SCIENCE CHINA
Mathematics

• ARTICLES •

January 2022 Vol. 65 No. 1: 153–168
https://doi.org/10.1007/s11425-019-1702-5

Stable correlation and robust feature screening

Xu Guo[1,*], Runze Li[2], Wanjun Liu[2] & Lixing Zhu[3,4]

[1]*School of Statistics, Beijing Normal University, Beijing* 100875*, China;*
[2]*Department of Statistics, Pennsylvania State University, University Park, PA* 16802-2111*, USA;*
[3]*Center for Statistics and Data Science, Beijing Normal University, Zhuhai* 519085*, China;*
[4]*Department of Mathematics, Hong Kong Baptist University, Hong Kong, China*

*Email: xustat*12*@bnu.edu.cn, rzli@psu.edu, wxl*204*@psu.edu, lzhu@hkbu.edu.hk*

Received November 27, 2019; accepted May 23, 2020; published online April 30, 2021

Abstract In this paper, we propose a new correlation, called stable correlation, to measure the dependence between two random vectors. The new correlation is well defined without the moment condition and is zero if and only if the two random vectors are independent. We also study its other theoretical properties. Based on the new correlation, we further propose a robust model-free feature screening procedure for ultrahigh dimensional data and establish its sure screening property and rank consistency property without imposing the subexponential or sub-Gaussian tail condition, which is commonly required in the literature of feature screening. We also examine the finite sample performance of the proposed robust feature screening procedure via Monte Carlo simulation studies and illustrate the proposed procedure by a real data example.

Keywords feature screening, nonlinear dependence, stable correlation, sure screening property

MSC(2020) 62H12, 62H20

Citation: Guo X, Li R Z, Liu W J, et al. Stable correlation and robust feature screening. Sci China Math, 2022, 65: 153–168, https://doi.org/10.1007/s11425-019-1702-5

1 Introduction

It is fundamental to characterize the dependence between two random vectors. Developing measures to characterize nonlinear dependence between two random vectors receives more and more attention in the recent literature. The classical correlation coefficients such as the Pearson correlation, Spearman [18]'s ρ and Kendall [8]'s τ cannot be used for measuring general dependence between two random vectors. The distance correlation (DC) [20] can be used to characterize the nonlinear dependence between two random vectors of different dimensions. The DC is well defined only when certain moment conditions are imposed on the random vectors. To remove the moment condition, Heller et al. [5] suggested using ranks of distances, but its practical implementation requires choosing several tuning parameters. Zhu et al. [24] proposed a projection correlation (PC) for any two random vectors, and they demonstrated that the DC may be less efficient than PC in detecting nonlinear dependence when the moment conditions are violated. Weihs et al. [21] developed a generalized framework for nonparametric measure of dependence. Spearman [15] proposed a generic measure of dependence in the Banach space.

* Corresponding author

In this paper, we first propose a new correlation for measuring dependence between two random vectors of arbitrary dimensions. The proposed correlation shares the same spirit of the DC, which is defined via a weighted L_2 distance between the joint characteristic function and the product of the marginal characteristic functions. By taking a different weight function from the one used in the DC, the new correlation can be well defined without moment conditions. We study properties of the new correlation, and show that the new correlation retains virtues of the DC, and can be used to measure nonlinear dependence between two random vectors with heavy-tailed distributions.

As an application of the proposed new correlation, we use it to develop a robust model-free feature screening procedure for ultrahigh dimensional data. Since the seminal work by Fan and Lv [2], a number of feature screening procedures have been developed under various model settings. For a recent review, please refer to [13]. However, only a few works on robust feature screening have been proposed in the literature. Li et al. [10] proposed the robust rank correlation screening (RRCS) based on Kendall τ. It is known that Kendall τ can detect monotone relationship between the predictors and the response, but cannot detect arbitrary possible dependence between the predictors and the response. In fact, two random variables can still be dependent even when their Kendall τ is zero. Moreover the RRCS cannot directly handle grouped predictors and multivariate responses. Zhong et al. [22] developed a robust feature screening procedure for the ultrahigh dimensional single index model. Their procedure is limited to the univariate response. The authors applied the distance correlation screening procedure [11] to the rank statistic of the response and the predictors. Thus, their procedure is robust in the direction of the response, but not the direction of predictors. Recently, Liu et al. [12] applied the PC to the ultrahigh dimensional feature screening problem. Although the PC does not require any moment condition, its computation can be expensive.

Compared with the robust screening procedures developed in [10,22], our proposed screening procedure allows the multivariate response and grouped predictor, and it is model-free and robust in both directions of responses and predictors. The proposed procedure does not need the subexponential tail condition, a common condition imposed in the works related to feature screening.

The rest of this paper is organized as follows. In Section 2, we propose a new correlation and study its theoretical properties. In Section 3, we propose a new robust feature screening procedure based on the new correlation. Section 4 presents numerical studies and a real data example to illustrate the proposed methodology. Conclusion is given in Section 5. All the proofs are presented in Appendix A.

2 A new correlation

Suppose that V and W are two random vectors with d_V and d_W dimensions. In this section, we propose a new correlation to characterize both linear and nonlinear dependence between V and W.

Let $\phi_{V,W}(t,s) = \mathrm{E}[\mathrm{e}^{\mathrm{i}(t^{\mathrm{T}}V+s^{\mathrm{T}}W)}]$ be the joint characteristic function (CF) of V and W, and $\phi_V(t)$ and $\phi_W(s)$ be the marginal CFs of V and W, respectively, where t^{T} denotes the transpose of t. To introduce the new correlation, define

$$Q(V,W \mid \omega) = \int_{\mathbb{R}^{d_V+d_W}} |\phi_{V,W}(t,s) - \phi_V(t)\phi_W(s)|^2 \omega(t,s)dtds, \tag{2.1}$$

where $\omega(t,s) \geqslant 0$ is a weight function. Clearly, $Q(V,W \mid \omega)$ is zero if V and W are independent. In general, it is hard to evaluate $Q(V,W \mid \omega)$ with an arbitrary $\omega(t,s)$. Fortunately, with a careful choice of the weight function $\omega(t,s)$, we can derive a closed form of $Q(V,W \mid \omega)$. Székely et al. [20] advocated

$$\omega(t,s) = (c_{d_V} c_{d_W} \|t\|^{1+d_V} \|s\|^{1+d_W})^{-1}$$

with $\|\cdot\|$ being the Euclidean norm and

$$c_d = \pi^{(1+d)/2}/\Gamma((1+d)/2),$$

where $\Gamma(\cdot)$ is the Γ function. With this weight function, $Q(V,W \mid \omega)$ has a closed form, and the authors further defined the distance covariance and distance correlation. In order to ensure that $Q(V,W \mid \omega)$ is

well defined and finite, moment conditions are required on V and W. In this paper, we propose to use another class of weight functions so that we may avoid the moment conditions imposed on V and W.

For a complex number z, let $\bar{z}$ be its conjugate and $\mathrm{Re}(z)$ be its real part. Then

$$\begin{aligned}&|\phi_{V,W}(t,s)-\phi_V(t)\phi_W(s)|^2\\&\quad=|\phi_{V,W}(t,s)|^2+|\phi_V(t)\phi_W(s)|^2-2\mathrm{Re}(\phi_{V,W}(t,s)\overline{\phi_V(t)\phi_W(s)}).\end{aligned}$$

Further notice that

$$\begin{aligned}&|\phi_{V,W}(t,s)|^2=\mathrm{E}[\cos(t^{\mathrm{T}}(V-V_1)+s^{\mathrm{T}}(W-W_1))],\\&|\phi_V(t)\phi_W(s)|^2=\mathrm{E}[\cos(t^{\mathrm{T}}(V-V_1))]\mathrm{E}[\cos(s^{\mathrm{T}}(W-W_1))],\\&\mathrm{Re}(\phi_{V,W}(t,s)\overline{\phi_V(t)\phi_W(s)})=\mathrm{E}[\cos(t^{\mathrm{T}}(V-V_1)+s^{\mathrm{T}}(W-W_2))].\end{aligned}$$

Here, (V_k,W_k), $k=1,2$ are independent copies of (V,W).

From [14], the CF of a spherical stable law is given by

$$\phi_Z(t)=\int_{\mathbb{R}^q}\cos(t^{\mathrm{T}}z)f_{a,q}(z)dz=\mathrm{e}^{-\|t\|^a},$$

where $f_{a,q}(\cdot)$ denotes the density of a spherical stable law in $\mathbb{R}^q$ with characteristic exponent $a\in(0,2]$. The density function takes the following form:

$$f_{a,q}(z)=c_2\int_0^1 g_{a,q}(\|z\|u)(1-u^2)^{(q-3)/2}du,$$

where $g_{a,q}(u)=\int_0^\infty\cos(ur)r^{q-1}\mathrm{e}^{-r^a}dr$, and $c_2=2c_1(2\pi)^{-q}$ with $c_1=2\pi^{(q-1)/2}/\Gamma((q-1)/2)$. The spherical stable family includes the multivariate Gaussian and Cauchy distributions as special cases with $a=2$ and $a=1$, respectively.

