li et al. (2012) then proposed to rank predictors based on ${\mathrm{dCor}^2}_n$.

For two univariate normal random variables with the Pearson correlation coefficient r, we can show that dCor is strictly increasing in $|r|$ (Székely et al. (2007)). This property implies that the distance correlation based feature screening procedure is equivalent to the marginal Pearson correlation screening for linear and Gaussian cases. Meanwhile, it is well defined for multivariate variables, and could be more effective than the marginal Pearson correlation screening in the presence of nonlinear relationship, which makes it a good measure to use for independence screening. However, the magnitude of dCor does not always indicate the importance of the predictors when predictors have different types and dimensions. We, therefore, propose to combine a frequent-voting method with the test of independence based on the dCov measure, avoiding a direct comparison of the dependence measure between different variables.

Now consider the variable selection problem for $X_k, 1 \leq k \leq p$. Consider the condi-

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0070

tional distribution function $F(y \mid X) = P(Y \leq y \mid X)$. We have the following definition of index sets:

$$\mathcal{A} = \{k : F(y \mid X_k) \text{ functionally depends on } X_k \text{ for some } y\},$$

$$\mathcal{I} = \{k : F(y \mid X_k) \text{ does not functionally depends on } X_k \text{ for any } y\}.$$

Predictors in $\mathcal{A}$ are called *active* predictors. Predictors in $\mathcal{I}$ are called *inactive* predictors. We use p^* to denote the cardinality of $\mathcal{A}$. With a sample of size n, our aim is to construct $\hat{\mathcal{A}}_n$ that contains all active predictors and contains none or very few inactive predictors.

For $k = 1, ..., p$, let $\rho_k = \mathrm{dCov}^2(X_k, Y)$. We use the notation $z_{ki} = (X_{ki}, Y_i)$ to denote the paired data. For any given data $Z = (z_{ki_1}, ..., z_{ki_m})$ with size m, $\hat{\rho}_{k;m}$ is the empirical estimate based on m pairs of data. Let $\tau_{k,\alpha}$ as the α-level critical value of the limiting null-distribution of $m\hat{\rho}_{k;m}$, which is a mixture of χ^2 distributions (Székely et al., 2007). For a given sub-sample size m, we consider all ordered subset of m different integers chosen from $\{1, ..., n\}$, and its corresponding sub-samples $(z_{ki_1}, ..., z_{ki_m})$. A statistical decision based on $(z_{ki_1}, ..., z_{ki_m})$ is expressed as

$$h(z_{ki_1}, ..., z_{ki_m}; \tau_{k,\alpha}) = I(m\hat{\rho}_{k;m} > \tau_{k,\alpha}).$$

A decision that takes value 1 is called a vote for the k-th variable. The selection of the k-th variable is based on the frequency of the vote it gets from multiple sub-samples of size m.

Remark 1. When employing the dCov-based test of independence, a permutation procedure is usually used to approximate the critical value of the null distribution. Several recent results show that the permutation critical value converge to the critical value of the null distribution for a class of independence tests whose test-statistics take a U-statistics form; see Theorem A.1 of Kim et al. (2020), Theorem 1 of Xu and Zhu (2022), and Proposition 18 of Berrett et al. (2021) and also Rindt et al. (2021).

Formally, the stability score of X_k from all possible m tuples is defined as

$$V^k_{n;m} = \binom{n}{m}^{-1} \sum_{i_1<\ldots<i_m} h(z_{ki_1}, \ldots, z_{ki_m}; \tau_{k,\alpha}), \tag{2.1}$$

where the summation is taken over all sets of m different integers chosen from $\{1, \ldots, n\}$. For a pre-specified frequency $\theta \in (\alpha, 1)$, the selected set of variables is then

$$\hat{\mathcal{A}}_n = \{k : V^k_{n;m} > \theta\}. \tag{2.2}$$

In practice, it is desirable to approximate $V^k_{n;m}$ with some large enough number of subsamples rather than computing all n-choose-m combinations.

While the numerical experiments and data analysis in this paper focuses on the use of the distance covariance test and the distance correlation measure, the frequent-voting framework and the following asymptotic analysis does not tie to a particular method of independence test as long as it satisfies the technical conditions.

The following assumptions are needed to prove sure screening properties.

(A1) There exist some constant c_1 and $0 < \kappa < \frac{1}{2}$ such that

$$\min_{k\in\mathcal{A}} \rho_k \geq c_1 n^{-\kappa}.$$

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0070

(A2) For any $\epsilon > 0$, there exists some constant c_2 such that

$$P(|\hat{\rho}_{k;n} - \rho_k| > \epsilon) \leq \exp(-c_2 n \epsilon^2).$$

Assumption (A1) states that the dependency between each active predictor and the response variable should not be too weak. It is similar to (C2) of Li et al. (2012) and (C1) of Pan et al. (2019). Assumption (A2) ensures that the difference between the population dependency and the estimated statistic is bounded by an exponential function. Most of the dependence coefficients in the literature can be estimated by a U-statistic. Assumption (A2) can be established using the concentration inequality for U-statistics with some technical conditions. In particular, Li et al. (2012) showed that this holds for the distance covariance when the variables are bounded. The boundedness assumption can be relaxed to some Bernstein type tail moment conditions with recent developments in U-statistics concentration inequalities. In the case where the dimension of the variable grows to infinity with sample size n or even faster than sample size n, the convergence rate of $\hat{\rho}_k$ might differ. In our application of variable selection, we allow the total number of variables under consideration to grow with n, but the dimension of each variable is considered as a fixed dimension.

Theorem 1. *(Sure screening property) Assume (A1)-(A2). Let $m = c_0 n^{\gamma}$ for $\gamma \in (2\kappa, 1]$, and a constant $c_0 > 0$. If $\log(p^*) = o(n^{(\gamma - 2\kappa) \vee (1-\gamma)})$, we have $P(\mathcal{A} \subset \hat{\mathcal{A}}_n) \to 1$.*

Remark 2. Theorem 1 guarantees that all active predictors are selected in the model with probability converging to one, when p^*, the cardinality of $\mathcal{A}$, grows at a rate slower than certain exponential function of n, which is easy to satisfy. Here γ controls the sub-sampling ratio, whose lower limit is determined by the minimum signal strength as specified in (A1). The optimal error bound is $p^* \exp(-n^{1-2\kappa})$, which is the same as the rate for the corresponding scale-based method (Li et al., 2012; Pan et al., 2019).

