Acta Mathematicae Applicatae Sinica, English Series
Vol. 37, No. 3 (2021) 590–601
https://doi.org/10.1007/s10255-021-1012-1
http://www.ApplMath.com.cn & www.SpringerLink.com

Acta Mathematicae Applicatae Sinica, English Series

Gini Correlation for Feature Screening

Jun-ying ZHANG[1], Xiao-feng LIU[2,†], Ri-quan ZHANG[3,4], Hang-WANG[1]

[1]Department of Mathematics, Taiyuan University of Technology, Taiyuan 030024, China
(E-mail: zhangjunying-zjy@163.com)
[2]College of Data Science, Taiyuan University of Technology, Taiyuan 030024, China
(E-mail: fengzhangj68@163.com)
[3]School of Finance and Statistics, East China Normal University, Shanghai 200241, China
(E-mail: zhangriquan@163.com)
[4]Department of Mathematics, Shanxi Datong University, Datong 037009, China

Abstract In this paper we propose the Gini correlation screening (GCS) method to select the important variables with ultrahigh dimensional data. The new procedure is based on the Gini correlation coefficient via the covariance between the response and the rank of the predictor variables rather than the Pearson correlation and the Kendall τ correlation coefficient. The new method does not require imposing a specific model structure on regression functions and only needs the condition which the predictors and response have continuous distribution function. We demonstrate that, with the number of predictors growing at an exponential rate of the sample size, the proposed procedure possesses consistency in ranking, which is both useful in its own right and can lead to consistency in selection. The procedure is computationally efficient and simple, and exhibits a competent empirical performance in our intensive simulations and real data analysis.

Keywords ultrahigh dimension; Gini correlation coefficient; variable screening; feature ranking
2000 MR Subject Classification 73L05; 41A60

1 Introduction

With rapid advances in information and scientific technology, we have witnessed an explosive growth in capabilities to generate and collect data in the last two decades. Ultra-high dimensional data are more frequently encountered in a large variety of areas such as finance, biomedical sciences, geological studies and many more areas. Consequently, statistical data analysis methods have to deal with large volume of data containing considerably many features; see Buhlmann and van de Geer[1], Hastie, Tibshirani and Friedman[23] and Fan and Lv[8] for overviews. A fundamental objective of statistical analysis on those ultra-high dimensional data is to identify relevant features, so that effective models can be subsequently constructed and applied to solve practical problems.

Recently, independence feature screening methods have been considered to be efficient for Ultra-high dimensional data. The sure independent screening (SIS) ranks the importance of individual variable by its correlation with the response and reduces the dimensionality form high to a moderate scale that is below the sample size. In fairly general framework, the procedure has been shown to have the sure screening property, meaning that all the important variables to be selected after variable screening with probability tending to 1[8]. The most common used

Manuscript received June 28, 2018. Accepted on October 13, 2018.
This paper is supported by the National Natural Science Foundation of China (Nos. 11171112, 11201190, 11101158) and Doctoral Fund of Ministry of Education of China (20130076110004) and the 111 Project of China (B14019).
†Corresponding author.

correlation is the Pearson correlation. It is most efficient in normal distributions, and easy to compute. However, it is sensitive to outliers and inefficient in heavy-tailed distributions. Also the Pearson correlation can not discover nonlinear relationships. To overcome those disadvantages, Li, et al.[13] utilized the Kendall tau correlation and proposed robust rank correlation screening (RRCS). The Kendall tau correlation is invariant under monotonic transformation and robust against heavy tailed distributions. The sure screening property of RRCS has been established. However, the computation of RRCS is heavy with the computation complexity $O(n^2p)$, where n is the sample size and p is the dimension. This may hinder its applications on huge data sets.

In this paper, we propose to use the Gini correlation in the screening procedure (GCS). Gini correlation is based on the covariance between one variable and the rank of the other. Depending on which variable taken in its variate values and which in ranks, there are two versions of Gini correlations. If the response is taken in its ranks, the Gini correlation screening method can be applied to transformation regression models that establish nonlinear relationship between the response and predictors. If a predictor variable is taken in its ranks, then the resulting GCRS is robust against x-outlyingness, which might be more serious than y-outlyingnesss in the regression setting. Comparing two sets of important variables selected from two Gini correlations, we are able to gain insights and interpretations from the degree of overlapping. Other benefits include high efficiency in the normal case and good efficiency in heavy-tailed distributions. Also importantly, its computation is manageable with the complexity $O(pn \log n)$. Using U-statistics theorem, we are able to establish the sure screening properties and model selection consistency under weaker assumptions than SIS.

Other independent feature screening methods include Fan and Song[6] for generalized linear models, Fan, Feng and Song[9] for nonparametric additive models. Fan and Song[6] also discussed independence screening by examining the magnitudes of the likelihood ratios. Wang[27] considered a sure independence screening by a factor profiling approach; see also Zhu et al.[28], Li et al.[13], Zhang et al.[29][30] for recent development using model-free approaches for feature screening.

The remainder of the paper is organized as follows. Section 2 introduces two Gini correlations and proposes Gini screening procedure (GCS). The sure screening property of GCS is explored. Section 3 presents iterative Gini screening method (IGCS) to overcome potential issues of GCS that only uses the marginal information. Section 4 provides some empirical evidences for the proposed procedures and all theoretical proofs are provided in Section 5.

2 Gini Screening Procedure

We first introduce Gini correlations and Gini regression, and then propose the screening procedure based on Gini correlations.

2.1 Two Gini Correlations and Their Relationship with the Pearson Correlation

Let X and Y be two non-degenerate random variables with marginal distribution functions F and G, respectively, and a joint distribution function H. To describe dependence correlation between X and Y, the Pearson correlation (denoted as ρ_p) is probably the most frequently used measure. This measure is based on the covariance between two variables, which is optimal for the linear association between bivariate normal variables. However, the Pearson correlation performs poorly for variables with heavily-tailed or asymmetric distributions, and may be seriously impacted even by a single outlier (e.g., Shevlyakov and Smirnov[21]). As a robust alternative, the Spearman correlation is a multiple (twive) of the covariance between the cumulative functions (or ranks) of two variables. The Gini correlation complements the Pearson and Spearman

correlations. It is based on the covariance between one variable and the cumulative distribution of the otherz[1]. Due to different roles of X and Y, two Gini correlations are defined as

$$\gamma(X,Y) := \frac{\mathrm{cov}(X,G(Y))}{\mathrm{cov}(X,F(X))}, \qquad \gamma(Y,X) := \frac{\mathrm{cov}(Y,F(X))}{\mathrm{cov}(Y,G(Y))}.$$

The representation of Gini correlation $\gamma(X,Y)$ indicates that it has mixed properties of those of the Pearson and Spearman correlations. It is similar to Pearson in X (the variable taken in its variate values) and similar to Spearman in Y (the variable taken in its ranks).