In this paper, we choose

$$\omega(t,s)=f_{a,d_V}(t)f_{a,d_W}(s).$$

Then it follows that

$$\begin{aligned}Scov^2(V,W)&=\int_{\mathbb{R}^{d_V+d_W}}|\phi_{V,W}(t,s)-\phi_V(t)\phi_W(s)|^2f_{a,d_V}(t)f_{a,d_W}(s)dtds\\&=:E_1+E_2-2E_3\end{aligned}\tag{2.2}$$

with $E_j, j=1,2,3$ being defined as

$$\begin{aligned}&E_1=\mathrm{E}[\mathrm{e}^{-\|V-V_1\|^a-\|W-W_1\|^a}],\\&E_2=\mathrm{E}[\mathrm{e}^{-\|V-V_1\|^a}]\mathrm{E}[\mathrm{e}^{-\|W-W_1\|^a}],\\&E_3=\mathrm{E}[\mathrm{e}^{-\|V-V_1\|^a-\|W-W_2\|^a}].\end{aligned}$$

Since the spherical stable law plays an important role in the definition of $Scov^2(V,W)$, we refer the nonnegative square root of $Scov^2(V,W)$ as the **stable covariance** between V and W. Similarly, we can define $Svar^2(V)=Scov^2(V,V)$ and $Svar^2(W)=Scov^2(W,W)$. The **stable correlation** (SC) between V and W is defined as the nonnegative square root of

$$SC^2(V,W)=\frac{Scov^2(V,W)}{\sqrt{Svar^2(V)Svar^2(W)}}.$$

The form of $Scov^2(V,W)$ is similar to that of $dcov^2(V,W)$, the distance covariance between V and W. Indeed, $Scov^2(V,W)$ replaces the Euclidean distance $\|\cdot\|$ in $dcov^2(V,W)$ by $\exp\{-\|\cdot\|^a\}$. This change is crucial since $\exp\{-\|\cdot\|^a\}$ is always bounded by 1, and as a result, no moment conditions on V and W are required. The stable correlation has the following desirable properties.

Proposition 2.1. *The SC has the following properties*:

(A) $SC(V,W)$ *exists for any two random vectors* V *and* W.

(B) $0 \leqslant SC(V,W) \leqslant 1$. $SC(V,W)=0$ *if and only if* V *and* W *are independent.*

(C) *Let* D_1 *and* D_2 *be two orthogonal matrices, and* a_1 *and* a_2 *be two vectors. Then*

$$SC(V,W) = SC(a_1 + D_1V, a_2 + D_2W).$$

The Svar has the following properties:

(D) $Svar(V)=0$ *if and only if* $X=\mathrm{E}(X)$ *almost surely.*

(E) $Svar(c+DV)=Svar(V)$ *for all constant vectors* c*, and orthogonal matrices* D.

(F) $Svar(V+W) \leqslant Svar(V)+Svar(W)$ *for independent random vectors* V *and* W.

The proof of this proposition follows the same lines as those of [20, Theorems 3 and 4] and thus is omitted here. From Proposition 2.1, we know that the SC is defined for arbitrary two random vectors with any dimension. Thus it can handle the multivariate response and grouped predictor. Furthermore, it is robust against outliers or heavy distributed data. Moreover, it is nonnegative and is zero only if the two random vectors are independent and thus it can efficiently detect not only linear but also nonlinear dependence.

As noted by the anonymous reviewers, there is a connection with the work of Sejdinovic et al. [16]. In fact, the stable covariance can be seen as special members of the Hilbert-Schmidt independence criterion (HSIC) statistics (see also [3]) defined with the class of kernel functions $\exp(-\|X-X'\|^a)$ and $\exp(-\|Y-Y'\|^a), a \in (0,2]$. In particular, $a=1$ and $a=2$ correspond to the Laplacian kernel and the Gaussian kernel widely known in the machine learning community, respectively. In this paper, by the virtue of Proposition 2.1(B), we show that the product kernel

$$k((X,Y),(X',Y')) = \exp(-\|X-X'\|^a)\exp(-\|Y-Y'\|^a)$$

does induce an HSIC that characterizes independence. From another point of view, the stable covariance is a weighted L_2 distance between the joint characteristic function and the product of the marginal characteristic functions. The weight is chosen to be the density of a spherical stable law.

As a natural extension of the correlation in [6], Zhu et al. [24] recently proposed a projection correlation to measure the dependence between two random vectors. However, two random vectors may not be independent even if their projection correlation equals zero (see the counterexample in [6]). Kim et al. [9] overcame this issue by extending the Blum-Keifer-Rosenblatt correlation coefficient [1] via projection pursuit. Their new correlation is zero if and only if the two random vectors are independent. Moreover, the computation complexity of SC is $O(n^2)$, which is the same as that of DC. The computation of projection correlation and [9]'s correlation is more expensive and the complexity is $O(n^3)$.

We next propose an estimator of $SC(V,W)$ by using its sample counterpart. Suppose that $\{(v_i,w_i), i=1,\ldots,n\}$ is a random sample from the population (V,W). We can estimate $E_j, j=1,2,3$ by using their moment estimators. Specifically,

$$\begin{aligned}
\widehat{E_1} &= \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n} \mathrm{e}^{-\|v_i-v_j\|^a-\|w_i-w_j\|^a},\\
\widehat{E_2} &= \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n} \mathrm{e}^{-\|v_i-v_j\|^a}\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n} \mathrm{e}^{-\|w_i-w_j\|^a},\\
\widehat{E_3} &= \frac{1}{n(n-1)(n-2)}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\sum_{l\neq i,j}^{n} \mathrm{e}^{-\|v_i-v_j\|^a-\|w_i-w_l\|^a}.
\end{aligned} \tag{2.3}$$

Thus, a natural estimator of $Scov^2(V,W)$ is given by

$$\widehat{Scov^2}(V,W) = \widehat{E_1} + \widehat{E_2} - 2\widehat{E_3}.$$

Similarly, we can define the sample covariances $\widehat{Scov^2}(V,V)$ and $\widehat{Scov^2}(W,W)$. Accordingly, the sample SC between V and W can be defined by

$$\widehat{SC^2}(V,W)=\frac{\widehat{Scov^2}(V,W)}{\sqrt{\widehat{Svar^2}(V)\widehat{Svar^2}(W)}}.$$

We establish the tail probability bound for the difference between $\widehat{SC^2}(V,W)$ and $SC^2(V,W)$ as follows.

Theorem 2.2. *There exists a positive constant $C>0$, for any $\epsilon>0$, such that*

$$\Pr(|\widehat{SC^2}(V,W)-SC^2(V,W)|\geqslant\epsilon)\leqslant O(\exp\{-Cn\epsilon^2\}). \tag{2.4}$$

The inequality (2.4) indeed is an exponential inequality of $\widehat{SC^2}(V,W)$, and this inequality implies that $\widehat{SC^2}(V,W)$ converges to $SC^2(V,W)$ almost surely. Based on the theory of U-statistic, the asymptotic distributions of $\widehat{SC^2}(V,W)$ depend on the value of $SC^2(V,W)$. When $SC^2(V,W)=0$, $\widehat{SC^2}(V,W)$ is a degenerate U-statistic, and it converges to a non-degenerate distribution at the n^{-1} convergence rate, while when $SC^2(V,W)\neq 0$, $\widehat{SC^2}(V,W)$ is a non-degenerate U-statistic, and then $\widehat{SC^2}(V,W)-SC^2(V,W)$ converges to a normal distribution at the $n^{-1/2}$ convergence rate. The details are omitted in this paper to save space.

3 Robust feature screening procedure

Based on the SC, we propose a robust feature screening procedure for ultrahigh dimensional data. Let $Y=(Y_1,\ldots,Y_q)^{\mathrm{T}}$ be the response vector and $X=(X_1,\ldots,X_p)^{\mathrm{T}}$ be the predictor vector. We will concentrate on the setting in which q is fixed, but p may increase in an exponential order of n. Denote by $F(Y\mid X)$ the conditional distribution function of Y given X. We define the index set of the active and inactive predictors as follows:

$$\begin{aligned}&\mathcal{A}=\{k:F(Y\mid X)\text{ depends on }X_k\text{ for some }y\in\Omega_Y\},\\&\mathcal{I}=\{k:F(Y\mid X)\text{ does not depend on }X_k\text{ for any }y\in\Omega_Y\}.\end{aligned} \tag{3.1}$$

Here, Ω_Y is the support of Y. We further denote $X_{\mathcal{A}}=\{X_k:k\in\mathcal{A}\}$ and $X_{\mathcal{I}}=\{X_k:k\in\mathcal{I}\}$. Clearly, $X_{\mathcal{A}}$ can be regarded as an active predictor vector and its complement $X_{\mathcal{I}}$ as an inactive predictor vector. We aim to identify the index subset $\mathcal{A}$ of all the active predictors.