Theorem 2. *Assume (A1)-(A2). Let $m = c_0 n^\gamma$ for $\gamma \in (2\kappa, 1)$ and a constant $c_0 > 0$. If $\log(p - p^*) = o(n^{1-\gamma})$, we have $P(\hat{\mathcal{A}}_n \subset \mathcal{A}) \to 1$.*

Theorem 2 requires m to be in a smaller order than n. When $m = c_0 n$, we can compute an upper bound of $P(k \in \hat{\mathcal{A}}_n | k \in \mathcal{I})$ using the Chebyshev inequality;

$$P(V_{n;m}^k > \theta) = P(V_{n;m}^k - E[V_{n;m}^k] > \theta - E[V_{n;m}^k]) \leq \frac{Var[V_{n;m}^k]}{2(\theta - E[V_{n;m}^k])^2},$$

which holds for sufficiently large m and n. Based on the U-statistic result, $Var[V_{n;m}^k] \leq m/n \cdot Var[h^k]$ (Serfling (2009)), the bound asymptotically becomes $c_0\alpha(1-\alpha)/\{2(\theta-\alpha)^2\}$. When $\alpha = 0.05$, $c_0 = 0.8$ and $\theta = 0.8$, this error bound is approximately equal to 0.03.

2.1 Selection of the frequency threshold θ

In practice, one may select θ based on the desired level of error control and the desired level of stability across sub-samples. Meinshausen and Bühlmann (2010) suggests that θ can be chosen from the interval (0.6, 0.9) for stability consideration. In our simulations, we show results for $\theta = 90\%$, 80%, 70% and 50%. The proposed frequent-voting method outperforms the scale-based method across all of the different choices of θ under consideration.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0070

Recall that, for any $k \in \mathcal{I}$, the stability score $V^k_{n;m}$ follows a distribution F_0 with asymptotic mean equal to α and the asymptotic variance bounded by $m/n \cdot \alpha(1-\alpha)$. For any $k \in \mathcal{A}$, the stability score $V^k_{n;m}$ will have asymptotic mean equal to 1 and the asymptotic variance shrinking to zero. According to the upper bound derived after Theorem 2, the level α and the frequency cut-off value θ collectively control the individual type I error bound. For $\alpha = 0.05$ and $m/n = 0.8$, θ can be chosen to control the individual type I error at 0.05, which corresponds to $\theta = 0.67$ in our simulations. In some recent research, data-driven method has been used to select the frequency threshold aiming at controlling the False Discovery Rate (**?**Du et al., 2023; Dai et al., 2023). Motivated by these, we also developed a data-driven approach to approximately control the FDR. Figure 2 is a histogram of $V^k_{n;m}$ in one simulation (from simulation 1), for all 200 variables with $n = 200$.

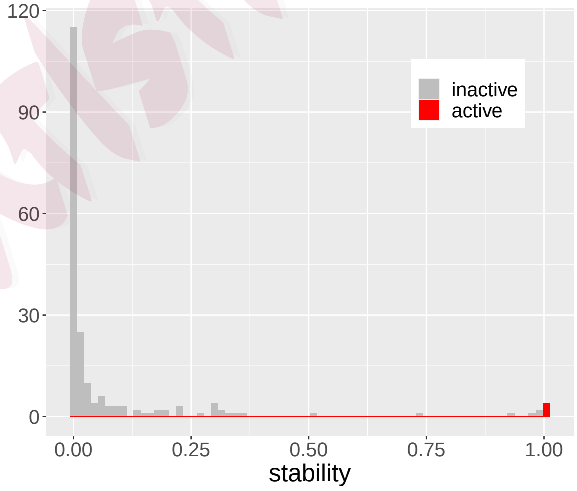


Figure 2: Histogram of the stability score generated from a simulated data.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0070

Assuming that the active predictors are sparse, we can consider fitting a distribution to the observed stability scores whose values are less than 1. In particular, we employ a zero-inflated beta distribution with a support range [0,1), characterized by a point mass η at $x = 0$ and a density function $f_0(x) = (1-\eta)\cdot\text{Beta}(\beta_0, \beta_1)$ for $0 < x < 1$. Here, $\text{Beta}(\beta_0, \beta_1)$ is a beta density function with parameters β_0, β_1, which is a common distribution used to model proportions. Once $\hat{f}_0(x)$ is estimated, an upper bound for FDR at the given θ can be computed as $\frac{p\cdot\int_\theta^1 \hat{f}_0(x)dx}{|\hat{\mathcal{A}}_n(\theta)|}$, where $|\hat{\mathcal{A}}_n(\theta)|$ is the number of selected predictors for a given θ, and p is the total number of candidate predictors. Then the frequency θ might be chosen to control the FDR at a desired level. We note that this method reasonably controls FDR if the correlation among candidate predictors are negligible and active predictors are sparse. If candidate variables are largely correlated, the actual FDR is often inflated.

3. Simulation

We consider $p = 200$ covariates. $X_1, ..., X_{100}$ are one-dimensional and generated from a normal distribution with zero mean and a covariance matrix $\Sigma = (\sigma_{ij})_{100\times100}$ where $\sigma_{ij} = 0.5^{|i-j|}$. The rest $X_{101}, ..., X_{200}$ are 11-dimensional, where variables in each dimension are normally distributed with the same covariance matrix Σ. We conduct simulations for three sample sizes $n = 100, 150, 200$.

Examples 1 - 3 are for a continuous response Y, and designed to demonstrate different types of relationships. The notation $\overline{X}$ denotes the average of the 11-dimensional components of X and $\overline{X^2}$ denotes the average of the squared dimensions.

Example 1. $Y = 1.2X_1 + 1.2X_{11} + 8\overline{X_{101}} + 8\overline{X_{111}} + \epsilon$, $\epsilon \sim N(0,1)$.

Example 2. $Y = 0.8X_1 + X_{11}^2 + 4\overline{X_{101}} + 3I(\overline{X_{111}} > 0) + \epsilon$, $\epsilon \sim N(0,1)$.

Example 3. $Y = 1.2X_1 + X_{11}^2 + 5\overline{X_{101}} + 5\overline{X_{111}^2} + \epsilon$, $\epsilon \sim N(0,1)$.

Example 4 is designed to evaluate the performance of the methods when the response variable Y is binary. Y follows a Bernoulli distribution with probability $1/(1+e^{\pi})$ where π is generated as follows:

Example 4. $\pi = X_1 + X_2 + 5\overline{X_{101}} + 5\overline{X_{111}}$.

The results from 500 simulations are summarized in Tables 1 - 4.