Two Gini correlations are equal if X and Y are exchangeable up to a linear transformation. That is, there exist a, b, c, and d $(a, c > 0)$ such that $(X,Y)^T$ and $(aY+b, cX+d)^T$ are equally distributed. Particularly, if $(X,Y)^T$ are elliptically distributed with linear correlation parameter ρ, then X and Y are exchangeable up to a linear transformation, hence $\gamma(X,Y) = \gamma(Y,X)$ and they are equal to ρ, the linear correlation parameter. And in the same time, they are equal to the Pearson correlation if the second moment of distribution exists. Note that Gini correlations exist only require the finiteness assumption of the first moment comparing with the one on the second moment for the Pearson correlation. Hence the Gini correlations are more suitable for heavy-tailed distributions than Pearson correlation. More details are referred to Schechtman and Yitzhaki[17, 18], Yitzhaki and Schechtman[19].

We present properties of $\gamma(X,Y)$, and results on $\gamma(Y,X)$ can be obtained similarly. $\gamma(X,Y)$ can be written in the form as below

$$\gamma(X,Y) = \frac{\mathrm{Cov}(X,G(Y))}{\mathrm{Cov}(X,F(X))} = \frac{\mathbb{E}h_1((X_1,Y_1),(X_2,Y_2))}{\mathbb{E}h_2((X_1,Y_1),(X_2,Y_2))}, \tag{2.1}$$

where $(X_1,Y_1)^T$ and $(X_2,Y_2)^T$ are independent copies of $(X,Y)^T$, $h_1((x_1,y_1),(x_2,y_2)) = \frac{1}{4}[(x_1-x_2)I(y_1 > y_2) + (x_2-x_1)I(y_2 > y_1)]$ and $h_2((x_1,y_1),(x_2,y_2)) = \frac{1}{4}|x_1-x_2|$. Note that $\mathrm{Cov}(X,F(X)) = \mathbb{E}|X_1 - X_2|/4$ is a quarter of the Gini mean difference of X.

Then given a sample data $(x_i,y_i)^T$, $i = 1, \cdots, n$, the sample Gini correlation is

$$\widetilde{\gamma}(X,Y) = \frac{U_1}{U_2} = \frac{\binom{n}{2}^{-1} \sum\limits_{1\le i<k\le n} [(x_i-x_k)I(y_i > y_k) + (x_k-x_i)I(y_k > y_i)]/4}{\binom{n}{2}^{-1} \sum\limits_{1\le i<j\le n} |x_i - x_k|/4}. \tag{2.2}$$

It is a ratio of two U-statistics with the kernel of U_1 being h_1 and the kernel of U_2 being h_2. The asymptotic normality of $\widetilde{\gamma}(X,Y)$ in (2.2) can be easily established through U-statistics theory. For computation, however, more efficient algorithm is based on

$$\widetilde{\gamma}(X,Y) = \frac{\sum\limits_{i=1}^{n}(2i-1-n)x_{y_{(i)}}}{\sum\limits_{i=1}^{n}(2i-1-n)x_{(i)}}, \tag{2.3}$$

where $x_{y_{(i)}}$ is the x value that belongs to $y_{(i)}$ with $x_{(i)}$ and $y_{(i)}$ being the i^{th} order statistic of x sample and y sample, respectively. This makes its computation complexity as $O(n \log n)$.

2.2 Gini Regression

Gini correlation has been widely used for simple and multiple regression settings. See Yitzhaki and Schechtman[19] for details. Here we look at the simple Gini regression problem that motivates us to use Gini correlation to rank the importance of covariates. Assume that the following model is given:

$$Y = \alpha + \beta X + \varepsilon. \tag{2.4}$$

Taking covariance with respective to X in both sides of (2.4), we obtain

$$\text{cov}(Y, X) = \text{cov}(\alpha + \beta X + \varepsilon, X) = \beta \text{cov}(X, X) + \text{cov}(\varepsilon, X).$$

Imposing the orthogonality restriction $\text{cov}(\varepsilon, X) = 0$, we have the ordinary least square coefficient,

$$\beta = \frac{\text{cov}(Y, X)}{\text{cov}(X, X)} = \rho_p \sqrt{\frac{\text{cov}(Y, Y)}{\text{cov}(X, X)}} = \frac{\sigma_Y}{\sigma_X} \rho_p,$$

where ρ_p is the Pearson correlation between X and Y and σ_Y and σ_X be the standard deviation of Y and X, respectively. If $\sigma_Y = \sigma_X = 1$, $\beta = \text{cov}(Y, X) = \rho_p$.

Taking covariance with respective to $F(X)$ and along with the condition $\text{cov}(\varepsilon, F(X)) = 0$, we obtain

$$\beta = \frac{\text{cov}(Y, F(X))}{\text{cov}(X, F(X))} = \gamma(Y, X) \frac{\text{cov}(Y, G(Y)}{\text{cov}(X, F(X)} = \frac{\sigma_{gY}}{\sigma_{gX}} \gamma(Y, X),$$

where σ_{gY} and σ_{gX} are Gini mean difference of Y and X, respectively. When $\sigma_{gY} = \sigma_{gX} = 1$, $\beta = \text{cov}(Y, F(X)) = \gamma(Y, X)$.