For ease of presentation, we write

$$\omega_k=SC^2(X_k,Y)\quad\text{and}\quad\hat{\omega}_k=\widehat{SC^2}(X_k,Y)$$

for $k=1,\ldots,p$, based on a random sample $\{x_i,y_i\}$, $i=1,\ldots,n$. We consider using ω_k as a marginal utility to rank the importance of X_k. The SC naturally serves as a good marginal screening utility due to its several merits: (i) it is model-free, i.e., the corresponding screening procedure does not require to impose a specific model structure on the regression function of Y on X. (ii) The SC allows multivariate responses and grouped predictors. (iii) It is robust against potential outliers since no moment condition is needed for the response as well as the predictors. A large value of $\hat{\omega}_k$ indicates that the predictor X_k is more correlated with the response. We select a set of important predictors with large $\hat{\omega}_k$, i.e., we define

$$\hat{\mathcal{A}}=\{k:\hat{\omega}_k\geqslant cn^{-\kappa},\ \text{for }1\leqslant k\leqslant p\},$$

where c and κ are pre-specified threshold values, which will be defined in the following condition:

(C1) The minimum SC of active predictors satisfies

$$\min_{k\in\mathcal{A}}\omega_k\geqslant 2cn^{-\kappa}$$

for some constants $c>0$ and $0\leqslant\kappa<1/2$.

Condition (C1) requires that for the active predictor X_k, $\omega_k = SC^2(X_k, Y)$ is not too close to 0. Such an assumption is commonly imposed in works on the marginal screening (see, for example, [2, Condition 3] and [11, Condition (C2)]). The sure screening property for the proposed robust screening procedure is established in the following theorem.

Theorem 3.1. *There exists a positive constant $C > 0$, such that*

$$\Pr\Big(\max_{1\leqslant k\leqslant p}|\hat{\omega}_k - \omega_k| \geqslant cn^{-\kappa}\Big) \leqslant O(p\exp\{-Cn^{1-2\kappa}\}). \tag{3.2}$$

Under Condition (C1)*, we have*

$$\Pr(\mathcal{A}\subseteq\hat{\mathcal{A}}) \geqslant 1 - O(s_n\exp\{-Cn^{1-2\kappa}\}), \tag{3.3}$$

where $s_n = |\mathcal{A}|$ is the cardinality of $\mathcal{A}$.

Due to the sure screening property, we write the proposed SC-based robust sure independence screening procedure as the SC-SIS for short. Compared with many existing methods, the subexponential tail condition can now be totally removed due to the boundedness of $Scov^2$. The sure screening property holds for the SC-SIS under milder conditions than that for the DC-SIS [11] in that the SC-SIS does not require any moment conditions for both X and Y. We also achieve the NP dimensionality $\ln p = o(n^{1-2\kappa})$, $0 \leqslant \kappa < 1/2$. The DC-SIS generally cannot handle such a rate unless X_k and Y are bounded uniformly in p. However, this condition is too strong to be satisfied in practice.

We can characterize the size of the reduced model after screening in the following theorem.

Theorem 3.2. *We have*

$$\Pr\left(|\hat{\mathcal{A}}| \leqslant 2c^{-1}n^{\kappa}\sum_{k=1}^{p}\omega_k\right) \geqslant 1 - O(p\exp\{-Cn^{1-2\kappa}\}). \tag{3.4}$$

This theorem implies that if $\sum_{k=1}^{p}\omega_k = O(n^b)$ for some $b > 0$, the model after screening is of polynomial size with probability approaching to 1.

We further note that given $X_{\mathcal{A}}$, Y is independent of $X_{\mathcal{I}}$. This implies that Y should be more dependent upon $X_{\mathcal{A}}$ than $X_{\mathcal{I}}$. In other words, ω_k for $k \in \mathcal{A}$ is larger than ω_k for $k \in \mathcal{I}$. This is formulated as follows:

(C2) $\min_{k\in\mathcal{A}}\omega_k - \max_{k\in\mathcal{I}}\omega_k = \Delta > 0$.

It is noted here that Δ can tend to zero as $n \to \infty$, which is discussed later. With (C2), we have the ranking consistency property for the SC-SIS procedure.

Theorem 3.3. *Under the conditions* (C1) *and* (C2)*, we have*

$$\Pr\Big(\min_{k\in\mathcal{A}}\hat{\omega}_k > \max_{k\in\mathcal{I}}\hat{\omega}_k\Big) \geqslant 1 - O(p\exp\{-Cn\Delta^2\}). \tag{3.5}$$

This theorem shows that ω_k can always rank an active variable above an inactive variable with probability approaching 1 provided that $p\exp\{-Cn\Delta^2\} = o(1)$. This implies that for the ranking consistency property, it is unnecessary to assume Δ to be a constant. Instead, Δ is allowed to tend to zero as $n \to \infty$. In fact, if we set $\Delta = O(n^{-\gamma})$, $0 < \gamma < 1/2$ and $\ln p = o(n^{1-2\gamma})$, we still have $p\exp\{-Cn\Delta^2\} = o(1)$ and thus the ranking consistency property still holds.

4 Numerical studies

In this section, we conduct Monte Carlo simulation studies to examine the performance of the SC-SIS and compare its performance with its competitors including the DC-SIS (distance correlation sure independence screening) [11], DC-RoSIS (distance correlation robust sure independence screening) [22], RRCS (robust rank correlation screening) [10], PC-SIS (projection correlation sure independence screening) [12], SIRS (sure independence and ranking screening) [23] and SIS (sure independence screening) [2].

To examine the robustness of the SC-SIS, we generate features X from a mixture distribution $(1-\alpha)X_n + \alpha X_t$ with α being equal to 0, 0.1 or 0.2, where X_t is a p-dimensional random vector with each component being independent t_1-distribution (i.e., the Cauchy distribution) and $X_n \sim N(0,\Sigma)$ is a p-dimensional multivariate normal distribution with mean 0 and covariance matrix Σ. Throughout the simulation, we set $\Sigma = (\sigma_{ij})_{p\times p}$ with entries $\sigma_{ij} = 0.75^{|i-j|}$, $i,j = 1,\ldots,p$. We consider two types of error ε: the standard normal distribution and the t_1-distribution. Let $\hat{\mathcal{A}}_\delta = \{k : \hat{\omega}_k \geqslant \delta\}$ and define the minimum model size (MMS) to be the cardinality of the smallest $\hat{\mathcal{A}}_\delta$ that includes all the active predictors, i.e.,

$$\text{MMS} = \min\{|\hat{\mathcal{A}}_\delta| : \mathcal{A} \subseteq \hat{\mathcal{A}}_\delta\}.$$

A screening method with a small value of the MMS suggests that it is more powerful in detecting the dependence between features and responses. We report the 25%, 50%, 75% and 90% quantiles of the MMS to compare the performances of different screening methods based on 500 replications. In particular, in Subsection 4.1, we study what value of a to use in practice. In Subsections 4.2–4.4, we apply the SC-SIS and other existing methods to different models including the linear model, the generalized linear model, the nonlinear model and the model with the grouped predictor. In Subsection 4.5, we demonstrate the proposed SC-SIS by a real data example.

For implementation of the stable covariance, we employ the estimation strategy in [20] to reduce the computation complexity. In that way, the stable covariance can be implemented in $O(n^2)$ computations. To further improve the performance, the U-centering idea in [19] is adopted. With the technique developed in [7], it would even be implemented in $O(n\log n)$ computations.

4.1 Choice of parameter a

The proposed stable correlation involves a unspecified parameter a. Though our theory holds for any choice $a \in (0,2]$, in this section, we study how the parameter a effects the performance of the SC-SIS empirically and suggest what value of a to use in practice. To this end, we set $a = 0.1, 0.2, \ldots, 2.0$ and apply the SC-SIS to Models 1.a, 2.a and 2.b defined in Subsections 4.2 and 4.3. Figure 1 reports the 25%, 50% and 75% quantiles of the MMS for the SC-SIS when a varies in Models 1.a, 2.a and 2.b. Figure 1 clearly shows that the quantiles of the MMS first decrease as a increases and then increase as a further increases. Therefore, the SC-SIS is more powerful in capturing active features when a is neither too small nor too large. According to Figure 1, we suggest that any value between 0.3 and 0.7 is a reasonable choice for a and we set $a = 0.5$ in the rest of the numerical studies.