For the scale-based method, standardized versions dCor(X_k, Y), for $k = 1, ..., p$, are used. There is no universal way to determine the model size of the scale-based method and previous papers (Li et al. (2012); Pan et al. (2019)) have suggested to use $(n/\log n)$, which are much larger than the true model size in our simulations. Later, **?** suggest a data-driven approach to select a cut-off value based on a sample splitting. For comparison purpose, we list the selection results for multiple model sizes that match the size used in the frequent-voting method. We hope to retain the true variables with high probability with a reasonable size of the model. For the frequent-voting method, the results are based on 500 sub-samples with a significance level of 0.05 for the independence dCov test. Sub-sampling ratio 0.8 is used. The l_2 distance is used in calculating the distance covariance. Critical values are approximated by a permutation procedure on the whole data set. The total

Table 1: The proportions that each or all active predictors are selected in Example 1 are calculated from 500 simulations. Here "ratio" is the average fraction of falsely selected high-dimensional variables among all falsely selected predictors, and "size" is the average number of selected predictors.

		n=100				n=150				n=200			
		90%	80%	70%	50%	90%	80%	70%	50%	90%	80%	70%	50%
Freq	$\mathcal{P}_1$	0.52	0.62	0.70	0.79	0.76	0.83	0.86	0.91	0.88	0.92	0.93	0.97
	$\mathcal{P}_{11}$	0.51	0.61	0.68	0.77	0.71	0.79	0.83	0.89	0.87	0.89	0.93	0.96
	$\mathcal{P}_{101}$	0.85	0.89	0.92	0.95	0.98	0.98	0.99	0.99	1.00	1.00	1.00	1.00
	$\mathcal{P}_{111}$	0.82	0.88	0.91	0.95	0.97	0.98	0.99	0.99	1.00	1.00	1.00	1.00
	$\mathcal{P}_{all}$	0.23	0.33	0.41	0.56	0.51	0.64	0.70	0.80	0.76	0.81	0.86	0.93
	ratio	0.48	0.49	0.50	0.50	0.54	0.51	0.51	0.51	0.62	0.60	0.59	0.57
Scale	$\mathcal{P}_1$	0.28	0.31	0.34	0.38	0.49	0.52	0.55	0.59	0.68	0.72	0.74	0.78
	$\mathcal{P}_{11}$	0.27	0.30	0.32	0.37	0.46	0.49	0.51	0.56	0.69	0.72	0.74	0.77
	$\mathcal{P}_{101}$	0.92	0.96	0.97	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	$\mathcal{P}_{111}$	0.90	0.95	0.97	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	$\mathcal{P}_{all}$	0.07	0.09	0.11	0.15	0.24	0.27	0.30	0.35	0.47	0.52	0.55	0.60
	ratio	0.92	0.95	0.97	0.97	0.93	0.94	0.95	0.96	0.94	0.95	0.95	0.97
Size		4.5	6.4	8.5	13.8	5.1	6.5	7.8	11.3	6.8	8.5	10.3	15.1

number of variables selected, some of which could be multi-dimensional, is determined by a frequency cutoff; variables are selected if they achieve 90%, 80%, 70% or 50% of the vote from sub-sampled data. The results are based on the implementation using R packages "dcov".

In each table, the twelve columns represent the results under four different frequency

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0070

Table 2: The proportions that each or all active predictors are selected in Example 2 are calculated from 500 simulations. Here "ratio" is the average fraction of falsely selected high-dimensional variables among all falsely selected predictors, and "size" is the average number of selected predictors.

		n=100				n=150				n=200			
		90%	80%	70%	50%	90%	80%	70%	50%	90%	80%	70%	50%
Freq	$\mathcal{P}_1$	0.48	0.59	0.64	0.77	0.76	0.83	0.88	0.93	0.90	0.94	0.96	0.97
	$\mathcal{P}_{11}$	0.47	0.55	0.64	0.72	0.81	0.88	0.91	0.95	0.89	0.92	0.95	0.97
	$\mathcal{P}_{101}$	0.52	0.66	0.70	0.76	0.85	0.89	0.93	0.96	0.92	0.96	0.96	0.98
	$\mathcal{P}_{111}$	0.69	0.78	0.84	0.90	0.98	0.98	0.99	1.00	1.00	1.00	1.00	1.00
	$\mathcal{P}_{all}$	0.14	0.24	0.32	0.45	0.51	0.63	0.71	0.84	0.76	0.84	0.88	0.92
	$ratio$	0.47	0.47	0.48	0.49	0.49	0.47	0.47	0.47	0.5	0.48	0.48	0.49
Scale	$\mathcal{P}_1$	0.24	0.28	0.31	0.34	0.55	0.58	0.59	0.64	0.68	0.72	0.75	0.78
	$\mathcal{P}_{11}$	0.13	0.14	0.17	0.23	0.20	0.26	0.30	0.38	0.48	0.49	0.53	0.59
	$\mathcal{P}_{101}$	0.63	0.72	0.78	0.84	0.92	0.93	0.96	0.98	0.95	0.98	1.00	1.00
	$\mathcal{P}_{111}$	0.80	0.86	0.89	0.95	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00
	$\mathcal{P}_{all}$	0.02	0.04	0.06	0.09	0.08	0.12	0.15	0.24	0.29	0.33	0.38	0.45
	$ratio$	0.89	0.94	0.95	0.97	0.94	0.97	0.98	0.98	0.95	0.98	0.98	0.98
Size		4.4	7.2	9.8	17.0	5.8	8.1	10.5	17.0	5.8	7.6	9.5	14.5

cutoffs and with three levels of samples sizes. We evaluate the performance through $\mathcal{P}_1$, $\mathcal{P}_{11}$, $\mathcal{P}_{101}$, $\mathcal{P}_{111}$ and $\mathcal{P}_{\text{all}}$, which are the proportions of retaining each active variable and all of the active variables in 500 replications. As expected, we observe that the average size of the model increases as the frequency cutoff becomes lower. The performance of the frequent-voting method and the scale-based method improves as the sample size

Table 3: The proportions that each or all active predictors are selected in Example 3 are calculated from 500 simulations. Here "ratio" is the average fraction of falsely selected high-dimensional variables among all falsely selected predictors, and "size" is the average number of selected predictors.

		n=100				n=150				n=200			
		90%	80%	70%	50%	90%	80%	70%	50%	90%	80%	70%	50%
Freq	$\mathcal{P}_1$	0.70	0.78	0.81	0.89	0.93	0.95	0.96	0.98	0.99	0.99	0.99	1.00
	$\mathcal{P}_{11}$	0.19	0.27	0.33	0.45	0.45	0.55	0.61	0.74	0.74	0.81	0.85	0.90
	$\mathcal{P}_{101}$	0.44	0.55	0.61	0.71	0.75	0.83	0.87	0.92	0.93	0.96	0.97	0.98
	$\mathcal{P}_{111}$	0.16	0.23	0.32	0.44	0.37	0.45	0.54	0.66	0.65	0.76	0.81	0.87
	$\mathcal{P}_{all}$	0.03	0.07	0.10	0.20	0.18	0.26	0.34	0.49	0.49	0.62	0.68	0.78
	ratio	0.52	0.52	0.52	0.54	0.49	0.47	0.47	0.47	0.49	0.53	0.57	0.58
Scale	$\mathcal{P}_1$	0.51	0.53	0.57	0.61	0.80	0.82	0.83	0.86	0.92	0.92	0.94	0.95
	$\mathcal{P}_{11}$	0.03	0.03	0.04	0.07	0.06	0.07	0.08	0.10	0.16	0.20	0.24	0.30
	$\mathcal{P}_{101}$	0.58	0.68	0.76	0.84	0.85	0.92	0.94	0.97	0.97	0.99	0.99	1.00
	$\mathcal{P}_{111}$	0.22	0.33	0.43	0.57	0.52	0.64	0.74	0.83	0.81	0.89	0.92	0.97
	$\mathcal{P}_{all}$	0.00	0.00	0.01	0.03	0.03	0.04	0.05	0.07	0.12	0.16	0.21	0.27
	ratio	0.93	0.96	0.97	0.98	0.94	0.97	0.98	0.98	0.89	0.94	0.96	0.97
Size		2.5	4.1	5.7	10.2	3.7	5.1	6.5	10.8	5.0	6.6	8.3	12.9