Unusually $\mathbb{E}(\varepsilon) = 0$ is assumed. The traditional assumption of independence X and ε implies both $\text{cov}(\varepsilon, X) = 0$ and $\text{cov}(\varepsilon, F(X)) = 0$. Note that weaker than the independence assumption, $\mathbb{E}(\varepsilon|X) = 0$ also imply both $\text{cov}(\varepsilon, X) = 0$ and $\text{cov}(\varepsilon, F(X)) = 0$. If $\text{cov}(\varepsilon, X) = 0$ and $(\varepsilon, X)^T$ follows an elliptical distribution, then $\mathbb{E}(\varepsilon|X) = 0$ and $\text{cov}(\varepsilon, F(X)) = 0$. Hence, under some assumptions, the regression coefficient can be obtained from either correlation multiplying with a scale ratio. That motivates us to extend the sure independent screening (SIS) which is based on the Pearson correlation to the Gini SIS (GIS).

2.3 Gini Correlation Screening

Let Y is the response, $X = (X_1, \cdots, X_p)^T$ is a covariate vector. Four models are considered. For $i = 1, 2, \cdots, n$ and $j = 1, 2, \cdots, p$,

Model 1: $Y_i = \sum_{j=1}^{p} \beta_j X_{ij} + \varepsilon_i$;

Model 2: $Y_i = \sum_{j=1}^{p} \beta_j g_j(X_{ij}) + \varepsilon_i$;

Model 3: $h(Y_i) = \sum_{j=1}^{p} \beta_j X_{ij} + \varepsilon_i$;

Model 4: $F_y(y|x) = F_0(y|\beta_1 X_1, \cdots, \beta_p X_p)$,

where h and g_j are unspecified strictly monotone functions. Model 1 is the regular linear model. Model 3 is commonly used to stabilize the variance and to normalize error distribution. With different forms of h, this model generates many different parametric families of models. For instance, if h takes the form of a power function, Model 3 is the family of Box-Cox transformation models. If h is taken to be the logarithm function and the response is the time to event of interest. Model 3 includes many survival models with different choices of error distributions. Model 2 is the additive model which has been extensively studied in a nonparametric approach without specifying the function forms of g_j's. Model 4 includes the other three models as special cases. It is considered by Zhu, et al.[13].

We define the notion of active predictors in the above models as follows. Let the two index sets be

$$\mathcal{A} = \{1 \le j \le p : \beta_j \ne 0\},$$

$$\mathcal{I} = \{1 \le j \le p : \beta_j = 0\}.$$

If $j \in \mathcal{A}$, X_j is referred to as an active predictor, whereas if $j \in \mathcal{I}$, X_j is referred to as an inactive predictor. For each of above models, under some assumptions (listed later), we can rewrite the active variable set as

$$\mathcal{A} = \{1 \le j \le p : \gamma(Y, X_j) \ne 0\} \text{ or } \mathcal{A} = \{1 \le j \le p : \gamma(X_j, Y) \ne 0\},$$

which are equivalent to

$$\{1 \le j \le p : u_j = \text{cov}(Y, F_j(X_j)) \ne 0\} \text{ or } \{1 \le j \le p : v_j = \text{cov}(X_j, G(Y)) \ne 0\},$$

respectively, where F_j is the distribution of X_j.

Lemma 2.1. *For random variable X_j and Y, $\text{cov}(X_j, Y) = 0$ if and only if $\text{cov}\big(X_j, G(Y)\big) = \text{cov}\big(Y, F_j(X_j)\big) = 0$.*

This implies that u_j may be used for feature screening in this setting.

With sample data $\{\mathbf{x}_i, y_i\}$, $i = 1, \cdots, n$ available, u_j and v_j are estimated by

$$\widetilde{u}_j = \binom{n}{2}^{-1} \sum_{1 \le i < k \le n} \frac{1}{4}[(y_i - y_k)I(x_{ij} > x_{kj}) + (y_k - y_i)I(x_{kj} > x_{ij})],$$

$$\widetilde{v}_j = \binom{n}{2}^{-1} \sum_{1 \le i < k \le n} \frac{1}{4}[(x_{ij} - x_{kj})I(y_i > y_k) + (x_{kj} - x_{ij})I(y_k > y_i)].$$

To select active variables, we propose to rank all the candidate predictors X_j, $j = 1, \cdots, p$, according to $|\widetilde{u}_j|$ or $|\widetilde{v}_j|$ from the largest to smallest. Later we establish a thresholding rule for obtaining the cutoff value a_n, b_n that separates the active and inactive predictors. The selected variable sets are following

$$\widetilde{\mathcal{A}}_{a_n} = \{1 \le j \le p : |\widetilde{u}_j| > a_n\},$$

$$\widetilde{\mathcal{A}}_{b_n} = \{1 \le j \le p : |\widetilde{v}_j| > b_n\},$$

where a_n or b_n are predefined threshold values.

2.4 Theoretical Properties

In this section we consider the screening method by Gini correlation $\gamma(Y, X)$ and the theoretical properties of u_j in detail, not Gini correlation $\gamma(X, Y)$ and v_j which have similar properties by Lemma 2.1. As fitting marginal regressions to a joint regression is a type of model misspecification. The following theorem reveals that under some conditions, the marginal model pertains to the information about the important variables in the joint model.

To ensure the feature screening property, the following conditions are necessary.

C1. $\max_{j \in \mathcal{I}} |\text{cov}(Y, F_j(X_j))| \le \min_{j \in \mathcal{A}} |\text{cov}(Y, F_j(X_j))|$.

C2. There exists a positive constant t_0 such that

$$\max_{1 \le j \le p} E\{\exp(tX_j)\} < \infty \text{ for } 0 < t \le t_0.$$

C3. $\max_{j \in \mathcal{A}} |\text{cov}(Y, F_j(X_j))| > Cn^{-k}$ for some positive constant C and $0 < \kappa < 1/2$.

We next present the main theoretical results on feature ranking in terms of the utility measure $\widetilde{u}_j$.