4.2 Linear model and generalized linear model

We consider the linear model and the generalized linear model as follows:

Model 1.a. (Linear model) $Y = X^{\mathrm{T}}\beta + \varepsilon$.

Model 1.b. (Poisson regression) $Y \mid X \sim \text{Poisson}(\lambda(X))$, where $\lambda(X) = \exp\{X^{\mathrm{T}}\beta\}$ and Poisson denotes the Poisson distribution.

We set $n = 100$, $p = 2{,}000$, and $\beta = (\mathbf{1}_5^{\mathrm{T}}, \mathbf{0}_{p-5}^{\mathrm{T}})^{\mathrm{T}}$ and thus we include 5 active predictors in the two models. The 25%, 50%, 75% and 90% quantiles of the MMS are reported in Table 1. First, for Model 1.a with $\varepsilon \sim N(0,1)$ and $\alpha = 0$, the 25%, 50%, 75% and 90% quantiles of the MMS are exactly 5 for all the methods. This implies that in the setting where both predictors and errors are not heavy-tailed, all the methods can capture active predictors efficiently. However, as α changes from 0 to 0.1 and 0.2, the DC-SIS, DC-RoSIS, SIRS, and SIS start to perform poorly, especially for the 90% quantile while our proposed SC-SIS, RRCS and PC-SIS can still perform reasonably well even when $\alpha = 0.2$. When ε follows the t_1-distribution for Model 1.a, we observe a similar phenomenon: the SC-SIS, RRCS and PC-SIS outperform the other methods especially when $\alpha \neq 0$. This implies that the SC-SIS, RRCS and PC-SIS are more robust to outliers. Among all the methods, the RRCS has the best performance for the linear model (Model 1.a). This is not surprising since the RRCS is designed to capture the monotonic relationship. For Model 1.b, the proposed SC-SIS and PC-SIS outperform the other methods and can effectively find

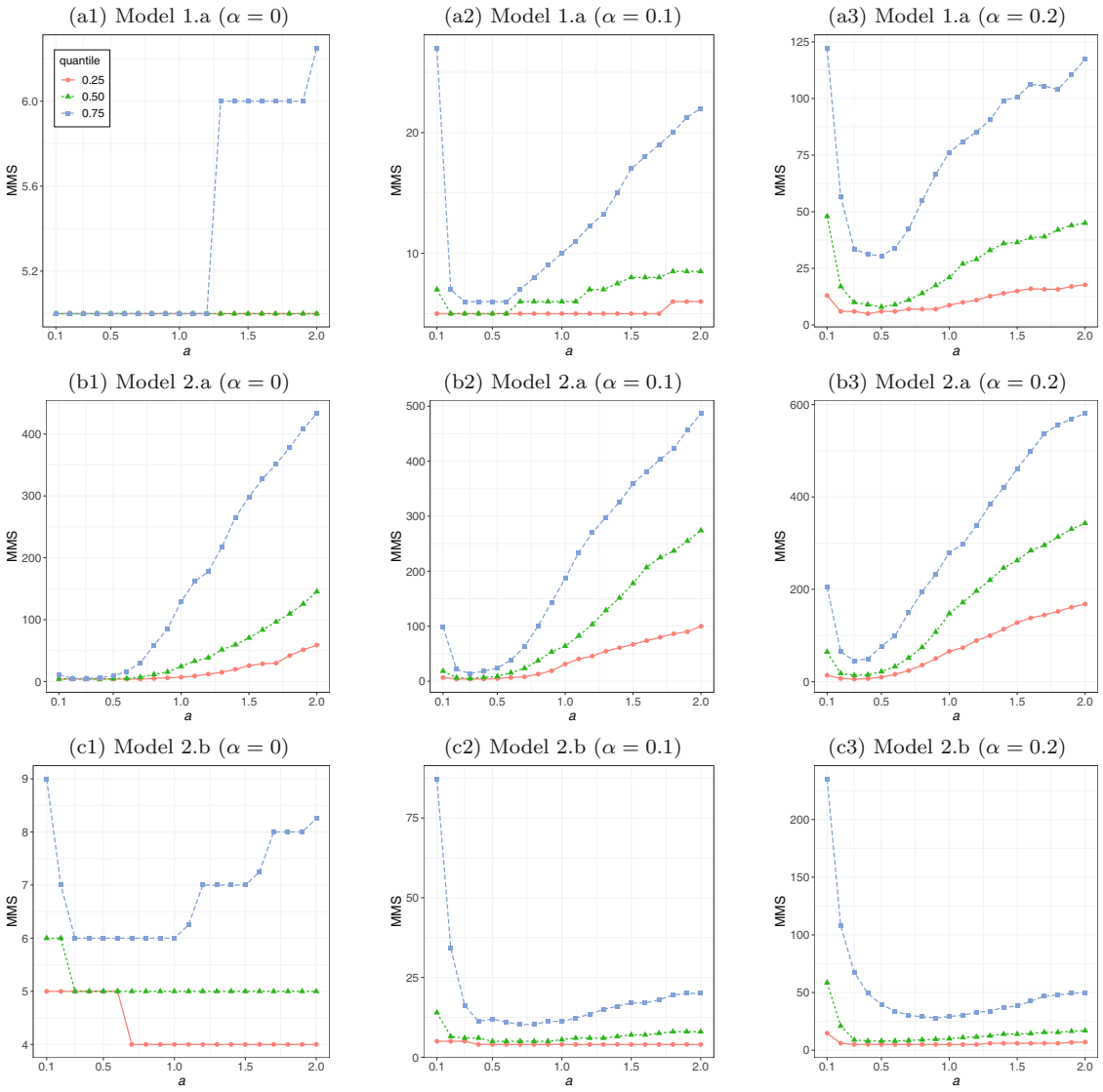

Figure 1 (Color online) The 25%, 50% and 75% quantiles of the MMS for $\varepsilon \sim t_1$. The top, middle and bottom panels are for Models 1.a, 2.a and 2.b with α being equal to 0, 0.1 and 0.2, respectively

all the active predictors. In this case, the RRCS completely fails since it cannot detect non-monotonic relationship. The DC-SIS, DC-RoSIS, SIRS and SIS still perform very poorly when $\alpha \neq 0$.

4.3 Nonlinear model

In this subsection, we consider two nonlinear models.

Model 2.a. $Y = 5X_1X_2 + 5X_3I(X_3 > 0) + 5\sin(2\pi X_4) + \varepsilon$.

Model 2.b. $Y = X_1 + X_2 + 2/\exp(X_3 + X_4) + \varepsilon$.

Here, $I(\cdot)$ is the indicator function. We set $n = 200$ and $p = 2{,}000$ and simulation results are summarized in Table 2. As shown in Table 2, for $\varepsilon \sim N(0, 1)$ and $\alpha = 0$, the RRCS, SIRS and SIS perform poorly. This suggests that these methods cannot detect nonlinear relationship between predictors and responses. On the contrary, the SC-SIS, DC-SIS, DC-RoSIS and PC-SIS are able to efficiently detect nonlinear relationship. When $\alpha = 0.1$ or $\alpha = 0.2$, the performances of the DC-SIS and DC-RoSIS are not satisfactory. Again, this implies that these two methods are not robust against outliers or heavy-tailed distributions. The SC-SIS also outperforms the PC-SIS in this example. Overall, in this nonlinear case, our method SC-SIS has the best performance in all the settings, indicating that our method is not only able to detect any possible dependence but also insensitive to outliers.