increases. Overall, the performance of the frequent voting-method is very satisfactory with a sufficient sample size $n = 200$, 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, in the sense that the probability of retaining all true variables is higher with the frequent-voting method when the size of the

Table 4: The proportions that each or all active predictors are selected in Example 4 are calculated from 500 simulations. Here “ratio” is the average fraction of falsely selected high-dimensional variables among all falsely selected predictors, and “size” is the average number of selected predictors.

		n=100				n=150				n=200			
		90%	80%	70%	50%	90%	80%	70%	50%	90%	80%	70%	50%
Freq	$\mathcal{P}_1$	0.37	0.45	0.52	0.63	0.60	0.68	0.74	0.84	0.76	0.83	0.87	0.92
	$\mathcal{P}_{11}$	0.38	0.46	0.53	0.63	0.60	0.69	0.75	0.84	0.79	0.83	0.87	0.92
	$\mathcal{P}_{101}$	0.31	0.42	0.49	0.62	0.63	0.74	0.79	0.87	0.86	0.90	0.92	0.96
	$\mathcal{P}_{111}$	0.27	0.36	0.43	0.57	0.64	0.72	0.78	0.85	0.86	0.92	0.95	0.97
	$\mathcal{P}_{all}$	0.00	0.01	0.03	0.12	0.12	0.20	0.31	0.49	0.41	0.56	0.65	0.78
	ratio	0.4	0.44	0.46	0.49	0.44	0.45	0.48	0.52	0.4	0.41	0.41	0.42
Scale	$\mathcal{P}_1$	0.32	0.36	0.40	0.44	0.51	0.56	0.61	0.65	0.68	0.72	0.74	0.78
	$\mathcal{P}_{11}$	0.34	0.38	0.41	0.44	0.53	0.57	0.60	0.63	0.72	0.76	0.78	0.81
	$\mathcal{P}_{101}$	0.37	0.50	0.59	0.72	0.69	0.81	0.86	0.93	0.89	0.93	0.95	0.98
	$\mathcal{P}_{111}$	0.31	0.45	0.55	0.70	0.69	0.80	0.84	0.90	0.89	0.95	0.97	0.99
	$\mathcal{P}_{all}$	0.00	0.01	0.03	0.08	0.11	0.18	0.22	0.31	0.36	0.47	0.51	0.60
	ratio	0.53	0.67	0.76	0.84	0.63	0.74	0.79	0.85	0.64	0.74	0.79	0.86
Size		2.1	3.1	4.3	7.4	3.6	4.9	6.2	9.5	4.6	5.8	7.2	10.3

model is kept the same. We also recorded the fraction of falsely selected 11-dimensional variables among all falsely selected variables, and the average fractions in all simulations are reported in the tables as “ratio”. We can see that the scale-based method tends to miss the one-dimensional active variables X_1 and X_2, and include inactive variables with higher dimensions. For the scale-based method, the fractions of falsely selected high-dimensional variables are close to 1 in Example 1 - Example 3. The fraction is a bit more balanced in Example 4, because the discrepancy in $dCor_n(X,Y)$ for X variables with varying dimensions is reduced when Y is binary.

We also explored the data driven approach for the selection of θ. For $n = 200$, to control FDR at 0.3 or 0.5, the average θ selected is around 78% and 69%, in all simulations. For a target value 0.5, the empricial FDR values are 0.57, 0.47, 0.43, 0.51 for simulations 1-4 respectively. For a target value 0.3, the empirical FDR values are 0.49, 0.37, 0.34, 0.41 for simulations 1-4 respectively. The results demonstrate that the empirical FDR are controlled within a reasonable range, with a certain degree of inflation. This inflation is mainly due to the correlation among predictor variables. We repeated the same experiments with uncorrelated predictor variables, and the FDR values are well controlled.

4. Application: ADHD 200

In this section, our method is employed to identify important variables related to Attention Deficit Hyperactivity Disorder (ADHD). ADHD is a common neurological disorder prevalent among school-aged children (Willcutt (2012)), characterized by difficulties in attention and impulse control. We use ADHD-200 consortium data set (ADHD-200-Consortium (2012); Bellec et al. (2017)), which was publicly released to support the development of scientific tools for diagnosing the condition. As this data set was made available as part of the ADHD-200 Global Competition, it naturally underwent a rigorous pre-processing step, making it highly suitable for methodological research as evidenced by previous publications.

The dataset contains phenotype variables and resting-state fMRI data written into MNI space at 4 mm x 4 mm x 4 mm voxel resolution. In each voxel, the mean blood-oxygen-level dependent (BOLD) signal was recorded at equally spaced time points. The data is processed using the Athena pipeline and aggregated over functionally parcellated regions of interest (ROIs) called “CC200” (Craddock et al. (2012)). The “ADHD Rating Scale IV” measurement is used as the response variable, which represents the severity of symptoms on a continuous scale. This analysis focuses on data collected at the NYU site, resulting in a final sample of 215 observations with seven phenotype variables and time course fMRI data from 190 ROIs. The phenotype variables include gender, age, handedness, verbal IQ, full-scale IQ, performance IQ, and medication status, all of which are one-dimensional.

To use the distance covariance test dCov or the distance correlation measure dCor in screening, one needs to employ an appropriate distance measure. For the phenotype data, l_2 distance is used. For the rs-fMRI data, we choose not to apply the l_2 distance directly on the time-series data, because the original time-series data are noisy and contain individual level horizontal shift, which could lead to spurious distance between pairs of ROIs. Following some previous papers (Biswal et al. (1995); Yu-Feng et al. (2007)) in ADHD studies with fMRI data, we transformed each time-series data to *Global Wavelet Power Spectrum* (GWPS) and then applied l_2 distance on the coefficients. The GWPS

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0070

transformation effectively summarizes the data by indicating the average power at specific frequency levels. The frequency of interest in this data is from 0 to 0.25 Hz, on an equally spaced grid of 60. Morlet's wavelet (R package "biwavelet") is used for our analysis. In addition to the curve analysis, we also consider a five-dimensional summary for the fMRI data, where data in each ROI is summarized into average GWPS values in five predefined frequency bins (0-0.0117 Hz, 0.0117-0.0273 Hz, 0.0273-0.0742 Hz, 0.0742-0.1992 Hz, 0.1992-0.25 Hz) (Zhang et al. (2015); Wang et al. (2015); Luo et al. (2020)). Then l_2 distance is used for this five-dimensional data when applying the dCov or dCor methods.