Theorem 2.2. *Under the conditions (C1)–(C2), we assume that* $p = o\{\exp(bn)\}$ *for any fixed* $b > 0$, *then, for any* $\eta > 0$, *there exists a sufficiently small constant* $s_\eta \in (0, 2/\eta)$, *such that*

$$P(\max_{1\le j\le p} |\widetilde{u}_j - u_j| > \eta) \le 2p\exp\{n\log(1 - \eta\log(1 - \eta s_\eta/2))/3\}.$$

In addition, if we write $\nu = \max_{j\in\mathcal{I}} |\mathrm{cov}(Y, F_j(X_j))| \le \min_{j\in\mathcal{A}} |\mathrm{cov}(Y, F_j(X_j))|$, *then there exists a sufficiently small constant* $s_\nu \in (0, 4/\nu)$ *such that*

$$P\big(\max_{j\in\mathcal{I}} |\widetilde{u}_j| < \min_{j\in\mathcal{A}} |\widetilde{u}_j|\big) \ge 1 - 4p\exp\{n\log(1 - s_\nu/4)/3\}.$$

Theorem 2.3. *Under the condition of (C1)–(C3), then*

$$P(\widetilde{\mathcal{A}} \subset \mathcal{A}) \ge 1 - 4s_n p\exp\{n\log(1 - s_\nu/4)/3\},$$

where s_n *is the number of the true model.*

Note in Theorem 2.3, we can handle the NP-dimensionality

$$p = o(\exp(an)), \qquad \text{for} \quad a > 0.$$

3 Iterative Screening

We screen variables with GCS method to get a smaller submodel which the dimension is smaller than the sample size with a probability tending to one. The procedure which uses the marginal utility measure may miss those predictors which are marginally unrelated but jointly related to the response. To overcome this problem, we develop an iterative version of our proposed screening method(IGCS). The iterative procedure are as follows:

Step 1. Apply our screening method, we first compute $\widetilde{u}_j$ for sample $\{Y, X_j\}$ $(j = 1, \cdots, p)$, then select p_1 predictors where p_1 is usually chosen to be smaller than $[n/\log n]$, a.e. if $\widetilde{u}_j > \widetilde{u}_{p_1}$, X_j is selected. We denote the set of indices of the selected predictors by $\mathcal{A}_1$ and the associated $n \times p_1$ data matrix by $X_{\mathcal{A}_1}$, then continue selecting a subset of m_1 predictors $M_1 = \{X_{i_1}, \cdots, X_{i_{m_1}}\}$ by a model selection method such as the nonconcave penalized M-estimation proposed by Li, Peng and Zhu[26] for model (2.4).

Step 2. Let I_1 denote the complement of $X_{\mathcal{M}_1}$, X_{I_1} denote the remaining $n \times (p - m_1)$ data matrix and $Y^* = Y - X_{\mathcal{M}_1}\widetilde{\beta}_{M_1}$. Next, for the data $\{Y^*, X_{\mathcal{M}_1^c}\}$, select a subset M_2 by the method of Step 1.

Step 3. Repeat Step 2 k times, until the total selected number of predictors $m_1 + \cdots + m_k$ exceeds the pre-specified number $[n/\log n]$. The final selected predictor set is $\mathcal{M}_1 \cup \cdots \cup \mathcal{M}_k$.

4 Simulation

We compare the performance of the eight methods: SIS, ISIS[8], RRCS, IRRCS[13] and the proposed methods GCS, IGCS in Section 3 by computing the proportions of selected models containing all the variables in the true model. In each example and real data analysis, the size of variables selected are $n/(\log n)$.

4.1 Examples

Example 1. Consider Example 1 of Li et al.[13] with the linear model:

$$Y = 5X_1 + 5X_2 + 5X_3 + \varepsilon,$$

where $\mathbf{X} = (X_1, \cdots, X_p)^T$ is a p-dimensional predictor and ε is the noise that is independent of the predictors. In the simulation, ε is from three different distributions: the standard normal $\mathcal{N}(0,1)$, the standard normal with 10% of contamination from the Cauchy distribution, and the $\mathcal{T}(3)$ distribution. A sample of $\mathbf{X}$ with size n was generated from a multivariate normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ with the covariance matrix $\mathbf{\Sigma}$ having entries $\sigma_{ii} = 1,\ i = 1, \cdots, p$ and $\sigma_{ij} = \rho,\ i \neq j$. For the combinations with $p = 100, 1000,\ n = 20, 50, 70$ and $\rho = 0, 0.1, 0.5, 0.9$, the experiment is repeated 200 times.

Example 2. Consider Example III of Fan and Lv[8] with the underlying model:

$$Y = 5X_1 + 5X_2 + 5X_3 - 15\sqrt{\rho}X_4 + X_5 + \varepsilon,$$

where $X_4 \sim \mathcal{N}(0,1)$ has correlation coefficient $\sqrt{\rho}$ with all other $p-1$ predictors, X_5 is uncorrelated with all the other $p-1$ variables, $X_1, X_2, X_3, X_6, \cdots, X_p$ and noise ε are distributed identical to those in Example 1 above.

Table 1. The Proportion of Predictors Containing the True Model Selected by Each Screening Method (Example1)