Table 1 Simulation results for the linear model and generalized linear model

	$\alpha=0$				$\alpha=0.1$				$\alpha=0.2$			
	25%	50%	75%	90%	25%	50%	75%	90%	25%	50%	75%	90%
Model 1.a: $\varepsilon\sim N(0,1)$												
SC-SIS	5.0	5.0	5.0	5.0	5.0	5.0	5.0	6.0	5.0	5.0	7.0	15.2
DC-SIS	5.0	5.0	5.0	5.0	5.0	5.0	14.0	28.1	8.8	21.0	49.0	108.3
DC-RoSIS	5.0	5.0	5.0	5.0	5.0	5.0	6.0	27.1	5.0	5.0	13.0	127.4
RRCS	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0
SIRS	5.0	5.0	5.0	5.0	5.0	8.0	39.2	115.1	6.0	16.0	75.0	139.0
SIS	5.0	5.0	5.0	5.0	8.0	27.0	208.2	1134.1	46.0	136.0	659.5	1521.0
PC-SIS	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.2
Model 1.a: $\varepsilon\sim t_1$												
SC-SIS	5.0	5.0	5.0	6.0	5.0	5.0	8.0	27.1	6.0	13.0	57.5	240.5
DC-SIS	5.0	5.0	5.0	14.0	12.0	25.0	55.0	155.1	32.0	71.0	136.2	303.4
DC-RoSIS	5.0	5.0	5.0	5.0	5.0	5.0	7.0	61.0	5.0	7.0	28.0	197.6
RRCS	5.0	5.0	5.0	5.0	5.0	5.0	5.0	6.0	5.0	5.0	7.0	12.0
SIRS	5.0	5.0	5.0	6.0	6.0	13.0	50.0	134.0	9.0	29.0	85.2	157.1
SIS	7.0	266.0	1219.2	1790.3	98.8	460.0	1238.2	1758.0	200.8	526.0	1263.5	1682.3
PC-SIS	5.0	5.0	5.0	6.0	5.0	5.0	5.0	6.0	5.0	5.0	7.0	15.1
Model 1.b												
SC-SIS	5.0	5.0	5.0	5.0	5.0	5.0	5.0	6.0	5.0	5.0	7.0	14.0
DC-SIS	5.0	8.0	28.0	88.2	515.5	1214.0	1593.5	1840.7	904.5	1308.5	1629.5	1877.1
DC-RoSIS	5.0	5.0	5.0	5.0	5.0	5.0	6.0	33.0	5.0	6.0	24.0	189.1
RRCS	250.0	1089.5	1801.5	1990.0	944.0	1652.0	1966.2	1998.1	1633.0	1931.0	1997.0	2000.0
SIRS	5.0	5.0	5.0	5.0	5.0	10.0	60.0	372.3	9.0	44.5	257.2	683.5
SIS	14.0	37.0	107.2	247.5	736.8	1304.5	1763.8	1920.2	916.0	1332.0	1749.8	1914.2
PC-SIS	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.1

4.4 Model with grouped predictors

In this subsection, we apply the SC-SIS to select grouped predictors. Following [11, Example 2], we consider the following model:

Model 3. $Y=X_1+X_2+\{I(X_3<q_1)+I(q_1\leqslant X_3<q_2)+I(X_3>q_3)\}+X_4+\varepsilon$, where q_1,q_2 and q_3 are the 25%, 50% and 75% quantiles of X_3, respectively. In this kind of model, $I(X_3<q_1)$, $I(q_1\leqslant X_3<q_2)$ and $I(X_3>q_3)$ are considered as a grouped predictor in the sense that either all three of them are included in the model or none of them is included. We write

$$X_3^*=\{I(X_3<q_1),I(q_1\leqslant X_3<q_2),I(X_3>q_3)\}^{\mathrm{T}}.$$

Consequently X_3^* is a grouped predictor with three levels and the number of active predictors is 4.

Among these methods, only the SC-SIS, DC-SIS, DC-RoSIS and PC-SIS can be directly used for screening grouped predictors and thus we only focus on these four screening procedures. We set $n=200$ and $p=2{,}000$. The simulation results are given in Table 3. When $\alpha=0$, all these four methods perform very well, i.e., with high probability, the MMS which ensures the inclusion of all the active predictors is the same as the number of active predictors. However, when $\alpha=0.1$, the DC-SIS starts to perform poorly and the 90% quantile of the MMS becomes high especially when $\varepsilon\sim t_1$. Both SC-SIS and DC-RoSIS can still control the MMS very well and the SC-SIS performs slightly better than the DC-RoSIS. When $\alpha=0.2$, the DC-RoSIS also performs poorly and the 90% quantile of the MMS becomes much higher than that of the SC-SIS. In fact, except for the setting where $\alpha=0.2$ and $\varepsilon\sim t_1$, the 90% quantile of the MMS of the SC-SIS is 4, which is exactly the number of active predictors. Thus our proposed SC-SIS is also efficient and robust in detecting grouped predictors. The performance of the PC-SIS is similar to that of the SC-SIS.

Table 2 Simulation results for the nonlinear model

	$\alpha=0$				$\alpha=0.1$				$\alpha=0.2$			
	25%	50%	75%	90%	25%	50%	75%	90%	25%	50%	75%	90%
					Model 2.a: $\varepsilon\sim N(0,1)$							
SC-SIS	4.0	4.0	4.0	5.0	4.0	5.0	10.0	31.0	5.0	12.0	34.0	107.0
DC-SIS	4.0	4.0	5.0	10.1	32.0	90.0	349.2	792.4	183.2	441.0	1031.2	1636.4
DC-RoSIS	4.0	4.0	7.0	16.0	8.8	29.0	145.2	414.1	42.8	186.0	505.5	1036.7
RRCS	39.0	235.5	839.2	1520.6	147.8	579.5	1251.2	1738.6	398.0	962.5	1551.8	1813.1
SIRS	37.0	220.0	808.0	1509.4	615.8	1193.5	1655.2	1887.4	862.0	1395.0	1746.0	1942.3
SIS	6.0	40.0	442.5	1470.5	549.8	1131.0	1618.0	1868.1	680.2	1177.0	1627.5	1845.5
PC-SIS	4.0	4.0	8.0	22.0	6.0	14.0	40.3	94.7	26.0	76.5	173.3	346.5
					Model 2.a: $\varepsilon\sim t_1$							
SC-SIS	4.0	4.5	8.0	24.0	5.0	13.0	46.0	148.1	15.0	54.0	168.0	444.3
DC-SIS	4.0	6.0	25.0	105.9	64.0	170.0	473.2	1084.8	217.2	521.0	1062.5	1595.4
DC-RoSIS	4.0	6.0	16.0	45.0	12.0	45.0	187.8	513.7	71.5	222.5	614.8	1129.9
RRCS	65.8	246.5	939.2	1525.5	204.2	637.0	1291.0	1759.4	496.0	1027.5	1509.5	1827.7
SIRS	53.8	286.0	883.5	1594.1	685.2	1222.5	1646.0	1895.1	877.8	1363.0	1762.2	1926.2
SIS	272.5	964.5	1612.5	1895.0	734.8	1236.5	1688.8	1880.1	716.8	1193.0	1558.2	1843.2
PC-SIS	4.0	8.0	23.0	81.4	11.0	36.0	102.0	225.1	63.0	155.5	347.3	651.2
					Model 2.b: $\varepsilon\sim N(0,1)$							
SC-SIS	4.0	5.0	5.0	5.0	4.0	4.0	5.0	6.0	4.0	4.0	5.0	9.0
DC-SIS	5.0	6.0	7.0	8.0	486.8	1217.5	1652.5	1846.0	976.8	1368.0	1717.0	1888.0
DC-RoSIS	6.0	7.0	8.0	10.0	7.0	9.0	20.0	58.0	6.0	11.0	42.2	153.8
RRCS	478.8	991.0	1465.5	1811.8	706.0	1119.0	1547.0	1813.0	436.5	960.0	1547.0	1825.2
SIRS	7.0	8.0	11.0	17.0	42.0	120.5	265.5	650.3	24.8	86.5	189.5	451.7
SIS	6.0	8.0	19.0	86.2	988.0	1446.0	1761.5	1895.7	1047.8	1456.5	1769.2	1901.2
PC-SIS	6.0	7.0	8.0	11.0	7.0	9.0	13.0	22.0	8.0	14.0	28.0	52.0
					Model 2.b: $\varepsilon\sim t_1$							
SC-SIS	5.0	5.0	6.2	12.0	4.0	6.0	23.0	99.4	5.0	13.0	65.2	319.2
DC-SIS	6.0	6.0	10.0	33.0	476.8	1111.0	1553.2	1847.0	951.0	1387.0	1707.0	1869.3
DC-RoSIS	7.0	11.5	24.0	51.1	14.0	34.5	89.2	219.5	15.0	37.0	106.5	390.5
RRCS	147.8	506.5	1144.0	1583.4	721.2	1189.0	1591.5	1823.0	788.2	1273.0	1683.5	1876.1
SIRS	9.0	15.0	33.2	77.0	93.8	216.0	516.8	1065.2	44.0	146.0	363.2	914.1
SIS	7.0	24.5	210.2	986.4	900.8	1366.5	1721.0	1890.9	1002.0	1400.0	1750.5	1906.0
PC-SIS	8.0	13.0	28.0	54.0	16.0	38.0	89.3	170.0	34.8	77.5	152.5	281.4

Table 3 Simulation results for grouped variable screening

	$\alpha=0$				$\alpha=0.1$				$\alpha=0.2$			
	25%	50%	75%	90%	25%	50%	75%	90%	25%	50%	75%	90%
					$\varepsilon\sim N(0,1)$							
SC-SIS	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0
DC-SIS	4.0	4.0	4.0	4.0	4.0	5.0	11.0	28.1	12.0	27.5	71.2	325.9
DC-RoSIS	4.0	4.0	4.0	4.0	4.0	4.0	4.0	6.1	4.0	4.0	6.0	50.7
PC-SIS	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0
					$\varepsilon\sim t_1$							
SC-SIS	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	7.0
DC-SIS	4.0	4.0	4.0	4.0	11.0	25.0	53.0	233.4	43.0	100.0	273.5	926.0
DC-RoSIS	4.0	4.0	4.0	4.0	4.0	4.0	5.0	13.0	4.0	4.0	9.2	172.4
PC-SIS	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0

4.5 Real data example

We now apply the proposed SC-SIS procedure to a cardiomyopathy microarray dataset and compare it with the DC-SIS procedure [11]. The cardiomyopathy microarray dataset was collected from a study based on a transgenic mouse model, in which the aim was to determine which genes were influential for overexpression of a G protein-coupled receptor, designated Ro1, in mice. The research is related to understanding types of human heart diseases. The response is the expression level Ro1 and was measured for $n = 30$ mice. There are in total 6,319 genetic expression levels ($p = 6{,}319$) for each mouse.