Both the frequency-voting method and the scale-based method are applied to select variables using distance covariance/distance correlation measures. The implementation details are the same as specified in the simulation section. For the phenotype variables, the frequency vote of each variable from the sub-samples is summarized in Table 5. The ranking of variables based on the frequency vote exactly matches the ranking based on the magnitude of dCor (the scale-based method). The order of variables is also aligned with the p-values from an individual test of independence between each predictor and the response variable using dCov. Several studies reveal an association between "Full-scale IQ" and ADHD (Bridgett and Walker (2006); Fabio et al. (2022)) and the p-value from the test of independence also supports this result. Therefore, the "Full-scale IQ" is used as a cutoff for the rest of the analysis. This means that variables, phenotypes and brain ROIs, are selected if their votes exceed 54% for the frequent-voting method and if their dependence measure (dCor) exceed that of "Full-scale IQ" for the scale-based method.

Table 5: Summary of the association between phenotype variables and the response.

Variable	Frequency vote	dCor	P-value
Medication Status	100%	0.49	0.00
Gender	100%	0.27	0.00
Verbal IQ	76%	0.20	0.01
Age	74%	0.18	0.02
Full-scale IQ	54%	0.17	0.04
Performance IQ	11%	0.14	0.12
Handedness	4%	0.12	0.17

Table 6: Selection path of brain ROIs using Full-scale IQ as a cutoff with CC200 labels. Numbers of selected ROIs are inside the parenthesis.

Method	Data	ROI selection path
Frequent-voting	5-dim	138 44 33 112 1 81 35 36 31 (9)
	Curve	138 44 33 1 35 189 31 112 81 (9)
Scale-based	5-dim	138 33 44 112 1 35 31 81 102 160 175 90 119 122 36 (15)
	Curve	138 33 35 112 1 31 44 102 160 90 32 189 122 119 81 175 108 173 47 (19)

The results of the brain ROI selection are summarized in Table 6 and Figure 3. The regions are ordered by the frequency vote or the magnitude of dCor. Each number represents a brain region labeled by CC200 parcellation. Both the curve data and its five-

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0070

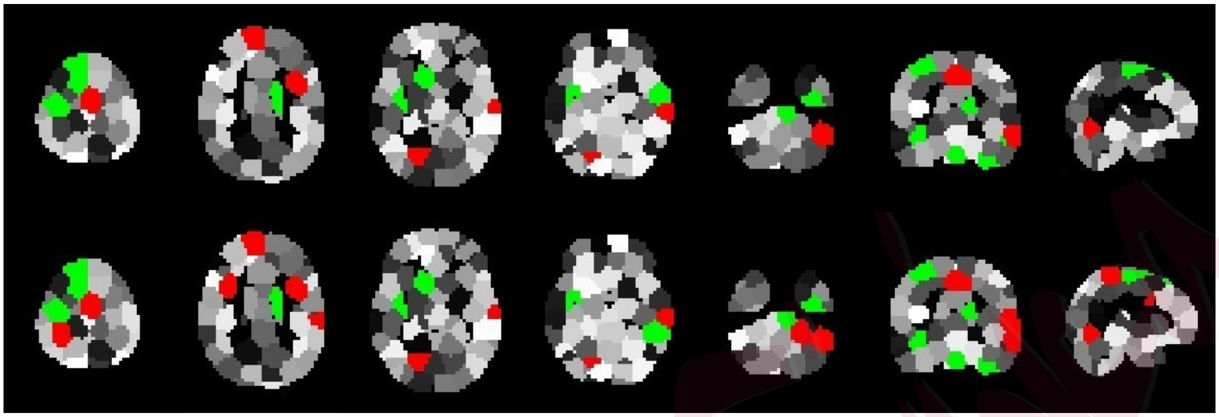


Figure 3: The selected brain regions are highlighted in green if selected by both the scale-based and the frequent-voting methods, and in red if only selected by the scale-based method. The top row displays the results using five-dimensional summary data; while the bottom row shows the results using the curve data. The brain images include five different horizontal slices, one sagittal slice, and one coronal slice.

dimensional summary data are used. The result from the frequent-voting method remains stable for both versions of the data, where eight out of the nine selected brain regions are the same. The scale-based method selects more brain regions. If the curve data are used, the scale-based method selects an even larger number of brain regions comparing to the result based on the five-dimensional version. We find that the ROIs selected by the frequent-voting method are a subset of ROIs selected by the scale-based method. Our simulations under a similar setting showed that the scale-based method tends to rank high-dimensional inactive variables higher than some low-dimensional active variables. While it is difficult to directly verify the correctness of the selection in this data analysis, we found that the frequent-voting method provides more parsimonious and stable results.

The top ROIs selected by both methods have been identified and well discussed in ADHD literature. For example, ROI 138 is mainly comprised of *Right Cerebellum* and *Fusiform Gyrus* which has been confirmed as related to ADHD with different data and theories (Wolf et al. (2009); Lei et al. (2014); Stoodley (2016); Chiang et al. (2020)). ROI 33 is largely a part of *left superior temporal gyrus* and *left supramarginal gyrus* which are found in Rubia et al. (2007); Wolf et al. (2009); Sidlauskaite et al. (2015); Zhang et al. (2020). ROI 44 contains a part of *lingual gyrus*, as discussed in An et al. (2013); Zhao et al. (2017); Lan et al. (2021).

5. Discussion

The marginal variable selection framework has a long history and has gained more attention after the work of sure screening (Fan and Lv, 2008), where Pearson's correlation coefficient is employed to filter out variables in a linear regression setting. In some applications where strong inter-dependencies present among variables, conventional marginal screening methods may lead to spurious (or overlooked) discoveries. Our study extends the marginal sure screening framework to accommodate scenarios involving multivariate (grouped) variables and complex relationships. It retains a marginal selection aspect in that we refrain from imposing a joint model on all variables. Nevertheless, it is not entirely marginal, as we can first identify relevant groups of variables that may exhibit high correlation and then study their collective relationship with y. A successful application of this grouping procedure in biological data can be found in Wang et al. (2012). We consider the proposed method a useful addition to the literature on variable selection.