		$\varepsilon \sim \mathcal{N}(0,1)$				$N(0,1)$ with %10 outliers				$\varepsilon \sim \mathcal{T}(3)$			
(p,n)	Method	$\rho = 0$	0.1	0.5	0.9	0	0.1	0.5	0.9	0	0.1	0.5	0.9
(100,20)	SIS	0.713	0.783	0.617	0.584	0.693	0.761	0.605	0.433	0.537	0.565	0.522	0.233
	RRCS	0.643	0.590	0.530	0.360	0.710	0.755	0.643	0.510	0.790	0.753	0.685	0.495
	GCS	0.620	0.563	0.535	0.362	0.704	0.757	0.650	0.623	0.525	0.795	0.635	0.691
	ISIS	1	1	0.968	0.968	0.852	0.845	0.850	0.840	0.885	0.910	0.864	0.835
	IRRCS	0.818	0.895	0.790	0.895	0.930	0.953	0.928	0.905	0.931	1	0.944	0.920
	IGCS	0.820	0.882	0.793	0.893	0.935	0.951	0.921	0.915	0.926	0.991	0.946	0.912
(100,50)	SIS	1	1	1	1	0.950	0.940	0.943	0.935	0.965	0.0.970	0.960	0.910
	RRCS	1	1	1	0.975	0.970	0.955	0.944	0.940	1	0.990	0.980	0.950
	GCS	1	1	1	0.970	0.975	0.960	0.949	0.940	1	0.993	0.975	0.955
	ISIS	1	1	1	1	0.975	0.970	0.970	0.935	1	1	0.970	0.948
	IRRCS	1	1	1	0.970	1	1	1	1	1	1	1	0.990
	IGCS	1	1	1	0.975	1	1	1	1	1	1	1	0.985
(1000,50)	SIS	1	0.985	0.935	0.835	0.950	0.985	0.845	0.655	0.985	0.985	0.810	0.620
	RRCS	0.990	0.970	0.825	0.570	0.945	0.990	0.755	0.555	1	0.990	0.930	0.750
	GCS	0.980	0.980	0.830	0.600	0.945	0.980	0.760	0.560	1	0.990	0.940	0.755
	ISIS	1	1	1	0.995	0.955	0.990	0.940	0.850	1	0.990	0.935	0.850
	IRRCS	1	1	0.990	0.995	0.980	0.995	0.950	0.865	1	1	1	0.985
	IGCS	1	1	0.995	1	0.985	1	0.955	0.870	1	1	1	0.970
(1000,70)	SIS	1	1	0.985	0.965	0.960	0.950	0.925	0.875	1	0.990	0.950	0.850
	RRCS	1	1	0.990	0.870	0.945	0.990	0.965	0.835	1	1	0.980	0.860
	GCS	1	1	0.990	0.905	0.950	0.995	0.960	0.850	1	1	0.985	0.870
	ISIS	1	1	1	1	0.970	0.960	0.950	0.940	1	1	0.980	0.960
	IRRCS	1	1	1	1	1	1	0.975	0.965	1	1	1	1
	IGCS	1	1	1	1	1	1	0.980	0.970	1	1	1	1

The proportions of predictors selected by each screening method containing the first 3 predictors X_1, X_2, X_3 are reported in Table 1.

We take $\rho = 0.5$ for simplicity. We generate 200 data sets for this model and report in Table 2.

Example 3. Consider the following generalized Box-cox transformation model:

$$H(Y_i) = X_i\beta + \varepsilon_i, \tag{4.1}$$

where the transformation functions are unknown. In the simulations, we consider the following forms:

- Box-Cox transformation, $\frac{|Y|^\lambda \mathrm{sgn}(Y)-1}{\lambda}$, where $\lambda = 0.25, 0.5, 0.75$;

- Logarithm transformation function, $H(Y) = \log Y$.

The linear regression model and the logarithm transformation model are special cases of the generalized Box-Cox transformation model with $\lambda = 1$ and $\lambda = 0$, respectively. Again, noise ε_i follows the distribution as $N(0,1)$, $\beta = (3, 1.5, 2, 0, \cdots, 0)^T$ and $\beta_k/\|\beta_k\| = (0.7682, 0.3841, 0.5121, 0, \cdots, 0)^T$ is a $p \times 1$ vector, and a sample of $(X_1, \cdots, X_p)$ with size n is generated from a multivariate normal distribution $N(0, \Sigma)$ whose covariance matrix $\Sigma = (\sigma_{ij})_{p\times p}$ has entries $\sigma_{ii} = 1, i = 1, \cdots, p$, and $\sigma_{ij} = \rho, i \neq j$. The replication time is again 200, and $p = 100, 1000, n = 20, 50, 70$ and $\rho = 0, 0.5, 0.9$, respectively.

Table 2. The Proportion of Predictors Containing the True Model Selected by Each Screening Method (Example2)

	$\varepsilon \sim$	$\mathcal{N}(0,1)$			$N(0,1)$ with 10% outliers			$\mathcal{T}(3)$		
p	Method	n=20	n=50	n=70	n=20	n=50	n=70	n=20	n=50	n=70
100	SIS	0	0.285	0.535	0	0.195	0.525	0	0.240	0.535
	RRCS	0	0.305	0.595	0	0.220	0.575	0	0.305	0.575
	GCS	0	0.300	0.560	0	0.225	0.580	0	0.310	0.580
	ISIS	0	0.465	0.855	0	0.415	0.805	0	0.405	0.775
	IRRCS	0	0.500	0.820	0	0.495	0.815	0	0.530	0.805
	IGCS	0	0.440	0.800	0	0.450	0.820	0	0.540	0.820
1000	SIS	0	0	0	0	0	0	0	0	0
	RRCS	0	0	0	0	0	0	0	0	0
	GCS	0	0	0	0	0	0	0	0	0
	ISIS	0	0.045	0.090	0	0.015	0.035	0	0	0.020
	IRRCS	0	0.035	0.085	0	0.030	0.055	0	0.030	0.020
	IGCS	0	0.030	0.080	0	0.035	0.060	0	0.040	0.030

Based on the simulation results as summarized in Tables 1–3, we can see that GCS procedure works well in screening out insignificant predictors comparing with SIS and RRCS for large covariate dimension.

4.2 Cardiomyopathy Microarray Data

In this subsection, we apply the GCS procedure to the cardiomyopathy microarray data and compare it with the SIS procedure[8]and RRCS. The cardiomyopathy microarray data are from a transgenic mouse model of dilated cardiomyopathy[15], which have been used by Hall and Miller[10]. This dataset consists of a $n\times p$ matrix of gene expression values $X = \{X_{ij}\}$, where X_{ij} is the expression level of the j-th gene ($j = 1, \cdots, p = 6319$) for the i-th mouse ($i = 1, \cdots, n = 30$). Each mouse also provides an outcome (Ro1) measure Y_i. The prediction dimension is high and the sample size is relatively small. Our aim is to reduce the dimensionality and determine which genes were influential for overexpression of a G protein-coupled receptor, designated Ro1. We first apply GCS, RRCS[13] and SIS[8] procedures to reduce the dimensionality from $p = 6319$ to a dimension $d_n < n$ so we can find the respective submodels to include the important genes. Training set of 15 observations and a test set of 15 observations is randomly partitioned to compute MS and PE respectively by repeating 100 times. Table 4 presents the average values and their associated robust standard deviations over 100 replications. We give the results of the simulation in Table 4. From Table 4 average model size (MS) and prediction error (PE) have little difference by using this three methods and our proposed methods is a competitive variable selection method comparing with RRCS and SIS, which could be very useful in subsequent studies.