In this empirical analysis, we focus on the SC-SIS and DC-SIS and apply them to rank the features. Both SC-SIS and DC-SIS rank Msa.2134.0 as the most important feature. As shown in Figure 2(a), there is a clear nonlinear relationship between Msa.2134.0 and Ro1. The SC-SIS identifies Msa.1024.0 as the second most important feature while the DC-SIS identifies Msa.1024.0 as the 19th most important feature. Figure 2(c) shows the scatter plot of Msa.1024.0 and Ro1. We detect 3 potential

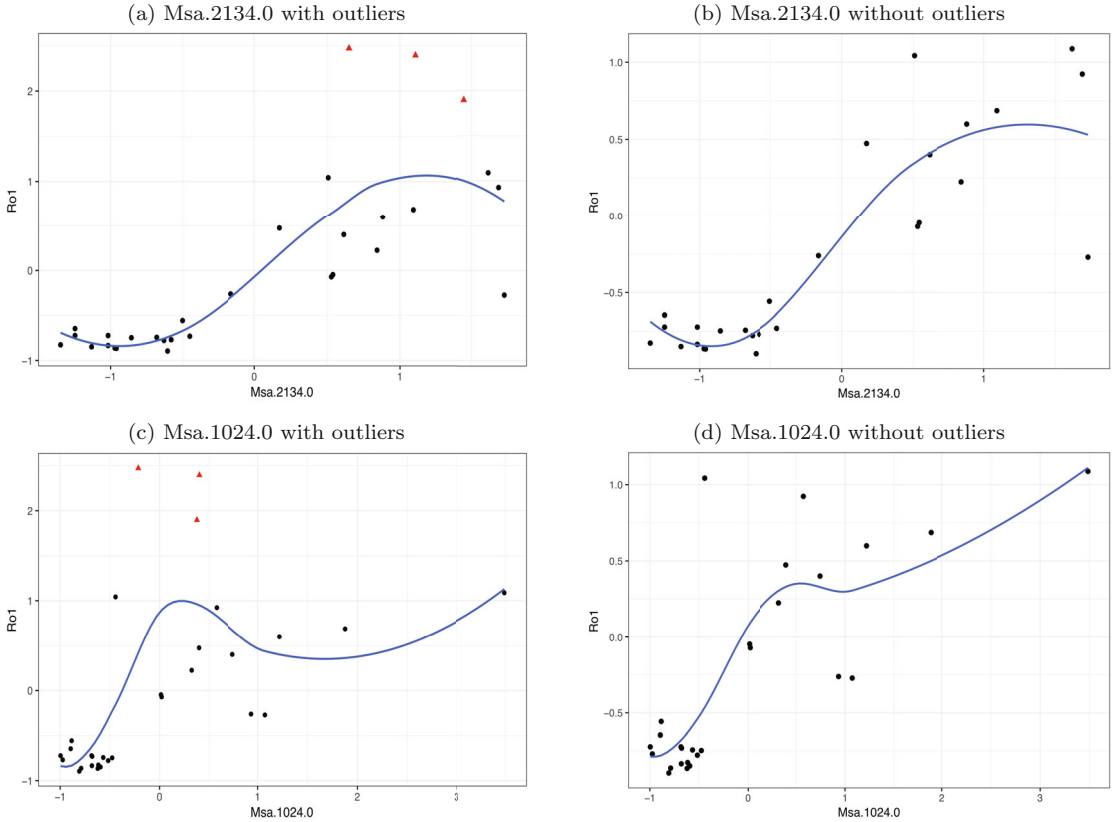

Figure 2 (Color online) Scatter plots of Msa.2134.0 and Msa.1024.0 versus Ro1. The left two panels are the scatter plots with the potential outliers and the potential outliers are marked by '△'. The right two panels are the scatter plots after removing the potential outliers. The solid curves are fitted by local linear regression

Table 4 Rankings determined by the SC-SIS and DC-SIS before and after removing the potential outliers

		Msa.2134.0	Msa.1024.0	Msa.5727.0	Msa.42131.0
SC-SIS	Before	1	2	3	4
	After	1	2	4	3
DC-SIS	Before	1	19	11	25
	After	1	9	6	4

outliers that are marked by '△' (see Figures 2(a) and 2(c)). These 3 potential outliers are detected by standardized residuals after fitting marginal linear regression. To examine the robustness of the SC-SIS, we remove the 3 potential outliers from the original dataset and the scatter plots are shown in Figures 2(b) and 2(d). As shown in the bottom panel of Figure 2(d), we can see a strong monotonic relationship between Msa.1024.0 and Ro1. The DC-SIS fails to capture such a monotonic relationship due to the presence of outliers while the SC-SIS can still capture it despite the potential outliers. After removing the potential outliers, the SC-SIS still ranks Msa.1024.0 as the second most important feature and the DC-SIS ranks Msa.1024.0 as the 9th most important feature instead of 19th. This observation indicates that the SC-SIS is more robust than the DC-SIS and is less sensitive to outliers. The top 4 features selected by the SC-SIS are $\{\text{Msa.2134.0}, \text{Msa.1024.0}, \text{Msa.5727.0}, \text{Msa.42131.0}\}$. Table 4 shows the rankings of these 4 features before and after removing the potential outliers determined by the SC-SIS and DC-SIS, respectively. The rankings given by the SC-SIS are very stable while the rankings given by the DC-SIS become much smaller when outliers are removed.

To further examine the robustness of the proposed SC-SIS, we apply the bootstrap approach proposed in [4] to this dataset. We describe the approach as follows:

• Apply the SC-SIS/DC-SIS to the original dataset and obtain the rankings $r_1, \ldots, r_p$ for all the features, where r_j is the ranking of the feature X_j.

• For $b = 1, \ldots, B$, draw a bootstrap sample $\{(x_1^{(b)}, y_1^{(b)}), \ldots, (x_n^{(b)}, y_n^{(b)})\}$ from the original dataset and apply the SC-SIS/DC-SIS to the bootstrap sample. The resulting rankings are $r_1^{(b)}, \ldots, r_p^{(b)}$.

• For each $j = 1, \ldots, p$, compute the 2.5% quantile r_{j-} and the 97.5% quantile r_{j+} based on $r_j^{(1)}, \ldots, r_j^{(B)}$.

Following [10], we set $B = 200$. Figure 3 shows the original ranking, 2.5% quantile r_{j-} and 97.5% quantile r_{j+} for the top 10 features selected by the SC-SIS, ordering in terms of increasing r_{j+}. In Figure 3, Msa.2134.0 and Msa.1024.0 emerge strongly as the top 2 features based on the SC-SIS and this result coincides with the result obtained from the original dataset. We can also see that for the same feature, the 95% interval of the ranking given by the SC-SIS is much shorter than that given by the DC-SIS. For example, the 95% intervals of Msa.2134.0 given by the SC-SIS and DC-SIS are $[1, 19]$ and $[1, 63]$, respectively. This indicates that the SC-SIS is a robust feature screening procedure, while the DC-SIS is not.