Another potential approach to mitigate the limitations of marginal selection in some applications is by employing a conditional dependence measure. Some work in this direction already exists. For instance, Wang et al. (2015) utilized the conditional distance covariance measure for variable selection. The proposed notion of frequent voting can be integrated with conditional independence tests, which are anticipated to outperform scale-based methods when dealing with variables of varying types or dimensions. Nevertheless, developing a conditional dependence measure for variables of different types and dimensions is itself an intricate challenge that is still under development.

Another point worth of discussion is the use of the sub-sampling approach instead of the bootstrapping approach (sub-sampling with replacement). The standard bootstrap, which involves constructing a resample that is of approximately the same size as the original sample, works well if the statistics under consideration have an asymptotic normal distribution and the function satisfies some continuity condition around the population value. The statistics for the independence test often violate these conditions. Under the null case, we have a degenerate U-statistic which follows a mixture of χ^2 distributions, while under the alternative case, the statistics have a normal distribution with a different scaling factor. Empirically, we found that bootstrapping with $m = 0.8n$ and $n = 200$ generated a spurious relationship between X and Y, because some pairs (X_i, Y_i) are sampled multiple times. This leads to an over selection of irrelevant variables. A commonly used approach to overcome this difficulty is to use an “m-out-of-n” bootstrapping method with a small m relative to n, which is essentially the same as a sub-sampling approach with a small m. Whether there is a better resampling approach is a question might be worth of further research.

Acknowledgement

We thank the editor, associate editor and three anonymous referees for their helpful comments and suggestions. Kehui Chen’s work is partly supported by NSF DMS 2210402

References

ADHD-200-Consortium (2012). The adhd-200 consortium: a model to advance the translational potential of neuroimaging in clinical neuroscience. *Frontiers in systems neuroscience 6*, 62.

An, L., Q.-J. Cao, M.-Q. Sui, L. Sun, Q.-H. Zou, Y.-F. Zang, et al. (2013). Local synchronization and amplitude of the fluctuation of spontaneous brain activity in attention-deficit/hyperactivity disorder: a resting-state fmri study. *Neuroscience bulletin 29*(5), 603–613.

Barber, R. F. and E. J. Candès (2015). Controlling the false discovery rate via knockoffs. *Annals of statistics 43*(5), 2055–2085.

Barut, E., J. Fan, and A. Verhasselt (2016). Conditional sure independence screening. *Journal of the American Statistical Association 111*(515), 1266–1277.

Bellec, P., C. Chu, F. Chouinard-Decorte, Y. Benhajali, D. S. Margulies, and R. C. Craddock (2017). The neuro bureau adhd-200 preprocessed repository. *Neuroimage 144*, 275–286.

Bergsma, W. and A. Dassios (2014). A consistent test of independence based on a sign covariance related to kendall's tau. *Bernoulli 20*(2), 1006–1028.

Berrett, T. B., I. Kontoyiannis, and R. J. Samworth (2021). Optimal rates for independence testing via u-statistic permutation tests. *Annals of Statistics 49*(5), 2457–2490.

Biswal, B., F. Zerrin Yetkin, V. M. Haughton, and J. S. Hyde (1995). Functional connectivity in the motor cortex of resting human brain using echo-planar mri. *Magnetic resonance in medicine 34*(4), 537–541.

Bridgett, D. J. and M. E. Walker (2006). Intellectual functioning in adults with adhd: a meta-analytic examination of full scale iq differences between adults with and without adhd. *Psychological assessment 18*(1), 1–14.

Bühlmann, P. and B. Yu (2002). Analyzing bagging. *The Annals of Statistics 30*(4), 927–961.

Chiang, C.-T., C.-S. Ouyang, R.-C. Yang, R.-C. Wu, and L.-C. Lin (2020). Increased temporal lobe beta activity in boys with attention-deficit hyperactivity disorder by loreta analysis. *Frontiers in Behavioral Neuroscience 14*(85).

Craddock, R. C., G. A. James, P. E. Holtzheimer III, X. P. Hu, and H. S. Mayberg (2012). A whole brain fmri atlas generated via spatially constrained spectral clustering. *Human brain mapping 33*(8), 1914–1928.

Dai, C., B. Lin, X. Xing, and J. S. Liu (2023). A scale-free approach for false discovery rate control in generalized linear models. *Journal of the American Statistical Association 118*(543), 1551–1565.

Du, L., X. Guo, W. Sun, and C. Zou (2023). False discovery rate control under general dependence by symmetrized data aggregation. *Journal of the American Statistical Association 118*(541), 607–621.

Fabio, R. A., G. E. Towey, and T. Caprì (2022). Static and dynamic assessment of intelligence in adhd subtypes. *Frontiers in Psychology 13*(712).

Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association 96*(456), 1348–1360.

Fan, J. and J. Lv (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70*(5), 849–911.

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

Gretton, A., K. Fukumizu, C. Teo, L. Song, B. Schölkopf, and A. Smola (2007). A kernel statistical test of independence. *Advances in neural information processing systems 20*, 585–592.

Han, Y., X. Guo, and C. Zou (2022). Model-free variable selection in sufficient dimension reduction via fdr control. *arXiv preprint arXiv:2210.12382*.

Heller, R., Y. Heller, and M. Gorfine (2012). A consistent multivariate test of association based on ranks of distances. *Biometrika 100*(2), 503–510.

Kim, I., S. Balakrishnan, L. Wasserman, et al. (2020). Robust multivariate nonparametric tests via projection averaging. *Annals of Statistics 48*(6), 3417–3441.

Lan, Z., Y. Sun, L. Zhao, Y. Xiao, C. Kuai, and S.-W. Xue (2021). Aberrant effective connectivity of the ventral putamen in boys with attention-deficit/hyperactivity disorder. *Psychiatry Investigation 18*(8), 763–769.

Lei, D., J. Ma, X. Du, G. Shen, X. Jin, and Q. Gong (2014). Microstructural abnormalities in the combined and inattentive subtypes of attention deficit hyperactivity disorder: a diffusion tensor imaging study. *Scientific Reports 4*(1), 1–7.

Li, G., H. Peng, J. Zhang, and L. Zhu (2012). Robust rank correlation based screening. *Annals of statistics 40*(3), 1846–1877.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0070

Li, R., W. Zhong, and L. Zhu (2012). Feature screening via distance correlation learning. *Journal of the American Statistical Association 107*(499), 1129–1139.

Liu, Y., J. Xu, and G. Li (2021). Sure joint feature screening in nonparametric transformation model for right censored data. *Canadian Journal of Statistics 49*(2), 549–565.

Luo, F.-F., J.-B. Wang, L.-X. Yuan, Z.-W. Zhou, H. Xu, S.-H. Ma, Y.-F. Zang, and M. Zhang (2020). Higher sensitivity and reproducibility of wavelet-based amplitude of resting-state fmri. *Frontiers in neuroscience 14*(224).

Lyons, R. (2013). Distance covariance in metric spaces. *Annals of Probability 41*(5), 3284–3305.