Table 3. The proportion of predictors containing the true model selected by each screening method (Example3)

		$\varepsilon \sim$	$\mathcal{N}(0,1)$			$N(0,1)$ with 10% outliers			$\mathcal{T}(3)$		
(p,n)	ρ	Method	$\lambda = 0.25$	$\lambda = 0.50$	$\lambda = 0.75$	$\lambda = 0.25$	$\lambda = 0.50$	$\lambda = 0.75$	$\lambda = 0.25$	$\lambda = 0.50$	$\lambda = 0.75$
(100,20)	0	SIS	0.154	0.320	0.415	0.145	0.265	0.380	0.190	0.360	0.420
		RRCS	0.435	0.435	0.440	0.425	0.450	0.430	0.560	0.590	0.525
		GCS	0.440	0.440	0.445	0.425	0.465	0.435	0.565	0.595	0.530
		IGCS	0.970	0.985	0.985	0.910	0.900	0.930	0.915	0.940	0.965
	0.5	SIS	0.090	0.155	0.190	0.085	0.160	0.190	0.175	0.325	0.355
		RRCS	0.345	0.400	0.400	0.365	0.390	0.370	0.385	0.355	0.450
		GCS	0.440	0.440	0.445	0.365	0.395	0.375	0.390	0.360	0.455
		IGCS	0.965	0.950	0.975	0.875	0.885	0.880	0.890	0.920	0.930
	0.9	SIS	0.0025	0.005	0.030	0.0015	0.005	0.005	0.005	0.090	0.200
		RRCS	0.225	0.225	0.225	0.220	0.195	0.220	0.185	0.285	0.220
		GCS	0.230	0.230	0.230	0.225	0.200	0.225	0.190	0.290	0.225
		IGCS	0.890	0.865	0.855	0.685	0.750	0.760	0.790	0.820	0.845
(100,50)	0	SIS	0.815	0.935	0.935	0.680	0.795	0.875	0.760	0.855	0.890
		RRCS	0.965	0.965	0.965	0.955	0.950	0.965	0.900	0.955	0.960
		GCS	0.970	0.970	0.970	0.960	0.955	0.965	0.900	0.955	0.960
		IGCS	1	1	1	1	1	1	1	1	1
	0.5	SIS	0.680	0.810	0.855	0.585	0.740	0.795	0.720	0.730	0.850
		RRCS	0.955	0.950	0.955	0.955	0.950	0.945	0.945	0.930	0.910
		GCS	0.960	0.955	0.955	0.955	0.950	0.950	0.945	0.930	0.915
		IGCS	1	1	1	0.980	0.960	0.970	0.980	0.960	0.965
	0.9	SIS	0.370	0.380	0.850	0.260	0.355	0.385	0.305	0.390	0.415
		RRCS	0.865	0.845	0.875	0.885	0.880	0.870	0.900	0.890	0.890
		GCS	0.905	0.895	0.890	0.880	0.880	0.870	0.870	0.845	0.870
		IGCS	0.975	0.980	0.980	0.920	0.920	0.925	0.915	0.935	0.915
(1000,50)	0	SIS	0.200	0.490	0.615	0.160	0.370	0.490	0.155	0.455	0.530
		RRCS	0.755	0.760	0.750	0.665	0.655	0.650	0.755	0.745	0.710
		GCS	0.760	0.765	0.750	0.670	0.665	0.655	0.760	0.750	0.710
		IGCS	1	1	1	0.950	0.955	0.945	0.960	0.980	0.930
	0.5	SIS	0.035	0.110	0.145	0.020	0.080	0.130	0.055	0.150	0.130
		RRCS	0.470	0.465	0.485	0.440	0.440	0.430	0.375	0.430	0.435
		GCS	0.475	0.450	0.485	0.445	0.445	0.435	0.380	0.435	0.440
		IGCS	1	1	1	0.950	0.935	0.945	0.940	0.945	0.940
	0.9	SIS	0	0	0	0	0	0	0	0	0.005
		RRCS	0.240	0.815	0.230	0.215	0.215	0.215	0.215	0.170	0.180
		GCS	0.245	0.820	0.235	0.220	0.215	0.215	0.220	0.175	0.185
		IGCS	0.785	0.820	0.850	0.730	0.770	0.785	0.730	0.750	0.715
(1000,70)	0	SIS	0.435	0.775	0.860	0.365	0.550	0.670	0.440	0.760	0.840
		RRCS	0.875	0.885	0.880	0.830	0.865	0.880	0.835	0.915	0.915
		GCS	0.875	0.890	0.875	0.835	0.870	0.885	0.845	0.920	0.920
		IGCS	1	1	1	0.965	0.960	0.965	0.960	0.960	0.970
	0.5	SIS	0.010	0.275	0.375	0.075	0.230	0.270	0.010	0.280	0.370
		RRCS	0.725	0.715	0.725	0.710	0.670	0.695	0.655	0.610	0.700
		GCS	0.730	0.720	0.730	0.715	0.670	0.705	0.660	0.615	0.705
		IGCS	1	1	1	0.940	0.940	0.940	0.930	0.920	0.935
	0.9	SIS	0	0.0015	0.005	0	0	0.015	0	0.0015	0.105
		RRCS	0.490	0.470	0.515	0.500	0.515	0.510	0.410	0.440	0.395
		GCS	0.495	0.475	0.515	0.505	0.520	0.515	0.415	0.445	0.400
		IGCS	0.920	0.955	0.975	0.905	0.905	0.915	0.890	0.880	0.920

Table 4. Average Model Size (MS) and Prediction Error (PE)

screening method	MS	PE
SIS	5.5	0.50
RRCS	3.5	0.53
GCS	4.7	0.48

5 Proof of Theorem

In this section we give the proof of the Lemma and Theorems.