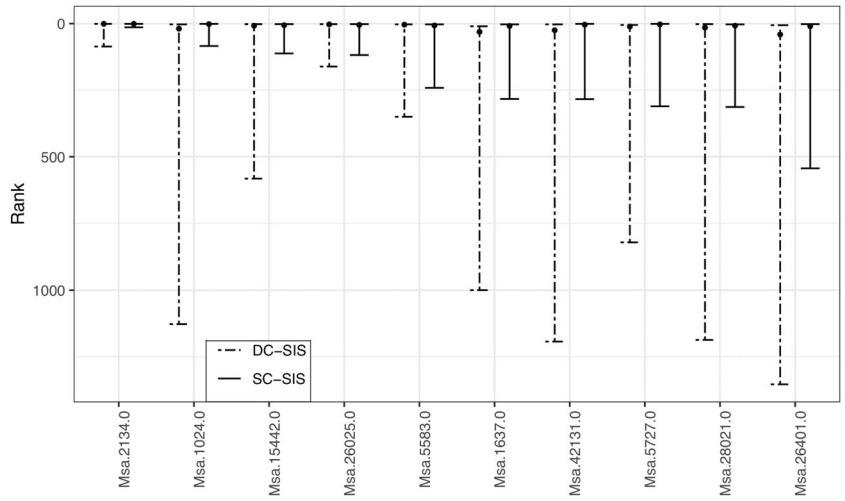

Figure 3 95% intervals of ranking obtained from the DC-SIS and SC-SIS based on 200 bootstrap samples for the top 10 genes selected by the SC-SIS. Variables are ordered in increasing order of r_{j+} by using the SC-SIS method. The black dots are the ranking obtained from the original dataset

5 Conclusion

In this paper, we propose the stable correlation (SC) to measure the dependence between two random vectors. SC is well defined for any pair of random vectors and no moment condition is required. One of its desirable properties is that it equals zero if and only if the two random vectors are independent. Motivated by this nice property, we further propose the SC-SIS, a robust feature screening procedure without specifying a regression model. For the SC-SIS, we establish its sure screening property and rank consistency property without imposing the subexponential tail condition. The numerical studies demonstrate that this SC-SIS is able to capture both linear and nonlinear dependence between predictors and responses and is robust against outliers.

Acknowledgements The first author was supported by National Natural Science Foundation of China (Grant No. 11701034). The second author was supported by National Science Foundation of USA (Grant No. DMS-1820702). The authors are grateful to the two anonymous referees for the constructive comments and suggestions that led to significant improvement of an early manuscript.

References

1 Blum J R, Kiefer J, Rosenblatt M. Distribution free tests of independence based on the sample distribution function. Ann Math Statist, 1961, 32: 485–498
2 Fan J Q, Lv J C. Sure independence screening for ultrahigh dimensional feature space. J R Stat Soc Ser B Stat Methodol, 2008, 70: 849–911
3 Gretton A, Fukumizu K, Teo C H, et al. A kernel statistical test of independence. In: Proceedings of the Twenty-First Annual Conference on Neural Information Processing Systems. Advances in Neural Information Processing Systems. Cambridge: MIT Press, 2008, 585–592
4 Hall P, Miller H. Using generalized correlation to effect variable selection in very high dimensional problems. J Comput Graph Statist, 2009, 18: 533–550
5 Heller R, Heller Y, Gorfine M. A consistent multivariate test of association based on ranks of distances. Biometrika, 2013, 100: 503–510
6 Hoeffding W. A non-parametric test of independence. Ann Math Statist, 1948, 19: 546–557
7 Huo X M, Székely G J. Fast computing for distance covariance. Technometrics, 2016, 58: 435–447
8 Kendall M G. A new measure of rank correlation. Biometrika, 1938, 30: 81–93
9 Kim I, Balakrishnan S, Wasserman L. Robust multivariate nonparametric tests via projection-averaging. Ann Statist, 2020, in press
10 Li G R, Peng H, Zhang J, et al. Robust rank correlation based screening. Ann Statist, 2012, 40: 1846–1877
11 Li R Z, Zhong W, Zhu L P. Feature screening via distance correlation learning. J Amer Statist Assoc, 2012, 107: 1129–1139
12 Liu W J, Ke Y, Li R Z. Model-free feature screening and FDR control with Knockoff features. J Amer Statist Assoc, 2020, in press
13 Liu W J, Li R Z. Variable Selection and Feature Screening. Macroeconomic Forecasting in the Era of Big Data, vol. 52. Cham: Springer, 2020
14 Nolan J P. Multivariate elliptically contoured stable distributions: Theory and estimation. Comput Statist, 2013, 28: 2067–2089
15 Pan W L, Wang X Q, Zhang H P, et al. Ball covariance: A generic measure of dependence in Banach space. J Amer Statist Assoc, 2020, 115: 307–317
16 Sejdinovic D, Sriperumbudur B, Gretton A, et al. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. Ann Statist, 2013, 41: 2263–2291
17 Serfling R J. Approximation Theorems of Mathematical Statistics. New York: Wiley, 1980
18 Spearman C. The proof and measurement of association between two things. Amer J Psych, 1904, 15: 72–101
19 Székely G J, Rizzo M L. Partial distance correlation with methods for dissimilarities. Ann Statist, 2014, 42: 2382–2412
20 Székely G J, Rizzo M L, Bakirov N K. Measuring and testing dependence by correlation of distances. Ann Statist, 2007, 35: 2769–2794
21 Weihs L, Drton M, Meinshausen N. Symmetric rank covariances: A generalized framework for nonparametric measures of dependence. Biometrika, 2018, 105: 547–562
22 Zhong W, Zhu L P, Li R Z, et al. Regularized quantile regression and robust feature screening for single index models. Statist Sinica, 2016, 26: 69–95

23 Zhu L P, Li L X, Li R Z, et al. Model-free feature screening for ultrahigh-dimensional data. J Amer Statist Assoc, 2011, 106: 1464–1475
24 Zhu L P, Xu K, Li R Z, et al. Projection correlation between two random vectors. Biometrika, 2017, 104: 829–843

Appendix A

Lemma A.1. *Let $h(Y_1,\ldots,Y_m)$ be a kernel of the U-statistics U_n and $\theta = \mathrm{E}h(Y_1,\ldots,Y_m)$. If $a \leqslant h(Y_1,\ldots,Y_m) \leqslant b$, then for any $t>0$ and $n \geqslant m$,*

$$\Pr(U_n - \theta \geqslant t) \leqslant \exp\{-2[n/m]t^2/(b-a)^2\},$$

where $[n/m]$ denotes the integer part of n/m.

This lemma indeed is [17, Theorem 5.6.1.A].

Proof of Theorem 2.2. It is sufficient to establish the exponential inequality for the denominator and the numerator of $\widehat{SC^2}(V,W)$, respectively. Because the denominator of $\widehat{SC^2}(V,W)$ has a similar form to the numerator, we deal with the numerator only below. Throughout the proof, the notations C and c are generic constants, which may take different values at each appearance.

Recall the definitions of $\widehat{E}_j, j=1,2,3$ as follows:

$$\widehat{E}_1 = \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n} \mathrm{e}^{-\|v_i-v_j\|^a-\|w_i-w_j\|^a},$$

$$\widehat{E}_2 = \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n} \mathrm{e}^{-\|v_i-v_j\|^a}\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n} \mathrm{e}^{-\|w_i-w_j\|^a},$$

$$\widehat{E}_3 = \frac{1}{n(n-1)(n-2)}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\sum_{l\neq i,j}^{n} \mathrm{e}^{-\|v_i-v_j\|^a-\|w_i-w_l\|^a}.$$

It is easy to see that $\widehat{E}_1$ is a U-statistic with $m=2$ and $h(v_i,w_i;v_j,w_j) = \mathrm{e}^{-\|v_i-v_j\|^a-\|w_i-w_j\|^a}$ being the kernel. A significant feature of this kernel is that $0 < h(v_i,w_i;v_j,w_j) \leqslant 1$. By applying Lemma A.1, we obtain

$$\Pr(\widehat{E}_1 - E_1 \geqslant \epsilon) \leqslant \exp\{-n\epsilon^2\}.$$

As a result, we have

$$\Pr(|\widehat{E}_1 - E_1| \geqslant \epsilon) \leqslant 2\exp\{-n\epsilon^2\}.$$

For the term $\widehat{E}_2$, we write $\widehat{E}_2 = \widehat{E}_{2,1}\widehat{E}_{2,2}$, where

$$\widehat{E}_{2,1} = \{n(n-1)\}^{-1}\sum_{i=1}^{n}\sum_{j\neq i}^{n} \mathrm{e}^{-\|v_i-v_j\|^a}$$

and

$$\widehat{E}_{2,2} = \{n(n-1)\}^{-1}\sum_{i=1}^{n}\sum_{j\neq i}^{n} \mathrm{e}^{-\|w_i-w_j\|^a}.$$

Similarly, we write $E_2 = E_{2,1}E_{2,2}$, where $E_{2,1} = \mathrm{E}[\mathrm{e}^{-\|v_i-v_j\|^a}]$ and $E_{2,2} = \mathrm{E}[\mathrm{e}^{-\|w_i-w_j\|^a}]$. We can similarly show that

$$\Pr(|\widehat{E}_{2,1} - E_{2,1}| \geqslant \epsilon) \leqslant 2\exp\{-n\epsilon^2\} \quad \text{and} \quad \Pr(|\widehat{E}_{2,2} - E_{2,2}| \geqslant \epsilon) \leqslant 2\exp\{-n\epsilon^2\}.$$