Meinshausen, N. and P. Bühlmann (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 72*(4), 417–473.

Meinshausen, N., L. Meier, and P. Bühlmann (2009). P-values for high-dimensional regression. *Journal of the American Statistical Association 104*(488), 1671–1681.

Pan, W., X. Wang, W. Xiao, and H. Zhu (2019). A generic sure independence screening procedure. *Journal of the American Statistical Association 114*(526), 928–937.

Pan, W., X. Wang, H. Zhang, H. Zhu, and J. Zhu (2019). Ball covariance: A generic measure of dependence in banach space. *Journal of the American Statistical Association*.

Pan, W., X. Wang, H. Zhang, H. Zhu, and J. Zhu (2020). Ball covariance: A generic measure of dependence in banach space. *Journal of the American Statistical Association 115*(529), 307–317.

Rindt, D., D. Sejdinovic, and D. Steinsaltz (2021). Consistency of permutation tests of independence using distance covariance, hsic and dhsic. *Stat 10*(1), e364.

Roth, V. (2004). The generalized lasso. *IEEE transactions on neural networks 15*(1), 16–28.

Rubia, K., A. B. Smith, M. J. Brammer, and E. Taylor (2007). Temporal lobe dysfunction in medication-naive boys with attention-deficit/hyperactivity disorder during attention allocation and its relation to response variability. *Biological psychiatry 62*(9), 999–1006.

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

Sidlauskaite, J., K. Caeyenberghs, E. Sonuga-Barke, H. Roeyers, and J. R. Wiersema (2015). Whole-brain structural topology in adult attention-deficit/hyperactivity disorder: Preserved global–disturbed local network organization. *NeuroImage: Clinical 9*, 506–512.

Stoodley, C. J. (2016). The cerebellum and neurodevelopmental disorders. *The Cerebellum 15*(1), 34–37.

Székely, G. J. and M. L. Rizzo (2013). The distance correlation t-test of independence in high dimension. *Journal of Multivariate Analysis 117*, 193–213.

Székely, G. J., M. L. Rizzo, and N. K. Bakirov (2007). Measuring and testing dependence by correlation of distances. *The annals of statistics 35*(6), 2769–2794.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological) 58*(1), 267–288.

Tibshirani, R. (1997). The lasso method for variable selection in the cox model. *Statistics in medicine 16*(4), 385–395.

Wang, H., S.-H. Lo, T. Zheng, and I. Hu (2012). Interaction-based feature selection and classification for high-dimensional biological data. *Bioinformatics 28*(21), 2834–2842.

Wang, J., Z. Zhang, G.-J. Ji, Q. Xu, Y. Huang, Z. Wang, Q. Jiao, F. Yang, Y.-F. Zang, W. Liao, et al. (2015). Frequency-specific alterations of local synchronization in idiopathic generalized epilepsy. *Medicine 94*(32), e1374.

Wang, X., W. Pan, W. Hu, Y. Tian, and H. Zhang (2015). Conditional distance correlation. *Journal of the American Statistical Association 110*(512), 1726–1734.

Willcutt, E. G. (2012). The prevalence of dsm-iv attention-deficit/hyperactivity disorder: a meta-analytic review. *Neurotherapeutics 9*(3), 490–499.

Wolf, R. C., M. M. Plichta, F. Sambataro, A. J. Fallgatter, C. Jacob, K.-P. Lesch, M. J. Herrmann, C. Schönfeldt-Lecuona, B. J. Connemann, G. Grön, et al. (2009). Regional brain activation changes and abnormal functional connectivity of the ventro-

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0070

lateral prefrontal cortex during working memory processing in adults with attention-deficit/hyperactivity disorder. *Human brain mapping 30*(7), 2252–2266.

Xu, C. and J. Chen (2014). The sparse mle for ultrahigh-dimensional feature screening. *Journal of the American Statistical Association 109*(507), 1257–1269.

Xu, K. and L. Zhu (2022). Power analysis of projection-pursuit independence tests. *Statist. Sinica 32*, 417–433.

Yang, G., Y. Yu, R. Li, and A. Buu (2016). Feature screening in ultrahigh dimensional cox's model. *Statistica Sinica 26*, 881–901.

Yu-Feng, Z., H. Yong, Z. Chao-Zhe, C. Qing-Jiu, S. Man-Qiu, L. Meng, T. Li-Xia, J. Tian-Zi, and W. Yu-Feng (2007). Altered baseline brain activity in children with adhd revealed by resting-state functional mri. *Brain and Development 29*(2), 83–91.

Zhang, H., L. Zhang, and Y. Zang (2015). Fluctuation amplitude and local synchronization of brain activity in the ultra-low frequency band: an fmri investigation of continuous feedback of finger force. *Brain Research 1629*, 104–112.

Zhang, H., Y. Zhao, W. Cao, D. Cui, Q. Jiao, W. Lu, H. Li, and J. Qiu (2020). Aberrant functional connectivity in resting state networks of adhd patients revealed by independent component analysis. *BMC neuroscience 21*(1), 1–11.

Zhao, Q., H. Li, X. Yu, F. Huang, Y. Wang, L. Liu, et al. (2017). Abnormal resting-state functional connectivity of insular subregions and disrupted correlation with working memory in adults with attention deficit/hyperactivity disorder. *Frontiers in psychiatry 8*(200).

Zhao, S. D. and Y. Li (2012). Principled sure independence screening for cox models with ultra-high-dimensional covariates. *Journal of multivariate analysis 105*(1), 397–411.

Zou, H. and T. Hastie (2005). Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology) 67*(2), 301–320.

Appendix A : Proof

Proof of Theorem 1

Proof. By a U-statistic theory (Serfling, 2009), we have

$$\mathbb{E}[V^k_{n;m}] = \mathbb{E}[h(z^k_{i_1}, ..., z^k_{i_m}; \tau_{k,\alpha})],$$

$$Var[V^k_{n;m}] \leq \frac{m}{n} Var[h(z^k_{i_1}, ..., z^k_{i_m}; \tau_{k,\alpha})]$$

For $k \in \mathcal{A}$, ρ_k is strictly positive. Therefore, there exists a positive integer n_0 and also a corresponding m_0 such that $\rho_k - \frac{1}{m}\tau_{k,\alpha} > 0$, for $m > m_0$. Then, for $n > n_0$ and $m > m_0$, we have

$$\begin{aligned}\mathbb{E}[V^k_{n;m}] &= P(m\hat{\rho}_{k;m} > \tau_{k,\alpha})\\ &= P(\rho_k - \hat{\rho}_{k;m} < \rho_k - \frac{1}{m}\tau_{k,\alpha})\\ &\geq P(|\rho_k - \hat{\rho}_{k;m}| < \rho_k - \frac{1}{m}\tau_{k,\alpha})\\ &\geq 1 - \exp(-c_2 m(\rho_k - \frac{1}{m}\tau_{k,\alpha})^2)\end{aligned}$$

where the last inequality is driven from (A2). As $m = c_0 n^\gamma$, we have

$$\exp(-c_2 m(\rho_k - \frac{1}{m}\tau_{k,\alpha})^2) = \mathcal{O}(\exp(-m\rho_k^2)) = \mathcal{O}(\exp(-n^{\gamma-2\kappa})) \to 0.$$