Proof of Lemma 2.1. We firstly prove the equality $\text{cov}(X_j, Y) = 0$ if and only if $\text{cov}(Y, F_j(X_j)) = 0$. It suffice to prove that for random variables Y and X_j a real-valued random variable and X_j with distribution function $F_j(X_j)$, we have

$$E(Y|X_j) = E(Y|F_j(X_j)), \qquad \text{a.s.} \tag{5.1}$$

For random variable X_j with distribution function $F_j(X_j)$ we have $P(T(F_j(X_j)) \neq X_j) = 0$

where $T(t) = \{h : F_j(h) \geq t\}$, for any $0 < t < 1$. (See, for instance Proposition 3, Chapter 1 in Shorack and Wellner[16]) From this and the properties of the conditional expectations, for any bounded measurable function q we have

$$\begin{aligned} E(q(X_j)E(Y|X_j)) =& E(q(X_j)Y) = E(q(T(F_j(X_j)))Y) \\ =& E(q(T(F(X_j)))E(Y|F_j(X_j))) = E(q(X_j)E(Y|F_j(X_j))). \end{aligned}$$

Since $E(Y|F_j(X_j))$ is a measurable function of X_j, the almost sure uniqueness of the conditional expectation implies the equality in equation (5.1).

The smiler method can show $\text{cov}(X_j, Y) = 0$ if and only if $\text{cov}(X_j, G(Y)) = 0$. □

Proof of Theorem 1. Note that $\widetilde{u}_j$ is a standard U-statistic. With Markov's inequality, we can obtain that, for fixed positive constant s_0 and any $0 < t < s_0 k^*$, where $k^* = [n/3]$,

$$P(\widetilde{u}_j - u_j \geq \eta) \leq \exp\{-t\eta\} \exp\{-tu_j\} E[\exp\{t\widetilde{u}_j\}\}].$$

Through 5.1.6 in the book by Serfling (1980), the U-statistic $\widetilde{u}_j$ can be represented as an average of independent and identically distributed random variables; that is, $\widetilde{u}_j = (n!)^{-1} \sum_{n!} w(X_{1j}, Y_1; \cdots; X_{nj}, Y_n)$, where each $w(X_{1j}, Y_1; \cdots; X_{nj}, Y_n)$ is an average of $k^* = [n/3]$ independent and identically distributed random variables, and $\sum_{n!}$ denotes summation over $n!$ permutations $i_1, \cdots, i_n$ of $(1, \cdots, n)$. We denote that $\varphi_h(s) = E[\exp\{sh(X_{ij}, Y_i; X_{kj}, Y_k)\}]$ for $0 < s < s_0$. Since the exponential function is convex, it follows by Jensen's inequality that

$$\begin{aligned} E[\exp\{t\widetilde{u}_j\}] =& E\big[\exp\{t(n!)^{-1} \sum_{n!} w(X_{1j}, Y_1; \cdots; X_{nj}, Y_n)\}\big] \\ \leq& (n!)^{-1} \sum_{n!} E[\exp\{tw(X_{1j}, Y_1; \cdots; X_{nj}, Y_n)\}] \\ =& \varphi_h^{k^*}(t/k^*). \end{aligned}$$

Combining the above two results, we obtain that

$$\begin{aligned} P(\widetilde{u}_j - u_j \geq \eta) \leq& \exp\{-t\}[\exp\{-tu_j/k*\}\varphi_h(t/k^*)]^{k*} \\ =& [\exp\{-s\eta\} \exp\{-su_j\}\varphi_h(s)]^{k^*}, \end{aligned} \tag{5.2}$$

where $s = t/k^*$. Note that $E\{h(X_{ij}, Y_i; X_{kj}, Y_k)\} = u_j$, and with Taylor expansion, $\exp\{sZ\} = 1 + sZ + s^2M/2$ for any generic random variable Z, where $0 < M < Z^2 \exp\{s_1 M\}$, and s_1 is a constant between 0 and s. It follows that

$$\exp\{-su_j\}\varphi_h(s) \leq 1 + s^2[E\{h^4(X_{ij}, Y_i; X_{kj}, Y_k)\} \times E \exp\{2s_1(h - u_j)\}]^{1/2}/2.$$

From Condition (C2) it follows that there exists a constant C (independent of n and p) such that $\max_{1 \leq j \leq p} \exp\{-su_j\}\varphi_h(s) \leq 1 + Cs^2$; that is,

$$\max_{1 \leq j \leq p} \exp\{-su_j\}\varphi_h(s) = 1 + O(s^2).$$

Recall that $0 < s = t/k^* < s_0$. For a sufficiently small s, which can be achieved by selecting a sufficiently small t, we have that $\exp(-s\eta) = 1 - \eta s + O(s^2)$ and therefore,

$$\max_{1 \leq j \leq p} [\exp(-s\eta) \exp(-su_j)\varphi_h(s)] \leq 1 - s\varepsilon/2. \tag{5.3}$$

Combining the results (5.2) and (5.3), we show that, for any $\eta > 0$, there exist a sufficiently small s_η such that $\max_{1 \leq j \leq p} \{P(\widetilde{u}_j - u_j) \geq \eta\} \leq (1 - \eta s_\eta/2)^{n/3}$. Here we use the notation s_η to emphasize

s depending on η. Similarly, we can prove that $\max_{1\le j\le p}\{P(\widetilde{u}_j - u_j \le -\eta)\} \le (1-\eta s_\eta/2)^{n/3}$. Therefore,

$$P\big(\sup_{1\le j\le p}|\widetilde{u}_j - u_j| > \eta\big) \le 2p\exp\{n\log(1-\eta s_\eta/2)/3\}.$$

This completes the proof of the first part.