Since $E_{2,1}$ and $E_{2,2}$ are both smaller than 1, we then get

$$\Pr(|(\widehat{E}_{2,1} - E_{2,1})E_{2,2}| \geqslant \epsilon) \leqslant \Pr(|\widehat{E}_{2,1} - E_{2,1}| \geqslant \epsilon) \leqslant 2\exp\{-n\epsilon^2\},$$
$$\Pr(|(\widehat{E}_{2,2} - E_{2,2})E_{2,1}| \geqslant \epsilon) \leqslant \Pr(|\widehat{E}_{2,2} - E_{2,2}| \geqslant \epsilon) \leqslant 2\exp\{-n\epsilon^2\}$$

and

$$\Pr(|(\widehat{E}_{2,1}-E_{2,1})(\widehat{E}_{2,2}-E_{2,2})|\geqslant\epsilon)\leqslant\Pr(|\widehat{E}_{2,1}-E_{2,1}|\geqslant\epsilon/2)\leqslant 2\exp\{-n\epsilon^2/4\}.$$

It follows from Bonferroni's inequality and the above results that

$$\begin{aligned}\Pr(|\widehat{E}_{2,1}\widehat{E}_{2,2}-E_{2,1}E_{2,2}|\geqslant 3\epsilon)&\leqslant\Pr(|\widehat{E}_{2,1}-E_{2,1}|\geqslant\epsilon)+P(|\widehat{E}_{2,2}-E_{2,2}|\geqslant\epsilon)\\&\quad+\Pr(|(\widehat{E}_{2,1}-E_{2,1})(\widehat{E}_{2,2}-E_{2,2})|\geqslant\epsilon)\\&\leqslant 4\exp\{-n\epsilon^2\}+2\exp\{-n\epsilon^2/4\}.\end{aligned}$$

Turn to the term $\widehat{E}_3$. First note that

$$\widehat{E}_3=\binom{n}{3}^{-1}\sum\sum\sum_{1\leqslant i<j<l\leqslant n}H^s(U_i,U_j,U_l).$$

Here, $U_i=(v_i,w_i)$ and $H^s(U_i,U_j,U_l)=(H_{ijl}+H_{ilj}+H_{jil}+H_{jli}+H_{lij}+H_{lji})/6$ is the kernel with $H_{ijl}=\mathrm{e}^{-\|v_i-v_j\|^a-\|w_i-w_l\|^a}$. Again $\widehat{E}_3$ is a U-statistic with $m=3$ and $H^s(U_i,U_j,U_l)$ being the kernel. It is easy to know that $0<H^s(U_i,U_j,U_l)\leqslant 1$. By using Lemma A.1, we obtain

$$\Pr(\widehat{E}_3-E_3\geqslant\epsilon)\leqslant\exp\{-2n\epsilon^2/3\}.$$

As a result, we have

$$\Pr(|\widehat{E}_3-E_3|\geqslant\epsilon)\leqslant 2\exp\{-2n\epsilon^2/3\}.$$

Thus we conclude that

$$\begin{aligned}\Pr(|(\widehat{E}_1+\widehat{E}_2-2\widehat{E}_3)-(E_1+E_2-2E_3)|\geqslant\epsilon)&\leqslant\Pr(|\widehat{E}_1-E_1|\geqslant\epsilon/4)+\Pr(|\widehat{E}_2-E_2|\geqslant\epsilon/4)\\&\quad+\Pr(|\widehat{E}_3-E_3|\geqslant\epsilon/4)\\&\leqslant O(\exp\{-Cn\epsilon^2\})\end{aligned}$$

for some positive constant C. The convergence rate of the numerator of $\widehat{SC^2}(V,W)$ is now achieved. Following similar arguments, we can obtain the convergence rate of the denominator. In effect, the convergence rate of $\widehat{SC^2}(V,W)$ has the same rate as that of the numerator. The details are omitted here. □

Proof of Theorem 3.1. From the result in Theorem 2.2, it follows that for any k,

$$\Pr(|\hat{\omega}_k-\omega_k|\geqslant\epsilon)\leqslant O(\exp\{-Cn\epsilon^2\}).$$

Let $\epsilon=cn^{-\kappa}$, where κ satisfies $0\leqslant\kappa<1/2$. We thus have

$$\Pr\Big(\max_{1\leqslant k\leqslant p}|\hat{\omega}_k-\omega_k|\geqslant cn^{-\kappa}\Big)\leqslant p\max_{1\leqslant k\leqslant p}\Pr(|\hat{\omega}_k-\omega_k|\geqslant cn^{-\kappa})\leqslant O(p\exp\{-Cn^{1-2\kappa}\}).$$

The first part of Theorem 3.1 is proved.

To prove the second part of Theorem 3.1, we consider the event

$$\mathcal{B}=\Big\{\max_{k\in\mathcal{A}}|\hat{\omega}_k-\omega_k|\leqslant cn^{-\kappa}\Big\}.$$

Since for all $k\in\mathcal{A}$, the condition (C1) ensures that $\omega_k\geqslant 2cn^{-\kappa}$, then for event $\mathcal{B}$, we have for all $k\in\mathcal{A}$, $\hat{\omega}_k\geqslant cn^{-\kappa}$. Hence we have $\mathcal{A}\subseteq\hat{\mathcal{A}}$. Thus we get

$$\Pr(\mathcal{A}\subseteq\hat{\mathcal{A}})\geqslant\Pr(\mathcal{B})=1-\Pr\Big(\max_{k\in\mathcal{A}}|\hat{\omega}_k-\omega_k|>cn^{-\kappa}\Big)=1-O(s_n\exp\{-Cn^{1-2\kappa}\}).$$ □

Proof of Theorem 3.2. Note that the number of $\{k:\omega_k\geqslant 2^{-1}cn^{-\kappa}\}$ cannot exceed $2c^{-1}n^{\kappa}\sum_k\omega_k$. Otherwise, we have

$$\sum_k\omega_k\geqslant\Big(1+2c^{-1}n^{\kappa}\sum_k\omega_k\Big)\times\frac{c}{2}n^{-\kappa}>\sum_k\omega_k.$$

Further note that on the set $\mathcal{C} = \{\max_{1\leqslant k\leqslant p}|\hat{\omega}_k - \omega_k| \leqslant 2^{-1}cn^{-\kappa}\}$, the number of $\{k : \hat{\omega}_k \geqslant cn^{-\kappa}\}$ cannot exceed the number of $\{k : \omega_k \geqslant 2^{-1}cn^{-\kappa}\}$. Thus we have

$$\Pr\left(|\hat{\mathcal{A}}| \leqslant 2c^{-1}n^{\kappa}\sum_k \omega_k\right) \geqslant \Pr(\mathcal{C}) \geqslant 1 - O(p\exp(-Cn^{1-2\kappa})). \qquad \square$$

Proof of Theorem 3.3. Recalling the condition (C2), we have $\Delta = \min_{k\in\mathcal{A}}\omega_k - \max_{k\in\mathcal{I}}\omega_k$. Thus we have

$$\begin{aligned}
\Pr\Big(\min_{k\in\mathcal{A}}\hat{\omega}_k \leqslant \max_{k\in\mathcal{I}}\hat{\omega}_k\Big) &= \Pr\Big(\Big[\min_{k\in\mathcal{A}}\omega_k - \max_{k\in\mathcal{I}}\omega_k\Big] - \Big[\min_{k\in\mathcal{A}}\hat{\omega}_k - \max_{k\in\mathcal{I}}\hat{\omega}_k\Big] \geqslant \Delta\Big)\\
&= \Pr\Big(\Big[\max_{k\in\mathcal{I}}\hat{\omega}_k - \max_{k\in\mathcal{I}}\omega_k\Big] - \Big[\min_{k\in\mathcal{A}}\hat{\omega}_k - \min_{k\in\mathcal{A}}\omega_k\Big] \geqslant \Delta\Big)\\
&\leqslant \Pr\Big(\max_{k\in\mathcal{I}}|\hat{\omega}_k - \omega_k| + \max_{k\in\mathcal{A}}|\hat{\omega}_k - \omega_k| \geqslant \Delta\Big)\\
&\leqslant \Pr\Big(\max_{1\leqslant k\leqslant p}|\hat{\omega}_k - \omega_k| \geqslant \frac{\Delta}{2}\Big) \leqslant O(p\exp\{-Cn\Delta^2\}).
\end{aligned}$$

This completes the proof. $\square$