Then

$$\mathbb{E}[V^k_{n;m}] = 1 - \mathcal{O}(\exp(-n^{\gamma-2\kappa}))$$

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0070

and

$$\begin{aligned} Var[h^k(z^k_{i_1}, ..., z^k_{i_m}; \tau_{k,\alpha})] &= \mathbb{E}[h^k](1 - \mathbb{E}[h^k]) \\ &= \mathcal{O}(\exp(-n^{\gamma-2\kappa})) \end{aligned}$$

We now establish an upper bound of $P(k \notin \hat{\mathcal{A}}_n)$. Since $\mathbb{E}[V^k_{n;m}] \to 1$, we can find a large enough n_0 and a corresponding m_0 such that $\mathbb{E}[V^k_{n;m}] - \theta > 0$, for $n > n_0$ and $m > m_0$. We consider $n > n_0$ and $m > m_0$ below.

(i) Using a Chebyshev inequality,

$$\begin{aligned} P(k \notin \hat{\mathcal{A}}_n) &= P(V^k_{n;m} < \theta) \\ &= P(\mathbb{E}[V^k_{n;m}] - V^k_{n;m} > \mathbb{E}[V^k_{n;m}] - \theta) \\ &\leq P(|\mathbb{E}[V^k_{n;m}] - V^k_{n;m}| > \mathbb{E}[V^k_{n;m}] - \theta) \\ &\leq \frac{Var(V^k_{n;m})}{2(\mathbb{E}[V^k_{n;m}] - \theta)^2} = \mathcal{O}(n^{\gamma-1}\exp(-n^{\gamma-2\kappa})) \end{aligned}$$

where the last inequality is derived from

$$Var(V^k_{n;m}) \leq \frac{m}{n} Var[h(z^k_{i_1}, ..., z^k_{i_m}; \tau_{k,\alpha})] = \mathcal{O}(n^{\gamma-1}\exp(-n^{\gamma-2\kappa})).$$

(ii) Using a Bernstein bound for U-statistic,

$$P(|V^k_{n;m} - \mathbb{E}[V^k_{n;m}]| > \epsilon) < \exp(-\frac{n/m \cdot \epsilon^2}{2(\sigma^2_{h^k} + \epsilon/3)})$$

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0070

where $\sigma^2_{h^k} = \mathbb{E}[h^k](1 - \mathbb{E}[h^k]) = \mathcal{O}(\exp(-n^{\gamma-2\kappa}))$. Then,

$$\begin{aligned} P(k \notin \hat{\mathcal{A}}_n) &= P(V^k_{n;m} < \theta) \\ &= P(\mathbb{E}[V^k_{n;m}] - V^k_{n;m} > \mathbb{E}[V^k_{n;m}] - \theta) \\ &\leq P(|\mathbb{E}[V^k_{n;m}] - V^k_{n;m}| > \mathbb{E}[V^k_{n;m}] - \theta) \\ &\leq \exp(-\frac{n/m \cdot (\mathbb{E}[V^k_{n;m}] - \theta)^2}{2(\sigma^2_{h^k} + (\mathbb{E}[V^k_{n;m}] - \theta)/3)}) = \mathcal{O}(\exp(-n^{1-\gamma})). \end{aligned}$$

Combining (i) and (ii), we have $P(k \notin \hat{\mathcal{A}}_n) = \mathcal{O}(\exp(-n^{(\gamma-2\kappa)\vee(1-\gamma)}))$. Since

$$P(\mathcal{A} \subset \hat{\mathcal{A}}_n) > 1 - p^* P(k \notin \hat{\mathcal{A}}_n),$$

and $\log(p^*) = o(n^{(\gamma-2\kappa)\vee(1-\gamma)})$, $P(\mathcal{A} \subset \hat{\mathcal{A}}_n) \to 1$ is derived.

The variance $\sigma^2_{h^k}$ converges to zero, and therefore the Chebyshev inequality provides a better bound than the Bernstein method for some values of γ.

□

Proof of Theorem 2

Proof. If $k \in \mathcal{I}$, we use a Bernstein bound for U-statistic;

$$P(|V^k_{n;m} - \mathbb{E}[V^k_{n;m}]| > \epsilon) < \exp(-\frac{n\epsilon^2/m}{2(\sigma^2_{h^k} + \epsilon/3)})$$

where $\sigma^2_{h^k} = \mathbb{E}[h^k](1 - \mathbb{E}[h^k])$. Since $\mathbb{E}[V^k_{n;m}] \to \alpha$ and $\theta > \alpha$, there exists some positive integer n_0 and a corresponding m_0 such that $\theta - \mathbb{E}[V^k_{n;m}] > 0$, for $n > n_0$ and $m > m_0$.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0070

Then

$$\begin{aligned}
P(k \in \hat{\mathcal{A}}_n) &= P(V_{n;m}^k > \theta) \\
&= P(V_{n;m}^k - \mathbb{E}[V_{n;m}^k] > \theta - \mathbb{E}[V_{n;m}^k]) \\
&\leq P(|V_{n;m}^k - \mathbb{E}[V_{n;m}^k]| > \theta - \mathbb{E}[V_{n;m}^k]) \\
&\leq \frac{1}{2}\exp(-\frac{n/m \cdot (\theta - \mathbb{E}[V_{n;m}^k]))^2}{2(\sigma_{h^k}^2 + (\theta - \mathbb{E}[V_{n;m}^k])/3)}) = \mathcal{O}(\exp(-n^{1-\gamma}))
\end{aligned}$$

In a last step, we used $\mathbb{E}[V_{n;m}^k] = \mathbb{E}[h(z_{i_1}^k, ..., z_{i_m}^k)] \to \alpha$ and $\sigma_{h^k}^2 \to \alpha(1-\alpha)$ so that the order only depends on a term n/m. Since

$$P(\hat{\mathcal{A}}_n \subset \mathcal{A}) \geq 1 - (p - p^*)P(k \in \hat{\mathcal{A}}_n),$$

given $\gamma < 1$ and $\log(p - p^*) = o(n^{1-\gamma})$, we have $P(\hat{\mathcal{A}}_n \subset \mathcal{A}) \to 1$. □

Haeun Moon

Departments of Statistics & Data Science, Carnegie Mellon University, PA 15213

E-mail: haeunm@andrew.cmu.edu

Kehui Chen

Department of Statistics, University of Pittsburgh, PA 15260

E-mail: khchen@pitt.edu