Now we show the second part of Theorem 1.

$$\begin{aligned}
&P\big(\min_{j\in\mathcal{A}}|\widetilde{u}_j| \le \max_{j\in\mathcal{I}}|\widetilde{u}_j|\big)\\
=&P\big(\min_{j\in\mathcal{A}}|\widetilde{u}_j| - \min_{j\in\mathcal{A}}|u_j| + \nu \le \max_{j\in\mathcal{I}}|\widetilde{u}_j| - \max_{j\in\mathcal{I}}|u_j|\big)\\
\le&P\big(\sup_{j\in\mathcal{A}}\big|\widetilde{u}_j - u_j\big| \ge \nu/2\big) + P\big(\sup_{j\in\mathcal{I}}\big|\widetilde{u}_j - u_j\big| \ge \nu/2\big).
\end{aligned}$$

By using the result of the first part with $\nu = 2\eta$, the second result holds. □

Proof of Theorem 2. In Theorem 1, let $\eta = 2Cn^{-\kappa}$. By Condition (C3), it follows that

$$\min_{j\in\mathcal{A}}\widetilde{u}_j \ge Cn^{-\kappa}.$$

Let the event

$$B_n = P(\max_{1\le j\le p}|\widetilde{u}_j - u_j| \le Cn^{-\kappa}).$$

Hence by choice of $v_n = Cn^{-\kappa}$, we have $\mathcal{A}\subset\widetilde{\mathcal{A}}$. The result now follows from a simple union bound

$$P(B_n^c) \le 2ps_n\exp\{n\log(1-\eta s_\eta/2)/3.$$

This completes the proof. □

References

[1] Bhlmann, P., van de Geer, S. Statistics for High-Dimensional Data Methods. Theory and Applications, Springer, Heidelberg, Dordrecht, London, New York, 2011

[2] Chen, J.H., Chen, Z.H. Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*, 95: 759–771 (2008)

[3] Fan, J., Gijbels, I. Local Polynomial Modeling and Its Applications. Chapman and Hall, New York, 1996

[4] Fan, J., Li, R. Variable Selection via Nonconcave Penalized Likelihood and it Oracle Properties. *Ann. Statist. Assoc.*, 96: 1348–1360 (2001)

[5] Fan, J., Ren, Y. Statistical analysis of DNA microarray data. *Em Clin. Cancer Res.*, 12: 4469–4473 (2006)

[6] Fan, J., Song, R. Sure independence screening in generalized linear models with NP-dimensionality. *Ann. Statist. Assoc.*, 38: 3567–3604 (2010)

[7] Fan, M., Ma, Y., Dai, W. Nonparametric Independence Screening in Sparse Ultra-High Dimensional Varying Coefficient Models. *Ann. Statist. Assoc.*, 109: 1270–1284 (2013)

[8] Fan, J., Lv, J. Sure independence screening for ultrahigh dimensional feature space (with discussion). *J. Roy. Statist. Soc.B.*, 70: 849–911 (2008)

[9] Fan, J., Feng, Y., Song, R. Nonparametric independence screening in sparse ultra-highdimensional additive models. *J. Am. Statist. Assoc.*, 106: 544–557 (2011)

[10] Hall, P., Miller, H. Using generalized correlation to effect variable selection in very high dimensional problems. *Journal of Computational and Graphical Statistics*, 18: 533–550 (2009)

[11] Hastie, T., Tibshirani, R. Generalized additive models, *Statistical Science*, 3: 297–318 (1986)

[12] Li, R., Liang, H. Variable Selection in Semiparametric Regression Model, *The Ann Statist.*, 36, 261–286 (1999)

[13] Li, G., Peng, H., Zhang, J., Zhu, L. Robust Rank Correlation Based Screening, *Ann. Statist.*, 40: 1846–1877 (2012)

[14] Luo, S., Chen, Z. Sequential Lasso Cum Ebic For Feature Selection With Ultra-High Dimensional Feature Space, *J. Am. Statist. Assoc.*, 109: 1229–1240 (2014)

[15] Redfern, C.H., Coward, P., Degtyarev, M.Y., Lee, E. K., Kwa, A.T., Hennighausen, L., Bujard, H., Fishman, G.I., Conklin, B.R. Conditional expression and signaling of a specifically designed Gi-coupled receptor in transgenic mice. *Nat. Biotechnol*, 17: 165–169 (1999)
[16] Shorack, G., Wellne, J. Empirical Processes with Applications to Statistics. Wiley, New York, 1986
[17] Schechtman, E., Yitzhaki, S. A measure of association based on Gini's mean difference, *Comm. Statist.*, 16(1): 207–231 (1987)
[18] Schechtman, E., Yitzhaki, S. On the proper bounds of the Gini correlation. *Econom. Lett.*, 63: 133–138 (1999)
[19] Schechtman, E., Yitzhaki, S. A Family of Correlation Coefficients Based on the Extended Gini Index. *J. Econ. Inequal.*, 12: 129–146 (2003)
[20] Schechtman, E., Yizhaki, S., Artsev, Y. The similarity between mean-variance and mean-Gini: Testing for equality of Gini correlations. *Advances in Investment Analysis and Portfolio Management (AIAPM)*, 3: 103–128 (2007)
[21] Shevlyakov, G.L., Smirnov, P.O. Robust Estimation of the Correlation Coefficient: an Attempt of Survey. *Austrian Journal of Statistics*, 40: 147–156 (2011)
[22] Storey, J.D., Tibshirani, R. Statistical significance for genome-wide studies. *Proc. Natn. Acad. Sci. USA*, 100: 9440–9445 (2003)
[23] Hastie, T., Tibshirani, R., Friedman, J. Elements of statistical learning: data mining. Inference and Prediction,. 2nd Edition, Springer, Berlin, 2009
[24] Tibshirani, R. Regression Shrinkage and Selection via LASSO. *Journal of the Royal Statistical Society, Series B*, 58: 267–288 (1996)
[25] Wang, H., Xia, Y. Shrinkage Estimation of the Varying Coefficient Model. *J. Am. Statis. Assoc.*, 104: 747–757 (2009)
[26] Li, G., Peng, H., Zhu, L. Nonconcave penalized M-estimation with a diverging number of parameters. *Statist. Sinica*, 21: 391–419 (2011)
[27] Wang, H. Factor profiled sure independence screening, *Biometrika*, 99: 15C-28 (2012)
[28] Zhu, L., Li, X., Li, Z., Zhu, X. Model-free feature screening for ultrahigh-demensional data. *J. Amer. Statist. Assoc.*, 106: 1464–1474 (2011)
[29] Zhang,J., Zhang, R., Lu, Z. Quantile-adaptive variable screening in ultra-high dimensional varying coefficient models. *Journal of Applied Statistics*, 43: 643–654 (2016)
[30] Zhang, J., Zhang, R.,Zhang, J. Feature Screening for Nonparametric and Semiparametric Models with Ultrahigh-dimensional Covariates. *J. Syst. Sci. Complex*, 31: 1350–1361 (2018)